doi: 10.22363/2521-442X-2020-4–4-9-22
rudn.tlcjournal.org

Original Research

Standard setting and internal validation of a novel approach adopted for assessing speaking

by Zoltán Lukácsi

Zoltán Lukácsi Euroexam International zoltan.lukacsi@euroexam.org
Received 14.08.2020 | **Revised** 23.11.2020 | **Accepted** 10.12.2020
Recommended citation format: Lukácsi, Z. (2020). Standard setting and internal validation of a novel approach adopted for assessing speaking. *Training, Language and Culture, 4*(4), 9-22. Doi: 10.22363/2521-442X-2020-4-4-9-22

In 2016, Euroexam International launched a complex and ambitious project to redesign the rating process and practice of its subjectively scored papers. As part of the project, a series of level-specific performance checklists were developed for speaking and writing. This paper reports on the standard setting and internal validation procedures for B2 speaking in English as a foreign language (EFL) as executed in Spring, 2020. The participants of the study were 8 EFL teachers and oral examiners from international backgrounds with varying degrees of expertise in language testing. The method adopted for standard setting was the Body of Work method. The participants were allocated into either of two groups working independently on a random sample of 32 audio-recorded live speaking paper performances in a counterbalanced design. The study found that (a) a level-specific 30-item checklist for EFL speaking at B2 can adequately replace the operational rating scales for pass/fail decisions; (b) the granular nature of the checklist renders it more capable of covering the targeted content areas; and (c) the explicit checklist statements support fairness, transparency and accountability.

KEYWORDS: *assessing speaking, speaking assessment, checklist, standard setting, validation, transparency, EFL, English as a foreign language, Euroexam*

1. INTRODUCTION

Rating scales are the primary measure when it comes to assessing speaking or writing in high-stakes contexts. The false promise of transparency is easily challenged when considering that the score is supposed to reflect the intricate texture of candidate ability, assessment criteria, rater perception, and rating consistency (Hamp-Lyons, 1990). Rating on a scale can be used to classify test takers into a series of broad levels, but they are inadequate when fine-grained distinctions are required, as in level-specific testing. Euroexam International repeatedly redesigned its operational rating scales over the past two decades, but problems with score inconsistency and rater reliability still persevered. In an ambitious research agenda, a set of checklists was developed to help resolve the problems with human-mediated assessment. The study this paper reports on sets out to answer the following research questions.

(1) Can a checklist be used to score candidate performance at B2 speaking exams in place of the operational rating scales without distorting the construct?

(2) What is the performance standard for speaking general English at level B2 on the checklist?

2. THEORETICAL BACKGROUND

Research into the assessment of speaking a second or foreign language has for the most part focused on construct relevant issues, such as definition, operationalisation, and validation, and rating scale development (Fulcher, 2015). Fulcher (2003) provides a detailed overview of the historical development of the construct of speaking pointing it out that the term itself has lost much of its former appeal with the advance and general acceptance of the action-oriented approach in the CEFR in Europe and beyond (Council of Europe, 2001). The ability of speaking is a multi-faceted construct (Fulcher, 2003, p. 25) as operationalised by the principal components of (a) pronunciation and intonation; (b) fluency; (c) discourse management; (d) language resource range and accuracy; and (e) task achievement (Fulcher, 2003; Luoma, 2004). Apart from these fundamental components, a number of complementary elements have been suggested by research depending on context (Chapelle, 1999), domain specificity (Biber, 2006), level of integration (Lewkowicz, 1997), scoring method (Wang et al., 2018), mode of delivery (Nakatsuhara et al., 2016), test format (Ffrench, 1999), test purpose (Douglas, 2000), and topical knowledge (Huang et al., 2018). The decisions on these and many other variables inform the test specifications and translate into the rating criteria (Luoma, 2004, p. 114). Accordingly, Weir (2005) makes a primary distinction between a priori validity evidence based on theory and context, and a posteriori validity evidence nested in scoring, criterion measures and consequences.

Luoma (2004, p. 59) defines scores for speaking as an expression of *'how well the examinees can speak the language'*. Scores are expected to arrive from the assessment criteria and reflect the operations the candidate performs under the conditions described in the test design (Weir, 2005, p. 191). However, there is a plethora of research discussing threats to the validity of scores for speaking. Rating scales are known to be laden with unfavourable properties including score inconsistencies (Hamp-Lyons, 1990; McNamara, 1995) that necessitate post-test adjustment (Bonk & Oakey, 2003; Eckes, 2005; Lumley & McNamara, 1995; Weigle, 1998; Wigglesworth, 1993). Unreliable scores can be the result of criteria described in vague terms that leave ample room for divergent construct interpretation (Alderson, 1991, 2007; Alderson et al., 1995; Luoma, 2004; Upshur & Turner, 1995). Further, rating scales have been criticised for lack of reliance on empirical research (Chalhoub-Deville, 1997; Fulcher, 1996; Harsch & Martin, 2012; North, 1997; North & Schneider, 1998; Turner & Upshur, 2002). Over and above the problems with rating scales, much of the research to date has focused on differences between rater cognition and behaviour in comparison with candidate performance (Ang-Aw & Meng Goh, 2011; Kormos, 1998; Lumley & McNamara, 1995; McNamara, 1995; Orr, 2002; Wigglesworth, 1993).

Recent attempts at counterbalancing threats to scoring validity have increasingly relied on technology. Developments in IT coupled with sophisticated statistical data processing enabled analysts to control for unmodeled rater variance in speaking performance assessment (Bachman et al., 1995). In a purposely partial review, McNamara and Knoch (2012, p. 570) declare that *'by the 2000s, then, the Rasch wars were essentially over'* claiming that many-facet Rasch measurement (MFRM) provided the adequate response to the noise in test scores.

Apart from the criticism for statistical inconsistency (De Jong & Linacre, 1993), MFRM works on the level of the scores but leaves the underlying construct unaddressed (Lumley & McNamara, 1995). Automated assessment offers a radically different but highly resource-intensive method to ensure scoring validity (Xi, 2010). Research has shown, however, that some candidates might repudiate natural language processing algorithms (Wood, 2020).

Few studies have attempted to design and develop an alternative assessment strategy to rating scales. Fulcher et al. (2011) proposed the Performance Decision Tree of a series of binary decisions based on a rich description of empirical data. In accordance with the recommendation that level-specific performance be judged in relation to a list of points deemed relevant (Council of Europe, 2001), Kim (2010, 2011), Struthers et al. (2013), and subsequently Lukácsi (2020) developed writing assessment checklists with a diagnostic merit.

Criterion-references assessment necessarily involves standard setting to classify examinees into attainment levels (Kaftandjieva, 2004). The essence of standard setting is the establishment of one or more cut-off points on the score scale. Kane (2001) views these cut scores as operationalisations of performance standards. The theory and practice of standard setting is well-documented in the literature (Cizek & Bunch, 2007; Kaftandjieva, 2004, 2010). Despite its scientific nature, however, standard setting relies on personal judgement (Reckase, 2009) which might in turn lead to different standards (Kaftandjieva, 2004).

The review of the relevant literature in the field suggests that while rating scales may be the most frequently applied instrument when assessing spoken performance, they are innately fraught with a number of issues from imprecise wording through lack of empirical support to inconsistent rater behaviour. One solution proposed to control for much of the unfavourable noise within the judgements is rating on a checklist, where rater judgement is limited to noticing construct elements and indicating their presence (or absence).

3. EUROEXAM CHARACTERISTICS

3.1. The speaking paper

Established in 2002, Euroexam was designed to test candidates' spoken production and interaction in a paired format to reflect the professional preference for dyadic oral performance assessment (Swain, 2001). The test specifications (Euroexam International, 2019) provide a detailed account of the contents defined. The fundamentals of the speaking paper at level B2 are listed in Table 1.

Table 1
Speaking paper tasks, target activities, agents, and time

TEST TASK	TARGET ACTIVITY	AGENTS	TIME
Interview	Interaction: conversation	Interlocutor interviews each candidate in turn	2 min
Picture story	Production: narration	Individual candidate performance in succession	5 min
Transactional dialogues	Interaction: scripted functional role-play	Interlocutor talks to each candidate in turn	5 min
Discussion	Interaction: negotiation	Candidates discuss topic together	5 min

As Table 1 shows, the speaking paper contains four tasks at level B2. Candidates have 10 minutes preparation time for the picture story with a cartoon strip with a starting sentence printed as prompts before the examination. During preparation, candidates can use a dictionary. In a paired format where candidates can choose their partners, the speaking exam starts with a short, scripted interview on a pre-selected topic. The interlocutor asks two questions from each candidate

helping them get accustomed to the test situation. The second task aims to test candidates' independent spoken production through a 2-minute picture story, where they can rely on their notes but cannot read out from them. The third task is a series of three scripted dialogues starting with a cue card, followed by candidate response, then interlocutor spoken cue and complete with candidate response. Finally, in the discussion task, candidates have 3 minutes to try and agree on a given topic. The allocated time in Table 1 contains time for setting up the tasks and rounding off the examination.

In a live examination setting, candidate performance is assessed using the operational rating scales (Euroexam International, 2020). The scales have five components: (a) range and accuracy; (b) fluency and coherence; (c) pronunciation; (d) communication strategies; and (e) overall impression. Each component is scored 0 through 5. Score band 3 is a description of the minimally competent speaker at the level. The examiner scores performance on components (a) to (d), and the interlocutor scores performance on component (e). The rating scales are not task-specific, even though some tasks might lend themselves more to certain components.

The results are expressed on the reporting scale ranging from 0 to 100. The observed scores are converted by a simple linear transformation. The examination applies a combination of the conjunctive approach and compensation when reporting results. Any candidate with a reported score of 60 or more is regarded as successful in speaking. However, only candidates below a reported score of 40 are viewed as failing test takers, because the other test paper results might compensate for the relatively weak performance.

3.2. Reported scores for speaking

The frequency distribution in Figure 1 shows a tendency with reported scores for speaking. First, scores lower than 40, the minimum requirement, were rare, hence a negatively skewed distribution. Second, the most frequent score was 60, the paper level equivalent of a pass. In December 2019, test paper intercorrelations with speaking were always weaker (varying between $r = 0.411$, $p = 0.000$ and $r = 0.518$, $p = 0.000$) than among the other three papers in the monolingual exam (varying between $r = 0.470$, $p = 0.000$ for writing and listening, and $r = 0.645$, $p = 0.000$ for reading and writing). Reported scores for speaking correlated strongest with listening ($r = 0.518$, $p = 0.000$) rather than writing ($r = 0.498$, $p = 0.000$) despite the similarities in language activities. Such relationships among test papers were regular. Finally, even though appeals for rescoring were rare as a rule, if initiated, the outcome tended to be a modification of the reported score.

The reliability of the reported scores using the rating scales was $\alpha = .934$ with scale component correlations ranging from $r = 0.659$, $p = 0.000$ to $r = 0.779$, $p = 0.000$, potentially pointing to covert overlaps within the measurement traits. Reported scores for speaking seemed to be somewhat inconsistent with the rest of the test particularly with regard to low achievers.

3.3. The checklist

Developing a level-specific checklist for the assessment of speaking was part of a complex research agenda at Euroexam International that aimed to review and redesign current operational practice. The reasons for launching the research programme were manifold.

First and foremost, as explained in the review of relevant literature in the field, language testing theory and practice suggested that rating scales were inherently laden with challenges to construct representation and consistent construct interpretation.

Second but still within the realm of professional trends and traditions, rating scales that spanned over a number of proficiency bands were not ideally suited for level testing.

Third, as outlined in Reported scores for speaking, empirical data revealed unfavourable characteristics resulting from rating scale use rather than reflecting genuine tendencies within the population.

Fourth, there was an increasing intention at Euroexam International to improve transparency and accountability.

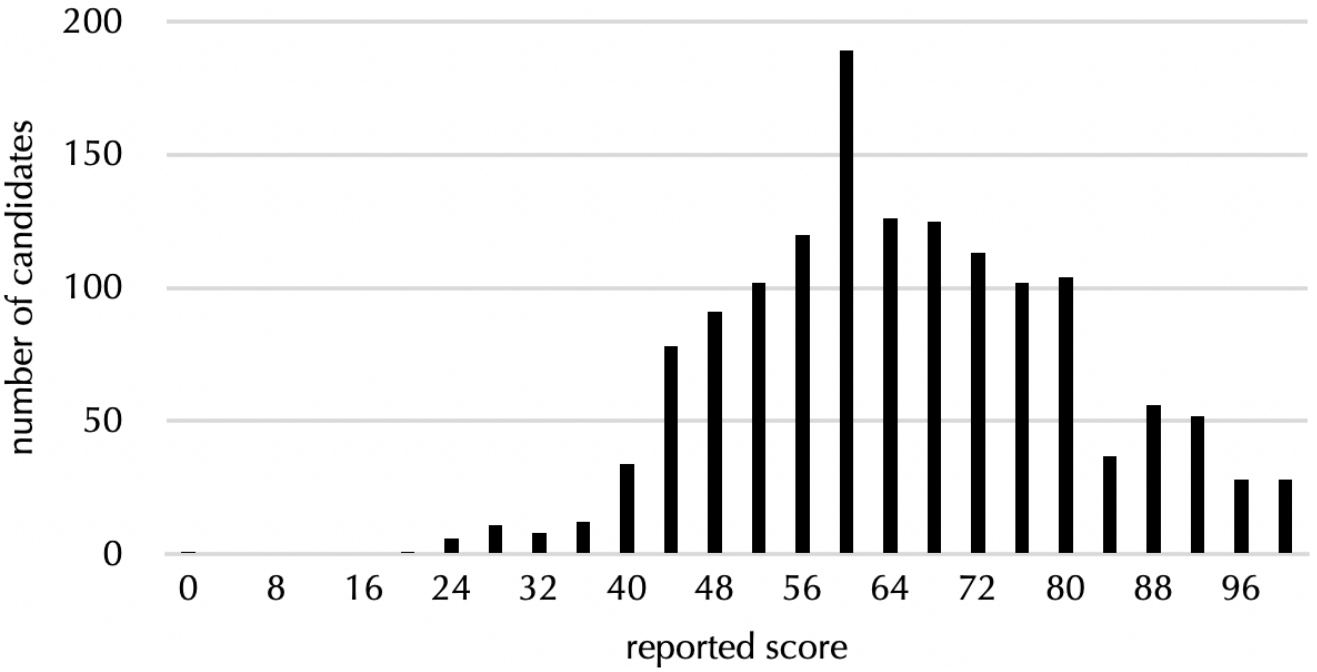

Figure 1. The frequency distribution of reported scores for speaking in December 2019 (*N* = 1424, *M* = 65.14, *SD* = 15.64)

Encouraged by the professional success of the set of checklists for assessing writing (Lukácsi, 2020), the research and development team developed a checklist for assessing speaking at B2 over the course of 19 months between May 2018 and November 2019. The 30-item checklist relied heavily on CEFR band descriptors but was at the same time developed in accordance with the test specifications and our history and ethos of language testing. A detailed account of the development of the checklist for assessing speaking falls outside the score of this research report. In brief, it entailed a review of theory and research findings, construct specification and analysis of sample video footage, piloting a pool of items under operational conditions, and rescoring audio recordings of live spoken exams.

4. THE STUDY

The participants in the study were eight experienced EFL teachers with over 20 years of teaching practice each (*M* = 27.50, *SD* = 7.52). They were also skilled interlocutors and oral examiners from international backgrounds with varying degrees of expertise in language testing (*M* = 13.12, *SD* = 7.60). As such, they were all familiar with the test specifications, the construct, and level requirements. For the purposes of the study, they were all independently contracted, briefed and asked to sign an informed consent and confidentiality form. In the data collection design, the participants were randomly assigned either to Group A or to Group B but worked individually.

The materials of the study were 32 audio-recorded live B2 speaking performances. Materials selection was informed by the following factors: (a) available panel; (b) available time; (c) the detailed test specifications; (d) the paired format of the speaking paper; (e) audio-recording, where same sex candidates could be difficult to differentiate; and (f) the fact that there was no link between test administrations in this part of the test. Within the constraints of these factors, the sample was limited to a specific exam location under the code name BP, the December 2019 administration, and different sex dyads. Further, for reasons of consistency, two sets of tasks were selected under the codes A12GM and B34HM, where A and B were two in-

terview topics; 1, 2, 3, and 4 were four picture stories; G and H were two sets of transactional dialogues; and M was a topic for discussion. Eight pairs of candidates were tested on tasks A12GM and another eight pairs on B34HM. The descriptive statistics for the sampled performances were: *M* = 64 raw score points, *Mdn* = 62, Mode = 56, *SD* = 15.6447.

The Body of Work method (Cizek & Bunch, 2007; Kingston et al., 2001) was used for standard setting. In essence, this method requires judges to evaluate complete sets of examinee performance and to render a single holistic judgement about each performance sample. First, the performance samples are ordered by total score, and the panellists are informed about which samples received high scores and which received low scores. Later, samples that are clearly at a given level and judges agree in their classification are systematically removed between rounds of data collection. Such samples are replaced by additional ones with scores near the score points at which there is greater disagreement. Kingston et al. (2001) refer to the activities as *rangefinding* (Round 1) and *pinpointing* (Round 2).

Data were collected in a counterbalanced design. The three stages panellists' tasks were divided into are represented in Figure 2.

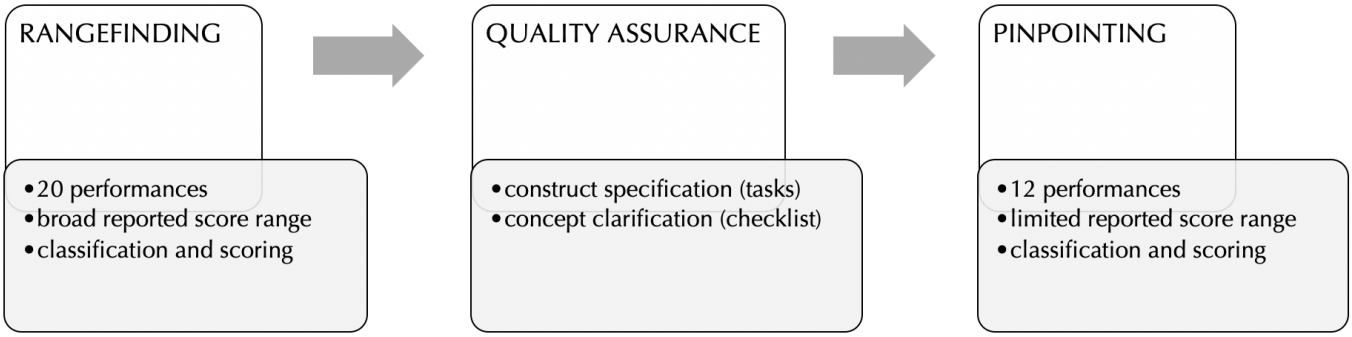

Figure 2. A flowchart of the stages of panellists' tasks

At the rangefinding stage, panellists in Group A listened to five candidate dyads and evaluated their speaking performances with regard to level B2 as a fail (0), a pass (1), or a pass with distinction (2). Group B listened to the same sample and scored the performances using the checklist items. As a next step within rangefinding, the panellists were given another set of five candidate dyads, but they swapped their roles. At the quality assurance stage between rangefinding and pinpointing, four panellists were assigned to work on construct specification in the tasks, while the other four worked on concept clarification in the checklist. Then, informed by the results from rangefinding, the pinpointing stage worked much the same way as regards data collection with two differences. First, clearly off-target performances were not sampled for judgement so that the precision of setting the standard could be increased by a more restricted score spread. Second, only six recordings, i.e. 12 candidate performances, were evaluated in a counterbalanced design.

The analytical procedure used to calculate the cut score was logistic regression. The operational definition of the cut score was the raw score where the probability of classification into either of two adjacent classes of success was the same.

5. RESULTS AND DISCUSSION

5.1. Applying the checklist

The checklist for assessing speaking at B2 can be found in full in Appendix A. Table 2 contains the item level descriptive statistics. Item responses were analysed using classical test theory (p-value, r_{it}, $r_{it}(w)$) and applying modern test theory (A, B, SE (B), M1, M2, M3), as well.

Table 2
Speaking checklist item descriptive statistics

LABEL	P-VALUE	R_{IT}	R_{IT}(W)	A	B	SE (B)	M1	M2	M3
Item_01	0.824	0.344	0.337	2	-0.978	0.141	0.74	-0.458	0.128
Item_02	0.278	0.258	0.212	1	0.917	0.218	1.519	-0.226	-0.139
Item_03	0.167	0.194	0.186	2	0.844	0.142	1.487	0.486	0.553
Item_04	0.194	0.468	0.513	3	0.545	0.104	-1.031	-0.777	-0.806
Item_05	0.009	0.149	0.144	1	4.684	0.973	99.999	-0.714	-0.876
Item_06	0.991	0.138	0.132	1	-4.862	0.973	99.999	-0.648	-0.844
Item_07	0.389	0.399	0.397	2	0.188	0.116	0.433	0.109	0.728
Item_08	0.519	0.294	0.235	1	-0.156	0.198	-0.396	-0.133	0.091
Item_09	0.185	0.148	0.126	1	1.46	0.249	-1.214	0.341	0.196
Item_10	0.296	0.479	0.493	3	0.31	0.095	0.706	0.54	-0.282
Item_11	0.056	0.253	0.254	3	1.086	0.153	1.137	-0.875	-0.365
Item_12	0.713	0.527	0.561	3	-0.486	0.097	0.171	-0.314	-0.868
Item_13	0.222	0.384	0.389	2	0.646	0.13	-0.417	-0.536	-0.411
Item_14	0.454	0.445	0.454	2	0.032	0.115	-1.001	-0.411	-0.74
Item_15	0.667	0.483	0.461	2	-0.486	0.12	-0.253	-0.414	-0.057
Item_16	0.62	0.549	0.58	3	-0.295	0.093	-0.454	-0.282	-0.143
Item_17	0.833	0.463	0.48	3	-0.795	0.11	0.22	0.09	0.079
Item_18	0.815	0.316	0.281	1	-1.633	0.25	-1.469	0.255	0.279
Item_19	0.769	0.442	0.426	2	-0.782	0.13	0.004	-1.284	-0.777
Item_20	0.731	0.394	0.388	2	-0.667	0.126	-0.599	1.268	0.319
Item_21	0.741	0.503	0.502	3	-0.549	0.099	-0.386	0.743	0.162
Item_22	0.75	0.361	0.349	2	-0.723	0.128	0.513	1.103	0.775
Item_23	0.12	0.171	0.15	1	1.979	0.294	99.999	0.236	-0.415
Item_24	0.5	0.425	0.417	2	-0.077	0.114	0.169	-0.623	0.16
Item_25	0.583	0.496	0.532	3	-0.224	0.091	0.168	-0.016	0.74
Item_26	0.139	0.283	0.261	2	0.961	0.151	-1.166	0.41	0.879
Item_27	0.75	0.433	0.416	2	-0.723	0.128	-0.283	-0.489	0.122
Item_28	0.389	0.455	0.454	2	0.188	0.116	0.433	0.549	-0.994
Item_29	0.778	0.375	0.368	2	-0.812	0.132	-1.465	1.092	1.14
Item_30	0.25	0.462	0.483	3	0.41	0.098	0.222	0.243	0.063

Note: p-value: item facility; r_{it}: item-test correlation, r_{it}(w): weighted item-test correlation, A: discrimination index, B: item difficulty parameter, SE (B): standard error associated with item difficulty parameter, M1, M2, M3: item fit statistics (Verhelst et al., 1995, p. 14-15).

The 30-item checklist covered a relatively broad span of candidate ability. P-values ranged from 0.991 (item_06) to 0.009 (item_05). Mean observed item difficulty was M = 0.49 (SD = 0.28). The literature is divided as to how to deal with items with unfavourable properties. Mehrens and Lehmann (1991, p. 259) recommend excluding items with extreme item facility values, whereas Crocker and Algina (1986, p. 336) argue in favour of domain representation. The decision taken by the research team following Bachman's (2004, p. 138) rule of thumb was to retain all items with non-negative discriminating power.

The geometric mean of the discrimination indices was 1.921. Item discrimination indices varied between 1 and 3. According to the M-tests, the individual items showed fit to the OPLM-model (Verhelst et al., 1995). The reliability of the checklist for assessing speaking was α = .803. The checklist as a coherent measurement tool was also found to fit the OPLM-model R1c(58) = 26.444, p = 0.999.

Based on these properties, and in response to Research Question 1 *(Can a checklist be used to score candidate performance at B2 speaking exams in place of the operational rating scales without distorting the construct?)*, the study found that the checklist for assessing speaking at B2 was accepted as a valid representation of the measurement principles and was ready to be applied in practice.

5.2. Setting the performance standard on the checklist

5.2.1. Rangefinding

The first comparison was drawn between panellists' judgements as to whether the candidate was successful in speaking or not with the pass or fail classification based on the reported score. Panellists showed complete agreement in 40% of all cases, and acceptable agreement with the exception of a single judge in 70% of all cases. Agreement was not reached on the candidate who received the lowest score (36, clear fail) in operational scoring. The most pronounced division was noted on a candidate who scored 56 in the live setting: an equal number of panellists saw this performance as a fail and as a pass. This result would later inform the pinpointing stage. Panellists' judgements were in a strong positive correlation with the pass or fail classification based on the reported score (r = 0.535, p = 0.015). These findings suggest that while panellists and examiners tended to classify candidate performance similarly, agreement was less than satisfactory not only among panellists but between panellists and examiners in a live test administration, as well.

'These findings suggest that while panellists and examiners tended to classify candidate performance similarly, agreement was less than satisfactory not only among panellists but between panellists and examiners in a live test administration, as well'

Next, the sums of panellists' checklist scores were compared with the reported scores. Low achievers in the live setting generally received lower checklist sum scores. Correlation varied between (r = 0.727, p = 0.017) and (r = 0.472, p = 0.168). One panellist's scores showed a negative albeit nonsignificant relationship with the reported scores (r = -0.111, p = 0.759). The conclusion drawn from these results was that even with the score discrepancies, as discussed in the section on reported scores for speaking and demonstrated in Figure 1, specification of construct coverage in the test tasks necessary, and potentially problematic concepts in the checklist items needed clarification. The information gleaned from this stage was that the score range for pinpointing could be reduced to between 50 and 80 on the reporting scale.

5.2.2. Quality assurance

In-between the rangefinding and the pinpointing stages, the participants were invited to complete two tasks focussing on quality assurance. First, the tasks needed to be analysed to see if the target construct was represented as specified in the

test specifications (Euroexam International, 2019). Second, because of the limited reliability of the panellists' checklist scores in particular, the operationalisation of its concepts needed clarification.

With regard to construct coverage, the four panellists assigned to this task found that the speaking paper provided test takers with opportunities to display their competences in the production activity of (a) sustained monologue: describing experience; in the production strategies of (b) planning, (c) compensating, and (d) monitoring and repair; in the interaction activities of (e) understanding the interlocutor, (f) informal discussion, (g) goal-oriented co-operation, (h) and interviewing and being interviewed; and the interaction strategies of (i) taking the floor, and (j) cooperating. However, the inspection revealed that despite explicit statement and intended targeting, candidates are not invited to participate in free conversation or ask for clarification.

Further, interaction around obtaining goods and services was fortuitous. This overview provided support for the checklist as a measurement tool and suggested that focused task selection and sequencing be applied in the transactional dialogues task if the construct were to retain the element of obtaining goods and services.

Concept clarification uncovered three problem areas. First, technical terminology posed a problem in cases such as subject-verb agreement or referencing. Second, basic notions needed to be operationalised with precision, as in the case of permissible deviation from the 2-minute timeframe in the picture story task. Third, awareness needed to be raised among panellists as to what the task of assessment entailed. Whereas scoring with the checklist meant mere noticing of a construct element in candidate output and registering its presence or absence, on occasion panellists were found to override this protocol and apply personal judgement.

To assist the scoring process, a glossary was compiled where potentially oblique terms were defined, and the layout of the checklist was modified to include a textbox of orientation for the examiner.

'In-between the rangefinding and the pinpointing stages, the participants were invited to complete two tasks focussing on quality assurance. First, the tasks needed to be analysed to see if the target construct was represented as specified in the test specifications. Second, because of the limited reliability of the panellists' checklist scores in particular, the operationalisation of its concepts needed clarification'

5.2.3. Pinpointing

As at the stage of pinpointing the range of speaking performance reported scores was limited to where most disagreement was observed with regard to pass or fail classification, agreement between panellists on the score level was not sought. Instead, concept clarification was relied on to maintain similar construct interpretation. The checklist as a measurement tool was designed to control for much of the noise that originated from idiosyncratic rater behaviour often by means of quantifiable phenomena, such as in item_08 (*The speaker can speak continuously for about 2 minutes without having to pause for language*), or by observing objective or objectifiable elements, as in the case of item_24 (*The speaker can invite their partner into the conversation*). Besides, some items left room for subjective reading giving way for score discrepancies, for instance item_06 (*The speaker is always clearly intelligible despite non-English accent*). There was a strong positive correlation between the reported scores for speaking and the mean ($r = 0.783$, $p = 0.003$) and median ($r = 0.801$, $p = 0.002$) values from checklist sum scores.

The cut score was calculated using logistic regression on the data from pinpointing. The aim was to find the checklist sum score where the probability of passing or failing the speaking paper was the same, $p = .50$.

Table 3
Logistic regression results on pinpointing data from checklist use

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1	Score	.447	.095	21.903	1	.000	1.563	1.297	1.885
	Constant	-6.920	1.569	19.446	1	.000	.001		

Note: R^2 = .36 (Hosmer & Lemeshow, 1999), .38 (Cox & Snell, 1971), .52 (Nagelkerke, 1991). Model χ^2 (1) = 40.25, $p < .001$

Table 4
Logistic regression results on pinpointing data from pass or fail classification

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1	Score	.063	.016	14.939	1	.000	1.065	1.031	1.099
	Constant	-3.779	1.090	12.022	1	.001	.023		

Note: R^2 = .18 (Hosmer & Lemeshow, 1999), .21 (Cox & Snell, 1971), .28 (Nagelkerke, 1991). Model χ^2 (1) = 19.76, $p < .001$

Table 3 contains the results from the calculations regarding the cut score on the pinpointing data using checklist sum scores as predictors of the pass or fail classification in the live examination setting. The probability of passing and failing was found equal at the checklist sum score of 15.48. Given that observed scores were always integers, a checklist sum score of 15 or less meant a fail, and 16 or more meant a pass.

By way of cross-validating classification consistency, in another set of calculations panellists' pass or fail classifications were used as predicted from reported scores on the speaking paper (Table 4).

Feeding the values from Table 4 into the formula, we found that the cut score on the reporting scale was 59.98, i.e. 59 reported score points or less meant a fail, and 60 or more meant a pass. This finding is in accordance with the operational standard, which supports the validity of panellists' judgements.

In response to Research Question 2 *(What is the performance standard for speaking general English at level B2 on the checklist?)*, the study found that a cut score of 15.48 on the checklist for assessing speaking at B2 represented the boundary between passing and failing this part of the test. Further, this cut score was found to be in line with current operational practice with regard to the success rate in speaking.

6. CONCLUSION

This study reports on research into the applicability of a novel method of assessing speaking and on the internal validation of the proposed measure. Despite their popularity, rating scales can lead to construct irrelevant variance, score inconsistencies, and occasional misclassification. As part of an ambitious research agenda, Euroexam International set out to explore the possibility of redesigning the scoring aspect of its speaking and writing papers. With a focus on speaking general English at B2, we found that the checklist as a measurement tool can effectively replace the operational rating scales without distorting the construct or having a radical impact on classification consistency.

APPENDIX. The checklist for assessing speaking at B2 **Overall:** below B2 / B2 / good B2 or above

BASIC USE OF ENGLISH (5)	✓
1. Mistakes don't lead to misunderstanding.	
2. The speaker can **correct their mistakes**.	
3. The speaker can **correct false starts**.	
4. The speaker can use **complex structures**, e.g. *I'd rather / used to / tend to.*	
5. The speaker can use **question tags**.	
PRONUNCIATION (2)	✓
1. The speaker is always **clearly intelligible** despite non-English accent.	
2. The speaker can pronounce the sounds /θ/, /ð/, and /w/	
TASK 2: PICTURE STORY (11)	✓
1. The speaker can **speak continuously for about 2 minutes** without having to pause for language.	
2. The speaker can relate a story with a **non-linear** development.	
3. The speaker can use **conjunctions** over and above *and, but, so, because, then*, etc.	
4. When pausing, the speaker can use language to **fill in the silence** instead of repeating, e.g. *erm.*	
5. The speaker can tell a story using **prosody** to indicate where one sentence ends, and another begins.	
6. The speaker can use **reported speech** when relating past events (e.g. *He thought…, She said to her mum…*).	
7. The speaker can use **irregular verbs** consistently correctly.	
8. The speaker can use the **tenses** to indicate the time of the verb: the timeline is consistently applied.	
9. There is consistent **subject-verb agreement**.	
10. The speaker can use **reference** instead of repeating lexical items.	
11. The speaker consistently limits themselves to **existing lexical items**.	
TASK 3: DIALOGUES (5)	✓
1. The speaker can always **respond to the aural cue** of the interlocutor beyond *Thank you, Good-bye*, etc.	
2. The speaker can consistently manipulate **word order** to ask questions.	
3. The speaker can speak in **complete sentences** observing the SV rule.	
4. The speaker can always **greet and address** their **partner appropriately**.	
5. The speaker can use **indirect questions** to sound more formal, polite, and less direct.	
TASK 4: DISCUSSION (7)	✓
1. The speaker can **invite their partner** into the conversation.	
2. The speaker can indicate (partial) **agreement** with functional exponents beyond *yes and* or *yes but*.	
3. The speaker can indicate **disagreement** with functional exponents.	
4. The speaker can **respond** to what their partner has said and doesn't speak alongside them.	
5. The speaker can **integrate** what their partner has said into the conversation.	
6. There is **no L1 / L3 interference** that requires extra effort from the partner or interlocutor: the candidate always restrains themselves to L2.	
7. The speaker can use **modal** verbs correctly where necessary.	

Note: With items containing **consistently** or **always**, only tick the box if there are no mistakes with regard to the criterion [items 6, 14, 15, 16, 18, 19, 20, 22, 30]. If the item doesn't specifically say consistently or always, **a single instantiation** suffices to tick the box [items 1, 2, 3, 4, 5, 7, 9, 10, 11, 12, 13, 17, 21, 23, 24, 25, 26, 27, 28, 30]. Item 8 cannot be placed into this dichotomy. This item aims at fluency within a time frame.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). London, UK: Macmillan.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663.

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

Ang-Aw, H. T., & Meng Goh, C. C. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal, 42*(1), 31-51. Doi: 10.1177/0033688210390226

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam, Netherlands: John Benjamins.

Bonk, W., & Oakey, G. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110. Doi: 10.1191/0265532203lt245oa

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1), 3-22. Doi: 10.1177/026553229701400102

Chapelle, C. (1999). From reading theory to testing practice. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading* (pp. 150-166). Cambridge, UK: Cambridge University Press.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

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

Cox, D. R., & Snell, E. J. (1971). On test statistics calculated from residuals. *Biometrika, 58*(3), 589-594. Doi: 10.1093/biomet/58.3.589

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

De Jong, J., & Linacre, J. M. (1993). Rasch estimation methods, statistical independence and global fit. *Rasch Measurement Transactions, 7*(2), 296-297.

Douglas, D. (2000). *Assessing speaking for specific purposes*. Cambridge, UK: Cambridge University Press.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221. Doi: 10.1207/s15434311laq0203_2

Euroexam International. (2019). *Euroexam detailed specifications*. Retrieved from https://rex.oh.gov.hu/FileDb/0015-AAAAAH/specifikacio/190206-003_EURO.doc

Euroexam International. (2020). *Euro B2 Webset. Speaking. Marking criteria*. Retrieved from http://www.euroexam.com/sites/network/files/private/practice-tests/euro_b2/72_web_b2_cri_speaking.pdf

Ffrench, A. (1999). *Study of qualitative differences between CPE individual and paired test formats (Internal UCLES EFL report)*. Cambridge, UK: University of Cambridge Local Examinations Syndicate.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238. Doi: 10.1177/026553229601300205

Fulcher, G. (2003). *Testing second language speaking*. New York, NY: Routledge. Doi: 10.4324/9781315837376

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching, 48*(2), 198-216. Doi: 10.1017/S0261444814000391

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29. Doi: 10.1177/0265532209359514

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, UK: Cambridge University Press.

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

Hosmer, D., & Lameshow, S. (1999). *Applied survival analysis: Time-to-event.* Hoboken, NJ: Wiley.

Huang, H.-T., Hung, S.-T., & Plakans, S. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27-49. Doi: 10.1177/0265532216677106

Kaftandjieva, F. (2004). Standard setting. 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 B)* (pp. 1-43). Strasbourg, France: Council of Europe/Language Policy Division.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests.* Arnhem, Netherlands: CITO, EALTA.

Kane, M. (2001). So much remains the same: Conception and status on validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.

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, Canada.

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

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the Body of Work method. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

Kormos, J. (1998). The use of verbal reports in L2 research: Verbal reports in L2 speech production research. *TESOL Quarterly, 32*(2), 353-358.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopaedia of language and education: Language testing and assessment* (Vol. 7) (pp. 121-130). Dortrecht, Netherlands: Kluwer.

Lukácsi, Z. (2020). Developing a level-specific checklist for assessing EFL writing. *Language Testing*. Advance online publication. Doi: 10.1177/0265532220916703

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71. Doi: 10.1177/026553229501200104

Luoma, S. (2004). *Assessing speaking.* Cambridge, UK: Cambridge University Press. Doi: 10.1017/CBO9780511733017

McNamara, T. F. (1995). Modelling performance: Opening Pandora's Box. *Applied Linguistics, 16*(2), 159-179. Doi: 10.1093/applin/16.2.159

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. Doi: 10.1177/0265532211430367

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692. Doi: 10.1093/biomet/78.3.691

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. D. (2016). *Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery. A preliminary comparison of test-taker and examiner behaviour.* IELTS Partners: British Council/Cambridge English Language Assessment/IDP: IELTS Australia. Retrieved from https://www.ielts.org/-/media/research-reports/ielts-partnership-research-paper-1.ashx

North, B. (1997). The development of a common framework scale or descriptors of language proficiency based on a theory of measurement. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 423-447). Jyväskylä, Finland: University of Jyväskylä.

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

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*(2), 143-154. Doi: 10.1016/S0346-251X(02)00002-7

Reckase, M. (2009). Standard setting theory and practice: Issues and difficulties. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 13-20). Arnhem, Netherlands: CITO, EALTA.

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing, 18*(3), 187-201. Doi: 10.1016/j.asw.2013.05.001

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing 18*(3), 275-302.

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. Doi: 10.2307/3588360

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12. Doi: 10.1093/elt/49.1.3

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: One-parameter logistic model*. Arnhem, Netherlands: CITO.

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing, 35*(1), 101-120. Doi: 10.1177/0265532216679451

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. Doi: 10.1177/026553229801500205

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave MacMillan. Doi: 10.1057/9780230514577

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

Wood, S. W. (2020). Public perception and communication around automated essay scoring. In D. Yan, A. A. Rupp & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 133-150). Boca Raton, FL: Chapman and Hall/CRC. Doi: 10.1201/9781351264808

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291-300. Doi: 10.1177/0265532210364643