Stefan O’Grady* and Özgür Taşkesen

# Developing a rating scale for integrated assessment of reading-into-writing skills

https://doi.org/10.1515/cercles-2022-2049
Received August 19, 2021; accepted October 1, 2021

**Abstract:** An important aspect of language assessment development is to create tasks that engage the competencies required in the target situation. For this reason, English-medium university entrance tests increasingly feature integrated reading-into-writing tasks as a way of enhancing target domain representation. Despite increased use of this task type, studies outlining the development of rating scales designed specifically for integrated assessments are rare. To this end, the study reports on the development of a rating scale to assess performance on an integrated reading-into-writing task as part of an English-medium university entrance test in Turkey. The authors discuss an iterative process of rating scale development involving examiner feedback, a focus group and many-facet Rasch measurement to validate the rating scale. The results indicate that the scale represents the integrated construct appropriately and reliably separates test candidates into different levels of ability.

**Keywords:** integrated skills assessment; many-facet Rasch measurement; rating scale development

## 1 Introduction

The importance of mediation skills in second language learning has recently been underscored by the Council of Europe (2018: 32) who argue “mediation language activities, (re)processing an existing text, occupy an important place in the normal linguistic functioning of our societies”. University admissions language tests increasingly assess mediation with integrated skills tasks that measure essential competences for undergraduate study (Chan et al. 2015; Dimova et al. 2020). To complete written assignments successfully, undergraduate students are expected to demonstrate uptake of course content and exhibit critical thinking skills. For this reason, researchers have argued that integrated

***Corresponding author: Stefan O’Grady**, University of St Andrews, St Andrews, UK, E-mail: so59@st-andrews.ac.uk
**Özgür Taşkesen,** Bilkent University, Ankara, Turkey, E-mail: taskesen@bilkent.edu.tr

Open Access. © 2022 Stefan O’Grady and Özgür Taşkesen, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

reading-into-writing tasks draw upon domain relevant cognitive processes, relating to discourse synthesis and textual representation, which are not engaged in independent writing tasks (Chan 2013, 2018; Weir et al. 2013). It follows that language ability profiles generated by language assessments may be made more informative by using combinations of integrated and independent writing tasks. Developing assessment criteria, i.e. a rating scale, for integrated tasks is complicated because the assessment involves two skills and criteria must describe the ways in which reading comprehension is evident in written production (Gebril and Plakans 2014; Ohta et al. 2018). A further challenge confronting test developers is that raters may misinterpret or even disregard criteria and score according to their own intuition, regardless of the training they receive (Fulcher 2003; Lumley 2005; Turner and Upshur 2002; Wisniewski 2017). Tailoring a rating scale to a local context by engaging raters in the scale development process represents an opportunity to resolve these issues because the specific pedagogical values and concerns that hold in the educational environment may be reflected in scale content (Dimova et al. 2020). To explore these issues further, the current study applied an iterative process of scale development that involved quantitative and qualitative methods to implement and validate a rating scale for an integrated reading-into-writing task in an English-medium university in Turkey.

## 2 Literature review

Integrated reading-into-writing assessment requires a clearly defined construct involving a model of proficiency that is comprehensively articulated in the rating scale. In a series of studies, Chan (2013, 2017, 2018) outlines the cognitive processes involved in reading-into-writing and presents five phases of integrated task completion relating to task conceptualisation, meaning construction, organising ideas, generating texts, and monitoring and revising. In Chan's integrated writing model, test takers develop and revise a macro level plan of the task, read the source text at global and local levels, select and organise relevant ideas, compose and edit their responses. Successful completion of integrated tasks involves each phase described in the model, but Chan's (2018) research findings indicate that proficient writers more comprehensively engage these processes when completing reading-into-writing tasks.

Despite the centrality of reading comprehension in Chan's model, research findings demonstrate that scores on integrated reading-into-writing tasks tend to reflect writing processes more than reading processes. Delaney (2008) correlated independent measures of writing and reading with scores on an integrated reading-into-writing task and concluded that the integrated scores were largely affected by variation in writing skills and only marginally by reading skills. In a mixed-methods study, Plakans and Gebril (2012: 32) reach a similar conclusion and report that for successful completion of the integrated reading-into-writing tasks used in their study "only a certain degree of comprehension is needed". These findings highlight the necessity of emphasising evidence of source text comprehension in rating scales to the extent that test candidates are unable to attain high scores without demonstrating it.

In addition to a theoretical basis, representation of the integrated skills construct requires rating scale development to draw upon empirical research findings (Knoch et al. 2021). Recent studies applying empirical methods of scale development for integrated skills assessment illustrate the value of basing scale content on rater analysis of task samples using appraisals, focus groups, and questionnaires (Chan et al. 2015; Shin and Ewert 2015). Rating scales based on rater analysis of task samples strengthen the correspondence between task performance, raters' expressed assessment criteria and scale content with the proposed effect that scores represent assessment constructs reliably and transparently. Although this is a promising direction for scale development, there is a significant caveat. Rating scales based on theory and empirical findings may still encourage rater effects such as central tendency (avoidance of the top and bottom scale levels) and halo effects (scores awarded to one category on an analytic rating scale influence ostensibly separate criteria) that distort measurement and create construct irrelevant variance (Messick 1989; Myford and Wolfe 2003, 2004).

In a recent scale development study, Ewert and Shin (2015) sought to improve the process of scoring integrated assessments by adapting Turner and Upshur's (2002) Empirical Binary Boundaries (EBB) approach that relies on empirically derived scales requiring binary decisions between score boundaries. Integrating rank-ordering procedures and rater analysis of test samples into binary criteria, EBB scales describe a construct that is referenced to a specific test taking population and task type and reflect raters' expressed criteria for making evaluative decisions. The binary format prevents central tendency and halo effects, thereby reducing construct irrelevant variance associated with conventional scales. Ewert and Shin (2015) argue that EBB scales bring raters' internal scoring criteria to the fore and hence promote assessment reliability.

The extent to which EBB content can simultaneously represent large groups of test takers exhibiting a range of performance features and serve as a useful point of reference for large groups of raters with different performance expectations has been debated (Ducasse 2009; Fulcher 2012). For rating scales to be applied as test developers intend, raters should recognise the content as relevant. To date, the potential to develop binary scales with larger groups of teachers has been unexplored in integrated skills research. This represents an important gap in the literature because the binary approach advocated by Turner and Upshur (2002) may enhance the validity of decisions based on integrated skills assessment (Ewert and Shin 2015). The current study seeks to fill this gap by adapting the binary approach to develop a scale for integrated assessment and investigating raters' perceptions of scale content.

# 3 Research questions

The literature indicates that scales developed according to Turner and Upshur's (2002) binary approach have potential in assessing the integrated skills construct reliably. However, questions remain about the suitability of the approach for large scale assessment involving larger groups of raters. To explore this issue, the current study seeks to answer the following questions.

1. What aspects of integrated task performance do different groups of raters identify as important for rating scale development?
2. Can these aspects be integrated into a rating scale to reliably discriminate between test candidates?

# 4 Method

## 4.1 Participants

The study involved 68 student participants registered in foundation courses in the English language preparatory program of a university in Turkey. The foundation course targets the B2 level on the Common European Framework as the university English language entry requirement (Council of Europe 2001). Student participants were between 18 and 20 years old and had been receiving English language tuition in classroom settings for nine years (gender information was not collected). English is taught as a foreign language in this context and there is little need to use the language outside classrooms.

In addition, 45 English language instructors were recruited to participate. Instructors had received specialised in-service examiner training and had over two years of examiner experience. The instructor group contained both international and local staff, all of whom had taken in-service and/or external teaching qualifications (Cambridge English ICELT, CELTA and DELTA) and MA degrees in fields related to language teaching.

## 4.2 Instruments

The instruments for data collection consisted of a reading-into-writing test featuring a source text, a rubric and a prompt (see Appendix) and a rating scale.

### 4.2.1 Reading-into-writing test

The reading-into-writing test was constructed according to specifications by a group of test developers working in the institution. The test features a source text, written by the test development group, consisting of approximately 320 words that presents two opposing views on a specific topic. Decisions about the topic and length of the source text were made by the university administration. A response of 150 words was deemed sufficient to explain two opposing views on a topic. A general topic that was deemed accessible to all students was selected and a text was composed by referring to online sources for background information. The text was analysed using CohMetrix version 3.0 (http://141.225.61.35/cohmetrix2017, accessed 1.06.2021) to establish the reading level. The text had a Flesch-Kincaid level of 9.3, which was deemed appropriate for a test targeting the B2 level (O'Sullivan et al. 2020).

Sample responses were collected during class hours as part of the students' preparation for the university entrance test. At the time of the study, classes had been in session for three weeks and students were becoming familiar with the test requirements. Participants were given 50 min to read the text and write responses. They did not have access to dictionaries or the internet during the task.

Representing academic writing at undergraduate level within the constraints of an English language proficiency test necessarily involves a level of compromise between construct coverage and practicality. Assessments must be completed within time limits and test content should not favour students with background knowledge (Chan 2013; Plakans 2012). Decisions to block external support such as dictionary and internet use also raise questions about domain representation because students typically have access to such resources when completing written tasks during undergraduate study. Written samples collected from integrated reading-into-writing tasks should be regarded as indicative of individuals' ability to comprehend, synthesise and interpret information and not fully representative of these abilities.

### 4.2.3 The rating scale

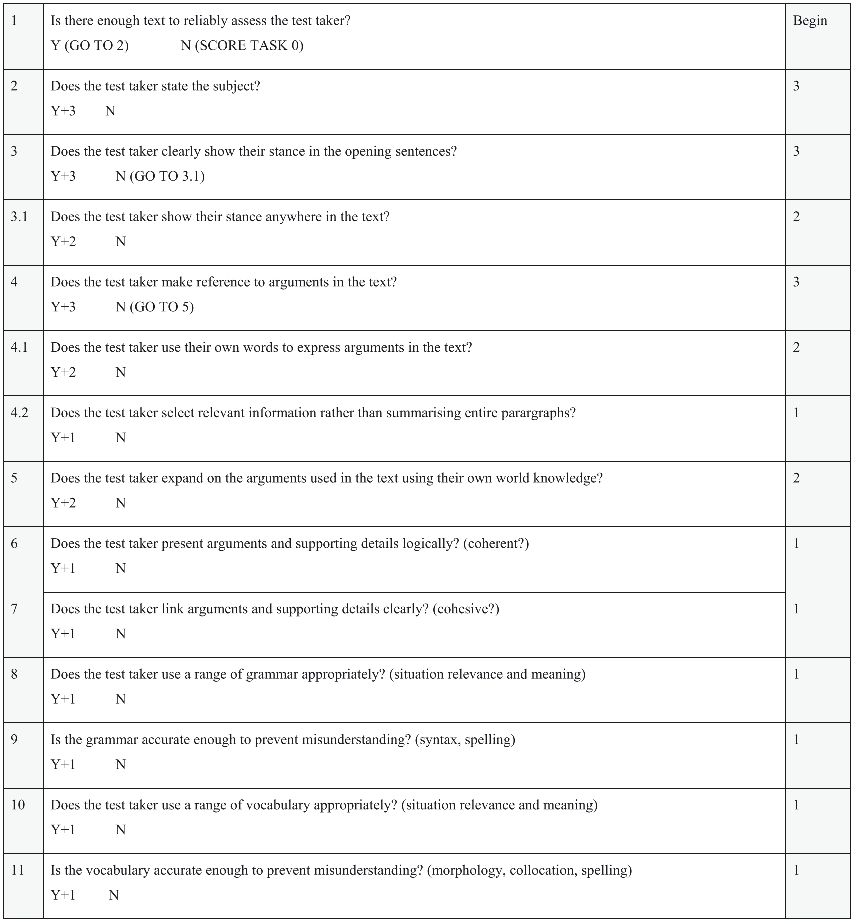
During the first stage of development, the researchers collaborated with the test development team to identify the necessary steps for successful task completion. In a group discussion, the following elements were identified:

- Students state the specific focus of the input text and express a stance toward the focus in the opening sentences to facilitate the rater’s interpretation of the argument
- Students support the stance with paraphrased ideas from the input text
- Students interpret and expand upon the ideas selected to support the stance
- Students organise the response coherently without summarising the whole text
- Students use language accurately
- Students use an appropriate range of language

Having identified the elements of successful task completion, the researchers wrote a series of binary questions and allocated points to each question ranging from one to three to reflect the relative importance of each element (see Figure 1).

The resulting binary scale was trialled by three examiners using three written responses selected to represent different levels of ability. Upon comparison of grades and discussion between the examiners, it became evident that while the responses exhibited clearly different levels of task success, each received the highest grade possible. The examiners suggested that it would be unusual to encounter responses in which necessary task elements were completely absent, as the binary scale implied, because foundation courses developed a high level of test awareness among the students. Raters would therefore rarely answer “no” to any of the binary questions. Rather than revising the criteria identified by the test development team, an additional level of performance was added to each category, which allowed raters to award partial grades. It was agreed that the rationale for awarding a partial score would be outlined during latter stages of the study. The resulting scale was presented in a 1-h training session during which a group of 12 raters discussed the integrated task, their expectations from the students, the scale content and three sample responses. Following this, each rater awarded a score to eight responses and provided feedback by annotating the scale. The annotations were collected, and amendments were made to the scale (see Figure 2). To exemplify, raters requested:

- a combination of range and accuracy of grammar and of vocabulary categories
- a statement format rather than a question format

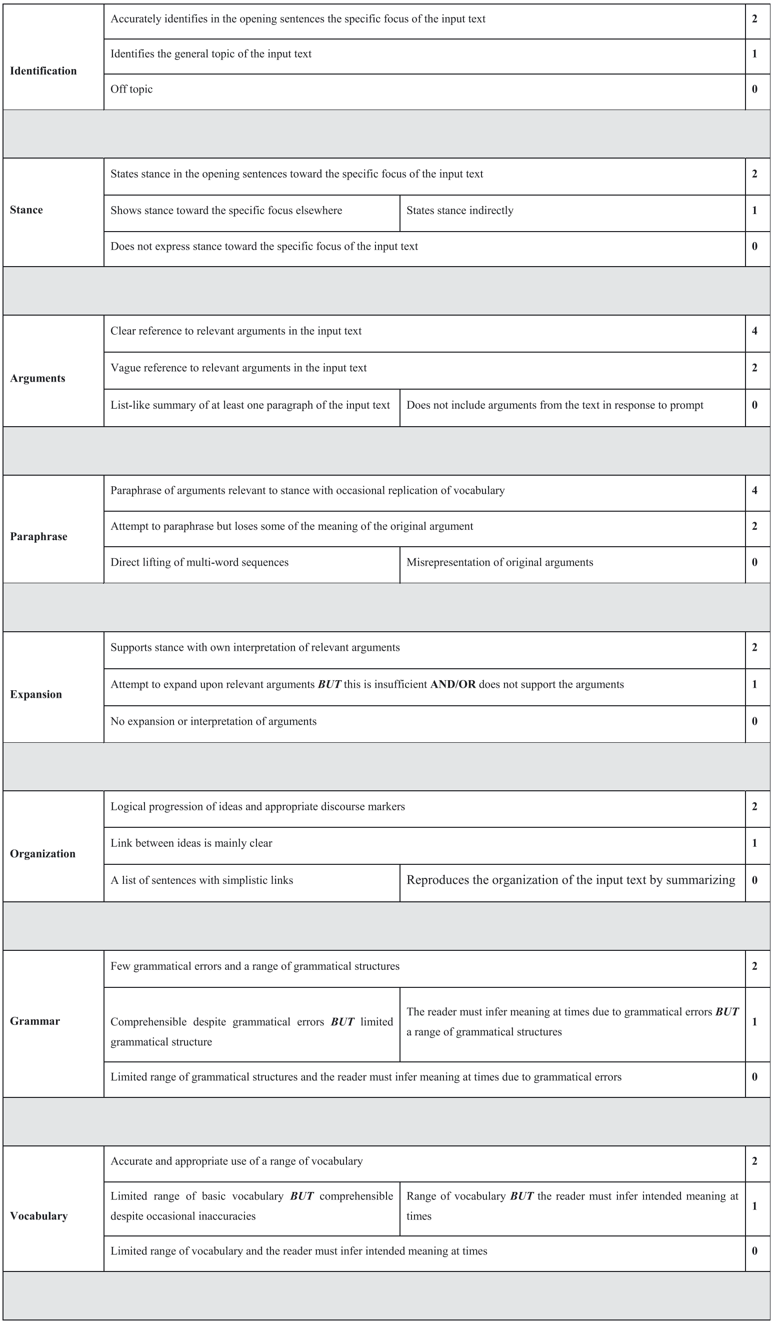


| | | |
|---|---|---|
| 1 | Is there enough text to reliably assess the test taker?<br>Y (GO TO 2) N (SCORE TASK 0) | Begin |
| 2 | Does the test taker state the subject?<br>Y+3 N | 3 |
| 3 | Does the test taker clearly show their stance in the opening sentences?<br>Y+3 N (GO TO 3.1) | 3 |
| 3.1 | Does the test taker show their stance anywhere in the text?<br>Y+2 N | 2 |
| 4 | Does the test taker make reference to arguments in the text?<br>Y+3 N (GO TO 5) | 3 |
| 4.1 | Does the test taker use their own words to express arguments in the text?<br>Y+2 N | 2 |
| 4.2 | Does the test taker select relevant information rather than summarising entire parargraphs?<br>Y+1 N | 1 |
| 5 | Does the test taker expand on the arguments used in the text using their own world knowledge?<br>Y+2 N | 2 |
| 6 | Does the test taker present arguments and supporting details logically? (coherent?)<br>Y+1 N | 1 |
| 7 | Does the test taker link arguments and supporting details clearly? (cohesive?)<br>Y+1 N | 1 |
| 8 | Does the test taker use a range of grammar appropriately? (situation relevance and meaning)<br>Y+1 N | 1 |
| 9 | Is the grammar accurate enough to prevent misunderstanding? (syntax, spelling)<br>Y+1 N | 1 |
| 10 | Does the test taker use a range of vocabulary appropriately? (situation relevance and meaning)<br>Y+1 N | 1 |
| 11 | Is the vocabulary accurate enough to prevent misunderstanding? (morphology, collocation, spelling)<br>Y+1 N | 1 |

**Figure 1:** Check list.

A group of 22 raters were standardised to the scale using three new samples and each rater was asked to mark 10 samples each after training. The raters were asked to annotate their copy of the scale by indicating the features they found irrelevant, misleading or difficult to use, and make recommendations. The recommendations were collected and categorised by the researchers and the scale was updated. At this stage, raters requested the following alterations:

- a combination of identification and stance categories
- a combination of arguments and paraphrase
- a total available score of 10



| | | | |
|---|---|---|---|
| **Identification** | Accurately identifies in the opening sentences the specific focus of the input text | | **2** |
| | Identifies the general topic of the input text | | **1** |
| | Off topic | | **0** |
| | | | |
| **Stance** | States stance in the opening sentences toward the specific focus of the input text | | **2** |
| | Shows stance toward the specific focus elsewhere | States stance indirectly | **1** |
| | Does not express stance toward the specific focus of the input text | | **0** |
| | | | |
| **Arguments** | Clear reference to relevant arguments in the input text | | **4** |
| | Vague reference to relevant arguments in the input text | | **2** |
| | List-like summary of at least one paragraph of the input text | Does not include arguments from the text in response to prompt | **0** |
| | | | |
| **Paraphrase** | Paraphrase of arguments relevant to stance with occasional replication of vocabulary | | **4** |
| | Attempt to paraphrase but loses some of the meaning of the original argument | | **2** |
| | Direct lifting of multi-word sequences | Misrepresentation of original arguments | **0** |
| | | | |
| **Expansion** | Supports stance with own interpretation of relevant arguments | | **2** |
| | Attempt to expand upon relevant arguments ***BUT*** this is insufficient **AND/OR** does not support the arguments | | **1** |
| | No expansion or interpretation of arguments | | **0** |
| | | | |
| **Organization** | Logical progression of ideas and appropriate discourse markers | | **2** |
| | Link between ideas is mainly clear | | **1** |
| | A list of sentences with simplistic links | Reproduces the organization of the input text by summarizing | **0** |
| | | | |
| **Grammar** | Few grammatical errors and a range of grammatical structures | | **2** |
| | Comprehensible despite grammatical errors ***BUT*** limited grammatical structure | The reader must infer meaning at times due to grammatical errors ***BUT*** a range of grammatical structures | **1** |
| | Limited range of grammatical structures and the reader must infer meaning at times due to grammatical errors | | **0** |
| | | | |
| **Vocabulary** | Accurate and appropriate use of a range of vocabulary | | **2** |
| | Limited range of basic vocabulary ***BUT*** comprehensible despite occasional inaccuracies | Range of vocabulary ***BUT*** the reader must infer intended meaning at times | **1** |
| | Limited range of vocabulary and the reader must infer intended meaning at times | | **0** |
| | | | |

**Figure 2:** Rating Scale 2.

To achieve this, the expansion and organisation categories were combined with the understanding that this feature of the scale would be discussed in the focus group (see Figure 3). This iteration of the scale was used in the main study.

## 4.3 Procedure

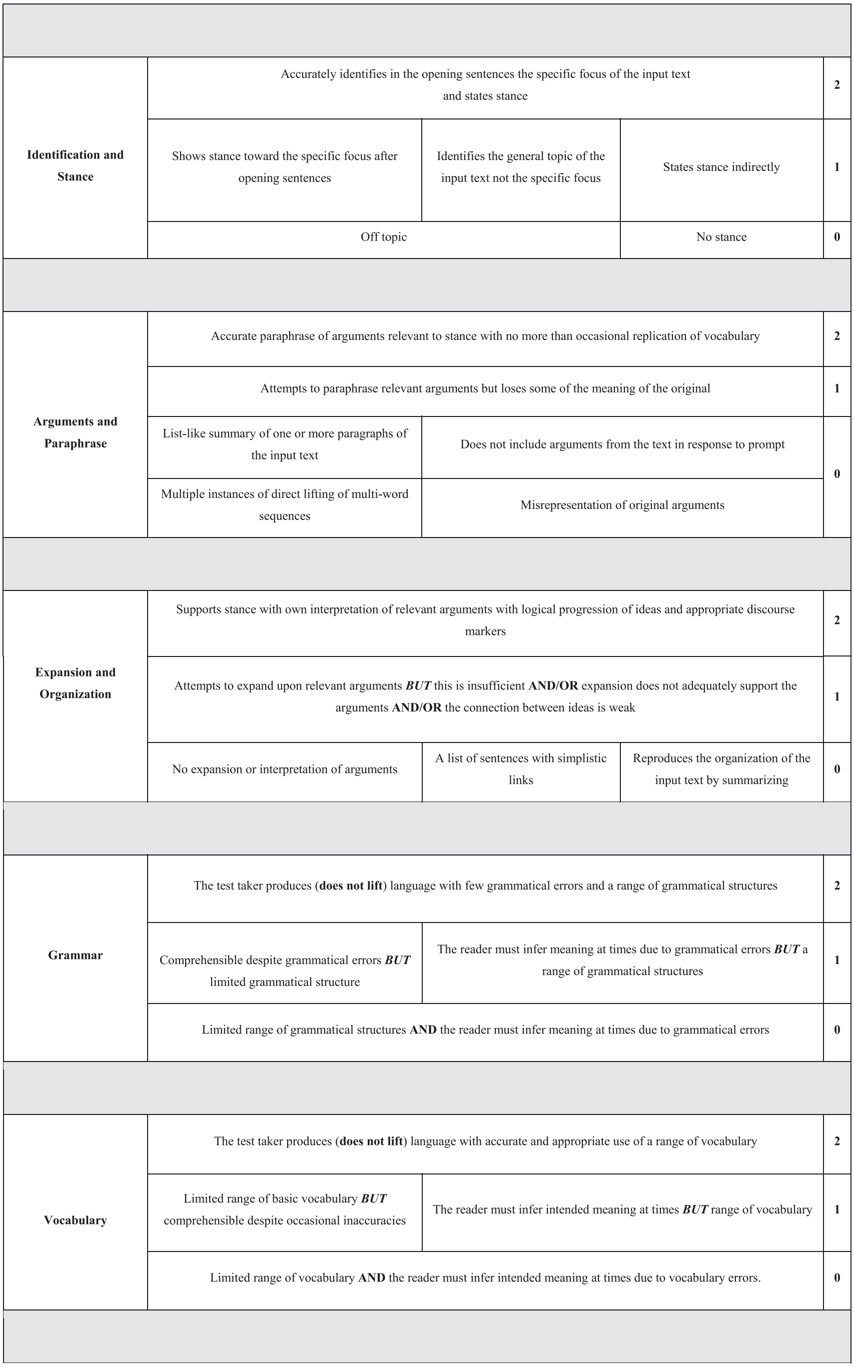
The rating scale was presented to 35 raters in a 1-h training session in which the researchers described scale development and gave them the opportunity to ask questions. Following training, a data matrix involving the 35 raters was created. The matrix ensured that 35 responses, which had not featured during development, were marked by at least six different examiners using the new scale. To ensure sufficient connectivity in the data to run many facet Rasch measurement (MFRM), the two researchers also marked the 35 papers independently. This resulted in a total of 455 grades.

## 4.4 Data analysis

Data analysis was conducted using the Facets programme, a software for conducting MFRM (Linacre 2019). The MFRM applied a “criterion-related three facet partial credit” model to investigate test taker scores, rater behaviour, and the functioning of the rating scale categories (for a detailed description see Eckes 2011: 128). This model was selected because the categories on the scale were not assumed to be equivalent; raters may demonstrate different levels of severity and consistency when assigning scores to each category. The partial credit model generates statistics describing the functioning of the five separate categories on the scale. These statistics can be used to make comparisons between the levels of difficulty associated with attaining a high score on each category, and the levels of consistency raters applied when assigning scores to each category (Wind 2020).

## 4.5 Focus group

Based on their availability, five rater participants agreed to be recorded in an hour-long focus group. Participants were Turkish nationals that were very familiar with the test taking population, had experience preparing students for the university entrance exam and were regarded as representing the views of the overall teaching faculty. In a semi-structured approach, the raters were asked to reflect on their experience of applying the rating scale. Upon completion



| | | | | |
|---|---|---|---|---|
| | | | | |
| **Identification and Stance** | Accurately identifies in the opening sentences the specific focus of the input text and states stance | | | **2** |
| | Shows stance toward the specific focus after opening sentences | Identifies the general topic of the input text not the specific focus | States stance indirectly | **1** |
| | Off topic | | No stance | **0** |
| | | | | |
| **Arguments and Paraphrase** | Accurate paraphrase of arguments relevant to stance with no more than occasional replication of vocabulary | | | **2** |
| | Attempts to paraphrase relevant arguments but loses some of the meaning of the original | | | **1** |
| | List-like summary of one or more paragraphs of the input text | Does not include arguments from the text in response to prompt | | **0** |
| | Multiple instances of direct lifting of multi-word sequences | Misrepresentation of original arguments | | |
| | | | | |
| **Expansion and Organization** | Supports stance with own interpretation of relevant arguments with logical progression of ideas and appropriate discourse markers | | | **2** |
| | Attempts to expand upon relevant arguments ***BUT*** this is insufficient **AND/OR** expansion does not adequately support the arguments **AND/OR** the connection between ideas is weak | | | **1** |
| | No expansion or interpretation of arguments | A list of sentences with simplistic links | Reproduces the organization of the input text by summarizing | **0** |
| | | | | |
| **Grammar** | The test taker produces (**does not lift**) language with few grammatical errors and a range of grammatical structures | | | **2** |
| | Comprehensible despite grammatical errors ***BUT*** limited grammatical structure | The reader must infer meaning at times due to grammatical errors ***BUT*** a range of grammatical structures | | **1** |
| | Limited range of grammatical structures **AND** the reader must infer meaning at times due to grammatical errors | | | **0** |
| | | | | |
| **Vocabulary** | The test taker produces (**does not lift**) language with accurate and appropriate use of a range of vocabulary | | | **2** |
| | Limited range of basic vocabulary ***BUT*** comprehensible despite occasional inaccuracies | The reader must infer intended meaning at times ***BUT*** range of vocabulary | | **1** |
| | Limited range of vocabulary **AND** the reader must infer intended meaning at times due to vocabulary errors. | | | **0** |
| | | | | |

**Figure 3:** Rating Scale 3.

of the focus group, the recording was analysed by the researchers to identify themes that would help interpret the quantitative findings and highlight potential scale revisions.

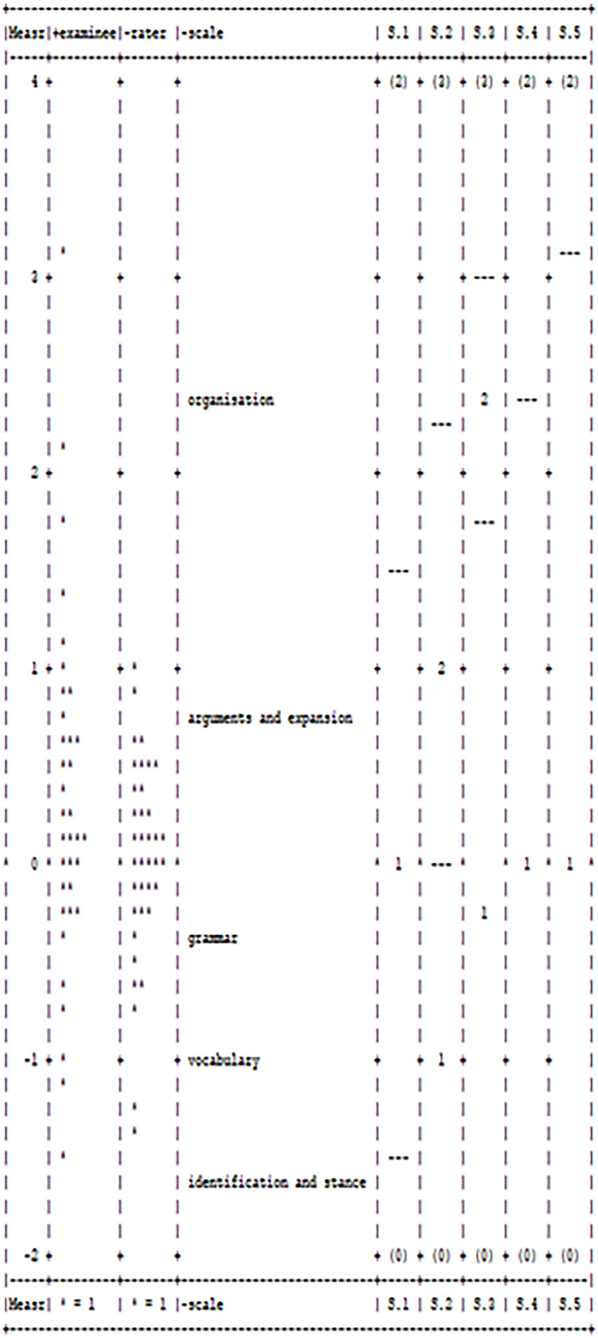
# 5 Results

## 5.1 Statistical analysis

Initial analysis indicated misfit of data to the Rasch model: one rater had recorded an infit mean square value that exceeded 2.00, which is a level of misfit that has a distorting effect on results. Although data deletion is not recommendable for scale validation studies (all instruments would be validated if researchers merely deleted inconvenient data), the single misfitting rater represented a very small proportion of the score data (4% of the total scores). Reliable measures could be attained for most raters. To resolve this issue, the misfitting rater was removed and a second analysis was completed. An anchor file was created from this second analysis to preserve all fitting data values. Upon this, the misfitting rater and the anchor file data were combined, and an analysis was completed successfully (for an explanation of this procedure please see Linacre 2020).

Results are first presented in a Wright Map (see Figure 4); a series of vertical histograms that calibrate the facets under investigation on a common logit scale (the first column). The second column rank orders test takers (each test taker is represented by an asterisk) according to ability measures; the most able test takers are located at the top of the map. Raters are arranged in order according to the level of severity they demonstrated when assigning scores in the third column: higher points on the map represent higher severity. The fourth column locates the rating scale categories on the logit scale by placing categories that were scored most severely (i.e. it was difficult to achieve high scores) at the upper end of the map. The following columns (S1, S2, S3, S4, S5) represent the categories on the rating scale, the values in parenthesis indicate scores that were rarely assigned. Horizontal lines in these columns represent thresholds between the scores on the categories (i.e. the point on the logit scale where a higher score is expected).

Rating scale category statistics are presented in Table 1, which demonstrates that the five categories recorded different average measure values on the logit scale. This indicates that the raters applied different levels of severity when assessing the different aspects of task performance described in the scale and may be taken as evidence that there was no cross over effect between the categories



```
+---------------------------------------------------------------------------------+
|Measr|+examinee|-rater |-scale                     | S.1 | S.2 | S.3 | S.4 | S.5 |
|-----+---------+-------+---------------------------+-----+-----+-----+-----+-----|
|   4 +         +       +                           + (2) + (3) + (3) + (2) + (2) |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     | *       |       |                           |     |     |     |     | --- |
|   3 +         +       +                           +     +     + --- +     +     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       | organisation              |     |     |  2  | --- |     |
|     |         |       |                           |     | --- |     |     |     |
|     | *       |       |                           |     |     |     |     |     |
|   2 +         +       +                           +     +     +     +     +     |
|     |         |       |                           |     |     |     |     |     |
|     | *       |       |                           |     |     | --- |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           | --- |     |     |     |     |
|     | *       |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     | *       |       |                           |     |     |     |     |     |
|   1 + *       + *     +                           +     +  2  +     +     +     |
|     | **      | *     |                           |     |     |     |     |     |
|     | *       |       | arguments and expansion   |     |     |     |     |     |
|     | ***     | **    |                           |     |     |     |     |     |
|     | **      | ****  |                           |     |     |     |     |     |
|     | *       | **    |                           |     |     |     |     |     |
|     | **      | ***   |                           |     |     |     |     |     |
|     | ****    | ***** |                           |     |     |     |     |     |
*   0 * ***     * ***** *                           *  1  * --- *     *  1  *  1  *
|     | **      | ****  |                           |     |     |     |     |     |
|     | ***     | ***   |                           |     |     |  1  |     |     |
|     | *       | *     | grammar                   |     |     |     |     |     |
|     |         | *     |                           |     |     |     |     |     |
|     | *       | **    |                           |     |     |     |     |     |
|     | *       | *     |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|  -1 + *       +       + vocabulary                +     +  1  +     +     +     |
|     | *       |       |                           |     |     |     |     |     |
|     |         | *     |                           |     |     |     |     |     |
|     |         | *     |                           |     |     |     |     |     |
|     | *       |       |                           | --- |     |     |     |     |
|     |         |       | identification and stance |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|     |         |       |                           |     |     |     |     |     |
|  -2 +         +       +                           + (0) + (0) + (0) + (0) + (0) |
|-----+---------+-------+---------------------------+-----+-----+-----+-----+-----|
|Measr| * = 1   | * = 1 |-scale                     | S.1 | S.2 | S.3 | S.4 | S.5 |
+---------------------------------------------------------------------------------+
```

**Figure 4:** Wright map.

**Table 1:** Rating scale category statistics.

| Total |  | Measure | Model S.E. | Infit |  | Outfit |  | Estim. Discrim | Correlation |  | Categories |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | Count |  |  | MnSq | Zstd | MnSq | Zstd |  | PtMea | PtExp |  |
| 728 | 453 | −1.68 | 0.10 | 1.08 | 1.1 | 1.24 | 2.4 | 0.84 | 0.32 | 0.42 | Identification and stance |
| 602 | 453 | 0.72 | 0.07 | 1.03 | 0.5 | 1.04 | 0.6 | 0.96 | 0.56 | 0.58 | Arguments and expansion |
| 384 | 453 | 2.33 | 0.13 | 0.95 | −0.4 | 0.99 | 0.0 | 1.03 | 0.38 | 0.33 | Organisation |
| 516 | 453 | −0.40 | 0.11 | 1.02 | 0.3 | 1.03 | 0.4 | 0.98 | 0.45 | 0.43 | Grammar |
| 528 | 453 | −0.97 | 0.12 | 0.97 | −0.3 | 1.02 | 0.2 | 1.00 | 0.44 | 0.40 | Vocabulary |
| 551.6 | 453 | 0.00 | 0.11 | 1.01 | 0.2 | 1.06 | 0.7 |  | 0.43 |  | Mean |
| 112.8 | 0.0 | 1.40 | 0.02 | 0.05 | 0.6 | 0.09 | 0.9 |  | 0.08 |  | SD (pop) |
| 126.1 | 0.0 | 1.57 | 0.02 | 0.05 | 0.7 | 0.10 | 1.0 |  | 0.09 |  | SD (sample) |

(Myford and Wolfe 2003, 2004). The infit mean square statistics, indicating the amount of randomness in the scores (values above 1.50 indicate unpredictability Linacre 2021), show that the allocation of scores on the five categories was consistent and the measure values associated with each category are reliable.

Test taker measure values ranged from −1.48 to 3.16 on the logit scale, indicating that there was a wide range of test taker ability in the sample. Test takers were separated into 3.23 statistically distinct levels of ability (the strata value was 4.64) and the reliability statistic was 0.91 (Wright and Masters 2002). This suggests that the raters were able to use the rating scale to place test takers into a low-level, mid-level and high-level of reading-into-writing ability with high levels of reliability.

Rater statistics demonstrating levels of severity and consistency are presented in Table 2. The range of rater severity by logit measure values was −1.33 to 0.98. At its most extreme, the difference between the most lenient and most severe raters was 2.31 logits. This suggests that there were substantial differences in rater severity measures. However, the separation statistic was 1.54 (strata = 2.38), which indicates that raters were grouped into less than two statistically distinct severity levels after accounting for measurement error (Wright and Masters 2002). This suggests that differences in rater severity played a negligible role in determining the test results.

Inconsistency in the levels of severity that individual raters applied may be examined using fit statistics (Myford and Wolfe 2003, 2004). Most raters ($n = 34$) recorded infit values within commonly applied thresholds of 0.50–1.50 (Linacre 2021). However, the true range of infit mean square statistics was from 0.59, indicating a slight overfit to the Rasch measurement model ("overfit": the scores are predictable, and the rater may not be using the full range of the scale; Eckes 2011: 102), to 2.24 indicating misfit ("misfit": the scores are unpredictable and the rater is scoring arbitrarily, e.g. the rater that was removed from the anchored analysis; Eckes 2011: 68). Overall, the results indicate that most raters are able to use the scale reliably but also require extra training to improve consistency.

Rating scale statistics representing the functioning of the rating scale categories are presented in Table 3, which reports the total number of test takers assigned to each band on each category and the proportion of the sample that this number represents. The rating scale categories recorded good levels of model fit. Regarding the functioning of the rating scale bands, Rasch-Thurstone thresholds showed that higher scores on the rating scale were associated with higher measure values and that the scale progressed monotonically (Eckes 2011). However, the lowest band levels in each category were infrequently used. For example, in the organisation category and arguments and expansion category,

**Table 2:** Rater severity and fit statistics.

| Rater | Measure | Model S.E. | Infit | | Outfit | |
|---|---|---|---|---|---|---|
| | | | MnSq | ZStd | MnSq | ZStd |
| 34 | −1.33 | 0.29 | 0.71 | −1.4 | 0.5 | −1.6 |
| 17 | −1.27 | 0.29 | 0.59 | −2 | 1.02 | 0.1 |
| 33 | −0.79 | 0.28 | 1.44 | 1.8 | 1.17 | 0.6 |
| 2 | −0.64 | 0.28 | 1.12 | 0.6 | 1.14 | 0.5 |
| 35 | −0.63 | 0.28 | 1.04 | 0.2 | 1.47 | 1.4 |
| 7 | −0.54 | 0.3 | 0.7 | −1.3 | 0.88 | −0.2 |
| 3 | −0.43 | 0.28 | 0.78 | −0.9 | 0.75 | −0.8 |
| 28 | −0.28 | 0.28 | 1.26 | 1.1 | 1.26 | 0.9 |
| 8 | −0.26 | 0.28 | 0.9 | −0.4 | 1.16 | 0.6 |
| 21 | −0.24 | 0.28 | 1.06 | 0.3 | 1.11 | 0.4 |
| 19 | −0.16 | 0.28 | 1.37 | 1.5 | 1.36 | 1.1 |
| 31 | −0.12 | 0.27 | 0.86 | −0.6 | 1.04 | 0.2 |
| 9 | −0.11 | 0.28 | 1.11 | 0.5 | 1.02 | 0.1 |
| 20 | −0.08 | 0.28 | 1.32 | 1.4 | 1.17 | 0.6 |
| 36 | −0.05 | 0.16 | 0.74 | −2.2 | 0.74 | −1.7 |
| 5 | −0.04 | 0.28 | 0.74 | −1.2 | 0.88 | −0.3 |
| 23 | −0.04 | 0.28 | 1.94 | 3.4 | 2.06 | 3 |
| 24 | 0 | 0.28 | 2.24 | 4.2 | 3.49 | 5.6 |
| 32 | 0.05 | 0.28 | 0.73 | −1.2 | 0.96 | 0 |
| 13 | 0.07 | 0.28 | 1.36 | 1.5 | 1.34 | 1.2 |
| 1 | 0.08 | 0.28 | 0.9 | −0.4 | 0.67 | −1.3 |
| 37 | 0.09 | 0.16 | 0.73 | −2.4 | 0.74 | −1.8 |
| 29 | 0.14 | 0.28 | 1.1 | 0.5 | 0.98 | 0 |
| 18 | 0.17 | 0.28 | 0.79 | −0.9 | 0.88 | −0.2 |
| 26 | 0.21 | 0.3 | 1.01 | 0.1 | 0.94 | 0 |
| 10 | 0.22 | 0.28 | 1.35 | 1.5 | 1.17 | 0.6 |
| 22 | 0.22 | 0.28 | 0.75 | −1.2 | 0.91 | −0.2 |
| 4 | 0.33 | 0.28 | 0.59 | −2.1 | 0.47 | −2.3 |
| 16 | 0.4 | 0.28 | 1.16 | 0.7 | 0.98 | 0 |
| 27 | 0.45 | 0.29 | 1.09 | 0.4 | 1.13 | 0.5 |
| 12 | 0.47 | 0.28 | 0.87 | −0.5 | 0.73 | −0.9 |
| 30 | 0.5 | 0.28 | 1.22 | 1 | 1.66 | 2.1 |
| 11 | 0.52 | 0.27 | 1.42 | 1.8 | 1.41 | 1.4 |
| 6 | 0.58 | 0.29 | 1.07 | 0.3 | 0.8 | −0.5 |
| 14 | 0.67 | 0.28 | 0.85 | −0.6 | 0.65 | −1.5 |
| 15 | 0.85 | 0.27 | 0.99 | 0 | 1.04 | 0.2 |
| 25 | 0.98 | 0.28 | 1.09 | 0.4 | 1.03 | 0.2 |
| Mean | 0.00 | 0.27 | 1.05 | 0.1 | 1.10 | 0.2 |

Inter-Rater agreement opportunities: 15,490 Exact agreements: 9,476 = 61.2% Expected: 9,453.8 = 61.0%.

**Table 3:** Rating scale category level statistics.

| | Score | Total used | % | Avge Meas | Exp. Meas | Outfit MnSq | Rasch-Thurstone Thresholds |
|---|---|---|---|---|---|---|---|
| Identification and stance | 0 | 13 | 3% | 0.44 | 0.87 | 0.87 | |
| | 1 | 152 | 34% | 1.74 | 1.51 | 1.5 | −1.34 |
| | 2 | 288 | 64% | 2.24 | 2.34 | 1.1 | 1.33 |
| Arguments and expansion | 0 | 68 | 15% | −1.20 | −1.26 | 1.0 | |
| | 1 | 203 | 45% | -0.68 | -0.65 | 1.1 | −2.14 |
| | 2 | 147 | 32% | 0.09 | 0.08 | 1.0 | 0.00 |
| | 3 | 35 | 8% | 0.95 | 1.04 | 1.2 | 2.13 |
| Organization | 0 | 73 | 16% | −2.91 | −2.72 | 0.9 | |
| | 1 | 377 | 83% | −1.82 | −1.87 | 0.9 | −3.98 |
| Grammar | 0 | 28 | 6% | -0.06 | -0.16 | 1.2 | |
| | 1 | 334 | 74% | 0.56 | 0.58 | 1.0 | −2.39 |
| | 2 | 91 | 20% | 1.65 | 1.58 | 1.0 | 2.38 |
| Vocabulary | 0 | 10 | 2% | 0.48 | 0.36 | 1.3 | |
| | 1 | 358 | 79% | 1.11 | 1.13 | 0.9 | −3.06 |
| | 2 | 85 | 19% | 2.26 | 2.26 | 1.0 | 3.05 |

16 and 15% of the scores were assigned to the lowest band. In the remaining categories, identification and stance, grammar, and vocabulary, the lowest band constituted 3, 6, and 2% of the scores respectively. These values indicate problems with the band level descriptors that required clarification in the focus group.

## 5.2 Focus group

At the beginning of the session, the participants were asked to describe their general impressions of the scale. The participants commented on how the scores they awarded reflected their own evaluations of the quality of the students’ texts:

- *Participant 2. Overall I felt that I was giving higher grades than I should be – it’s still a pass paper but still I would have given it lower if I had not assessed it analytically*
- *Participant 3. I found the opposite with this criteria … the grades would be lower*
- *Participant 4. The good papers I gave higher, the fail papers I gave lower*

During this discussion it became evident that the participants felt reluctant about awarding the zero band and some suggested that they had ignored the scale to avoid giving a zero:

- *Participant 1: I don't feel like giving a zero. It's very hard to give a zero.*
- *Participant 2: I very rarely gave a zero – maybe the paper deserved it*
- *Participant 3: I felt that way about grammar and vocabulary because the student made mistakes but it didn't feel like a zero*
- *Participant 2: I was reluctant to give a zero.*
- *Participant 4: Yes. I felt the same but I used the criteria*
- *Participant 2: işte* (Turkish) *I didn't* (use the criteria)

These comments offer insights into the rating scale statistics as scores of zero were very rarely awarded by the raters (see Table 3). In addition, participants commented that the descriptors associated with the zero band in the identification and stance category did not reflect their own observations:

- *Participant 2: For identification and stance nobody gets a zero because it's never completely off topic*

The grammar and vocabulary categories contained descriptors at the zero-band stating that the reader *must* infer meaning, which the participants regarded as too severe. This may explain why the zero band was so infrequently awarded in these categories:

- *Participant 5: we have to have a slightly weaker word* (than "must")
- *Researcher: for vocabulary?*
- *Participant 5: it's the same actually*

The participants suggested that they might have been less reluctant to award a score of zero on the grammar and vocabulary categories if the scale had allowed for more flexibility. In the same way, discussing the lack of variation in the organisation scores (only 16% of the scores were zero), a participant commented that the descriptor "reproduces the organisation of the input text by summarising" was confusing because summarising was regarded as evidence of successful text comprehension and not a limitation of organisation. Another participant made the comment that attaining full marks on the organisation category was simple because:

- *Participant 4: we don't expect much in terms of organisation*

An important aim of the focus group was to gather evidence about how participants had assessed students' comprehension of the input text. The focus group leader explained that requiring students to state their stance in the opening statements was a way of gauging in the opening lines whether they had successfully understood the topic:

- *Researcher: we would like to see students' comprehension of the text… first of all can the student understand what we have in the text the specific focus…we want to be so rigid in terms of understanding students' comprehension in the very beginning sentences*
- *Participant 2: Teaching them to express their stance in the beginning I think is better*

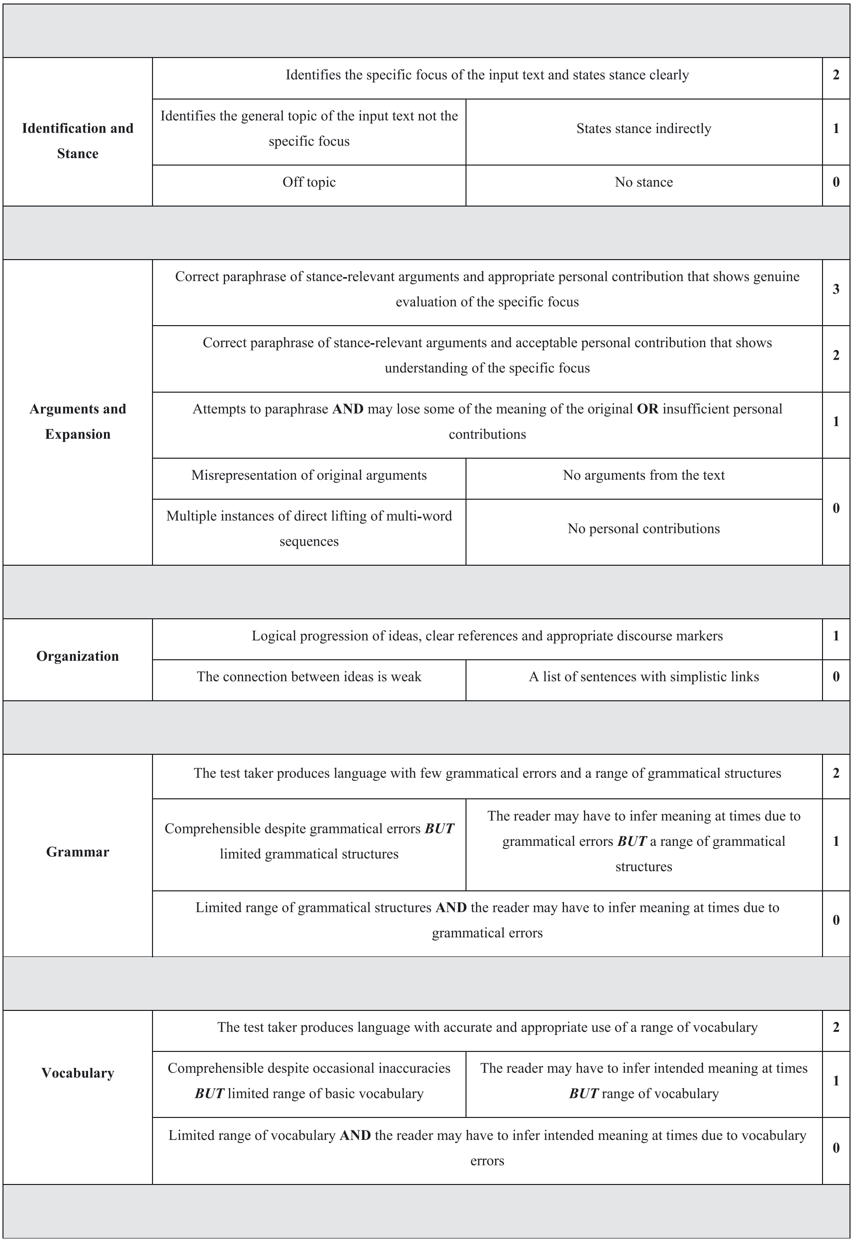
However, discussing the wording of the bands referring to comprehension, a participant explained:

- *Participant 1: I have a problem with the word interpretation. They might understand something completely different and they might support it.*

Throughout the discussion, participants explained that they were fully aware that the scale content would come under intense scrutiny from both teachers and students and that any ambiguous terminology required modification. The participants agreed that the term "misrepresentation" was more appropriate to describe evidence of the students' reading comprehension than "interpretation" because evidence of reading comprehension could only be found in students' texts. This led to a discussion about the possibility of lifting directly from the text and penalisation. One participant explained that the scale had made her more careful when checking for lifting, whereas others expressed concern about the level of lifting that could be tolerated, specifically with reference to the replication of vocabulary. On this, the raters agreed that teachers would need training:

- *Participant 1: while marking I had to re-read the original many times to see if the student is paraphrasing is lifting … it took me longer … it's being more meticulous actually*
- *Participant 2: sometimes depending on the topic students may have to take some vocabulary as they are*
- *Participant 5: look the student is repeating the same word and I have to penalise*
- *Participant 5: teachers should be taught what paraphrasing is*
- *Researcher: what do we understand by paraphrasing*
- *Participant 5: keeping the same meaning*

Based on the focus group comments, a final iteration of the scale was developed (see Figure 5).



| | | | |
|---|---|---|---|
| | | | |
| **Identification and Stance** | Identifies the specific focus of the input text and states stance clearly | | **2** |
| | Identifies the general topic of the input text not the specific focus | States stance indirectly | **1** |
| | Off topic | No stance | **0** |
| | | | |
| **Arguments and Expansion** | Correct paraphrase of stance-relevant arguments and appropriate personal contribution that shows genuine evaluation of the specific focus | | **3** |
| | Correct paraphrase of stance-relevant arguments and acceptable personal contribution that shows understanding of the specific focus | | **2** |
| | Attempts to paraphrase **AND** may lose some of the meaning of the original **OR** insufficient personal contributions | | **1** |
| | Misrepresentation of original arguments | No arguments from the text | **0** |
| | Multiple instances of direct lifting of multi-word sequences | No personal contributions | |
| | | | |
| **Organization** | Logical progression of ideas, clear references and appropriate discourse markers | | **1** |
| | The connection between ideas is weak | A list of sentences with simplistic links | **0** |
| | | | |
| **Grammar** | The test taker produces language with few grammatical errors and a range of grammatical structures | | **2** |
| | Comprehensible despite grammatical errors ***BUT*** limited grammatical structures | The reader may have to infer meaning at times due to grammatical errors ***BUT*** a range of grammatical structures | **1** |
| | Limited range of grammatical structures **AND** the reader may have to infer meaning at times due to grammatical errors | | **0** |
| | | | |
| **Vocabulary** | The test taker produces language with accurate and appropriate use of a range of vocabulary | | **2** |
| | Comprehensible despite occasional inaccuracies ***BUT*** limited range of basic vocabulary | The reader may have to infer intended meaning at times ***BUT*** range of vocabulary | **1** |
| | Limited range of vocabulary **AND** the reader may have to infer intended meaning at times due to vocabulary errors | | **0** |
| | | | |

**Figure 5:** Rating Scale 4.

# 6 Discussion and conclusion

The aim of this study was to develop and validate a rating scale to assess test taker performance on an integrated reading-into-writing task as part of a university entrance test. The rating scale content was supplied by members of the rater population and feedback was collected at various stages of the project to inform development. The scale was initially conceived as a binary scale but as the project progressed, the scale evolved into a checklist with different performance descriptors for the separate categories raters had identified as important. This was an unexpected development that reflected an inability to discriminate sufficiently between test candidates using binary criteria. Ewert and Shin (2015) indicated that a limitation of the binary format in their study was insufficient representation of the range of criteria their participants identified as important when constructing the EBB scale. In contrast, in the current context reducing complex decisions about requisite elements of successful integrated task completion to binary responses may be impractical because test takers are familiar with task requirements and are unlikely to completely omit necessary steps during task completion.

Research question one asked about the aspects of task performance that raters consider when evaluating task samples. Reflecting on their expectations and experiences of scoring, raters described various characteristics of the samples to develop the scale. The band level descriptors were refined in a focus group in which participants identified ambiguous and misleading scale content. The raters identified aspects of performance relating to students' ability to identify the input text topic and express a stance, paraphrase arguments and expand on the ideas contained in the input text, organise the response, and use a range of grammar and vocabulary accurately. These criteria overlap substantially with Chan's (2013, 2017, 2018) model of integrated writing. The criteria involving identification of the topic, expression of a stance toward the topic, paraphrase and personal contribution to the argument are most relevant to meaning construction processes in the Chan model, such as careful reading and generating new meaning. The organisation criteria assess the coherent organisation of ideas in integrated writing tasks described by Chan (2013, 2017, 2018). With reference to the latter phases of the model, increased accuracy of language use results from the opportunity to monitor and revise (Ellis and Yuan 2004) and this relates to the scale categories describing accuracy of paraphrase, grammar and vocabulary. The scale reflects the integrated task processes described in Chan's model and hence represents the assessment construct appropriately.

The second research question asked whether the rating scale was applied reliably to discriminate between test candidates. MFRM demonstrated that the five categories in the rating scale were associated with different measure values on the logit scale and this was interpreted as evidence that the separate categories were assessing different aspects of writing ability (Eckes 2011). In addition, examination of Rasch-Thurstone thresholds showed that advances in the rating scale were associated with increases in ability in the latent trait. Further, the increases in scores on the rating scale proceeded at regular intervals along the logit scale. This was an encouraging finding that indicates raters interpreted the test construct in the way that the designers had intended (Linacre 2021). Test takers were separated into approximately three statistically distinct levels of ability: a weak group, a strong group and a mid-level group. The test may therefore be applied to separate students into those that require further tuition, those that may directly begin undergraduate study and those that may require extra language support during the freshman year. A tendency that emerged in the MFRM was that the zero band on the scale was rarely used. Comments in the focus group confirmed that raters were reluctant to award scores of zero. This limitation may have reduced discrimination between test takers and requires attention. In future training programmes, the value of the zero band as a deterrent against producing off-topic texts or direct lifting will need to be emphasised to raters. Furthermore, the adjustments to the vocabulary and grammar criteria (see Section 5.2) requested in the focus group may increase use of the zero band on these categories and this will be monitored when assessment data becomes available.

Reliability was investigated on two levels: the first was the internal consistency individual raters applied when scoring the samples, and the second was the degree to which different raters applied similar levels of severity when scoring. To examine the levels of consistency in the score data, the infit mean square statistics were consulted. Most raters demonstrated acceptable levels of model fit (infit mean square values were within a range of 0.50–1.50 Linacre 2021). However, the full range of infit statistics showed various levels of misfit and overfit. This is not surprising given the short amounts of training provided to raters and is indicative of the need for further training on the scale. Regarding the levels of between rater consistency, severity measures indicating the raters' locations on the logit scale were shown to vary but the separation statistic indicated that the raters had been separated into less than two statistically distinct levels of severity. This is promising and suggests that test takers would receive similar grades regardless of the rater that scored their papers. However, the differences in rater severity will need to be addressed before the rating scale is officially used as part of the university entrance test and further rater training is required, for instance by focusing on direct lifting and replication of key vocabulary (see Section 5.2).

Overall, the developmental approach outlined in this study has the potential to contribute to the assessment of integrated skills in local language tests but there are clear caveats. During the focus group, the raters described the difficulty of determining whether students had sufficiently comprehended the reading text. The difficulties raters experience in evaluating student levels of text comprehension is a recurrent theme in integrated skills research (Chan et al. 2015) and the approach applied in this study does not seem to have resolved this. In addition, the raters explained that they were reluctant to award zero on the scale even to the point of ignoring the scale completely to avoid the relevant descriptor. This represents an important limitation that indicates raters require further training. The developmental methodology involved rater comments, feedback and a focus group but the potential to define a construct with this methodology has limitations. As participants made clear, determining whether test takers had successfully "interpreted" the input text was not possible. Approaches such as discourse analysis of test taker texts and methods aimed at uncovering the cognitive processing involved in integrated task completion, such as stimulated recall, may offer important insights in this regard that could be incorporated into the rating scale (Golparvar and Rashidi 2021; Michel et al. 2020).

Despite these caveats, the study has implications for language testing practitioners seeking to develop rating scales for integrated assessments. Dimova et al. (2020) identify the benefits of local language assessment development as the potential to reflect achievement goals and institutional values in language tests. The current study represents an attempt to reflect local teacher values directly in a rating scale for integrated reading-into-writing assessment. As such, the study contributes to the incipient body of research documenting the development and validation of rating scales for integrated skills assessment.

To conclude, this study applied quantitative and qualitative methods to inform an iterative approach to rating scale development for an integrated reading-into-writing assessment. Several forms of analysis were completed to develop and investigate the content and scoring validity of the scale. The scale was shown to reflect and reliably assess the integrated writing construct; however, some important limitations relating to construct representation were outlined during the validation process. The objective of this developmental stage was primarily to help the institution collect informative and accurate information concerning raters' opinions about competency in integrated language skills. The findings indicate that a rating scale based on this analysis may be applied in the university entrance test given adequate rater training and standardisation.

# Appendix: Sample reading-into-writing input text

In the text below, the writer discusses social media companies' views on the content presented on them. In your opinion, are these social media companies right in their claim? Write a paragraph of approximately 150 words.

Your paragraph will be assessed on task completion, organisation, grammar and vocabulary. Copying sentences or chunks from the text is not acceptable and will be penalised.

Recently, Facebook and YouTube have appeared in the news. They are accused of unfair censorship by some and by others of publishing offensive or dangerous content. Are social media companies neutral platforms where users can share anything, or are they publishers of content like a newspaper?

Some, like Jeff Greene, an attorney for a social networking company, argue that social networks aren't responsible for the content that is published on them. They deny that YouTube, Facebook, and Twitter are publishers, like a newspaper company. To call these companies publishers is to say that their task is to produce content. These companies claim the essential value of these networks is conversation, not content. Others agree, saying that the problem is not technology, but human behavior using technology, the bad acts of a small number of propagandists, trolls, and troublemakers. They are manipulating the platforms. And the platforms should not be held accountable for the bad behavior of a few. Social networks are considered a sort of free open market of ideas. For them, social media are and should remain places for public discussion where all people can share their opinions with no filter.

On the other side of the argument, Bill Jordan of *P.C. World* notes sarcastically, "these social media companies call themselves platforms, they don't want responsibility for what people say on them. Still they continue making advertising money from it." These are not platforms, but modern publishers, just like a newspaper company, he claims. In the U.S. alone 81% of Americans have at least one social media account, and almost 65% of them get their news from what is posted there. Given how many people get their news from these internet spaces, aren't the potential threats too great? If a newspaper prints false information, they are forced to correct their mistake. Not so with these sites. Do they not have a bigger responsibility to society?

# References

Chan, Sathena. 2013. *Establishing the validity of reading-into-writing test tasks for the UK academic context*. Luton: University of Bedfordshire PhD thesis.

Chan, Sathena. 2017. Using keystroke logging to understand writers’ processes on a reading-into-writing test. *Language Testing in Asia* 7(10), https://doi.org/10.1186/s40468-017-0040-5 (accessed 19 August 2021).

Chan, Sathena. 2018. Some evidence of the development of L2 reading-into-writing skills at three levels. *Language Education & Assessment* 1. 9–27.

Chan, Sathena, Chihiro Inoue & Lynda Taylor. 2015. Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing* 26. 20–37.

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

Council of Europe. 2018. *Common European framework of reference for languages: Learning, teaching, assessment companion volume with new descriptors*. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4 (accessed 18 August 2021).

Delaney, Yuly Ascension. 2008. Investigating the reading-to-write construct. *Journal of English for Academic Purposes* 7(3). 140–150.

Dimova, Slobodanka, Xun Yan & April Ginther. 2020. *Local language testing: Design, implementation, and development*. London: Routledge.

Ducasse, Ana Maria. 2009. Raters as scale makers for an L2 Spanish speaking test: Using paired test discourse to develop a rating scale for communicative interaction. In Annie Brown & Kathryn Hill (eds.), *Tasks and Criteria in performance assessment*, 15–39. Frankfurt: Peter Lang.

Eckes, Thomas. 2011. *Introduction to many-Facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang.

Ellis, Rod & Fangyuan Yuan. 2004. The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition* 26. 59–84.

Ewert, Doreen & Sun-Young Shin. 2015. Examining instructors’ conceptualizations and challenges in designing a data-driven rating scale for a reading-to-write task. *Assessing Writing* 26. 38–50.

Fulcher, Glenn. 2003. *Testing second language speaking*. London: Routledge.

Fulcher, Glenn. 2012. Scoring performance tests. In Glenn Fulcher & Fred Davidson (eds.), *The Routledge handbook of language testing*, 378–393. London: Routledge.

Gebril, Atta & Lia Plakans. 2014. Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing* 21. 56–73.

Golparvar, Seyyed & Fatemeh Rashidi. 2021. The effect of task complexity on integrated writing performance: The case of multiple-text source-based writing. *System* 99. 1–11.

Knoch, Ute, Bart Deygers & Apichat Khamboonruang. 2021. Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing* 38(4). 602–626.

Linacre, John Michael. 2019. *Facets computer program for many-facet Rasch measurement, Version 3.83.0*. Beaverton, Oregon. winsteps.com.

Linacre, John Michael. 2020. Facets noncentre clarification [Discussion post]. *Rasch Measurement Forum*. https://raschforum.boards.net/post/8093/quote/3085 (accessed 18 August 2021).

Linacre, John Michael. 2021. A user’s guide to FACETS Rasch-model computer programs. http://www.winsteps.com/winman/copyright.htm (accessed 18 August 2021).

Lumley, Tom. 2005. *Assessing second language writing*. Frankfurt: Peter Lang.

Messick, Samuel. 1989. Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher* 18(2). 5–11.

Michel, Marijie, Andrea Révész, Xiaojun Lu, Nektaria-Efstathia Kourtali, Minjin Lee & Lais Borges. 2020. Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research* 36. 307–334.

Myford, Carol & Edward Wolfe. 2003. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement* 4(4). 386–422.

Myford, Carol & Edward Wolfe. 2004. Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement* 5(2). 189–227.

Ohta, Renka, Lia Plakans & Atta Gebril. 2018. Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing* 38. 21–36.

O'Sullivan, Barry, Jamie Dunlea, Richard Spiby, Carolyn Westbrook & Karen Dunn. 2020. *Aptis technical manual Version 2.2*. Council. www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf (accessed 18 August 2021).

Plakans, Lia. 2012. Writing integrated items. In Glenn Fulcher & Fred Davidson (eds.), *The Routledge handbook of language testing*, 249–261. London: Routledge.

Plakans, Lia & Atta Gebril. 2012. A close investigation into source use in integrated second language writing tasks. *Assessing Writing* 17. 18–34.

Shin, Sun-Young & Doreen Ewert. 2015. What accounts for integrated reading-to-write task scores? *Language Testing* 32. 259–281.

Turner, Caroline & John Upshur. 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. 49–70.

Weir, Cyril, Ivana Vidakovic & Evelina Galaczi. 2013. *Measured constructs: A history of Cambridge English examinations, 1913–2012*. Cambridge: Cambridge University Press.

Wind, Stefanie. 2020. Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing* 43. 1–14.

Wisniewski, Katrin. 2017. Empirical learner language and the levels of the common European Framework of reference. *Language Learning* 67. 232–253.

Wright, Ben & Geoff Masters. 2002. Number of person or item strata: (4*Separation + 1)/3. *Rasch Measurement Transactions* 16. 888.