Original Research

Issues of practicality and impact in checklist-based scoring of writing

by Zoltán Lukácsi and Borbála Fűköh

Zoltán Lukácsi Euroexam International, Hungary zoltan.lukacsi@euroexam.org
Borbála Fűköh Euroexam International, Hungary fukoh.borbala@euroexam.org
Article history Received October 13, 2022 | Revised January 29, 2023 | Accepted March 6, 2023
Conflicts of interest The authors declared no conflicts of interest
Research funding No funding was reported for this research
doi 10.22363/2521-442X-2023-7-1-9-20
For citation Lukácsi, Z., & Fűköh, B. (2023). Issues of practicality and impact in checklist-based scoring of writing. *Training, Language and Culture, 7*(1), 9-20.

In 2016, Euroexam International launched an extensive research campaign with the aims of increasing transparency and accountability while preserving our professional values and achievements. Reviewing and renewing our scoring methodology was uppermost on the agenda, especially as the diverse potential in checklist-based scoring was particularly appealing in our context. Research and development had to conform to the administrative frameworks including time and financial constraints, as well as recognise stakeholder needs. In the present study, we report on the issues we had to tackle when introducing checklist-based scoring into practice. The participants of this mixed-methods research were 12 examiners, six EFL teachers and 28 language learners in four study groups. We used a total of 600 scripts by stratified random sampling from live examinations at levels B2 and C1. In the quantitative study, two division schemes were trialled to decide how to share the workload between examiner pairs. In the qualitative inquiry, teachers and language learners were invited to share their views on checklist-based feedback for formative purposes. With mutually beneficial compromises that enabled the successful introduction of our checklists for writing, the most important implication of this study is that evolving stakeholder needs can be fulfilled with adequate flexibility.

KEYWORDS: *practicality, checklist-based scoring, writing, scoring methodology, practical implementation, English as a foreign language*

1. INTRODUCTION

Qualities of useful tests include reliability, construct validity, authenticity, interactivity, impact, and practicality (Bachman & Palmer, 1996) as reflecting the priorities of the context of application. While there is ample research into the theory and operationalisation of construct validity and reliability, there seems to be a relative lack of interest in what it entails to introduce sound research results manifested as language test instruments into practice. This study reports on issues of practicality when introducing checklist-based scoring of writing in an international language testing system. Our main aim with this discussion is to throw light on the diversity of the obstacles to be tackled when implementing the findings of seemingly completed applied linguistic research.

doi: 10.22363/2521-442X-2023-7-1-9-20
Volume 7 Issue 1, 2023, pp. 9-20
https://rudn.tlcjournal.org

2. THEORETICAL BACKGROUND

2.1. Practicality in language testing

Practicality is loosely equated with a response to resource requirements in the literature. Bachman and Palmer (1996, p. 35) claim that *'if the resources required for implementing the test exceed the resources available, the test will be impractical'*. The resources could be human, material, time, or the all-pervading financial. Taking a broader and consequently less precise approach, Brown and Abeywickrama (2019, p. 28) define practicality as *'the logistical, down-to-earth, administrative issues involved in making, giving, and scoring an assessment instrument'*. Mousavi (2009) emphasises the costs associated with each step of language test construction and administration. Similarly, Fulcher (2010) zooms in on budgetary issues in the two questions he proposes to consider under practicality.

Viewed from the vantage point of who is affected, resource-related issues can be allocated to the different stakeholders involved. If reading were to be tested as humans read outside of an exam environment, Rouet (2012) argues, the measurement construct could resemble authentic behaviour, but the time and effort required would repel candidates. By contrast, practicality can generate resistant interest despite fundamental conceptual flaws (Grishechko et al., 2015). Hambleton and Pitoniak (2006) point out that the Angoff-Tucker method of standard setting is erroneous, but its cheap and simple administration along with the speed of completion makes it popular among panellists. Similarly, measurement professionals use software on the basis of availability even though the underlying psychometric modelling is invalid (Verhelst, 2004).

Rater involvement is a prerequisite for the quality of practicality in scoring (McNamara et al., 2002). In the development of any method of scoring, East (2009) recommends a discussion-based approach where scores are discussed as evidence of salient features in the script. Harsch and Martin (2012) report on a scale revision where raters were involved in the trial and revision of the instrument including analysis and modification of wording and clarification, as well as assessment. Harsch and Rupp (2011) highlight the importance of gaining insight into what aspect of a descriptor examiners focus on when scoring via a pass or fail judgement. Apart from content, rater involvement is necessary to ensure that the sheer volume of the decisions to be made in assessment does not overburden examiners' working memory (Fulcher et al., 2011) or provoke fatigue (Ling et al., 2014).

Assessment literacy is vital when communicating scores to stakeholders (Zenisky & Hambleton, 2012). Ryan (2006) identifies the following eight characteristics for score reports to include: audience, scale or metric for reporting, reference for interpretation, assessment unit, reporting unit, error of measurement, mode of presentation, and reporting medium. However, Vogt and Tsagari (2014, p. 383) point out that the majority of the test preparation teachers they sampled (60.9%) received no training in 'using statistics'. Therefore, much of the literature recommends using clear and concise language and avoiding jargon and highly technical terms or statistics (Goodman & Hambleton, 2004; Hambleton & Meara, 2000; Wainer et al., 1999; Zhuk & Ivanov, 2022). Even though summative information takes priority in high-stakes proficiency testing, often in the form of pass or fail, Zenisky et al. (2009) emphasise the importance of tailored reports to meet the needs of different stakeholder groups, particularly as these tests contain diagnostic information. One example Zenisky and Hambleton (2012) give in this regard is identifying areas for development with subscores so that the candidate understands subject weaknesses.

2.2. Impact on application

Impact is the effect a particular test has on the individual (micro level) or as a social group (macro level) (Bachman & Palmer, 1996). Impact, in the sense of 'test effect', is also referred to as 'washback' in the literature (Weigle, 2002). It also works at two levels. At the macro level of society, decisions based on test scores have a washback effect on education systems, whereas at the micro level of the individual, washback affects learners and teachers as to how assessment instruments affect practices and beliefs (Cohen, 1994).

Although the notion of washback is widely used in the context of test development and language testing, washback intentions of test designers are difficult to study and are *'often based on assumptions rather than on empirical evidence'* (Taylor, 2005, p. 154). Education environments are very complex, so it is not easy to find the reasons for a change in classroom practices, but there is empirical evidence that supports that based on test results, students can identify their strengths and weaknesses (Brown & Abeywickrama, 2019), whereas teachers may set specific goals for students, as their perceptions are usually influenced by examinations and students' test results (Cheng, 2004). A major implication for large-scale high-stakes testing concerns results reporting and examiner training. Examiner training has

been proven to enhance consistency and reliability and reduce the impact of specific and unwanted examiner behaviour (Eckes, 2008; Van Moere, 2013; Weigle, 1998). Increasing the objectivity of the scoring procedure and making the most of the compulsory training and re-training hours is in the interest of exam providers in general (Weigle, 1994). In addition to this, involving the examiners and asking for their opinion in connection with the scoring process is key (Inoue et al., 2021; Rossi & Brunfaut, 2020). Hamp-Lyons (2007) points out that examiners need to feel included and valued, which also means that their contribution and their being part of the process must be appreciated.

It is important to bear in mind that in the case of subjectively scored tasks, examiners have special added value, as it is the examiner who *'lies at the centre of the process'* (Lumley, 2002, p. 267).

2.3 Checklist-based scoring

The CEFR (Council of Europe, 2001, p. 189) defines rating on a checklist as *'judging a person in relation to a list of points deemed to be relevant for a particular level or module'*. The checklist as a measurement instrument is particularly suited for level testing, when the ultimate aim of the assessment is to decide how much of the target learning goals have been embraced by the language learner. There is little theoretical guidance and literally no practical assistance in the CEFR as to how to construct a checklist beyond stating that (a) it may look like a questionnaire; (b) it may be a set of yes or no questions; and (c) it may take the form of a series of steps. The very term 'rating on a checklist' is something of a misnomer itself, for in the case of dichotomous items, the unit of measurement leaves no room for levels of completion or mastery, even if the cumulative result can be meticulously detailed depending on the number of items on the list. It is also important to point out that the presence of checks or ticks does not suffice to identify a rating system as a checklist. Some of the successful research projects to develop a checklist include Kim (2010, 2011) and Struthers et al. (2013).

3. CONTEXT

Institution appears in the context of the international language testing traditions, i.e., it offers language exams which are levels tests and are linked to the CEFR (Council of Europe, 2001). Being embedded in this tradition also means that Institution followed the trend of using scale-based tools for assessing writing. The shortcoming of rating scales is a prevailing topic in assessment literature, and we may find various suggestions to counteract these (Bachman & Palmer, 1996; Eckes, 2008; Hamp-Lyons, 1990; Harsch & Martin, 2012; Lukácsi, 2021; Wiggelsworth, 1993).

Earlier research conducted locally (Fűköh, 2020) unveiled that there is no consensus among Institution writing examiners on the interpretation of descriptors, and that the concepts of the construct need to be clarified in order to reduce the impact of specific and unwanted examiner behaviour. It became clear that no matter how experienced the examiners are, they have diverse and contrasting perceptions of the task and the rating scale. Moreover, their understandings are often contradictory and inconsistent. In response to this challenge, we launched a research project aiming at developing a level specific checklist-based scoring tool for the assessment of writing skills. The research and development project observed the following Institution values: consistency, flexibility, fairness, objectivity, and finally transparency. We believe that the checklist-based tool helps the Institution meet the challenges and keep their values at the same time.

Our research agenda spanned a total of six years starting in early 2016 and ending with administrative implementation in September 2022. Details of the development have been published elsewhere (Fűköh, 2020; Lukácsi, 2021), only the features relevant to the discussion of the present study will be listed here. Our system of checklists for assessing writing comprises a set of level-specific measurement instruments of varying lengths at levels B1, B2, and C1. As directed by initial discussions with examiners, the checklist items were never intended to reflect degrees of ability individually, but rather they were designed to determine the presence or absence of a construct element. Therefore, the items were first developed as yes or no questions, which were then transformed into binary statements scored dichotomously. Originally planned to be genre-specific, the checklists were later modified to be genre-inclusive where specific textual features are implied so that the integrity of a checklist is retained. As the checklists directly reflect the construct, their length in terms of the number of items is different. The checklist at level B1 has 15 items, level B2 has 36 items, and level C1 has 30 items. Besides, regardless of their volume, the items are grouped into the same component structure throughout the three levels. With varying numbers of items, our components as item bundles are clarity, complexity, effect, precision, and text structure.

doi: 10.22363/2521-442X-2023-7-1-9-20
Volume 7 Issue 1, 2023, pp. 9-20
https://rudn.tlcjournal.org

4. METHOD

The present study was composed of a first phase focusing exclusively on issues of practicality, and a second phase targeting areas of impact and practicality combined. In the first phase, we set out to explore how to introduce checklist-based scoring within the types and amount of resources available while preserving our level of professionalism. In the second phase, stakeholders were invited to share their impressions and opinions of checklist-based scoring as a potentially unknown method comparing and contrasting it with previous experience with scale-based rating weighing up advantages and disadvantages of each. After checklist-based scoring was introduced as the administrative method in live examinations, the examiners were asked to provide feedback on the use of the instrument. Figure 1 is a visual representation of the major phases of the research design.

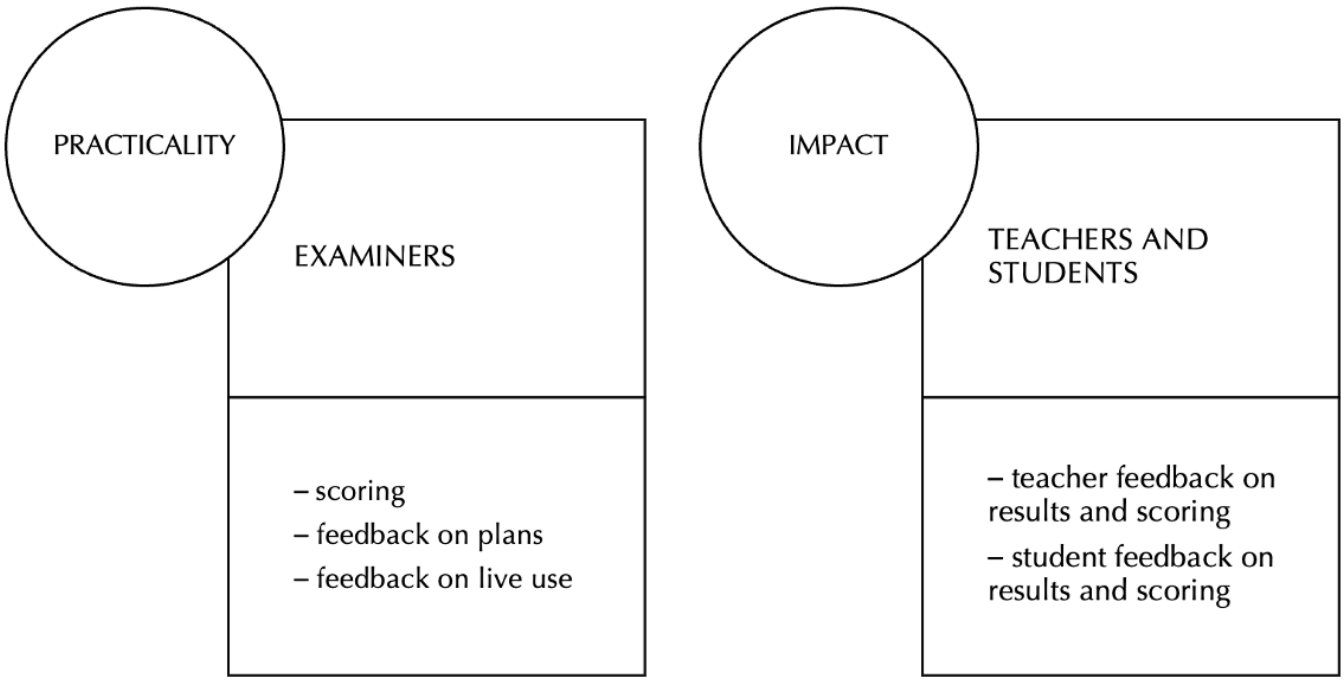

Figure 1. Schematic outline of the research design

The research questions in the study were: 1. How can scoring on the checklist be introduced so that (a) the results reflect professional values, and (b) concerns of practicality are mitigated at the same time? 2. How do stakeholders respond to scoring on the checklist?

The participants in the present study were 12 examiners, six EFL teachers and 28 language learners in four study groups. The examiners had considerable experience in both teaching (M = 30.68 years, SD = 10.24) and testing (M = 21.16 years, SD = 2.64).

In the first phase, 400 scripts from the October 2021 level B2 examination, and 200 scripts from the December 2021 level C1 examination were used. The scripts were selected by stratified random sampling. The examiners were asked to complete a questionnaire about the use and practicalities of the checklist. It was also important to see how the other stakeholder groups – language learners (or future test takers) and teachers – respond to checklist-based scoring. Learners in four study groups and their teachers were approached in schools, where the language learners completed Institution writing tasks at B2 and C1 level.

The writing products were marked with both the scale-based and the checklist-based rating tool, and further to the results, the teachers and learners were provided with an explanation of the scores and an evaluation of the products using both rating tools. To collect feedback on the reception of the results and the rating tool, we conducted semi-structured interviews with the learners and the teachers.

The interviews conducted focused on (a) the format of the results, (b) the content of the evaluation of the writing product, and also (c) its usefulness in language learning.

5. RESULTS AND DISCUSSION

5.1. Phase 1: Practicality

The rationale behind the first phase of the study was to reduce the tension between the interests of sound research on one hand and those of practical implementation on the other. The research was deemed successful, yielding three coherent instruments trialled extensively as units. However, administrative considerations including time allocated for scoring along with concerns of examiner fatigue dictated that the workload should be curtailed. In order to maintain routine double-marking, where mostly for reasons of test security each product is scored by two examiners, we agreed to divide the checklists between the two examiners so that each script was scored by both, but they would work with different items. The research leaders outlined a total of 10 different plans to divide the checklist at level B2. The plans were different in terms of (a) how even the item distribution was between the examiners and (b) the principles of the division. Eventually, two plans were tested in the first phase of data collection. In the first one (Plan A), the items were halved based on results obtained from factor analysis. In the second (Plan B), the items were classified into two groups based on expert judgement of content analysis. The two plans classified 20 items in the same way.

As the scripts were written in live test settings, the observed scores allocated in live scoring with the operational rating scales were available for comparison. These were completed in the datasets with the checklist-based scores collected in a fully-crossed design. The data collection design is presented in Table 1.

Table 1
Quantitative data collection design in phase 1

WAVE 1	WAVE 2	WAVE 3
Live test scores from operational rating scale use	Plan A – Checklist part 1 - Examiner 1 Plan A – Checklist part 1 - Examiner 2 Plan B – Checklist part 1 - Examiner 3 Plan B – Checklist part 1 - Examiner 4	Plan A – Checklist part 2 - Examiner 8 Plan B – Checklist part 2 - Examiner 1 Plan B – Checklist part 2 - Examiner 2 Plan A – Checklist part 2 - Examiner 3
	Plan A – Checklist part 2 - Examiner 5 Plan A – Checklist part 2 - Examiner 6 Plan B – Checklist part 2 - Examiner 7 Plan B – Checklist part 2 - Examiner 8	Plan A – Checklist part 1 - Examiner 4 Plan B – Checklist part 1 - Examiner 5 Plan B – Checklist part 1 - Examiner 6 Plan A – Checklist part 1 - Examiner 7

In Table 1, the quantitative data collection design is presented as composed of three waves. Wave 1 took place during the live administration in October 2021. At the time, six rating scales were used with a score range of [0,5] each: (a) task achievement; (b) appropriacy; (c) cohesion; (d) coherence; (e) grammatical range and accuracy; and (f) lexical range and accuracy. Both raters used the same set of scales, so the cumulative sum of the observed scores had a theoretical maximum of 120 score points. Simple linear transformation was used to convert the observed scores into reported scores ranging from 0 to 100. Pass on the Writing Paper was attained at 60% on the reporting scale, that is 72 observed score points or more. In Waves 2 and 3, Plans A and B indicate the two trialled checklist division schemes, Checklist parts 1 and 2 stand for the first and second halves of the checklist items, and Examiners 1 through 8 mean the research participants. The data collection design controlled for the effects of order both in terms of plan and part. In Waves 2 and 3, each examiner had a workload of 50 scripts per wave. The reported scores and the results from checklist-based scoring were closely correlated (r = .983, p < .001). The mean difference was statistically significant (M = 1.91, t(399) = 13.281, p < .001). Nevertheless, given that the pass or fail classification showed exact agreement in 94.75% of all cases, its practical significance was negligible. This result provided further evidence that the operational rating scales could be replaced by checklist-based scoring without any major detrimental effect.

When comparing the two checklist division plans, we found that the mean sums were not different (M = 0.17, t(398) = -0.348, p = .728). Further, the scores from Plan A yielded the exact same pass rate as those from Plan B. Following Glen (2021), scoring reliability was measured as examiner consistency (Table 2).

doi: 10.22363/2521-442X-2023-7-1-9-20
Volume 7 Issue 1, 2023, pp. 9-20
https://rudn.tlcjournal.org

Table 2
Scoring reliability

EXAMINER	PLAN	PART	α
1	Plan A	Checklist part 1	.831
1	Plan B	Checklist part 2	.920
2	Plan A	Checklist part 1	.820
2	Plan B	Checklist part 2	.881
3	Plan B	Checklist part 1	.894
3	Plan A	Checklist part 2	.870
4	Plan B	Checklist part 1	.680
4	Plan A	Checklist part 1	.904
5	Plan A	Checklist part 2	.926
5	Plan B	Checklist part 1	.767
6	Plan A	Checklist part 2	.875
6	Plan B	Checklist part 1	.830
7	Plan A	Checklist part 1	.900
7	Plan B	Checklist part 2	.784
8	Plan B	Checklist part 2	.844
8	Plan A	Checklist part 2	.821

Scoring consistency was excellent ($\alpha \geq .90$) in four cases, good ($.90 > \alpha \geq .80$) in nine cases, acceptable ($.80 > \alpha \geq .70$) in two cases, and questionable ($.70 > \alpha \geq .60$) in one case (Glen, 2021).

Based on the results from statistical analyses, we concluded that the checklist could be divided into two halves without jeopardising the scoring system. The theory behind the division did not have an impact on the scores. By way of cross-validation, the study was replicated on a sample of 200 at level C1 in a similar data collection design with four examiners. The replication study yielded similar results.

While the quantitative analysis uncovered no evidence against arranging checklist items into two so that the workload related pressure of time and fatigue on examiners could be eased, it provided no guidance as to which of the trialled plans to adopt. Therefore, the research participants were asked to complete an online poll past scoring to collect information regarding their preferences. The six open-ended questions were designed to tap: (a) the nature and stages of the scoring process; (b) the manageability of the workload; (c) the level of perceived objectivity in the scores; and (d) the transparency of item allocation. The polls revealed that while examiners' scoring behaviour and practices were not homogeneous, two major processes surfaced. In a script-based approach, they read the scripts slowly trying to associate text features with checklist items on the go, and then re-read the text for items they failed to score until finished. In an item-based approach, they studied the checklist along with the supporting material provided and grouped the items, sometimes colour coding them. Some even organised the scripts into more manageable batches of 10 and then they read the scripts for a group of items moving on to a different group on completion. In terms of the workload, they all managed to complete the task in time with 62.5% claiming it was a reasonable amount of work, 25% reported tiredness, and 12.5% said it was easier than expected. For the most part, the checklist proved useful in controlling unmodelled variance, while on a few occasions raters still relied on personal judgement. This unwanted behaviour was more typical with certain scripts than with specific items. In the polls, 62.5% of the participants reported that the script and the items directed their scoring, 12.5% said that some scripts evoked their personal opinion, and 25% said that there were some items where they were reluctant to limit their opinion. Looking back at scoring, half of the examiners reported that the selection of the items into Parts 1 and 2 was evident, whereas the other half said that generally it was clear, but some items seemed to stick out. They also expressed a preference for Plan B, i.e., expert judgement based on content analysis.

Based on the results of the qualitative data from the examiners, the checklist items were arranged under the five components outlined in Checklist-based scoring. The components were colour coded so that item selection was more apparent. Plan B was accepted but adapted so that all the items in the top-down components of effect and text structure were in Part 1, and all the items in the bottom-up components of complexity and precision were in Part 2. For the sake of a balanced workload, the items of clarity were divided between the two examiners.

As a result, by means of rearranging the items, creating an overt component structure, and dividing the checklist, we managed to balance research and administration easing examiners' workload while abiding by formal requirements.

5.2. Diagnostic potential as positive washback

Apart from increasing transparency and accountability as well as supporting scoring validity, the reasons for introducing checklist-based scoring included its strong diagnostic potential. By breaking down the construct into precisely defined elements, we intended to utilise the checklist outside of testing. Teaching and learning a foreign language could benefit from streamlined assistance as long as the concepts are unambiguously expressed. From the outset, there was general consensus that the linguistic jargon needed reformulation so that teachers and students could work efficiently and independently from the test centre. In May 2022, three educational specialists identified potentially problematic lexical items in the checklists at all three levels, and recommended ways of collapsing items in order to create a useful tool for language learning and test preparation. Table 3 offers a comparison of the checklist items of the clarity component and their corresponding features of good writing. As the table shows, complex linguistic and testing terms such as 'legible', 'lexicon', or 'comprehension' were replaced with easily accessible expressions like 'easy to read', 'words', or 'understand'. Some particularly oblique words, for example 'interference' were incorporated into another item with a largely similar content to create one combined feature.

Further, relatively transparent items with fairly similar content were also collapsed so that the number of features teachers and students would have to work with was more convenient. As a result, the measurement instruments were reformulated to create useful tools for language teaching and learning at the three levels.

Table 3
Checklist items for scoring and features for teaching and learning

CHECKLIST ITEMS FOR SCORING	FEATURES OF GOOD WRITING FOR TEACHING AND LEARNING
Item_15 This text is legible, i.e.., the reader doesn't have to guess what the writer is trying to say. *Item_16* This text is the required length as defined by the task.	The writing is easy to read and just as long as it should be.
Item_17 There is little or no irrelevant information in this text. *Item_18* This text is clear and concise.	This text is clear with no irrelevant information.
Item_19 The writer can use the English lexicon to express the intended meaning instead of periphrases or non-existent terms.	The writer can use English words to send a clear message instead of complicating terms.
Item_20 Grammatical or linguistic errors in this text do not impede comprehension. *Item_21* There is no L1 or L3 interference that makes reading difficult.	This text may contain errors, but the reader can easily understand it.

doi: 10.22363/2521-442X-2023-7-1-9-20
Volume 7 Issue 1, 2023, pp. 9-20
https://rudn.tlcjournal.org

In order to collect empirical data on the impact of the instrument, it was important to see how the other stakeholder groups perceive test results and the evaluation they receive about writing performance. The individuals who are directly affected by the results using the checklist-based and the scale-based instrument are the learners and their teachers. Both groups received the results of the students' writing performance, and they were explained what the scores and the results mean based on the two instruments. Following that, in the interviews, we asked them how they perceive the different evaluations of the writing products to see to what extent they find these positive or negative, i.e., to see the possible washback effect of the instrument.

Based on the interviews, it became clear that the teachers were familiar with the scale-based results, and they expressed views about its summative nature. Those who preferred the scale, referred to it as 'simple' and 'easy to understand'. They recognised that it is highly 'judgemental', but they also pointed out that this is a feature they prefer, as they want to judge their students' writing performance. One teacher also highlighted that by using the scale, 'it is easier to focus on the mistakes and the shortcomings of the writing product, which is difficult to do based on the positive statements of the checklist'. As opposed to this, it became clear that other teachers realised the diagnostic potential of the checklist as they could clearly see that the checklist items were much more 'detailed' and 'precise'. Another concept that occurred often through the interviews was the transparency of the checklist, which they identified as a clear indicator of formative assessment and a positive washback effect on teaching.

The emerging key words of teachers in connection with scale-based and checklist-based results are presented in Table 4.

Table 4
Keywords from teacher interviews

SCALE	CHECKLIST
familiar	innovative
ordinary	compelling
summative	formative
simple	detailed
judgemental	diagnostic
abstract	precise
vague	transparent
subjective	concrete, specific

As for the students, they were equally less familiar with both instruments, so the innovative nature of the checklist was not mentioned by any of them. Evidently, students who are at B2, and especially at C1 level, are used to receiving evaluation on their writing products, and thus they were able to tell the difference between the two results and the content of the evaluation. Based on the interviews, we identified two different learner types: those who expect test-centred development in language classes, and others who are more interested in further developing their language skills, in other words, more learning-centred. The former expressed that they preferred 'simple' and 'brief' evaluation, whereas the latter praised the 'detailed' and 'transparent' nature of the checklist. One student said that they used the checklist items to list their strengths and weaknesses, and 'wrote up a list of areas where there is room for development'. Following these ideas, we can see that the main difference between the two student types is that those who are test-centred were only interested in their pass/fail classification and expected a critical evaluation of their mistakes based on the scale, while the learning-centred students were happy to receive instructions as to how to build on and complement their existing knowledge based on the checklist. Remarkably, the test-centred students' critical attitude towards the checklist was present in the above-mentioned teacher's comment, in which they pointed out that the checklist is not suitable for marking and listing the mistakes in the writing product. The emerging key words of students in connection with scale-based and checklist-based results are presented in Table 5.

Table 5
Keywords from student interviews

SCALE	CHECKLIST
critical	instructive
test-centred	learning-centred
broad	finely-grained
brief	detailed
general	specific
simple	precise
crude	transparent, fair

The research phase per se was finished after the large-scale trial and empirical data collection. Based on the results report and the feedback from the different stakeholders, we concluded that the checklist was a suitable scoring tool for level testing, and its practical implementation proved its manageability. In September 2022, the exam office decided to introduce checklist-based assessment of writing papers at B1, B2 and C1 levels.

The examiners of the live examinations took part in an extensive training and benchmarking session and were provided with various materials to support their work. In addition to this, after the exam sessions, they were invited to share their feedback on their scoring experience.

Preceding the date of the live examination, we arranged an online training session for all the examiners of the writing tasks. The training was divided into four parts: (a) practical information on the online marking platform, (b) information on the traits of the instrument and presentation of additional materials, (c) controlled practice using the instrument, and finally (d) free practice.

The examiners needed information on practical issues because the exam office decided to introduce a new online platform for marking the writing papers. Although the writing test is paper-based, examiners must record their scores on an online platform, which means that they had to learn to use a new platform and find a convenient way of reading the scrips on paper but reading the items and scoring the papers online.

In the second part of the training, the examiners listened to presentations about the instruments in general, and they familiarised themselves with the additional materials we provided to support the scoring process and to enhance the common understanding of the concepts and the construct elements of the checklists. It is important to stress that these materials, including concept check questions for each checklist item, glossaries and definitions of the key words in the checklist are not part of the instrument. We designed them to ensure the comprehension of the items and to reinforce common construct interpretation within the examiners.

The final two practical parts focused on providing the examiners with hands-on experience for using the checklists. The examiners were given writing products they had not seen before, and first they familiarised themselves with samples that had been pre-scored and displayed examples of evidence for the scores. Following this, they practised scoring previously unseen candidate performances and locating evidence for their scores. These scripts had also been pre-scored, and at the end of the training, the examiners got feedback on their scores. The examiner training was a prerequisite to act as an examiner of the live administration.

In addition to examiner training, the next step prior to the live examination was the organisation of benchmarking sessions at each level. These sessions are held before each live exam session, and their purpose is to standardise examiner scores. The focus of these sessions is finding evidence in candidate scripts and discussing and clarifying the problematic construct elements.

By way of complementing the already large amount of support materials, we decided to launch opinion polls after each live administration to enhance the examiners' insider status. The open questions of the online polls asked about the examiners' feelings about the scoring experience. The first questionnaire after the September administration targeted examiners' general impression about the scoring experience with four questions (allocated time, difficulty, attention span, quality of writing products), after that they were invited to share negative and positive feedback about the on-

line marking platform and were encouraged to express their opinion about the possibility of scoring scanned writing products. The second poll after the October administration had 10 questions which tried to tap on examiners' growing, and possibly changing experience with using the instrument, and also asked for suggestions in connection with possible improvement of the scoring routine.

Interestingly, the examiners have completely different feelings about the scoring experience. Nevertheless, they all expressed positive feelings towards the checklist, and they all prefer the new instrument to scale-based scoring. Their comments vary to a great extent in connection with the allocated time and difficulty, the aspects they seem to agree on are the manageability of the amount of work and meeting the deadlines. As for future changes, and the possibility of improvement, they agree that swapping paper-based batches is a huge burden, and they look forward to scoring scanned scripts online, but they also expressed their concerns about the quality of scanning and the possible illegibility digitalised handwriting.

6. CONCLUSION

This research reported on the practical implementation of a checklist-based scoring system of the Writing Paper of a high-stakes international proficiency examination in English. The essence of the applied linguistic problem the study aimed to respond to was captured in the complex tensions between professional rigour, the administrative framework, available resources, and evolving stakeholder needs. The result was a necessary compromise that eventually ameliorated the initial conflicting interests. With some flexibility, we were able to ease examiners' workload by dividing the checklist, meet time and financial constraints, and provide a useful tool of diagnostic information with a strong formative potential for language teachers and students.

References

Bachman, L. F., & Palmer, A. D. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Brown, D. H., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 147-170). Lawrence Erlbaum Associates.

Cohen, A. (1994). *Assessing language ability in the classroom*. Heinle & Heinle.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*, 88-115. https://doi.org/10.1016/j.asw.2009.04.001

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185. https://dx.doi.org/10.1177/0265532207086780

Fűköh, B. (2020). *Establishing the context and scoring validity of the writing tasks of Institution's English for academic purposes test* [Unpublished doctoral dissertation]. University of Szeged.

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

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

Glen, S. (2021). Cronbach's alpha: Simple definition, use and interpretation. *StatisticsHowTo.com*. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/cronbachs-alpha-spss

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220. https://doi.org/10.1207/s15324818ame1702_3

Grishechko, O. S., Akopova, A. S., & Grishechko, E. G. (2015). English linguistic purism: History, development, criticism. *Proceedings of Southern Federal University. Philology, 4,* 185-192. https://dx.doi.org/10.18522/1995-0640-2015-4-185-192

Hambleton, R. K., & Meara, K. (2000). Newspaper coverage of NAEP results, 1990 to 1998. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the national assessment of educational progress: Affirmation and improvements* (pp. 133-155). National Assessment Governing Board.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (pp. 433-470). American Council on Education and Praeger Publishing Group.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights from the classroom* (pp. 69-87). Cambridge University Press.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12*, 1-9. https://doi.org/10.1016/j.asw.2007.05.002

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

Harsch, C. & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly, 8*(1), 1-33. https://doi.org10.1080/15434303.2010.535575

Inoue, C., Khabbazbashi, N., Lam, D. & Nakatsuhara, F. (2021). *Towards new avenues for the IELTS Speaking Test: Insights from examiners' voices*. British Council, Cambridge Assessment English and International Development Program. https://www.ielts.org/teaching-and-research/research-reports

Kim, Y.-H. (2010). *An argument-based validity inquiry into the Empirically-derived Descriptor-based Diagnostic (EDD) assessment in ESL academic writing* [Unpublished doctoral dissertation]. University of Toronto.

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparametrized Unified Model. *Language Testing, 28*(4), 509-541. https://doi.org/10.1177/0265532211400860

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499. https://dx.doi.org/10.1177/0265532214530699

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing, 38*(1), 86-105. https://dx.doi.org/10.1177/0265532220916703

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(4), 246-276. https://doi.org/10.1191/0265532202lt230oa

McNamara, T, Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221-242. https://doi.org/10.1017/S0267190502000120

Mousavi, S. A. (2009). *An encyclopedic dictionary of language testing* (4th ed.). Rahnama Press.

Rossi, O., & Brunfaut, T. (2020). Raters of Subjectively-Scored Tests. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1-7). John Wiley & Sons. https://doi.org/10.1002/9781118784235.eelt0985

Rouet, J-F. (2012, June 2). *Reading skills in the information age: Cognitive processes and implications for assessment* [Paper presentation]. 9th EALTA Conference, Innsbruck, Austria.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677-710). Lawrence Erlbaum.

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing, 18*, 187-201. https://dx.doi.org/10.1016/j.asw.2013.05.001

Taylor, L. (2005). Washback and impact. *ELT Journal, 59*(2), 154-155. https://dx.doi.org/10.1093/eltj/cci030

Van Moere, A. (2013). Raters and ratings. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.1358-1374). John Wiley and Sons. https://doi.org/10.1002/9781118411360.wbcla106

Verhelst, N. D. (2004). Item response theory. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section G). Council of Europe.

Voigt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374-402.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*(4), 301-335.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223. https://dx.doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 253-287. https://dx.doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing*. CUP.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-319. https://dx.doi.org/10.1177/026553229301000306

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice, 31*(2), 21-26.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359-375. https://doi.org/10.1080/08957 340903221667

Zhuk N. V., & Ivanov V. D. (2022). The combination of question and exclamation marks (?!) in some emotive types of statements (based on the material of the German language). Bulletin of the Moscow Region State University. *Series: Linguistics, 6,* 101-108. https://dx.doi.org/10.18384/2310-712x-2022-6-101-108

ZOLTÁN LUKÁCSI
Euroexam International, Hungary zoltan.lukacsi@euroexam.org

BORBÁLA FŰKÖH
Euroexam International, Hungary fukoh.borbala@euroexam.org