ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/363861941

Implications of training university teachers in developing local writing rating scales

Article · September 2022

DOI: 10.58379/XVDF9070

CITATION
1

READS
66

4 authors, including:

Olga Kvasova
National Taras Shevchenko University of Kyiv
28 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Lyudmyla Hnapovska
Sumy State University
11 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Iuliia Budas
Vinnytsia State Pedagogical University
7 PUBLICATIONS **34** CITATIONS
SEE PROFILE

All content following this page was uploaded by Olga Kvasova on 27 September 2022.

The user has requested enhancement of the downloaded file.

Implications of training university teachers in developing local writing rating scales

Olga Kvasova, Taras Shevchenko National University of Kyiv, Ukraine
Lyudmyla Hnapovska, Sumy State University, Ukraine
Vira Kalinichenko, Vasyl' Stus Donetsk National University, Ukraine
Iuliia Budas, Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University, Ukraine

Language assessment literacy is currently in search of new, modern conceptualisations in which contextual factors have a growing significance and impact (Tsagari, 2020). This article presents an initiative to promote writing assessment literacy in a culture-specific educational context. Assessment of writing belongs to the under-researched areas in Ukrainian higher education, wherein teachers have to act as raters and as rating scale developers without being properly trained in language assessment. The gaps in writing assessment literacy prompted research into the strengths and weaknesses of using a local rating scale developed by university teachers. It was conducted within an Erasmus + Staff mobility project in 2016-2019 and followed up by dissemination events held in several universities in Ukraine. The current study aims to explore the impact of training in writing assessment on the processes and outcomes of university teachers' development and use of analytic rating scales. The paper analyses how three teams of teachers from different universities coped with the task, and whether the training they underwent enabled them to design well-performing rating scales. The nine participants in the study developed three local context-specific analytic rating scales following the intuitive method of scale design, detailed in the guidelines prepared by the trainer. Given the same context (ESP) and the same CEFR level (B1 ->B2), we managed to compare the three local rating scales. The study testifies to a positive impact of the training on teachers' literacy in writing assessment.

Key words: writing assessment literacy; local rating scales design; teacher rater training.

Email address for correspondence: olga.kvasova.1610@gmail.com

© The Author(s) 2022. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits the user to copy, distribute, and transmit the work provided that the original authors and source are credited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

The need to enhance teachers' language assessment literacy (LAL), which has been universally recognized by educators and assessment experts across the world (Taylor, 2013; Pill & Harding, 2013; Tsagari, 2020; Tsagari & Vogt, 2017), is crucial in the Ukrainian higher education context. As Bolitho & West (2017) argue, in Ukraine the quality of tests and exams has not been high enough, and that "[t]here is a pressing need for English teachers to be trained in methods of assessment and testing" (p. 81). As a reflection of this situation, Kvasova's (2018) survey of university teachers' assessment literacy revealed that not one of the respondent teachers (N=55) in her study had received training in language testing and assessment (LTA) organized formally by educational authorities in the past few years.

Since 2015, the niche of enhancing teachers' LAL has been occupied by the Ukrainian Association for Language Testing and Assessment (UALTA), a non-governmental organization bringing together university teachers of English from all regions of the country. The principal channel of expertise to UALTA members has been through workshops conducted as a result of UALTA's winning grants from the International Language Testing Association (ILTA) and the European Association for Language Testing and Assessment (EALTA)) or through UALTA members' participation in Erasmus+ mobility projects (https://erasmusplus.rs/home-eng/). These workshops led by international experts in LTA have been traditionally followed by dissemination events held at the participants' workplace, thus involving a much broader teaching community. In addition to international contributions to LAL development in Ukrainian higher education, a team of local trainers conducts workshops at the invitation of universities across the country. During these workshops, the trainers share the knowledge and practical skills they have gained while participating in the international training events and/or projects. Therefore, UALTA serves as a hub for dissemination activities, the effectiveness of which is testified by a steady increase of its membership.

Although university teachers credit their LAL enhancement to both UALTA activities and their department managers (Kvasova and Shovkovy, 2020), performance-based assessment remains an area that needs the particular attention of all those involved in Ukrainian teachers' professional development. While doctoral research studies into teaching L2 speaking and writing are not scarce in the country, the assessment issues have not been broadly addressed in practice. In respect of the assessment of writing in particular, teachers experience major difficulties, among which the absence of detailed guidelines on rating scale (RS) design and/or use seem to be critical causes (Kvasova et al., 2019). To bridge this gap, an Erasmus+ mobility project (2017-2019), with the team of UALTA trainers in LTA involved, aimed to explore the process and outcomes of university teachers' developing and applying an analytic RS (Kvasova et al., 2021). The project was followed by dissemination workshops held by UALTA at three universities in Ukraine, with a total of 47 participants. On being trained in face-to-face

sessions, the participants received a task to design their own RSs. In this article we will report the outcomes of this local project as well as implications for further training of teachers in LTA in the Ukrainian context.

Literature review

An RS is an important component of the rating context. As Weigle (2002) notes, the RS specifies what raters should look for in a written performance; it will have an ultimate influence on the validity of the inferences and the fairness of decisions based on writing test scores. The use of RSs is an essential prerequisite to accurately measuring candidates' skills, and if the scale is designed professionally, it will help ensure reliability of assessment.

It is not only the scale's authors, the designers, who are responsible for the performance of an RS. The lion's share of such responsibility falls on the scale users – raters whose behaviour and interpretation of the criteria may enhance the reliability of judgments or, on the contrary, drastically reduce it. Given the rater's significance, rater training has been well described in the literature (Deygers et al., 2013; Ducasse & Hill, 2015; Lim, 2011; Lumley 2002). RS design has also received meticulous attention from researchers, whereas the procedures for training future RS designers have not been broadly discussed.

RS design is a laborious process that proceeds from the definition of a writing *construct,* which will be represented *de facto* in the way that a RS is written (Knoch, 2011). The explication of the construct in a RS presupposes conceptualization of the most relevant level- and/or curricula-related aspects of writing and specifying them as *criteria.* If justified and adequate, the criteria provide a clear and credible basis for informed judgments and, therefore, promote validity of scoring (Weigle, 2002).

Conversely, potential weaknesses in the design of scale criteria may have a negative impact on the reliability of scoring based on the RS. Such weaknesses include: inadequate ordering of criteria that results from their inconsistency with SLA theory; irrelevance of criteria to tasks and content; incorrect grouping of criteria at different levels; and relativistic or vague wording (Turner & Upshur, 2002).

In standardized testing of writing, the qualifications of language testing professionals who design RSs seem to be unquestioned, whereas language teachers are generally assumed to lack such expertise (Taylor, 2013). There is evidence that the validity and reliability of teacher-developed tests are indeed rather low (Harlen, 2004; Gareis and Grant, 2015); and in regard to Ukrainian teachers in particular, a critical need to enhance their skills in test construction and use was voiced by Bolitho and West (2017). Based on this, we may suggest that the LAL of teachers is unlikely to be sufficient for them to undertake RS design. However, we assume that classroom teachers have other

strengths that may compensate for their lack of competence in LTA and serve as an advantage for them as grassroots RS developers.

According to Hamp-Lyons (1989), research on writing assessment must take a context-embedded approach. In line with this claim, Weir and Shaw (2006) assert that the *construct* "must be developed for each testing situation and it must take into account the test takers, the purpose of the test and the real-life situation the test is trying to 'simulate'" (p. 11). Standardized assessment is viewed by Fulcher and Davidson (2007) as lying outside any specific context, while classroom assessment is "directly relevant to assessment of learners in a particular setting" (p. 25). The role of impartial testers is different from the role of persons immediately involved in teaching a certain curriculum in a specific setting where "a known group of assessors rate a familiar population of learners" (North & Schneider, 1998, p. 220). This is the basis for our assumption that there are advantages in developing and using local, context-specific, RSs.

In a similar vein, Turner (2000) and Plakans (2013), who describe experienced teachers' participation in RS design, highly value teachers' knowledge about the curricula and ongoing developments in particular educational contexts. As Hill and Ducasse (2020) assert, teachers are most frequently viewed as mere recipients of expertise from scholarly supervisors, which underestimates their capacity to be proactive in research. Nevertheless, Green (2014) argues that practicing teachers may act effectively as rating scale designers, providing they possess an appropriate level of assessment literacy and are efficiently guided. Following this line of thought, we presume that teachers' knowledge of the context - curricula, learners' needs, cognitive abilities and L2 proficiency level, as well as their hands-on knowledge of language pedagogy and experience - will be crucial in enabling them to design effective RSs for their classroom.

A question about how to prepare teachers for the design of context-specific RSs prompted our analysis of the empirical research into procedures and outcomes of rater training. It appears that there are ample studies on training professional raters (e.g., Knoch, 2011; Lumley, 2002), and a limited number of studies that consider the involvement of raters (Barkaoui, 2010; Harsch & Martin, 2012) and teachers (Barkaoui, 2010; Plakans, 2013) in rating scale design, whereas no information is available on training teachers to act as RS designers *on their own*, to the best of the authors' knowledge. Therefore, in search of answers to the question of how to effectively train teachers in RS design, we undertook a synthesis of relevant research studies.

One such feasible approach to RS design is found in the study by Kim (2015). Having searched the literature for a detailed description of how to introduce a RS to raters-to-be, the researcher found it was addressed only in a limited number of papers (e.g., Barkaoui, 2011; Knoch, 2011; Lovorn & Rezaei, 2011). Wherever such description is provided, as Kim claims, "the common approach seems to be *Present and Clarify/Explain* with respect to the descriptors on the rating scale" (Kim, 2015, p.6). She

further points to the frustrating outcome that is typical of such hierarchical (or top-down) training: on the one hand, trainees are expected to accept and internalize the prescribed scale descriptors in their finished form, and yet even after the training raters continue to misinterpret the descriptors. Continuing misconceptions by raters are often referred to in the literature (e.g., Knoch, 2011).

Instead of training raters towards a complete understanding of the criteria and a subsequent ability to interpret them while rating, Kim advocates a bottom-up, or rater-centred approach. She exemplifies it by a training protocol which includes such steps as 1) activating existing knowledge about purposes of writing; 2) evaluating a writing sample based on existing knowledge without any rubric; 3) familiarization with rating scale descriptors; 4) matching current knowledge with the scale descriptors; 5) practice using the descriptors without giving scores; 6) familiarization with the complete version of the rating scale (with score indicators); and finally,7) practice using the complete scale with a familiar essay. The author argues that step-by-step scaffolding enables trainees to "develop an understanding of and the ability to apply the descriptors on the rating scale" as well as coping with "translating descriptors into numerical scores" (Kim, 2015, p. 9).

Although not empirically tested yet, the approach appears well-reasoned as far as the employment of trainees' cognitive abilities is concerned. The training proceeds from activating the existing knowledge and experience of trainees to engaging them in constructing new knowledge about rating, namely, by facilitating their hands-on experience in using RSs. In this, we observe a strong similarity with the constructivist approach to teacher training (Schcolnik et al., 2006; Vopel 2006), which is adopted in building LAL via workshops/webinars conducted by testing organizations such as EALTA, ILTA, the British Council, and the IATEFL Testing, Evaluation and Assessment (TEA) SIG. The authors of this article have participated in such training as trainees and have followed its tenets when acting as teacher trainers as well.

In this research, we argue that bottom-up training of raters as prospective RS designers is both feasible and constructive. The approach involves closely following a comprehensive algorithm of RS design proposed by Weigle (2002). Moreover, the development of a local RS will be based on and aligned with Weir's (2005) socio-cognitive framework of test development, administration and analysis, in that it will consider particular test takers, curriculum-related tasks, and a context-dependent scoring process, as well as post-test adjustment and scale modification. All these considerations seem critical to developing a well-performing rating scale.

Research rationale and questions

The overview of the sources on RS design and rater designer training corroborates our assumptions about the possibility to train practicing university teachers in the design and use of context-specific RSs. Based on our long-term teaching practice and

extensive experience of training teachers in LTA, we trust that practicing teachers have due competence in language pedagogy and in-depth theoretical knowledge as well as hands-on experience of delivering the curricula they develop and teach. University teachers have an invaluable understanding of their students' overall proficiency in L2 and their writing ability in particular. Since teachers and students in the Ukrainian context have a common national identity, they share social, educational and cultural perspectives and influence one another within the instructional collaboration.

Based on the results of a survey of typical assessment practices of 104 Ukrainian university teachers (Kvasova et al., 2019), we can state that the teachers are continually engaged in assessing students' writing. Some of them are reasonably familiar with the RSs used in commercial examination systems. Teachers also attempt to utilize these ready-made scales by tailoring them to their own context. The record of teachers' participation in UALTA-conducted workshops and further observation of their teaching/assessment practices allow us to describe a certain proportion of university teachers as hard-working and generally inclined to undertake life-long learning and professional development (Kvasova & Shovkovy, 2020).

Given the premises outlined above, we presume that the university teachers as well as students will benefit from the introduction and use of context-specific RSs in the summative assessment of writing. In our view, such an RS should be *analytic*, as this type of RS allows raters to thoughtfully consider all aspects of learners' writing and provide accurate feedback to them. The local RS should be developed in an *intuitive* rather than empirical way (Fulcher, 1996), as this is consistent with the professional background of the scale designers and is more practical in terms of utilizing assessors' precise knowledge of the context (curricula requirements, content of teaching writing, learners' abilities). The local RS should represent the writing test construct accurately, as it is based on and linked to the curriculum taught, with the most relevant features of target writing skills conceptualized as criteria for assessment.

In terms of organization, training teachers to act as RS developers should follow the constructivist approach which, according to McLeod (2019), promotes a sense of personal agency as they gain ownership of their learning and assessment. This learner-centred approach encourages trainees to draw on their prior experience, reflect on it and on the prior knowledge, and construct new knowledge/expertise through accomplishing meaningful tasks (Elliott et al., 2000; Honebein, 1996).

All these considerations need to be complemented by describing the conditions under which training teachers to act as RS developers is conducted by UALTA as a non-governmental, non-for-profit professional organization. As such, the current study differs fundamentally from RS design commissioned by governmental or other organizations that have provided ample funding to researchers (cf. Ducasse & Hill, 2015; Harsch & Martin, 2012; Plakans, 2013). The limited funding that may be allotted

by a local non-governmental organization has had a significant impact on the rationale for the current study and its evolution.

To begin with, reduced funding allows for just a single intensive face-to-face training session on site. The training is supposed to be continued in a blended format, with the independent work of teams being combined with periodic consulting of the trainer online. Second, to ensure efficient and effective RS design, the independent work needs to be conducted in compliance with detailed and carefully written guidelines. Additionally, the independent mode of RS design increases by far the importance of teamwork, which is to ensure the participants' full understanding of all tasks and/or ways to implement them, as well as their active and informed involvement in collaborative RS development.

Finally, RS developers' independence, which we view as autonomy regulated by the trainer's guidelines, should encourage ongoing reflection on the part of team members, whose voices should be heard in team discussions leading to decision making. In our case, regulated autonomy means chronicling the RS design in its entirety – including opinions expressed and reasons for accepting or rejecting them – thus enabling the trainer to monitor and investigate the process.

Thus, this study aims to address the following research questions:

1) How should the RS design process be organized to enable the development of a workable local RS? What are teacher perceptions of the process?

2) What are teacher perceptions of the usability of the RS they have designed?

Embarking on this investigation, we presumed that the factors affecting the rating scale development process might involve certain gaps in the linguistic and assessment competence of the assessors, which had been previously identified in the survey of teacher writing assessment literacy (Kvasova et al., 2019) Therefore, we expected that consideration of these issues would contribute to an understanding of best practice in scale design implemented by practicing teachers. The insights gained will be drawn on in the development of procedures and materials for training teachers in assessing writing in Ukrainian universities, which could be useful in other contexts, too. As a result, the research will contribute to filling the gap in the LAL literature about training teachers to act as RS designers and promote teacher writing assessment literacy on a more global level.

Method

Participants

The overall number of participants in the RS developer training was 47. These were university teachers of ESP working for three universities in Ukraine who signed up for

UALTA workshops aimed at disseminating the outcomes of the Erasmus+ mobility project. The participants engaged in face-to-face training in the form of a workshop conducted by UALTA trainers and then performed a post-workshop task.

For the purpose of this research, we decided to consider the data provided by one focus group from each of the three universities. The immediate participants in this study therefore were nine (three teams/focus groups made up of two to four teachers).

The members of the selected focus groups had homogeneous profiles: they worked for the same department, taught the same courses, had similar teaching experience (over 15 years) and had a record of participation in training events conducted by UALTA. We considered the internal homogeneity of the focus groups to be a factor in securing the quality of the research and the ensuing accuracy of its results. However, we were also aware that, despite the similarity of their profiles, the would-be RS designers possessed divergent linguistic and pedagogical competencies as well as assessment styles that would obviously result in variance in their scoring. Besides, we expected that the work of each focus group would be marked by institutional educational traditions and assessment culture that would also affect the developed RS and the scores obtained in the rating trial.

Materials and Instruments

The *Guidelines for RS developers* were introduced in the initial workshop and were intended to guide the work of the teams in developing their scales. The Guidelines were written by the principal trainer in the project, one of the authors of this article, for the participants in the training. They were based on the comprehensive algorithm for RS design proposed by Weigle (2002), which can briefly be reported as a sequence of stages, beginning with defining the construct or the components of ability to be measured through to producing the RS. The *Guidelines* contained eight sections describing the steps to take in RS design, as outlined in Table 1; they also included templates for each step's log (see Appendix A for the latter).

Table 1. The structure of *Guidelines for RS developers*

Steps	Activities
1 Defining the test construct	Teamwork: suggesting and justifying the number and names of criteria
2 Defining the criteria	Teamwork: articulating the descriptors for criteria
3 Defining the number of bands and descriptors of criteria	Teamwork: suggesting and justifying the number of bands and scores per band Teamwork: grading the criteria descriptors per each band
4 Developing an analytic RS	Teamwork: completing the template for the analytic RS
5 Learning to use the developed RS	Teamwork: learning to calculate a score using the RS
6 Trialling the RS	Individual rating of three sample essays of the type and on the topic prescribed by the curriculum Teamwork: discussing the assigned scores, achieving agreement on a common score for each essay

7 Rating essays	Individual rating of 10 essays of the type and on the topic prescribed by the curriculum, entering the scores in a common spreadsheet. Teamwork: discussing (comparing and justifying) the scores.
8 Rating essays and finalizing the RS	Individual rating of 20 essays using own scale Teamwork: discussing the RS use, suggesting improvements

The instrument for documenting the process of RS development was *a reflection log, or log* (Friesner and Hart, 2005), which is defined in education as a tool that is used to record someone's learning, experience and reflection (University College Dublin, n.d). As individuals note down or "log" what they have done, a log is intended to give them an accurate record of a learning process; it helps them to reflect on past actions and make better decisions for future actions. In this study, logs were used to 1) ensure precision in fulfilling step-by-step recommendations given in the *Guidelines for RS developers*; 2) obtain complete and detailed information about all actions taken, e.g., explanation of the choice of criteria and their wording; and 3) elicit reflections recapping the accomplishment of each step in terms of the process and its outcomes. In this research, the logs were filled in by team leaders. They chronicled the flow of the team's discussions, the procedure followed and the reasons for taking decisions, as well as recording ensuing actions regarding the RS design. The template for the log is found in Appendix A.

The *feedback questionnaire for RS developers* (see Appendix B) was intended to elicit the respondents' perceptions of the effectiveness of the procedure. It contained 10 questions, eight of which were selected-response, with an option to provide one's own view, and two were open-ended questions, recapping the overall impression of the experience and eliciting suggestions for further research.

Procedure

The study was implemented in three stages, as presented in Table 2.

Table 2. The overall study design

	Stage	**Materials**	**Purpose**	**Outcomes**
1-2. Training in writing assessment	RS developer training: face-to-face workshop (47 participants)	*Guidelines for RS developers*	Train teachers in writing assessment literacy	Preparation for independent RS design
	RS development and piloting: independent teamwork (9 participants in three teams)	Reflection log completed by the leaders of the three teams	Provide accurate record of RS development process	Development of three local RSs

3. Feedback on training	Collection of feedback on the RS design process (9 participants)	Questionnaire for RS developers	Elicit perceptions of the RS design process.	Insights into the design process

1. RS developer training

The RS developer training consisted of a four-hour face-to-face workshop, held at each of three participating universities. The workshop was developed by a principal trainer, one of the authors of this article, in collaboration with two other UALTA trainers, and these three trainers conducted the workshop sessions. During each session the trainees were led to:

- familiarize themselves with and understand RS-related concepts and terminology,
- reflect on the essential qualities of target writing,
- discuss the salient traits of written scripts, and
- become prepared to further articulate these as criteria.

Although the trainees were exposed to several RSs used in large-scale testing, they were encouraged to generate their own language to define the criteria and descriptors. This order of activities, similar to what Kim (2015) suggested, omitted the typical stage of rater training that aims to develop a full understanding of the criteria already formulated by the trainers (Knoch, 2011). We presumed that it would be more feasible for the trainees to use self-defined criteria/descriptors and utilize them in further scoring.

2. RS development and piloting

The post-workshop task involved 1) developing an analytic RS fit for a particular classroom setting and 2) piloting it in real-life teaching in line with the *Guidelines for RS developers* that consisted of the detailed description of eight consecutive steps (see Table 1 above).

RS development and piloting were conducted by the teams independently within 4-6 weeks following the workshops. As they proceeded through the steps, the team leaders kept *the logs* the template for which was provided in the *Guidelines for RS developers* (see Appendix A).

3. Feedback questionnaire

This stage consisted in completing the feedback questionnaire whose aim was to elicit teacher-rater impressions of the process of RS development and use. The questionnaire was compiled in Google Forms by the principal trainer, who had piloted it with the help of two other UALTA trainers. The survey was intended for the nine participants in the focus groups who served as RS designers. It was administered anonymously after the RS design was completed, in the participants' own time.

Data analysis

The outcomes of Stages 1-2, Training in writing assessment' were the three analytic RSs designed by the participants in the study. These RSs underwent comparative analysis regarding the number of bands, number of points in the grading, as well as criteria, their number, wording and interpretation. We also compared the mean scores given by the team to the ten student essays to see how similar they were. However, the study did not purport to establish accuracy of scoring and reliability of the developed RSs. Instead, we preferred to focus on the process of RS design and draw certain implications for similar kinds of training for teachers to act as RS developers.

The data entered in the logs were analysed against the following criteria: 1) accomplishment of the steps of RS design required by the *Guidelines for RS developers*; 2) completeness of the entries (provision of detailed information about team discussion leading to decision); and 3) reflection and critical evaluation of the accomplishments of each step. The entries, which allowed in-depth analysis of the processes of collaborative RS development, are further quoted by us in the Results and Discussion sections.

The data obtained via the questionnaire were analysed by means of simple frequencies of response.

Results

We will first present the results that are relevant to RQ1: "How should the RS design process be organized to enable the development of a working local RS? What are teacher perceptions of the RS design process?" The three RSs that were the final product, or the outcome of *1-2. Training in writing assessment* are shown in summary form in Table 3.

Table 3. Characteristics of the three teacher-developed rating scales

Scales	Criteria	Bands	Points per band
Scale 1	1 Task achievement	'band 5'	18-20
	2 Quality of writing	'band 4'	14-17
	3 Organization	'band 3'	10-13
	4 Language	'band 2'	6-9
		'band 1'	4-5
Scale 2	1 Content and organization	'excellent'	10
	2 Coherence and cohesion	'good'	9-8
	3 Lexical resource	'satisfactory'	7-6
	4 Grammar range and accuracy	'unsatisfactory'	5-0
	5 Task achievement		

Scale 3	1 Ideas and relevance	'excellent'	5
	2 Text organization	'good'	4
	3 Vocabulary range and accuracy	'satisfactory'	3
	4 Grammatical range and accuracy.	'poor'	2
		'very poor'	1

As is seen in Table 3, the number of bands and points in the RSs vary, whereas the criteria represent a variation of the most frequently used ones, such as task achievement, organization, vocabulary and grammar (Weigle, 2002).

Of more interest for us is the process of collaboration within the three teams while designing the RSs, based on the information provided in the logs. We will now go over the most significant entries, following the template for log entries in Appendix A.

1. Defining the test construct

Proceeding from what Weir and Shaw (2006) recommended, the *Guidelines* proposed that the teams define the construct for their "particular testing situation", i.e., taking into account their context (ESP), the test-takers' proficiency level and the curriculum. The teams were invited to conceptualize the expected writing ability of their students and try to capture the most relevant aspects of test-takers' writing. These discussions naturally led to the articulation of these aspects as criteria. For instance, having experience of using IELTS writing band descriptors, the RS2 team agreed on taking these descriptors as a basis. They initially wished to adopt the four criteria offered by IELTS (Task Achievement, Coherence and Cohesion, Lexical Resource, Grammatical Range and Accuracy) but in the course of discussion agreed to introduce the criterion Content and Organization. While assessing the presentation of ideas in an essay, they found it essential that logic and clarity of expression should be demonstrated by university students.

2. Defining the criteria

The way the criteria are worded reflects the teams' detailed vision of the features of written text they wished to assess from two perspectives: a) their linguistic and teaching competence, and b) their knowledge of and adherence to the curriculum they taught in a particular setting. The fact that they expressed their vision in their own terms encouraged their ownership of the scale and responsibility for the outcomes of their problem-solving collaboration.

As one RS2 log reads,

> initially we wanted to use terse and brief definitions [of criteria] but on doing so, we realized that concise definitions do not always reflect the intensity of a particular aspect of evaluation and sometimes make the wording too generalized and non-specific. Therefore, we decided to specify the wording and slightly extend their scope (Log 2).

Apart from exemplifying the critical description of the process, this comment makes one more important point. It suggests that the team agreed with the observations made by Alderson et al. (1995) and Fulcher (2010) that descriptors with a high degree of complexity and level of abstraction will influence the quality of the raters' judgments and may impede reliability of scoring. By recognizing the need for criteria that were characterized by reasonable brevity, absolute clarity as well as explicitness (Knoch, 2009), the team showed a commitment to avoid confusion and misinterpretation of the scale by its prospective users. We also view this as evidence of the RS designers' enhanced LAL and willingness to promote it in their setting.

3. Defining the number of bands and descriptors of criteria

The third step revealed controversies faced by the scale designers.

As is seen in Table 3, Scale 3 operated with a smaller range of scores (0-5). This is customary and convenient for teacher users as the points fully comply with the 5-point grading system long-established in the country. Even though summative assessment in universities is based on the European Credit Transfer and Accumulation System (ECTS) 100-point scale, in academic records the ECTS-based grade is usually complemented with the corresponding national grade, from "excellent" to "very poor/unsatisfactory".

In Scale 2, although the four bands were also labelled in compliance with the national tradition, the scores are distributed across a 10-point grading scale, which, in our view and the view of the team, makes them more compatible with the ECTS-related 100-point ECTS grading system, in which 90-100 stand for "excellent", 75-89 for "good", 60-74 for "satisfactory", and less than 60 for "poor". In their log, the developers of this RS noted that initially they had wanted to adjust their RS to the ECTS as much as possible and arrange the scale in 6 bands. However, they rejected the idea on the grounds of its impracticality and finally adopted a 4-band scale. In this scale the lowest passing grade – 6 (standing for "satisfactory") is compatible with 60 points, which is the lowest passing score within the ECTS scale.

Scale 1 complies the least of all with existing scales as it contains 20 points. The analysis of the log allowed us to reveal certain misconceptions about bands and score calibration that the team had. On the one hand, by adopting a 20-point grading system, the team seemed committed to award scores that were as accurate and precise as possible. However, they contradicted themselves by noting that "[N]either of us believed that we needed a very thorough and complicated description of each band as assessing writing is a part of our usual routine" (Log 1). Guided by the syllabus, according to which the top grade was '5', the team adopted five bands, which led them back to the traditional 5-point grading system. Piloting of the scale proved that the initial distribution of points within the bands contradicted the team's intention to be practical rather than overly scrupulous. This conclusion led them to convert the 20-point scale into a traditional 5-point one.

4. Developing an analytic RS

This step included completing the RS template with descriptors of the criteria for each band and, according to the logs, it proceeded more or less smoothly. However, the difficulty mentioned by all three teams lay in proper wording of graded descriptors. This resulted from somewhat subjective treatment of English evaluative adjectives, e.g., in the RS2 log for band "excellent", the range of vocabulary was interpreted as "sufficient"; for band "good" as "adequate"; for "satisfactory" as "basic" and for "unsatisfactory" as "very limited". These attributes seem quite inconsistent, especially when compared with some professionally designed RSs (cf., Jacob et al.'s scoring profile (1981), as cited in Weigle (2002, p. 116).

In respect of grammar, the RS1 log presented an attempt to quantify grammatical errors and their severity per band. The quantity of errors was not underpinned by theoretical evidence in the log, although it seemed to be supported by the team's practical experience. Despite the tradition to deduct points for grammar mistakes, which is deep-rooted in the Ukrainian L2 assessment of writing, the developers of RS 1 did not call for adhering to this practice. Additionally, the RS3 designers expressed doubts about the basis on which grammatical errors should qualify as impeding and not impeding communication. The answer to this, to our knowledge, is still unresolved in the LTA literature (Hyland & Anan, 2006; Touchie, 1986).

5. Learning to use the developed RS/calculate scores

This step is least commented on in the logs, which might have suggested that the information presented in the *Guidelines* was precise and clear. The RS2 team, for instance, noted that the use of a RS makes the calculation of the total score easier and more objective, since the average scores awarded across the entire scale reflect all the criteria, preventing subjective raising or lowering of scores. However, as it appeared later, during the analysis of awarded scores (Step 7 below) and online consultations that followed, some raters had overlooked the major advantage of analytic scoring: the opportunity to rate each particular trait of a script, expressed as a criterion, separately. On additional clarification, the misconception was corrected but the case provided quite meaningful food for thought for the trainer, which is further touched upon in the Discussion section.

6. Trialling the RS

After trialling their RS with three sample essays, the immediate response of the teams was that scoring was easier with an RS: as soon as you "come to grips with the procedure and are careful about the bands and points" (RS2 log); "colleagues help out with arithmetic and patiently explain calculations" (RS1 log); and "it helps a lot to discuss and justify the scores, compare your opinion with others', think and possibly accept their point of view" (RS3). However, trialling evoked in some designers the wish to "immediately reformulate some of the criteria as they did not work properly and were confusing". The quoted comments suggest that trialling performed its

function in helping RS designers to gain hands-on experience of rating against a scale and efficiently collaborate in teams.

7. Rating ten essays by their own students

The log entries showed that, after rating ten more essays, the teams found that their RSs facilitated scoring since "it is convenient to use for teachers and transparent for students to understand their scores. The use of the scale raised objectivity of assessing writing, resulting in scores that didn't vary considerably and that can be considered fair, valid and reliable enough" (RS3 log). Interestingly, the logs contained some discussion-provoking points, such as "What if the student exceeded the word count considerably or, on the contrary, produced a too short script but in general responded to the task? Which criterion do spelling mistakes fall into?" (RS3 log). These points should be definitely considered when preparing further training of RS designers.

8. Rating 20 essays of other teams' students

Rating 20 essays from the students of other teams was aimed at revealing more aspects of RS usability. Accepting that the RS they designed "proved efficient enough when applied to assessing the essays on different topics taken from different settings"(RS 1 log), RS designers raised some issues that needed additional consideration. These referred to the length of essays ("Should the word count be entered in the scale descriptors which automatically suggests revision of the criteria or should it be made a mandatory part of the task to ease rating?" RS 3 log) and academic integrity ("Should (allegedly) plagiarized chunks of text/paragraphs be assessed and how?" RS2 log). Rating the 20 additional scripts demonstrated that the issues identified in other teams' scripts should be considered and possibly integrated in the designed scales before finalizing them.

Apart from the logs, information about the process of RS development was elicited from the feedback questionnaire (items 2-3).

When asked about what resources the scale design procedure employed, all respondents emphasized the usefulness of the training materials and guidelines as well as additional readings. They also pointed that they had drawn on their teaching experience and/or intuition, knowledge of language testing, and their colleagues' experience and expertise. Additionally, seven respondents stated that they relied on their linguistic competence and knowledge of language pedagogy. Concerning the process of RS development, none of the respondents reported prompt accomplishment of the task, with all of them claiming that the development of the RS was quite demanding, yet feasible. The process involved a lot of discussion within the team while reviewing, rewording and/or significant revisiting of the scale (9/9). Interestingly, five of the respondents claimed that their collaborative effort was crucial for designing the RS, while four believed that the RS design could be done single-handedly, time permitting. One of the anonymous respondents wrote,

> It was very challenging to develop scales. I had never done it before and thus I had to turn to training materials. At the same time, this experience was rather useful as it gave me a chance to consider students' essays differently, and compare my opinion with my colleagues.

All in all, the respondents' evaluation of the experience was quite straightforward: all of them found the experience of scale design and piloting insightful and useful.

The analysis of the logs and questionnaire data confirm that first the RSs were designed in compliance with Weigle's (2002) algorithm of scale design reflected in the *Guidelines for RS developers;* they were therefore compatible with the scales commonly used in language testing (cf., Knoch, 2011) Secondly, the RSs were relevant to the contextual requirements in terms of the number of bands and scores as well as the choice and interpretation of the criteria Thirdly, the RSs were efficiently used in the rating of 30 student papers.

Moving now to Research Question 2**,** this was intended to find out teachers' perceptions of the RS usability as raters and their satisfaction with the developed product. To report these perceptions, we will draw on the data obtained from items 1 and 4-10 of the questionnaire (See Appendix B).

All respondents felt that their rating of essays became more informed, as "the process helped me reconsider the experience of scoring, taught how to distinguish what really matters and what is not so important". While seven out of the nine respondents claimed that the scores did not go either higher or lower when compared with the conventional scoring they implemented before the training, two of the respondents admitted assigning higher scores but no one reported assigning lower ones. As for the scoring time while using the RS, the opinions of raters were evenly divided: five respondents claimed that the rating was more time-consuming, whereas the other four thought they took less time than before.

Regarding the rating of the 10 scripts by their own students in Step 7, the respondents unanimously mentioned that they needed to re-read scripts more than twice and refer to the scale multiple times. They also reported some confusion while applying such criteria as "Task achievement" (in Scales 1 and 2), "Ideas and relevance" and "Test organization" (in Scale 3). It is worth noting that these issues were reflected on in the logs, thus corroborating the evidence obtained via the questionnaire.

Somewhat different responses were elicited with regard to rating the 20 scripts (from the two other teams) in Stage 8. Although all of the respondents admitted that re-reading each script more than twice was required, two of them mentioned "occasional rather than frequent referring to the scale". This might imply that the enhanced practice reduced the need for extra effort. Similarly, five respondents mentioned that less time was spent on scoring the last scripts in the batch as compared to the first ones. Besides, less confusion in the use of the criteria was reported in two cases, which

together with the previous statements suggests increased confidence among the raters. Instead, six out of nine respondents mentioned that they had needed to re-read the rubric of instructions for the writing tasks several times to really absorb what was stipulated. This is evidence of the diligent approach of the raters to the procedure and its documentation. Moreover, it emphasizes the significance of the writing task itself and the rater's familiarity with it.

Overall, the piloting of the RSs revealed an absolute need for all the raters to discuss and compare the scores with colleagues. Additionally and more specifically, the respondents reported: 1) the necessity to achieve better discrimination between levels (5/9), 2) a lack of focus on some essential features of the written product (6/9) and the need for additional criteria, such as "Task achievement" (in Scale 3) and "Logical presentation" or "Text integrity" (in Scale 2), 3) vagueness in the wording of descriptors for the criterion "Quality of writing" and 4) the need for subtler specification of the criterion "Organization" (in Scale 1). Importantly, they mentioned that effective rating presupposed aligning the scale with a particular curriculum (6/9) as well as considering task requirements (4/9). At the same time, the respondents were content with the number of points allocated to each of the criteria in their scales and the number of points per band.

Summarizing the advantages of relying on a RS in rating, all respondents emphasized enhanced thoughtfulness, as well as less subjectivity. What is more, they noted the possibility to provide informed feedback to test takers and see better informed implications for teaching writing. However, not all of them conceded that more accurate scoring was enabled when using the locally-developed scale (5/9).

The respondents' evaluation of RS use was generally very high, although some of them conceded that it was challenging to use the RS initially but it became easier with practice. Although only two out of nine respondents were fully satisfied with the quality of the developed RS and expressed that belief that it could be used accurately use in rating any type of texts, almost all (8/9) reported that they were going to apply it in the future and recommend it to other teachers (6/9). Fortunately, none of the respondents reported that it was difficult to use the RS all the time. Neither did anyone admit they could rate more effectively/accurately without a RS, i.e., intuitively, as they did before being trained.

In response to the question about their satisfaction with the training in RS design, the answers were entirely positive; in addition, they expressed the wish to participate in any other similar training sessions. The following quotation seems to express the participants' thoughts in a concentrated manner:

> My impressions are totally positive, I am thankful to the trainer and my colleagues from other Unis for obtaining useful experience in assessing student essays. Scoring an essay is not an easy procedure, there are some aspects that I still feel doubtful about, for example, the criterion "Task achievement" which

> is clear enough in theory but rather complicated when assessing a particular essay. Another thing that I need to clarify is whether one rating scale is appropriate for different types of essays or it is better to design a separate more precise scale for each essay type.

Moreover, among the impressions of the training experience, several suggestions were voiced about the necessity to continue the work and elaborate one common accomplished scale. This suggestion offers the prospect of future, larger-scale research into RS development, in particular RS developer training with a focus on feasibility of using the scales in the post-training period.

Discussion

As was stated in the Results section, the three developed RSs may be considered appropriate for the university classroom. They have been developed by teachers who know the teaching/learning context, including the curriculum and its requirements for writing in full detail - and are experienced in teaching and assessment as part of it. The RS design process was documented in detail in the reflection logs, which allowed a thorough examination of this process and the shaping of implications for organizing similar training in the future.

To begin with, some interesting implications arose in respect of the LAL developed by the trainee teams. The first one confirms our assumption about quite a high level of LAL that university teachers initially possessed, enabling them to design working RSs. This level of LAL, defined by Pill and Harding (2013) as "procedural and conceptual literacy", suggests that university teachers understand central concepts of the field and are able to use LTA knowledge in practice quite effectively. The way the criteria were worded during the RS design reflected the teams' clear and detailed vision of the features of written text (Weigle, 2002) from the perspectives of their linguistic and teaching competence appropriate for their specific, university, context. In line with the need for contextualizing LAL (Tsagari, 2017, 2020) it is important that the teachers utilized their in-depth knowledge of and adherence to the curriculum they taught in particular settings. Consequently, the LAL that they have gained may be considered contextually appropriate, being totally tailored to their professional and intellectual needs, educational level, conceptions, experience, and motivation to construct new knowledge and build new skills. Last but not least, the practitioners tended to value their intuition (Scarino, 2013) and well-proven experience (Berry et al., 2019) and trust them.

Apart from the mentioned conceptual issues, one more issue worth discussing is the treatment of grammatical errors in rating against a relevant criterion. How many errors are acceptable in a script within the existing bands? What kind of errors are likely to be impeding communication and to what extent? Interestingly, the degree of error gravity is viewed differently by native and non-native speaker teachers, as non-

native speaking raters have typically been harsher raters than native speakers (Kim & di Gennaro, 2012). Besides, based on our own working experience, we presume that there exists considerable variation in treating grammatical errors as impeding/not impeding communication within the Ukrainian teacher community. The implication is that subjective treatment of errors may be a serious threat to reliability of scoring and is worth further investigation by applied linguists to support language teachers in non-Anglophone contexts.

The most significant practical implication from our study concerns the organization of the training, namely a combination of face-to-face intensive workshops and independent RS development guided through detailed written recommendations and online consultations with the trainer. The quite high level of trainees' writing LAL achieved in the training suggests that it can be considered effective, but may improve by 1) increasing the number of contact hours allotted to workshops, with appropriate funding provided; 2) improving the *Guidelines by RS developers* in some sections by making them more specific and clear (e.g., how to calculate scores); 3) setting deadlines for doing tasks envisaged by the steps in the *Guidelines* and completing the logs; as well as 4) employing interviews rather than a questionnaire to achieve a clearer understanding of the RS design process.

Finally, although the study has made a certain contribution to promoting context-specific LAL and providing meaningful insights, it has some limitations. These refer primarily to the small-scale scope and number of participants as well as reliance on the questionnaire, known as a tool for eliciting subjective opinions and making judgements on them.

We see the prospects of future research in accomplishing the development of the training materials including the *Guidelines for RS developers,* improving the methodology of employing logs and conducting a large-scale study aimed at enhancing the writing assessment literacy of university language teachers.

Conclusion

The procedure of RS design described in this article was carried out in line with the algorithm proposed by Weigle (2002) and was based on the *Guidelines for RS developers* prepared by the trainer. Those appeared comprehensive in terms of providing all necessary recommendations for the steps in RS design. The logs confirmed that the guidelines were followed by the RS developers closely, thus promoting the overall effect of the intensive training. However, as was mentioned above, there is still room for improving the recommendations.

The RSs developed by the focus groups are influenced by the individual views of the RS developers, their teaching experience and the setting-specific assessment culture.

However, they are comparable with one another as being well-designed, clearly-worded, feasible and well-performing in practice.

The criteria included in the self-developed RSs appear to be consistent with the RS designers' beliefs in the essential features of student writing; they explicitly articulated the construct and in the final versions are devoid of vague wording that would need additional effort in understanding them. The teachers displayed full awareness of possible causes of confusion in interpreting some of the criteria and were meticulous about remedying the situation in Steps 6 and 7.

On the whole, the use of a self-designed RS brought confidence in teachers' rating and transparency to the assessment process, thus facilitating efficient feedback on students' writing, which is a clear advantage of using an analytic RS in scoring. The RS teacher developers acknowledged the efficacy of both the training they received and the RS development procedure. They expressed a strong commitment to using the developed scales with some modification in the future, adapting the scales to other curricula when needed as well as recommending RS use to other teachers. It appears that the methodology of enhancing writing assessment literacy in a particular educational context proved effective.

References

Alderson, J. C, Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54-74. https//doi.org/10.1080/15434300903464418

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293. https//doi.org/10.1080/0969594X.2010.526585

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers? *ELT Journal, 73*(2), 113-123. doi:10.1093/elt/ccy055

Bolitho, R., & West, R. (2017). *The internationalisation of Ukrainian universities: The English language dimension.* British Council Ukraine report on English for Universities Project. https://www.teachingenglish.org.uk/article/internationalisation-ukrainian-universities-english-language-dimension

Deygers, B., Van Gorp, K., Joos, S., & Luyten, L. (2013) Rating scale design: A comparative study of two analytic rating scales in a task-based test. In E. Galaczi & C. Weir (Eds.), *Exploring language frameworks* (pp. 271–287). Cambridge University Press.

Ducasse, A. M., & Hill, K. (2015). Development of a Spanish generic writing skill scale for the Colombian Graduate Skills Assessment (SaberPro). *Papers in Language Testing and Assessment*, *4*(2), 18-33.

Elliott, S.N., Kratochwill, T.R., Littlefield Cook, J. & Travers, J. (2000). *Educational psychology: Effective teaching, effective learning (3rd ed.)*. McGraw-Hill College.

Friesner, T. and Hart M. (2005). Learning logs: Assessment or research method? *Electronic Journal of Business Research Methods 3*(2), 117-122.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208-238.

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Routledge.

Gareis, C.R., & Grant, L.W. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning* (2nd ed.). Routledge.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 229-244). Gunther Narr.

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In *Research Evidence in Education Library*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=116

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*(4), 228-250. https//doi.org/10.1016/j.asw.2012.06.003

Hill, K., & Ducasse, A.M. (2020). Advancing written feedback practice through a teacher-researcher collaboration in a university Spanish program. In M. Poehner & O. Inbar-Lourie (Eds.), *Toward a reconceptualization of second language classroom assessment* (pp. 153-172). Springer.

Honebein, P. C. (1996). Seven goals for the design of constructivist learning environments. In B. G. Wilson (Ed.) *Constructivist learning environments*: *Case studies in instructional design* (pp. 11-24). Educational Technology Publications.

Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. *System, 34*(4), 509-519. https://doi.org/10.1016/j.system.2006.09.001

Kim, A., & di Gennaro, K. (2012). Scoring behavior of native vs. non-native speaker raters of writing exams. *Language Research, 48*(2), 319-342.

Kim C. (2015). Rater training for scoring rubrics: Rater-centered bottom-up approach. *MinneTESOL Journal*, *31*(2), 1-14.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304. https://doi.org/10.1177/0265532208101008

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Kvasova, O. (2018). Climbing the assessment ladder: Spotlight on Ukrainian practices. Paper presented at the IATEFL TEASIG Conference, Luton, UK, 28–29 October.

Kvasova, O., Kavytska, T. & Osidak, V. (2019). Investigation of writing assessment literacy of Ukrainian university teachers. *Ars Linguodidacticae, 4*, 10–16.

Kvasova, O., Kavytska, T., Osidak, V. & Green, A. (2021). Development of a writing rating cale for university classroom in Ukraine. Paper presented at 17th EALTA Online Conference, 4-5 June. https://www.ealta.eu.org/conference/2021/SIG%202021%20CBLA/Kvasova%20et%20al_CBLA_SIG.pdf

Kvasova, O. & Shovkovy, V. (2020). Reliability of classroom-based assessment as perceived by university managers, teachers and students. In S. Hidri (Ed.), *Perspectives on language assessment literacy: Challenges for Improved student learning* (pp. 176-195). Routledge.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543– 560. https://doi.org/10.1177/0265532211406422

Lovorn, M. G., & Rezaei, A. R. (2011). Assessing the assessment: Rubrics training for pre-service and new in-service teachers. *Practical Assessment, Research & Evaluation, 16*, 1-18. https://doi.org/10.7275/sjt6-5k13

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

McLeod, S. A. (2019). Constructivism as a theory for teaching and learning. *Simply Psychology.* www.simplypsychology.org/constructivism.html

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–263. https://doi.org/10.1177/026553229801500204

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing 30*(3), 381–402. https://doi.org/10.1177/0265532213480337

Plakans, L. (2013). Writing scale development and use within a language program. *TESOL Journal*, *4*(1), 151-163. https://doi.org/10.1002/tesj.66

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309–27.

Schcolnik, M., Kol, S., & Abarbanel, J. (2006). Constructivism in theory and in practice. *English Teaching Forum, 4*, 12-20.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403-412. https://doi. org/10.1177/0265532213480338

Touchie, H. Y. (1986). Second language learning errors: Their types, causes, and treatment. *JALT Journal, 8*(1), 75-80.

Tsagari, D. (2017). The importance of contextualizing language assessment literacy. Paper presented at the 39th Language Testing Research Colloquium, Bogotá, Colombia, 17–21 July.

Tsagari, D. (2020). The conceptualisations of language assessment literacy (LAL). In S. Hidri (Ed.) *Perspectives on language assessment literacy: Challenges for improved student learning* (pp. 13-32). Routledge.

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment, 6*(1), 41-63.

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review, 56*(4), 555–580. https://doi.org/10.3138/cmlr.56.4.555

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*(1), 49-70. https://doi.org/10.2307/3588360

University College Dublin (n.d.). *Learning journals and logs.* https://www.ucd.ie/teaching/t4media/learning_journals_logs.pdf

Vopel, K.W. (2006). *Wirksame Workshops: 80 Bausteine für dynamisches Lernen*. Iskopress.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach.* Palgrave McMillan.

Weir, C.J., & Shaw, S.D. (2006). Defining the constructs underpinning Main Suite Writing Tests: A socio-cognitive perspective. *Research Notes, 26*, 9-14. https://www.cambridgeenglish.org/Images/23145-research-notes-26.pdf

Appendix A

Log entry 1

A brief record of your:

- discussing expectations of your ideal RS,
- the criteria you have looked at
- the criteria you have selected.

Reason your choice

Log entry 2

A brief record of your:

- choice of wording (Which do you prefer: detailed or shorter?),
- discussing options of wording and order of importance;
- feedback from your colleagues.

Reason your decisions.

Log entry 3:

A brief record of your:

- discussing the number of bands,
- discussing descriptors of performance on each band
- deciding on number of points per band.

Reason your decisions

Log entry 4:

A brief record of your designing the scale.
Write down the issues you faced and how you solved them.

Log entry 5:

A brief record:

- of any confusion/difficulty/ clash of opinions when learning to calculate a score using the RS? In which cases?
- what was the solution?

Reason your decisions.

Log entry 6:

A brief record of your:

- discussing the procedure of scoring (Is the procedure feasible, infeasible; practical, too effort consuming; easying the rating or confusing it?),
- discussing the scores (How close/far was agreement/disagreement? in which cases? reasons for variance?),
- commenting on the scores supplying reasons for each essay.

Reason your decisions

Log entry 7:

A brief record of your:

- discussing the process of rating (Does it easy the scoring or confuses it?),

- discussing the scores (How close was agreement/disagreement between the scores assigned?),
- evaluating the efficacy of the RS (Does the scale reflect all salient features of writing? Is it convenient to use? Do you think the scores are fair/unfair?).

Log entry 8:
A brief description of your:

- discussing the process of rating essays of other teams using your RS,
- evaluating the performance of your RS when assessing other essays,
- suggesting improvements to your RS based on the two rounds of rating.

Appendix B
Questionnaire

1 Compare the difference in your **rating based on rating scale** (RS) and the way you **used to rate before this study**. Tick (✓) ONCE in each category.

a) the quality of assessment
increased
lowered
no difference

b) rating procedure
less time consuming
more time consuming
no difference

c) the awarded scores
got higher
got lower
no difference

2** Tick (✓) as many times as appropriate. The **scale development required your drawing on:
teaching experience and intuition
linguistic expertise
knowledge of language pedagogy
knowledge of language testing
training materials and guidelines
colleagues' experience and expertise
additional reading
YOUR ANSWER:

3** Tick (✓) ONCE in each category. **The scale development:

a) was carried out without excessive effort, with normal ease
was quite demanding but feasible

b) flowed smoothly, in overall agreement within team
involved a lot of discussion within team

c) could be done single-handedly time permitting

needed collaborative effort
d) was accomplished promptly
e) involved reviewing, rewording, major revision

4 *Tick (✓) as many times as appropriate. Rating the* ***ten (own) scripts*** *involved:*
reading the scripts once or twice
re-reading scripts (more than twice)
multiple referring to the scale
some confusion while applying criterion …. (please indicate) and criterion …. (please indicate)
a lot of problems when deciding on scores (please indicate papers' # ….)
less time spent on scoring last scripts as compared to first ones
YOUR ANSWER:

5 *Tick (✓) as many times as appropriate. Rating the* ***twenty scripts (of other teams)*** *involved:*
a need to frequently re-read the test task set to the testees
reading the scripts once or twice
re-reading scripts (more than twice)
multiple referring to the scale
occasional rather than frequent referring to the scale
some confusion while applying criterion …. (please indicate) and criterion …. (please indicate)
a lot of problems when deciding on scores (please indicate papers' # ….)
less time spent on scoring last scripts as compared to first ones
YOUR ANSWER:

6 *Tick (✓) as many times as appropriate. Applying* ***OWN scale*** *revealed:*
need in considering the task requirements
need in aligning the scale with a particular curriculum
vagueness in wording of descriptors for criteria … (please indicate)
lack of focus on some actual features of written product
need in subtler specification of criteria such as … (please indicate)
need in additional criteria such as … (please indicate)
need in better discrimination between levels
need in different points assigned to criteria (1-point, 2-points, 3-points??)
need in changing the grading (e.g. 5-point grading →?)
need to discuss/compare the scores with colleagues
YOUR ANSWER:

7 *Tick (✓) as many times as appropriate.* **Advantages** of using the scale:
thoughtful/analytical approach to rating
less subjective rating
more accurate scoring
informed feedback to testees

informed implications for teaching writing
YOUR ANSWER:

8 *Tick (✓) as many times as appropriate.* Which of the statements are **true for you?**
I found the experience of scale design and use insightful and useful
It was difficult to use RS initially but became easier with practice
It was difficult to use RS at all times
I am happy with the quality of my RS
I think the RS may be applied for rating ANY type of text
I modified the RS after rating 10 own scripts
I have modified the RS after rating 20 scripts
I think I will modify the RS as soon as I can
I will use this (or other) RS in my further practice
I would use the RS time permitting
I would recommend other teachers to use RS
I have made sure that I can rate quite accurately without RS, intuitively
I am happy with the received training and guidance
I would like to undergo some more special training on assessing writing
YOUR ANSWER:

9 *Please share your other impressions of the RS design and use.*

10 *Please make your suggestions to improve rater designer training.*

View publication stats