OLGA KVASOVA (Ukraine),
TAMARA KAVYTSKA (Ukraine),
VIKTORIIA OSIDAK (Ukraine), and
ANTHONY GREEN (UK)

ORCID: 0000-0002-1479-0811
ORCID: 0000-0002-1528-9439
ORCID: 0000-0003-2058-5547
ORCID: 0000-0003-4893-1798

DEVELOPMENT OF A WRITING RATING SCALE FOR CULTURE- AND CONTEXT-SPECIFIC EDUCATIONAL SETTING IN UKRAINE

Abstract

In the Ukrainian university context, the assessment of writing in language programs has traditionally been performed by the individual teacher, with no institutionally mandated scales or criteria to rely on. The use of rating scales from external proficiency tests cannot meet the requirement of total alignment with the curricula taught in context-specific settings. This is obvious in terms of frequent mismatches of course objectives and test construct as well as task types employed on the external and local tests. To address he tissue, this paper reports on efforts to introduce a rating scale for collective use by teachers across a department teaching English to students majoring in Linguistics. Although teacher raters can be trained to employ external rating scales consistently, we believe that it is both more valuable in terms of professional development and more likely to result in acceptance of the scales if they engage in scale development. The evidence we collected suggests that the collaborative approach adopted in this study can offer a sustainable template for improving the poor standards of tests and examinations in our country (and elsewhere), which often result from a lack of testing and assessment expertise among those who prepare assessment materials.

Keywords: *writing scale, summative writing test, context-specific settings of education.*

Introduction

The internationalization of the professional and educational domains observed globally over recent years has dramatically raised the significance of good writing skills in L2 English in countries like Ukraine. As a result, both large-scale placement exams and University proficiency tests include writing tasks as indicators of students' abilities to express themselves in writing. Additionally, growing numbers of candidates put their writing ability to test in IELTS which is the most popular external test of English in the country. However, most English language learners at Ukrainian universities are still building their writing skills in L2 classroom settings. Hence the teacher's ability to conduct fair measurements of learners' skills in their day-to-day work plays a crucial role.

Using a rating scale is a pre-condition to accurate measuring candidates' skills (Hamp-Lyons, 1995; Knoch, 2009; Upshur & Turner, 1995; Weigle, 2002). A survey of Ukrainian practices of assessing writing in universities (Kvasova et.al, 2019) provided interesting though somewhat contradictory findings concerning teachers' employment of rating scales. 79% of respondents claimed they used rating scales in their assessment of writing, although 28% of them revealed misconception of the rating scale type (analytic or holistic) that they supposedly developed. 42% of the respondents who used rating scales stated that they most often developed the scales individually and 25% collaborated in scale design with other colleagues, whereas 39% and 28% of respondents resorted to the scales offered by the textbooks or examination systems, such as Cambridge English or IELTS, respectively. The researchers explained the reason for teachers' disregarding the scales offered in textbooks by almost total absence of such in the provided teacher's books. When it comes to the scales developed by testing experts, they may look arcane to grassroots teachers, although they could make attempts to adapt them to the local context.

However, we maintain that the reason that hinders the use of external scales in classroom-based assessment is much more serious: such scales cannot fully match the objectives of particular courses and include the specific task types envisaged in the curricula. Obviously, writing tests that teachers tend to administer are summative assessments which implies that they, in principle, should be aligned with the curriculum in a way that external tests cannot ensure. This practice fulfils the longstanding recommendation in the language testing literature that tests should be based on precise and detailed specifications of the needs of the learners for whom they are constructed (Heaton, 1991).

We cannot claim, however, that individual teachers could cope with adjusting external scales to specific curricula, since most activities in language testing are viewed as collegiate. Likewise, their European counterparts surveyed by Vogt and Tsagari (2014), Ukrainian teachers did not feel confident in applying ready-made tests (Kvasova & Kavytska, 2014). Currently, however, as Bolitho and West (2017) observed, in Ukrainian higher education, the preparation of assessment materials is solely the responsibility of each instructor at a time when «there are generally poor standards of tests and examinations». The researchers conclude that, «[t]here is a pressing need for English teachers to be trained in methods of assessment and testing» (Bolitho & West, 2017, p. 81).

When it comes to the assessment of writing skills, Ukrainian educators face several challenges. First, the assessment and testing of writing is a relatively new area of concern for teachers. Partly because of a traditional emphasis on receptive skills in foreign language classrooms, teachers lack concepts of standards in assessing writing. They do not find it easy to choose and apply ready-made rating scales (e.g., IELTS or FCE) effectively, let alone develop scales for themselves. Therefore, when assessing writing, most FL teachers opt to rely on intuition, experience, and their own perceptions of quality assessment, even if they are unable to articulate these for test score users.

Another challenge of writing assessment is rooted in the complexity of L2 writing as a process and as a linguistic and rhetorical skill. The assessment of this skill requires an appropriate level of L2 writing competence on the part of the assessors. As Crusan et al. (2016) state, in many contexts, teachers have not been taught to write effectively in L2 themselves. This may become a serious source of difficulty when it comes to the assessment of their students' writing performance. Moreover, since generic writing conventions vary across cultures, the raters might be influenced by their own culturally embedded expectations. These concerns gain additional significance if teachers not only assess writing, but also develop the rating scales.

Unsurprisingly, inconsistencies in the assessment of writing, as well as other gaps in the assessment literacy of university teachers were revealed in the survey referred to above (Author 1 et al., 2019). They are rooted in the absence of regulation by any common standards either in the national or local dimensions. Worth mentioning are such drawbacks in assessment practices as the focus on grammar and vocabulary with less attention given to organization and expression, the pursuit of 'penalizing-for-any-error' approach, providing none or late feedback to test-takers. The selection of criteria of assessment appeared to be totally at the teachers' discretion, too. As a result, most teachers admitted they needed substantial teacher training in language assessment.

These and other findings of the survey have led us to the assumption that engaging teachers in scale development may be an effective means of bringing them together to improve their assessment literacy while working within the Ukrainian assessment culture. Although teachers can be trained to employ external rating scales consistently, we believe that it is both more valuable in terms of professional development and more likely to result in acceptance of the scales if they engage in scale development. In this way, the new scales can encourage discussion and express the teachers' own understanding of what makes a piece of writing successful, rather than imposing the view of an external agency.

The current study, therefore, investigates the processes and outcomes involved in the development and use of rating scales by university teachers who teach L2 writing in a culturally- and educationally specific context: namely a classical university in Ukraine.

Review of literature

Assessing L2 writing came to the forefront of language testing research in the 1990s, with a range of seminal studies (Hamp-Lyons, 1990, 1995; Upshur & Turner, 1995; Weigle, 2002, 2007) that share the development of valid and reliable tools as a major objective to operationalize written performance in L2 with the view to arriving at meaningful scores that represent the test-takers' writing abilities.

As of today, most published research studies investigating the processes of rating scale development and validation have addressed large-scale high-stakes assessments (e.g., Banerjee et al. 2007; Hawkey & Barker, 2004;). Several recent studies give insights into a local or specific context of writing assessment, highlighting the significance of context-based rating scales (Ducasse & Hill, 2015; Kkese, 2018, Mendoza & Knoch, 2018).

Fast growing attention to classroom-based assessment has prompted research into the assessment of writing carried out by teachers (Cho, 2008; Crusan et al., 2016; Jeong, 2015; Mellati & Khademi, 2018; Skar & Jølle, 2017). However, as Becker (2018) states, «there is virtually no research that investigates the validity of rating scales used for summative, classroom-based writing assessments, which are prominent in most L2 writing courses […] and which are often used to make moderate to high-stakes decisions» (Becker, 2018, p.2). The lack of research into the processes of developing and validating rating scales for local classroom-based contexts will be addressed by the current study.

Basically, there are two *types of rating scales* used in performance assessment. Holistic scales prove more practical in large-scale testing or placement tests (Hamp-Lyons, 1995; Weigle, 2002). Analytic scales, by contrast, are thought to be more informative in terms of identifying areas of test-takers' strengths and weaknesses. From this perspective, the ability to provide helpful diagnostic input about testees' skills is referred to as 'the major merit of analytic schemes' (Gamaroff, 2000; Vaughan, 1991, as cited in Aryadoust, 2010). Besides, as Yamanishi et al. (2019) argue, analytic scales may be effectively used not only for assessment purposes but also for enhancement of learning and teaching of writing, which makes this type of scale more useful in the classroom assessment of writing and therefore rightfully attracts our attention.

A scale development process can follow two major *approaches* that are well described in literature – intuitive and empirical (Fulcher, 2003; Hamp-Lyons, 1995; Knoch, 2007, 2009; Weigle, 2007). North and Schneider (1998) interpret intuitive scales as developed «pragmatically by appeal to intuition, the local pedagogic culture and those scales to which the author had access» (1998, p.220). While empirically-informed approaches to scale construction have become the norm in the large-scale performance-based assessments, intuitively developed scales are known to perform well in low-stakes contexts where «a known group of assessors rate a familiar population of learners» (ibid., p. 220).

Irrespective of the approach taken, an effective scale primarily reflects the writing construct. Given that a rating scale is viewed as a *de facto* representation of the test construct (Knoch, 2011), the most essential role in the development of rating scales is played by the principled and justified choice of traits or *criteria*. Overviews of analytic frameworks (Aryadoust, 2010; Banerjee et al., 2007; Weigle, 2002;) indicate that these reflect the scale authors' conceptualization of writing ability. Respectively, scale designers purport to capture the most relevant aspects of test takers' writing (discoursal and linguistic) and articulate them as criteria.

The choice of criteria and the fashion in which they are worded, therefore, is essential to the design of valid rating scales. As Weigle (2002) points out, the scoring criteria need to provide a clear and credible basis for judgment, differentiating effectively between levels of writing performance. Turner and Upshur (2002) guard against inadequate ordering of criteria, which may inconsistently reflect SLA theory, irrelevance of criteria to tasks and content, incorrect grouping of criteria at different levels, as well as relativistic wording.

Although scales convey the scale developers' conception of the test construct, ultimately it is «the rater, not the scale, [that] lies at the centre of the process» (Lumley, 2002, p.267). The objective of the current research – the development of a local classroom-based rating scale for a university setting – prompts our special focus on teacher-raters. Hill and Ducasse (2020) and Plakans (2013) emphasize teachers' knowledge about the curricula in particular settings and the ongoing changes in educational contexts. Acknowledging teachers' competence, Hill & Ducasse argue that «teachers are often positioned as the recipients of 'expertise', as occurs in many professional development programs wherein teachers are informed of what the research community has determined to be best practice» (Hill & Ducasse, 2020, p.7). They further refer to Black et al. (2002) noting that classroom teachers, even those with little or no formal theoretical training, may effectively determine the appropriate methods and criteria for assessing their students. These considerations imply that practicing teachers may effectively act as rating scale designers as well as users, providing they follow effective procedures (Author 4, 2014).

Concluding the theoretical review, the authors will mention that the studies by Ducasse and Hill (2015), Mendoza & Knoch (2018) and Plakans (2013) referred to above had been commissioned by either education authorities or testing systems, therefore they received robust support from motivated institutions and policy makers. Such collaboration between policy makers and practitioners, regretfully, cannot be observed in many other contexts where teachers have to grapple with assessment challenges by themselves.

Research Context and Objectives

The study is based on the summative test practice pursued at a department responsible for teaching English to students majoring in Linguistics (Oriental language as the first FL and English as the second FL) at a University in Ukraine. The curriculum in General English (30 hours of classroom practice per term) is built around *Global: Upper Intermediate* (Clandfield et al., 2010). The teacher's book contains unit and progress tests which predominantly consist of selected and limited-production response tasks. Test tasks to assess productive skills are not provided, neither are any criteria or guidelines for the assessment of oral or written performance. Under the circumstances, in this setting, as well as in many others in Ukraine, summative tests, aimed at measuring students' achievements in all skills, have to be developed by the instructors themselves.

According to the official university standards for summative assessment, the end-of-term exam should consist of written and spoken tests, with a total of 40 points to be awarded. These points can be distributed by each department as they consider relevant to the curricula. In the department described, 10 points are assigned to the spoken test, and 30 to the written one. The staff agreed that the written test includes Grammar, Vocabulary and Writing parts, with each part assigned 10 points.

Under the circumstances, a team of volunteering teachers developed a local writing rating scale (Version 1), drawing on their prior expertise in language assessment, as well as teaching and research experience. The scale included four criteria; each criterion was scored on a five-band scale (from 0 to 4). Scores on the four criteria were added to give a total score for Writing of 10 points (see Appendix A for a copy of the scale).

Table 1

Criteria and Scores in the Initial Scale

Criteria	Max. score
Textual features (TF)	3
Coherence and Cohesion (CC),	3
Vocabulary and Register (VR)	2
Grammar (Gr.),	2
Total	10

After Version 1 had been introduced to the staff in the department, it was accepted without objection, consideration and/or discussion. That was probably due to a culture-specific tradition not to question instructions, which tend to arrive from managers in written form without detailed practical guidance. Moreover, the staff did not undergo any training in using the scale. It is fair to note that at that moment, no one of the staff members was qualified to deliver such training.

Version 1 was used during the summative assessment at the end-of-term test for year 2 students and further discussed at the department meeting. It was revealed that not all assessors made accurate use of the scale, if any at all. We attribute this primarily to time pressure since in the setting described teachers have to assess not only written test but oral part as well coping with both parts during the allotted 4-5 hours on one day. In such conditions, using the scale appeared too demanding for some teachers, others reported difficulty in understanding the criteria and awarding scores against them. Finally, the staff came to conclusion that training in using the scale would be needed if it was to be used effectively.

The current study was initiated as a response to this need. We set out to investigate the culture of writing assessment in this context and to develop a rating scale that would take this into account, promoting more ecologically valid outcomes. The *research questions* addressed in the study were:

– What strengths and weaknesses of Version 1 were identified by the teachers?

– What modifications did the teachers believe should be introduced to make writing scores more practical, reliable, and valid?

Not being commissioned by any national educational authority, the research was conducted within Erasmus + Staff mobility KA1 programme thus involving, in the majority, volunteering participants – university teachers.

Methodology

The study focuses on two iterations of scale development. Each involved 1) preparation of a version of the scale; 2) teacher-rater training; 3) application of the scale (scoring of a writing test); 4) evaluation of the scale by the teachers involved; and 5) analysis of results by the project team. The first iteration involved the design and construction of Version 1 (the first attempt to create a shared scale). The second involved construction of Version 2 (modified based on the feedback from the first iteration). Conceptually*,* the scale development approach developed over time as the project team learned more about the process.

Stage 1: Preparation

Training of prospective expert raters

During this phase, three university teachers of English with prior expertise in language assessment underwent intensive training in the assessment of writing at a university research centre, UK. One purpose of the visit was to launch a research study into rating practices in Ukrainian universities with a special focus on the use of rating scales by in-service teachers. The preparation stage at the research centre concluded with designing the research's main phase including its methodology, procedure, participants, and time frames. From now on the three teachers trained at the research centre will be referred to in this paper as expert raters (ERs).

Preparation for evaluation of Version 1

This stage included such activities, as:

a) reviewing Version 1 (criteria, bands, and scores) by the ERs who drew on the insights gained at the research centre; b) the ERs' rating of 10 students' papers, comparing, discussing the scores, and taking down the explanation for each score. The scores were nearly identical showing the ERs' agreement in most incidences; c) preparation of materials for rater training. These included: copies of Version 1, guidelines for using Version 1, grids for scoring; d) planning of the rater training sessions and evaluation of Version 1 in the real-life instructional situation.

Stage 2: Teacher-rater training

Participants

In line with the decision of the department meeting, the use of a rating scale should be preceded by teacher-rater training. The ERs recruited colleagues at the department who volunteered to undergo rater training and rate the required number of students' scripts. They were 10 English teachers working for the department. Six of them were PhD holders in TEFL, another two were PhD candidates. The teaching experience of the prospective raters ranged from 7 to over 20 years, all were female, non-native speakers of English. Assessment of writing was immediately related to their job duties and routines and was among the research interests of four of the ten participants.

Training sessions

Training session 1 consisted of a 60-minute plenary during which the ERs presented Version 1 with detailed explanation of each criterion, and the features that differentiated the bands. During the following 60 minutes of the session, the trainees' scored the 10 student papers that had been previously rated by the ERs. The scores awarded by the trainees were compared in group discussion and justified by trainees wherever necessary. The ERs also presented the scores that they had awarded and explained their judgements, using the prior written notes. In some cases, raters were persuaded to change their scores based on the ERs' scores and commentaries.

Training session 2 consisted in ERs' instructing the trainees on the operational procedures. Additionally, raters were asked to complete two questionnaires: Questionnaire 1 – *while* rating, and Questionnaire 2 – *after* rating.

During this session, the trainees requested that several papers should be scored together under the ERs' guidance; a simple verbal protocol procedure was used to elicit comments justifying the ratings in terms of each of the criteria. Two or three papers were rated both collaboratively in teams and individually during the session, which ensured better understanding of Version 1 and the scoring procedures.

Materials for training

Version 1, Guidelines for its use, sample written texts.

Stage 3: Evaluation of Version 1

Instruments

Test tasks

According to the official examination procedure, students should write one writing task. Each task currently exists in two variants. Both variants involved writing a letter of complaint based on similar short input texts. The texts were on-line advertisements for gadgets: Task 1 for a smartwatch, and Task 2 for smart sunglasses. Reading and writing advertisements, as well as writing letters of complaint, were curriculum requirements for year 2, so the writing tasks were based on content covered during the term.

The papers, selected at random from among the submitted test papers, were written by 100 second year students, all aged 17-18, sharing a mother tongue (Ukrainian), learning English for at least 7 years. All had attained the level B1+ on the External School-leaving test taken one year earlier. All papers were anonymized.

Questionnaires

Questionnaire 1 was to be filled out by raters *while* scoring *each* paper to elicit immediate, hands-on perceptions of rating each script. This questionnaire elicited their overall impression of using the scale by responding to three questions in relation to each script. The questions addressed the overall ease of rating the script using the scale ('easy', 'quite easy', 'quite difficult' or 'very difficult'), and whether using each of the four criteria had been convenient or difficult.

Questionnaire 2 was to be filled out by the raters *after* scoring *all* papers. It consisted of two parts: 1) five questions with Likert-scale response options, covering the scale's overall convenience, representation of aspects of writing ability in the scale, differentiation of criteria within bands, fairness of weighting (graining of criteria) and reliability and 2) five open-ended questions. These focused on problems that might have resulted from use of the scale. The respondents were also asked to come up with their own suggestions for improving the scale (see Appendix B).

A follow-up interview was to be conducted upon ERs' analysis of responses to both questionnaires and intended to elicit more specific information or additional comments from the raters.

Application of Version 1

Scoring of a writing test. 100 scripts of letter of complaint were selected at random from among the submitted test papers, were written by second year students, all aged 17-18, sharing a mother tongue (Ukrainian), learning English for at least 7 years. All had attained the level B1+ on the External School-leaving test given one year earlier. All papers were anonymized.

Stage 4: Pilot of Version 2

Based on the raters' feedback gathered in writing and orally in Stage 3, Version 1 underwent several modifications (these are described in detail in the results section) to create Version 2. That new scale was then piloted. The procedure involved the same participants, same scripts, and instruments (*Questionnaires 1 and 2, an oral interview and Rasch measurement).*

Results

Stage 5 (1): Evaluation of Version 1 of rating scale

Questionnaire 1. The values of perceived 'overall ease-difficulty of using the scale' (Fig.

1) demonstrate that overwhelmingly the raters found it easy to apply the scale.

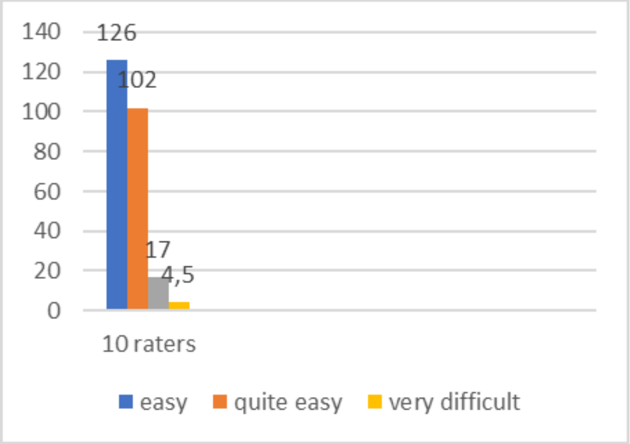

Fig. 1. Raters' Perceptions of Version 1 use

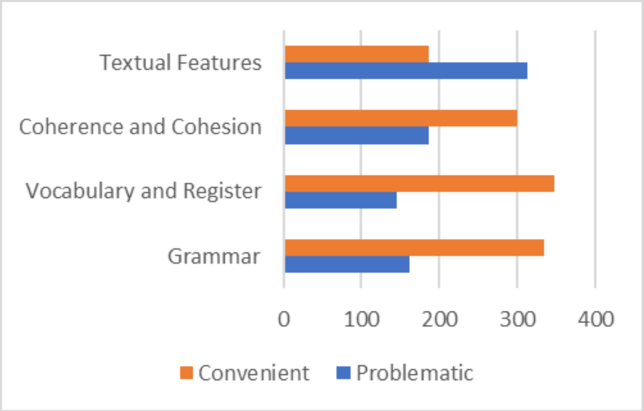

Fig. 2. Raters' Perceptions of Convenient/ Problematic Use of Criteria

Figure 2 indicates the raters' perceptions of the criteria as 'convenient/problematic to apply'. The most convenient criterion was *Vocabulary and Register* (VR), with its use perceived 2.4 times more convenient than problematic. Criterion *Grammar* (Gr) was second most convenient with the ratio 'convenient' : 'problematic' of 2.06. Criterion *Coherence and Cohesion* (CC) was ranked third with the index of convenience exceeding the value of being problematic by 1.67 times. It is in the only case, for criterion *Textual features* (TF), that the use of the criterion appeared more problematic than convenient to apply (by 1.7 times). The data suggest that three out of four criteria were rater-friendly. Nevertheless, we expected that responses to *Questionnaire 2* would reveal information to support or challenge the raters' hands-on impressions.

Questionnaire 2 was answered by seven out of ten raters. The first, Likert-type part of the questionnaire, yielded the responses that confirmed that the teachers found Version 1 usable, although they tended to give more *Tend to agree* than *Fully agree* responses.

In this respect, the perceptions elicited via *Questionnaire* 1 converge with those expressed by the raters in response to *Questionnaire 2.* This allowed us to assume that the opinions expressed by the raters the second time were better-considered and therefore more credible.

The comments that the raters provided by answering the open-ended questions were grouped around the major challenges faced during scoring.

Table 2

Raters' Perceptions of Version 1 Efficiency

		Fully agree	**Tend to agree**
1	The scale is rater-friendly and easy to use	2	5
2	The scale is comprehensive (it considers all relevant aspects of writing)	4	3
3	The criteria are appropriately grained		5
4	The weighting is fair	1	4
5	The scale seems a reliable tool to assess writing	2	5
	Total	9	26

The major challenge seems to have been the criterion TF. For example, one teacher commented, «*The Textual features criterion includes too many aspects like achieving a pragmatic purpose, relevant composition, appropriate register, and length. I didn't have a quite clear idea of what I was measuring in some of the students' writings, which might have influenced the objectivity of scoring*». As a result, four out of seven respondents made suggestions to split the criterion Textual features.

Other criteria that aroused misunderstanding and confusion were VR and Gr. For example, one respondent wrote, *'I had difficulty in differentiating between «wide» and «good» range of vocabulary. Similarly, problematic was defining the range of grammatical structures. Besides, which errors are considered crucial (hindering communication) and which are not?'*; '*I could not decide whether I had to give a lower score for the script with quite frequent grammar inaccuracies which did not hinder the achievement of its pragmatic purpose'.* Another respondent questioned '*the impact* of *range of vocabulary and grammar on the quality of task completion if the genre of writing is as highly conventionalized as is a letter of complaint.* '

The *follow-up interview* raised a few more issues that had not been apparent from the questionnaires. Raters were uncertain about how to grade the papers that had not been completed (e.g. due to inability to rewrite the final draft from the chore within the test time). Some raters awarded scores for incomplete papers, but, if completed, the papers could have deserved a high score. Similarly, awarding *any* score would not have a negative effect on the total score in the assessment. Such loyal raters' behaviour is truly culture-specific, reflecting the national tradition of teachers' understanding of learners' problems. However, an appropriate solution to the issue needs to be found.

The interview also revealed some of the reasons for the confusion experienced by the raters. The first one revealed the inability of the raters to award scores against the criteria that were weighted differently: for instance, in Version 1 the maximum score for criteria TF and CC was 3, whereas in the case of VR and Gr the maximum number of points was 2. The interviewees mentioned that the lack of granularity in the two latter criteria hindered efficient scoring. A more essential reason for confusion was difficulty in determining the level and degree of coherence and cohesion in the script. Additionally, lack of knowledge about the relevant features of a letter of complaint as a text genre was one more source of misconception of another criterion – TF. These facts once again confirmed the necessity of regular staff training and potential areas for professional development.

Rasch analysis. On Version 1, Task 1, the average scores awarded by the raters ranged from 5.20 to 7.98. The overall average score was 6.458 out of 10, with a standard deviation across raters of 0.925. There was 32.4% exact agreement between the raters on the scores they awarded across all criteria. On Task 2, scores ranged from an average of 5.24 awarded by the harshest to 7.76 by the most lenient rater. The average score was 6.743 out of 10, with a standard deviation across raters of 0.774. There was 36.6 % exact agreement between raters on the scores awarded across the criteria.

Rasch analysis revealed that none of the raters was misfitting (outfit mean square values greater than 1.3) on Task 1, but that two raters were misfitting on Task 2 (with outfit mean square values of 1.61 and 1.63).

Following the data obtained in Stage 3, the ERs introduced a number of modifications to Version 1 which led to Version 2.

Stage 5 (2): Creation of Version 2

1) Split of criterion 'Textual Features'

'Textual Features' appeared to be the most demanding for the raters to handle as it included too many features. A decision was therefore made to split this criterion in two: '*Task achievement'* and *'Genre conventions'* as was suggested by four respondents. The criterion *'Task achievement'* primarily was intended to indicate that the written text was completed, and the purpose of writing was achieved in a presentable way. If not, the script would be awarded 0 points and not considered for further rating. More importantly, this criterion accounted for the script's covering the necessary content points and the degree of their elaboration – from full and detailed coverage to merely mentioning some of the points or not mentioning them at all.

The criterion *'Genre conventions'* was intended to account for a script's meeting a certain number of genre-specific requirements. Hence, the raters were to check that a script's composition included the standard elements expected of texts of this genre, that rhetorical functions of each text component were fulfilled, that the register and tone of writing were culture-relevant, and also that the text length and layout were appropriate. Clearly, the application of this criterion was to be underpinned by detailed specification of text features typical of the genre concerned.

Specification of criterion 'Coherence and cohesion'

In Version 1, this criterion included checking for full/sufficient/insufficient coverage of content points in the writing along with other features, such as logical presentation of content points and appropriate use of cohesive devices typical of the text genre. In Version 2, the decisions about coverage of content points were rightfully transferred to the newly introduced criterion 'Task achievement'. Moreover, the descriptors, such as 'fully coherent text' or 'mostly coherent text' were specified in greater detail by offering differentiation of text, paragraph, and sentence level cohesion relevant to each band. Appropriate *paragraphing* of the text remained attributed to CC features whereas *punctuation* was found optional at this level of English proficiency.

Specification of descriptors for 'Vocabulary' and 'Grammar'

Rating against these criteria within Version 1 had caused confusion for the raters who found terms such as 'wide / good' range, as well as 'errors hindering/not hindering communication' ambiguous and allowing for subjectivity of interpretation. In order to reduce the effects of such subjective conceptions, it was decided to use more explicit terms and to standardize their wording wherever possible. Following this line of thought, the top bands for 'Vocabulary' and 'Grammar' were characterized as '*good range*', the next lower band – *'appropriate range'* whereas the two bottom descriptors contained the determining adjective *'limited'*. Additionally, the descriptors in each band were made more specific, for instance, *'good range'* with respect to Vocabulary was elaborated as, *'no inconsistencies in register; occasional inaccuracies in use of collocations'*, whereas *'good range of structures'* in respect to *Grammar* was clarified by *'used with a few minor inaccuracies in use of articles and prepositions. Errors occur as slips'.*

Equal weighting across all criteria

Splitting the criterion *'Textual Features'* in two entailed increasing the number of criteria from four to five. Keeping in mind that the raters were inclined towards equal weighting across all criteria, we faced a dilemma: whether to assign each criterion a maximum of 2 points, thus preserving the overall weighting of the scale at 10 points (which was initially determined for the reason of practicality of scoring the summative written test), or to assign each criterion a maximum of 3 points thus increasing the overall weighting from 10 to 15 points (which would ensue changes in scoring in other parts of the summative test). The dilemma was resolved by ERs' trialing both versions of the scale, which led to understanding that a finer grained, 15-point, band was better suited for accurate scoring which should not be compromised in the case of summative assessment.

As a result of modification, the new scale included five criteria: 'Task achievement', 'Genre conventions', 'Coherence and cohesion', 'Vocabulary', and 'Grammar, with each criterion having equal weighting (3 points) and the total score amounting to 15 points (see Appendix C). This scale contained more clearly articulated descriptors which were intended to be more granular.

During *Stage 4* Version 2 was piloted by the 10 raters who had also participated in the first pilot. Figure 3 presents mean values of the perceived 'overall easiness-difficulty' of Version 2 use of all raters. The bars demonstrate lower variance in values of 'easiness-difficulty' than in Figure 1. This might testify to raters' increased awareness of the criteria, their more reasonable perceptions that had been developed in the process of two rounds of rating.

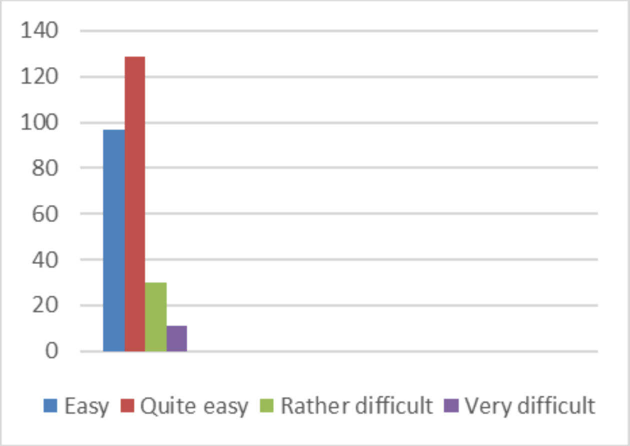

Fig. 3. Raters' Perceptions of Version 2 Use

This inference is supported by the evidence collected via *Questionnaire 2* from all 10 raters who participated in piloting of Version 2. The respondents expressed their satisfaction with all the modifications and noted that the rating scale became more comprehensible, and the scoring procedure became easier and more convenient. In the *follow-up interview,* seven raters maintained that while using Version 2 they grew more observant to all the rated aspects, which led to their awarding overall higher scores than during the evaluative pilot of Version 1. The interviewees also emphasized the need in conducting regular rater training to ensure sustained accuracy of scoring.

Rasch analysis. On Version 2, Task 1, scores awarded by the raters across criteria ranged from an average of 5.31 (out of 10) awarded by the harshest rater to 8.00 by the most lenient. The overall average across raters was 6.772 with a standard deviation of 0.846. There was 49.6% exact agreement between raters on the scores awarded across test takers and criteria. On Task 2, raters' average scores ranged from 4.35 to 8.14 with an overall average score of 6.617 out of 10. The standard deviation across raters was 0.997 points. There was 52.0 % exact agreement between raters on the scores awarded across test takers and criteria.

Again, Rasch analysis revealed that none of the raters was misfitting (outfit mean square values greater than 1.3) on Task 1, but that two raters were misfitting on Task 2 (with outfit mean square values of 1.35 and 1.61).

The results of the scale use and modification indicate that Version 2 produced higher levels of agreement between the raters on the scores awarded (50.8% across the two tasks on Version 2, compared with 34.5% on Version 1), suggesting more consistent interpretation of the scale categories even though there were now more to select from.

Conclusion

The study revealed the strengths and weaknesses of an intuitive analytic scale developed by teacher volunteers, and further applied in the summative assessment by the department staff without prior training in the use of the scale. Below are the implications that arose while modifying Version 1 and creating Version 2.

While both scales described in our study appeared to work for rating purposes, the Rasch measurement indicated that the agreement among raters on Version 2 was substantially higher than that of Version 1. We attribute this, among other factors, to a better specification of the descriptors in the modified scale. This claim is in line with preferences for more detailed scales expressed by professionally trained raters (Knoch, 2009), or a need in clear and complete understanding of «what assessment criteria really mean to the raters» (Lumley, 2002). In the case of teacher-raters' piloting of Version 2, as we exemplified above, the descriptors were made more concrete and specific in terms of reflecting curriculum requirements, as well as standardized in formulation so as to evade possible misconceptions in distinguishing confusable aspects of writing.

The current study revealed some misconceptions of the criteria and what they mean to teacher-raters. One of them regards, for instance, teachers' inability to interpret the criterion coherence and cohesion, although this has also proved to be a problematic criterion in other studies (e.g. Knoch 2007). Another source of difficulty reported by the teacher-raters was scoring a particular text type on 'textual features' (in Version 1). This evidence resonates with Mai's (2019) claim about textual features still remaining one of the main concerns of scale developers.

In our case, the reason for this may be twofold. First, teachers may fail to be clear of the genre conventions due to almost totally absent patterns of writing letters of complaint in the Ukrainian social culture. This has implications for the curriculum, raising questions about whether and how such letters should be taught. Second, rater training might usefully include (re)familiarization with and discussion of the conventions associated with the text types used on the test. Thus, the need for regular training in rating scale use that is explicitly stated in the recommendations by Author 4, 2014; Crusan et al, 2016; Jeong, 2015; Kkese, 2018; Lim, 2011; Plakans, 2013, and Skar and Jølle, 2017 gets yet another confirmation.

The study also provides evidence of effective involvement of practicing teachers in local scale' design. However, while existing research into rater training and its effects on professional raters is considerable, accurate descriptions of training for practicing teachers in the use, and particularly in the development and revision of rating scales is scarce. One step in bridging this gap may involve a detailed examination of teacher-raters' perceptions of scoring procedures, including the impact of rater training on their teaching and assessing practices as well as their teaching and assessment philosophy.

References

Aryadoust, V. (2010). Investigating Writing Sub-skills in Testing English as a Foreign Language: A Structural Equation Modeling Study. *TESL-EJ*, *13*(4). https://www.tesl-

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels [IELTS Research Report No. 7, the British Council/ University of Cambridge Local Examinations Syndicate]. https://www.ielts.org/research/research-reports/volume-07-report-5

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing*, *37*, *1–12*. https://doi.org/10.1016/j.asw.2018.01.001

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8–21. https://doi.org/10.1177/003172170408600105

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

Cho, D. (2008). Investigating EFL writing assessment in a classroom setting: Features of composition and rater behaviors. *The Journal of Asia TEFL*, *5*(4), 49–84. https://www.earticle.net/Article/A182195

Clandfield et al., (2010). *Global Upper Intermediate Student's Book*. Macmillan ELT.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, *28,* 43–56*.* https://doi.org/10.1016/j.asw.2016.03.001

Ducasse, A. M., & Hill, K. (2015). Development of a Spanish generic writing skills scale for the Colombian graduate skills assessment (SaberPro). *Papers in Language Testing and Assessment*, *4*(2), 18–33. https://arts.unimelb.edu.au/__data/assets/pdf_file/0009/1770606/2.Ducasse-And-Hill.pdf

Fulcher, G. (2003). *Testing second language speaking*. Pearson Longman.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System*, *28*(1), 31–53. https://doi.org/10.1016/S0346-251X(99)00059-7

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge. https://doi.org/10.4324/9781315889627

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing assessment issues and options*. Macmillan. https://doi.org/10.1017/CBO9781139524551.009

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, *29*, 759-762.Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159. https://doi.org/10.2307/3588173

Heaton, J.B. (1991). *Writing English language tests*. Longman.

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. Educational Linguistics* (vol. 41). Springer, Cham. https://doi.org/10.1007/978-3-030-35081-9_8

Jeong, H. (2015). Rubrics in the classroom: do teachers really follow them? *Language Testing in Asia*, *5*(6*)*. https://doi.org/10.1186/s40468-015-0013-5

Kkese, E. T. (2018). Assessing L2 Writing in the Absence of Scoring Procedures: Construction of Rating Scales in a Cypriot Greek EFL in Class Context. *Journal of English Education*, *3*(2), 46–56. https://doi.org/10.31327/jee.v3i2.809

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12,* 108–128. https://doi.org/10.1016/j.asw.2007.07.002

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,* 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Kvasova, O. & Kavytska, T. (2014). The assessment competence of university foreign language teachers: A Ukrainian perspective. *CerleS*, *4*(1), 159–177. https://doi.org/10.1515/cercles-2014-0010

Kvasova, O., Kavytska, T. & Osidak, V. (2019). Investigation of writing assessment literacy of Ukrainian University teachers. *Ars linguodidacticae*, *4*(2), 10–19. https://doi.org/10.17721/2663-0303.2019.4.02

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*, 543–560. https://doi.org/10.1177/0265532211406422

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

Mai, D. (2019). A review of theories and research into second language writing assessment criteria. V*NU Journal of Foreign Studies*, *35*(3). https://doi.org/10.25073/2525-2445/vnufs.4371

Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education*, *43*(6). http://dx.doi.org/10.14221/ajte.2018v43n6.1.

Mendoza, A., & Knoch, U. (2017). Examining the validity of an analytic rating scale for a Spanish test for academic purposes. *Assessing writing*, *35*, 41–55. https://doi.org/10.1016/j.asw.2017.12.003

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

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

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

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

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–126). Norwood, NJ: Ablex.

Vogt, K. & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, *11*(4), 374–402. http://dx.doi.org/10.1080/15434303.2014.960046

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Weigle, S.C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), *194–209.* https://doi.org/10.1016/j.jslw.2007.07.004

Yamanishi, H., Ono, M., & Hijikata, Y. (2019). Developing a scoring rubric for L2 summary writing: A hybrid approach combining analytic and holistic assessment. *Language Testing in Asia*, *9*(13). https://doi.org/10.1186/s40468-019-0087-6

Appendices

APPENDIX A

Version 1 of Rating Scale

Marks	Textual features max. 3 marks	Coherence & cohesion max. 3 marks	Vocabulary & register max. 2 marks	Grammar max. 2 marks
10	Meets all text types requirements	Fully coherent text; cohesive on sentence and paragraph level	Wide range of vocabulary, correct choice of words in compliance with register	Wide range of structures relevant to textual features, few minor inaccuracies
9				
8	Meets major text types requirements	Coherent text; appropriate sentence and paragraph-level cohesion	Good range of vocabulary with few cases of wrong choice of words; few inconsistencies in register	Good range of structures relevant to textual features, some inaccuracies that do not hinder communication
7				
6	Frequent inconsistencies in meeting text type requirements	Sentence-level cohesion noticeable, lack of paragraph-level cohesion	Limited range of vocabulary with frequent cases of wrong choice of words; frequent inconsistencies in register	Limited range of structures, frequent inaccuracies that hinder communication
5				
0 – 4	Does not meet text type requirements	Text not coherent	No range of vocabulary, wrong choice of words, no register requirements met	No range of structure, mostly inaccurate

APPENDIX B

Questionnaire 1 (while rating)

Please tick as appropriate commenting on rating EACH PARTICULAR PAPER

(See example and abbreviations)

Example	It was *overall* ☐ easy ☒ quite easy ☐ quite difficult ☐ difficult to rate	It was *quite convenient* to rate ☐Textual features (TF) ☒ Coherence-cohesion (CC) ☐Vocabulary & register (VR) ☒ Grammar (Gr)	It was *rather problematic* to rate ☒ Textual features (TF) ☐Coherence-cohesion (CC) ☒ Vocabulary & register (VR) ☐ Grammar (Gr)
Paper #	It was *overall … to use the scale*	It was *quite convenient* to rate …	It was *rather problematic* to rate …
VI. **01**	☐ easy ☐ quite easy ☐ quite difficult ☐ very difficult	☐ TF ☐ CC ☐ VR ☐ Gr	☐ TF ☐ CC ☐ VR ☐ Gr

Questionnaire 2 (after rating)

Please tick as appropriate:

#	Statement	*Strongly agree*	*Tend to agree*	*Tend to disagree*	*Strongly disagree*
1	The scale is rater-friendly and easy to use				
2	The scale is comprehensive (it considers all relevant aspects of writing)				

3	The criteria are appropriately grained
4	The weighting is fair
5	The scale seems a reliable tool to assess writing

Please give your comments on the scale:

The scale misses some important aspects of writing, such as … .

I suggest introducing such criteria as … .

Criteria … (please give names of criteria) should be grouped together.

Criteria … (please give names of criteria) should be split.

Criteria … (please give names of criteria) should be weighted differently: for instance, … .

3. Criteria … (please give names of criteria) are formulated in a confusing way.

I suggest reformulating them as … .

4. The biggest problem that I faced while rating was … .

I suggest the following: … .

5. My overall evaluation of the scale is … .

APPENDIX C

Version 2 of Rating Scale

Marks	Task achievement max. 3 marks	Genre conventions max. 3 marks	Coherence & cohesion max. 3 marks	Vocabulary max. 3 marks	Grammar max. 3 marks
15	All content points fully covered and elaborated; within the range of required length	Text fully complies with genre conventions	Fully coherent text; cohesive on sentence and paragraph level	Good range of vocabulary correct choice of words with no inconsistencies in register; occasional inaccuracies in use of collocations.	Good range of structures used with a few minor inaccuracies (article, prepositions). Errors occur as slips.
13-14					
12,5	Most content points are covered; required length inconsiderably violated	Text complies with major genre conventions	Mostly coherent text; appropriate sentence and paragraph-level cohesion	Appropriate range of vocabulary of general usage with a few cases of wrong choice of words and inconsistencies in register	Appropriate range of structures, some inaccuracies in verb tense forms, conditionals, modals. A few spelling errors.
11-12					
10	Some of content points covered or mentioned	The text violates many genre conventions	Sentence-level cohesion noticeable, some cases of inappropriate paragraph-level cohesion	Limited range of vocabulary with frequent cases of wrong choice of words; some inconsistencies in register	Limited range of structures, frequent inaccuracies and spelling errors
8-9					
7,5	Most content points are not covered	Text violates most genre convention	Text mostly incoherent: lack of paragraph-level cohesion	Limited range of vocabulary with frequent cases of wrong choice of words and inconsistencies in register that hinder understanding	Limited range of structures, frequent inaccuracies and spelling errors that hinder understanding
6-7					
0-7	Task is incomplete	Text does not meet genre requirements	Text not coherent	No range of vocabulary, wrong choice of words, no register requirements met	No range of structure, mostly inaccurate

СТВОРЕННЯ ШКАЛИ ОЦІНЮВАННЯ УМІНЬ ПИСЬМА З УРАХУВАННЯМ СПЕЦИФІКИ ТА КОНТЕКСТУ УМОВ ОСВІТИ УКРАЇНИ

Ольга Квасова (Україна), Тамара Кавицька(Україна),
Вікторія Осідак(Україна), Ентоні Грін (Велика Британія)

Анотація

Постановка проблеми: *У контексті української університетської освіти оцінка писемного мовлення на мовних програмах традиційно з дійснюється окремим викладачем; при цьому відсутні інституційно встановлені шкали або критерії, на які можна було б покластися і використовувати їх колегіально у процесі оцінювання письма. Використання шкал оцінювання із з овнішніх тестів з азвичай не може з адовольнити вимогу повної відповідності навчальним програмам, які реалізуються в конкретних умовах. Основна причина цього полягає у невідповідності цілей курсу та структури тесту, а також типів з авдань, що використовуються у зовнішніх та інституційних тестах. Водночас створення та з астосування власних шкал нерідко має несистемний характер через часозатратність процесу та брак з нань викладачів про етапи повного циклу створення шкали оцінювання.*

Метою *статті є дослідження процесу створення, впровадження та вдосконалення шкали оцінювання вмінь письма студентів мовних спеціальностей для колективного використання викладачами однієї кафедри (дослідницької групи, команди). Хоча викладачі можуть надавати перевагу використанню з овнішніх шкал оцінювання, автори вважають, що участь викладачів у розробленні шкал є більш цінним досвідом з точки з ору їхнього професійного розвитку.*

Методологія. *Дослідження має емпіричний характер і здійснюється на таких етапах створення та порівняння двох версій шкали оцінювання умінь письма: 1) створення шкали; 2) інструктування команди викладачів; 3) з астосування шкали (оцінювання письмових робіт); 4) оцінювання ефективності шкали викладачами; 5) аналіз результатів командою. Учасниками дослідження були 8 викладачів англійської мови, що добровільно взяли участь у процесі оцінювання письмових робіт студентів 2-го курсу, що вивчають англійську як другу іноземну мову .*

Результати. *Результати дослідження з асвідчили вищий ступінь узгодженості оцінок у версії 2 шкали. Зазначена версія була модифікована після критичних з ауважень щодо труднощів оцінювання робіт з а критеріями «когерентність/ когезія» та «текстові характеристики». Зібрані дані дозволяють припустити, що підхід, з апропонований у цьому дослідженні, може стати з разком для покращення не з авжди високих стандартів тестів та іспитів у нашій країні, які часто є результатом відсутності досвіду тестування та оцінювання у викладачів.*

Ключові слова: *шкала оцінювання письма, тест підсумкового оцінювання, специфіка та контекст процесу освіти .*

BIOS

Olga Kvasova, PhD, Associate Professor at the Department for teaching Ukrainian and foreign languages and literatures. Taras Shevchenko national University of Kyiv. Her research interests include teaching academic English, language testing and assessment, curricula and material design.

Email: olga.kvasova.1610@gmail.com

Tamara Kavytska, PhD, Associate Professor, Department of Teaching Methodology of Ukrainian and Foreign Languages and Literatures, Institute of Philology, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Her research interests lie within the areas of Translation Pedagogy, Cognitive and Rhetorical Grammar, Language Testing and Assessment.

Email: kawicka_t@ukr.net

Viktoriya Osidak, PhD, Associate Professor, Department of Teaching Methodology of Ukrainian and Foreign Languages and Literatures, Institute of Philology, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Her areas of research interests are self-assessment in FLT, autonomous learning, learning strategies.

Email: viktoriya_osidak@ukr.net.

Anthony Green, Professor of Language Assessment, director of Centre for Research in English Language Learning and Assessment, University of Bedfordshire, UK. His areas of research interests include various aspects of language assessment, test design, item writing as well as teaching and materials writing.

Email: tony.green@beds.ac.uk