Language Testing in Asia

RESEARCH Open Access

Development and validation of a rating scale for Iranian EFL academic writing assessment: a mixed-methods study

Nasim Ghanbari[1*] and Hossein Barati[2]

* Correspondence: btghanbari@pgu.ac.ir
[1]English Language and Literature Department, Faculty of Literature and Humanities, Persian Gulf University, Bushehr 75169, Iran
Full list of author information is available at the end of the article

Abstract

The present study reports the process of development and validation of a rating scale in the Iranian EFL academic writing assessment context. To achieve this goal, the study was conducted in three distinct phases. Early in the study, the researcher interviewed a number of raters in different universities. Next, a questionnaire was developed based on the results of the interview along with the related literature. Later, the questionnaire was sent to thirty experienced raters from ten major state universities in Iran. Results of the country-wide survey in this phase showed that there was no objective scale in use by the raters in the context. Therefore, in the second development phase of the study, fifteen of the raters who participated in the first phase were asked to verbalize their thoughts when each rating five essays. At the end of this phase, a first draft of the scale was developed. Finally, in the last validation phase of the study, ten raters were asked to each rate a body of twenty essays using the newly developed scale. Next, eight of the raters participated in a follow-up retrospective interview. The analysis of the raters' performance using FACE TS showed high profile of reliability and validity for the new scale. In addition, while the qualitative findings of the interviews counted some problems with the structure of the scale, on the whole, the findings showed that the introduction of the scale was well-received by the raters. The pedagogical implications of the study will be discussed. In addition, the study calls for further validation of the scale in the context.

Keywords: EFL academic writing assessment, Rating scale, Validation, FACETS, Think-aloud protocols, Raters

Introduction

In performance-based assessment, scoring rubrics (variously named as rating scales or marking schemes) are important as they show the construct to be performed and measured. Moreover, rubrics have a prominent role in instruction and evaluation (e.g., Osana & Seymour, 2004; Reitmeier, Svendsen, & Vrchota, 2006), evaluating student work (Campbell, 2005; Reddy & Andrade, 2009) and also as diagnostic instructional measures (Dunbar, Brooks, & Kubicka-Miller, 2006; Reddy & Andrade, 2009; Song, 2006). In addition, rubrics can "help explain terms and clarify expectations" (Crusan,

2010, p.43) and hence can reduce the long-recognized problem of rater variability (Bachman, Lynch, & Mason, 1995; McNamara, 1996). As a result, they promote reliable test scores and valid score inferences (Boettger, 2010; Crusan, 2010; Crusan, 2015; Dempsey, PytlikZillig, & Bruning, 2009; Knoch, 2009; Lukácsi, 2020; Rakedzon & Baram-Tsabari, 2017).

However, despite the crucial role that rubrics play in performance assessment, one might assume that the subject has been intensively researched and discussed. McNamara (1996, p.182) asserts that in the field of language assessment, "we are frequently simply presented with rating scales as products for consumption and are told little of their provenance and of their rationale. In particular, we too frequently lack any account of their empirical evidence for their validity." Actually, many teacher practitioners simply follow the strategy of "adopt and adapt" when seeking to develop an assessment rubric (Crusan, 2010, p.72). Although the ad hoc rubrics developed in this way can help teachers with the classroom assessment, for more high-stakes tests with significant impacts on the educational life of stake-holders, the rubrics grounded in theory are needed (Knoch, 2011; McNamara, 1996).

In performance assessment of writing "rating scale development is a ubiquitous activity" (Banerjee, Yan, Chapman, & Elliott, 2015, p.5). The current literature shows few writing scale development and revision attempts (e.g., Banerjee et al., 2015; Janssen, Meier, & Trace, 2015; Knoch, 2011; Lim, 2012; Sasaki & Hirose, 1999). The studies available in the literature mostly describe the development of speaking assessment scales (Ducasse, 2009; Fulcher, Davidson, & Kemp, 2011; Upshur & Turner, 1999) on the account that speaking and writing assessment is largely similar.

The local context of the present study was the particular Iranian EFL academic writing assessment. Two courses of advanced (paragraph) writing and essay writing form the core practice in this regard. A general survey of the context showed that there was no explicit assessment procedure in the context. Indeed, in the absence of a rating scale, teacher-raters did their own subjective assessment of the texts. Hence, in the absence of objective rubrics, the raters felt safe with their own rating which ignored significant aspects of reliability and validity of the assessment. Needless to say, the Iranian EFL academic writing assessment needed reliable and valid rating scales. Therefore, to fill this gap, the present study was conducted to develop rubrics for the assessment of academic writing. To achieve this goal the researchers first conducted a country-wide survey to find if the raters used any rating scales. Next, lack of a common rating scale in the context caused the development of a new rating scale based on both qualitative and quantitative measures. At the final stage, the new scale was validated in the context.

In fact, the development and validation of a rating scale in the Iranian EFL academic writing context is significant in a number of ways: first, the existence of reliable rating scales would improve the reliability and validity of writing assessment in the context. Next, development of a local rating scale as it considers the raters' sociocultural and disciplinary background is in line with the current concerns in the literature which call for redefining of the existing rating scales that legitimize the native speaker as the standard to judge the non-native writers' performance in this particular writing assessment context. Moreover, improving the raters' professional development through introducing a scale in the rater training programs which would eliminate the rater variability

would be another contribution of the present study (Weigle, 1998). Last but not the least, an analytic rating scale by providing diagnostic information on the strengths and weaknesses of the learners would help fulfill educational objectives in the academic context (Alderson, 2005).

Literature review

The literature on rating scale development shows two main approaches: the measurement-driven, intuitive and the performance data-driven, quantitative approaches (Fulcher et al., 2011). The measurement-driven approach as the most commonly used one by far relies on the intuition of language teaching and assessment experts (e.g., teachers, raters, SLA theorists) to develop the rating criteria (Hamp-Lyons, 1991). The focus of scale development here is on the clarity of the level descriptors, and in this way, it attempts to improve the usability of the rating scale. However, the intuitive approach has been criticized for the lack of precision, specificity, and scalability (Fulcher et al., 2011). Due to the impressionistic and abstract language of the level descriptors, it becomes difficult for the raters to distinguish between the students' performances across the levels (Knoch, 2009). In addition, the often a-theoretical nature of rating scales that are either based on no accepted model of language performance (Fulcher, 1996; North & Schneider, 1998) or even scales that are not themselves based on an empirical investigation of language performance (Young, 1995) creates scales with components that do not actually occur in the writing performances of the learners (Fulcher, 1996; Upshur & Turner, 1995). Moreover, development of the intuitive scales based on pre-existing scales might result in criteria that are irrelevant to the particular rating context (Turner & Upshur, 2002). The other criticism on these rating scales is the inconsistencies between the rating scales and the SLA findings (Brindley, 1998; North, 1995; Turner & Upshur, 2002; Upshur & Turner, 1995). While intuitive rating scales generally assume a linear development of language ability, the actual developmental route of language ability is less predictable at different stages of linguistic development. Although having the expert raters in the scale development enhances the usability of the measurement driven rating scales, no use of real performance prior to developing the descriptors threaten the reliability of the descriptors and also validity of the score inferences. As a remedy, the intuitive scales require post-hoc quantitative or qualitative analysis.

The other scale development approach—performance data-driven approach—constructs the rating scales by analyzing real language samples. In fact, in this approach, through analyzing real language traits or features that characterize and distinguish written texts or writers across proficiency levels are identified. This approach can be divided into two qualitative and quantitative strands. The qualitative methodologies provide a detailed analysis of the characteristic features of different levels of writing developed by the measurement-driven approach. The quantitative method by using empirical methodologies such as Rasch analysis quantifies and cross-validates the qualitative evidence on a large scale. According to Lim (2012), the two methods are complementary and should be used in combination. The main advantage of the performance data-driven approach is that the resulting scale reflects the real performance of the students. In other words, unlike measurement-driven approach in which the reliability and validity analysis is used to confirm the already developed scale, the analysis

of the real performance precedes the development of the scale. Nevertheless, the scale development procedure in data-driven approach is time-consuming. In addition, when using corpus-based tools, the resulting linguistic constructs emerged from the data might become difficult to operationalize for human raters (Fulcher, 2003). In the same vein, rater training programs should be structured in a way that individual criteria be considered equally and simultaneously but this does not bear pressure over the limited processing capacity of the raters in the real-time rating.

The literature shows that scholars have used a variety of empirical methods during the scale development procedure. Tyndall and Kenyon (1996) developed and validated a new holistic rating scale for scoring English essays in the context of a placement test offered to newcomers at George Washington University in the USA. Their study paved the way for the application of new measurement approaches (e.g., FACETS) in the validation of rating scales. As another study, Sasaki and Hirose (1999) developed an analytic rating scale for assessing L1 composition in the Japanese local context. Although the study by Sasaki and Hirose (1999) was motivated and conducted in an L1 context, the sound procedures and the results obtained made the study worthy of review among the studies concerned with developing and validating rating scales. Knoch (2007a) developed and validated a scale based on the features of the topic structure analysis in the context of DELNA (Diagnostic English Language Needs Assessment), a university-funded procedure designed to identify the English language needs of undergraduate students upon their admission to the University of Auckland. The study significantly expanded the Weigle's summary table (2002) by distinguishing two types of analytic scales: less detailed, a-priori developed scales and more detailed, empirically developed ones.

Janssen et al. (2015) using a mixed-methods approach examined how the ESL Composition Profile (Jacobs, et al., 1981) was functioning in the context of their study and in this way helped to revise the rubric in the context. The study showed an ongoing rubric analysis by the researchers and asked for similar rubric analysis in other contexts that use high-stakes performance assessment. In another scale revision project, Banerjee et al. (2015) used a combination of approaches (i.e., expert intuition, empirical analysis of performance data) to review and revise the rating scale for the writing section of a large-scale advanced level English language proficiency exam. The study shows the benefits of triangulating information from writing research, rater discussions, and real performances in the development and modification of the rating scales. In the same line, Chan, Inoue, and Taylor (2015) used the theoretical construct definition along with the empirical analyses of test-taker performances to develop rating rubrics for the reading-into-writing tests. As another empirical data-driven scale development attempt, Ewert and Shin (2015) probed into the conceptualizations and challenges of four ESL teachers when developing empirically derived binary choice, boundary definition (EBB) for reading-into-writing tasks. The findings showed that EBB scale can show the contribution of hybrid constructs to the overall quality of integrative reading-into-writing task performance and in this way it can enhance the connections among teaching, rating, and rater-training programs. In an attempt to improve the Michigan English Language Assessment Battery (MELAB) rating scale, Jung, Crossley, and McNamara (2019) used advanced computational tools to assess the linguistic features associated with lexical sophistication, syntactic complexity, cohesion, and text structure of texts rated by

the expert raters. The findings showed that linguistic features can significantly predict the raters' judgment of the essays. In the same line, Lukácsi (2020) developed a level-specific checklist of binary choice items to improve the discriminating power of Euro-exam International writing at B2 level. Drawing on an array of qualitative and quantitative methodologies including analyses of task materials, operational rating scales, reported scores, and candidate scripts, the author claimed that the newly developed checklist could provide a more coherent picture of the candidates' language ability.

In addition to intuitive and data-based approaches to scale development, some scholars (e.g., Fulcher, 1987; Knoch, 2007a, 2007b, 2011; McNamara, 2002) have also called for more theory-based scale development practices. The rationale here is that poor or no connection between the theories of L2 development and scale constructs underestimates the validity of the developed scale. However, there are no theory-based scale developments in the literature probably because there is not a unified theory of L2 development or language proficiency (Knoch, 2011).

With regard to Iranian EFL writing assessment context, there have been few scale development studies. As two early attempts in this regard, Farzanehnejad (1992) and Bateni (1998) used discourse analytic measures to develop their scales. Farzanehnejad (1992) propounded two measures called Measure of Cohesion (MC) and Maturity Index (MI) for assessing the writing proficiency of the learners. In the same spirit, Bateni (1998) developed his proposition-based measure for assessing EFL writing tasks involving the computer to accelerate the process. Later, Ahmadi-Shirazi (2008) developed and validated an Analytic Dichotomous Evaluation Checklist (ADEC) to enhance inter-and intra-rater reliability of the writing assessment. The analytic scale developed in her study consisted of 68 items comprising five subscales of content, organization, grammar, vocabulary, and mechanics. In another study, Maftoon and Akef (2010) developed and validated appropriate descriptors for different stages of the writing process, i.e., generating ideas, outlining, drafting, and editing in the Iranian writing assessment context. As another scale development attempt in the context, Khatib and Mirzaii (2016) conducted a three-strand mixed-methods study to develop a data-based analytic rating scale for assessing EFL descriptive writing.

This study

Despite the variety of scale development methodologies used in the above studies, it seems that the most effective approach to scale development and validation would be the one that combines theoretical bases of second/foreign language writing performance, expert intuition, and the empirical analysis of the performance data. Hence, the present study used a mixed-methods research strategy by triangulating three data sources: drawing the expert raters' intuition in their actual rating to build a scale, using the developed scale to rate a body of 100 texts and the quantitative and qualitative analysis of the rating data. Hence, to achieve these goals attempts were made to answer the following research questions in this study:

1. Is there any explicit rating scale in use in the present Iranian EFL writing assessment context?
2. How to develop a rating scale which is inclusive of the needs of the Iranian EFL writing assessors?

3. Does FACETS analysis indicate any validity evidence for the developed rating scale?
4. How do Iranian EFL raters perceive the rating scale developed in this study?

Method

Design of the study

In order to provide a multi-layered analysis of the research questions, the present study adopted a mixed methods design. The study proceeded in three distinct phases of exploration, development, and validation. Design of the study in the first phase which was concerned with the administration of a questionnaire was quantitative. The second development phase employed a qualitative design to analyze the rater's think-aloud protocols. Finally, in the last validation phase, ratings of ten raters were quantitatively analyzed and the interview data were qualitatively analyzed (Fig. 1).

Participants

Thirty experienced male and female raters participated in the exploration phase of the study. They were either PhD holders or PhD students in TEFL. They also varied in terms of age (M = 46) and the teaching background (M = 13). Only three of these people had attended rater training courses, but all of them had a minimum 5 years of experience in teaching and assessing writing (M = 7). The thirty raters were randomly selected from ten major state universities in Iran. Moreover, one hundred EFL BA students were asked to produce the texts for the study. There were both male and female students in the study (M = 67, F = 33) and they aged between 22 and 26.

In the development phase, a group of fifteen raters from those thirty raters who participated in the exploration phase contributed to the study. There was no randomization for the selection of the raters in this phase. Rather, they were selected following a convenience sampling approach. Due to the importance of this phase, only highly experienced raters who had beyond 15 years of writing assessment were asked to participate. Majority of the raters were males (86%) and they aged between 35 and 58.

Finally, a group of ten new raters (who had not participated in the study yet) agreed to participate in the validation phase. There was no randomization and the raters' participation was self-selected. All of them had at least 5 years of experience in teaching and assessing writing and there were equal numbers of male and female raters participating in this phase. These ten raters were asked to each rate a collection of twenty writing texts based on the new scale.

Fig. 1 Outline of the study

Instruments

Questionnaire

The researcher developed and trialed the questionnaire in the pilot study. For this aim, the researcher interviewed a group of ten experienced EFL raters in order to elicit their views on different aspects of rating. Findings of the interview along with the related EFL writing assessment literature helped the researcher to develop the questionnaire. Next, 50 raters were asked to take part in piloting the questionnaire. The piloting showed that the questionnaire was too long for the raters to fill out. Also, the wording of some items was ambiguous and some of them did not specifically address rating as the main concern of the questionnaire. In addition, the English version was found easier for the raters to respond. The reliability of the questionnaire estimated through Cronbach's alpha was 0.87. As for the validity of the questionnaire, the raters were asked to indicate what items make a cluster and are therefore related to one main concept. The high inter-rater reliability between the raters (0.95) indicated a high degree of agreement among the raters. The ultimate form of the questionnaire which included forty items aimed to probe the Iranian EFL writing assessment context regarding the existence of any rating scales in use (Additional file 1).

Writing texts

Under an exam-like condition, one hundred writing texts were collected from the participants. A brief piloting of the writing texts showed the desired topic (argumentative), the appropriate length of the essay (three-paragraph length), and the time needed for writing the text (45 min). The texts originally produced in hand-written form were all photocopied to be used in different phases of the study.

Think-aloud protocols (TAPs)

Audio-recorded verbal protocols obtained from the raters were the main instrument to develop the scale and also to validate the scale in last phase of the study. As each rater was supposed to rate three texts, overall, forty-five verbal protocols were collected from the raters in the development phase.

Training manual

In order to familiarize the raters with the new rating scale, the researcher developed a training manual for raters to study at home before attending the training session. In the manual, clear instructions were provided on how each trait was to be rated through examples.

Interview scheme

In the final validation phase, the researcher developed an interview scheme. It consisted of six questions which asked the raters to comment on the rating process and more importantly the structure of the rating scale.

Data collection procedure

Early in the study, the researcher randomly selected three raters from each of the ten major Iranian state universities included in the study. Then, the questionnaire was sent

to them. The raters could either complete the paper version of the questionnaire or an electronic version via email. The collection of the completed questionnaires lasted about 3 months.

Next in the development phase, a collection of forty-five writing texts was randomly selected from a pool of hundred essays collected early in the study. The writing texts, each bearing an ID number, were put into fifteen bundles of three and later were randomly assigned to each of 15 raters participating in this phase. Following a short instruction on the think-aloud procedure, the raters were asked to do their usual scoring while verbalizing their thoughts. There were no time constraints for the raters to do their scoring; however, most of the raters finished their thinking aloud between 30 and 45 min. Although the researcher was present in the think-aloud sessions, she did not interfere with the raters' flow of thoughts. Only at times when a silence period lasted more than 15 s (Green, 1998) or the raters continued their rating without thinking aloud, the researcher encouraged them to verbalize what they were processing. In sum, at the end of this phase, a collection of forty-five TAPs was obtained.

Finally, to validate the scale developed in the study, a group of ten raters agreed to participate in this phase. Each of the raters was also provided with a training manual to study prior to rating. Each rater was provided by the scale along with twenty rating sheets. The raters were asked to complete their ratings within 2 weeks. Once all the ratings were completed, the bundles were given back to the researcher. Finally, to know how the raters perceived the efficacy of the newly developed scale, the researcher conducted some semi-structured interviews with them. All ten raters who scored the texts based on the new scale took part in the retrospective interview sessions.

Data analysis

Findings of the questionnaire were analyzed using the descriptive statistics measure of frequency counts and percentages. In the development phase, the TAPs were initially transcribed by the researcher. In transcribing the data, the researcher was sensitive to all instances of verbal and non-verbal data (e.g., pauses, silence periods, etc.). Next, the transcriptions were checked against the original recordings to verify that all audio-data had been brought in the transcriptions. Through careful reading of the transcriptions, the parts indicating a rating criterion were segmented. Upon further readings, the segments were coded and then based on the commonalities among the codes; they were put to larger categories. The codes and categories were further refined to form the final categories and their components in the rating scale.

In the last validation phase of the study, the rating data were analyzed using Multi-Faceted Rasch Measurement (MFRM) program of FACETS version 3.67.1 (Linacre, 2010). For validating the scale developed in the study, a limited definition of validity—the one common in Rasch analysis—was adopted. Here, construct validity is investigated using the concept of model fit (Wright & Masters, 1982; Wright & Stone, 1979). Fit indices as one of the three pieces of information provided in a typical FACE TS output (other two being a logit measure and a standard error) determine how well the observed data for a specific facet fit the expectation of the measurement model. Finally, in this phase, the interviews were analyzed using the qualitative method of content analysis.

Results

Investigating the first research question

Through careful investigation of the items which explored the existence of a rating scale among the Iranian EFL raters, it was found that the raters who contributed to this study doubted the existence of an objective rating scale in their rating practice. While a substantial number of raters disagreed with an impressionistic approach to scoring (56.66%, item 15) and strongly believed that all raters had some criteria in their scoring (80%, item 17), they held differing attitudes about a common rating scale in their own rating practice. For example, when asked whether they solely relied on their own scoring or used any known scale, the pattern of their responses showed that although they had their own analytic scoring (item 11) which included the explicit criteria of word choice, structure, spelling, etc. (item 23), they were hesitant to reject the efficiency of the existing rating scales (item 24). In other words, their responses to items 36 and 19 revealed that they were aware of the existing scales but they followed their particular rating approach.

Moreover, the results of the analysis showed that the respondents had differing interpretations for the term scale. On the one hand, the reliability of the impressionistic rating in which no explicit scale is considered was strongly questioned. On the other hand, in both their holistic and analytic scoring (items 11 and 14), they expressed that they were not concerned with using available rating scales and believed that their own rating criteria fulfilled their aim quite well.

Furthermore, majority of the raters (about 67%) believed that rating scales had an important role in writing assessment (item 21). They strongly believed that the use of a rating scale considerably would improve the psychometric qualities of such an assessment (items 22, 28, and 37). As further evidence to item 22, the raters' responses to items 28 and 37 showed that they acknowledged the issues such as raters' consistency and/or bias. More importantly, the majority of the raters assumed that the existence of a rating scale in the context of Iranian writing assessment would greatly prepare the ground for a fair assessment of writing.

Overall, results of the country-wide survey showed that there was no objective rating scale in use among the raters in the Iranian EFL academic writing assessment.

Investigating the second research question

Results of analyzing the verbal protocols

At this stage, forty-five think-aloud protocols were carefully transcribed, analyzed, segmented, and categorized into certain codes. Therefore, following a preliminary categorization of the codes, they were further analyzed in order to develop more inclusive and general codes. Therefore, in each category, based on the commonalities among the codes, the higher-level codes were formed. For example, in the category of "grammar," the three codes of "inverted structure," "run-on sentences," and "garbled sentences" could be put under the sentence structure. Therefore, it was decided to put the three codes under the more general code of "sentence structure." Next, the researcher asked three experienced raters to code a portion of data and then categorize the codes. After coding and subsequent categorization of the codes by the raters, the high intercoder reliability among them showed that there was high agreement regarding the

nature and wording of the codes. However, there were some minor disagreements in the categorization and also weighting of the categories by the experts which was resolved through discussion.

Following the development of the components of the scale categories, it was decided to determine the frequency of each category and its constituting components in the whole data set. Therefore, first, the percentage of each category in the whole data was estimated. As Fig. 2 shows, "mechanics" (30.18%) and "grammar" (22.01%) were the two most frequent categories. "Organization" (13.20%) and "style" (10.69%) were the next two components that the raters considered in their rating. Finally, "cohesion and coherence" (8.50%), "vocabulary" (8.17%), and "content" (7.23%) were almost equally considered by the raters. Second, the percentage of each component within any single category was also estimated in a similar vein. This measure showed the weight carried by each of the components within any larger category as well.

Developing the first draft of the scale

Owing to pedagogical benefits of analytic scoring in the academic context (Knoch, 2009), and also its more reliability than other types (Hamp-Lyons, 1991, 1995; Huot, 1990), the scale was considered to be analytic. The weight of each category in the scale was determined by its associated percentage in the whole data. Therefore, the categories were assigned scores according to their correspondent percentage in the data. In addition, for the ease of scoring through rounding the numbers up and down decimals were dropped. Rounding the numbers was done in a way that they form a total hundred score. The reason was that when raters deal with a detailed scale, scoring according to a small number would complicate the scoring process. For example, for a category such as grammar with ten constituting components, a small number would create difficulties for the raters to score the sub-components.

The next step was to estimate the weight of the constituent components of each category. The constituent parts of each category were weighed similar to the way the

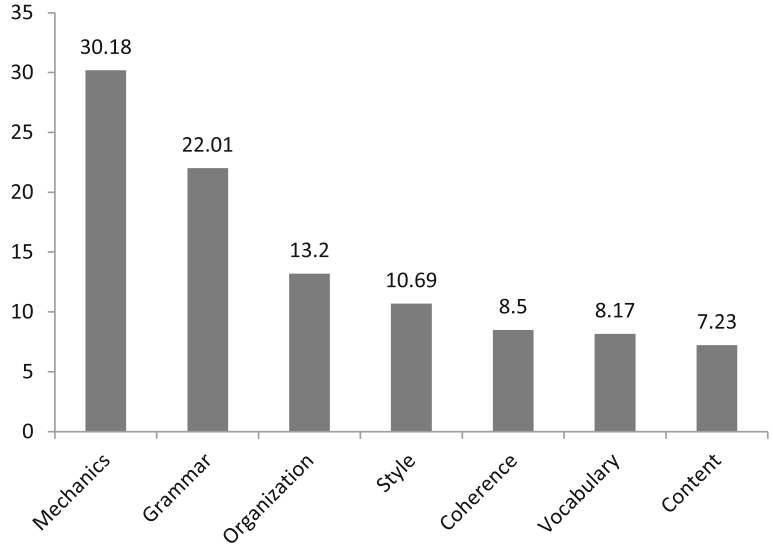

Fig. 2 Weight of the categories in the data

original categories were weighed. For example, while "punctuation" accounted for 25.33% of "mechanics," its corresponding weight was 8 out of 30. This weighing procedure was followed for other components in other categories as well. Decimals were also rounded for the ease of the calculation. Table 1 shows the weights assigned to the sub-components of the three categories of "mechanics," "grammar," and "organization."

Upon estimating the weight of the categories and their components in the scale, the first draft of the scale was developed.

Investigating the third research question

In this section, for each of the three facets in the study (i.e., rater, writer, and rating scale), the FACETS results are presented. Linacre (2010) believes that fit measures between 0.5 and 1.5 are productive for measurement. The digression of fit measures in either high (outfit) or low (infit) reduces the applicability of the measures in the measurement model. With regard to this, almost all the raters in this study performed within the range. It was just rater 10 (most severe) and rater 1 (most lenient) that were a little beyond the acceptable range. The mean infit (1.07) and outfit mean squares (0.96) also confirmed that the whole group of 10 raters fit the expectations of the model quite well. Thus, these raters were concluded to be able to consistently apply the new rating scale to the essays. Table 2 shows the raters in the order of their severity. As their logit measures show, rater 10 had the highest logit score (0.93) while rater 1 had the lowest (0.59).

As Table 2 shows, the point biserial (i.e., single rater vs. group of raters' correlation) was 0.92 which indicated to a strong agreement among the raters when using the scale. In addition, the measure of rater exact agreement showed that the raters chose the same score quite often (14.5% compared with the expected measure of 13.7% estimated by the model). Finally, the values of rater separation ratio and rater reliability showed that the raters were similar in terms of leniency or harshness when using the scale.

The second facet of interest was the writer. FACETS estimates to what extent the writers' performance matched the expectations of the model when rated by the scale. Table 3 shows those writers with misfit statistics. As shown, two of the writers (writers 1 and 6) had misfit values. The unexpected performance of the two writers indicated

Table 1 Sample weighing of the components of the scale

Mechanics		Grammar		Organization	
Punctuation	8	Sentence structure	8	Essay organization	7
Handwriting	5	Tense	4	Essay development	1
Capitalization	4	Preposition	3	Paragraph development	3
Spelling	4	Phrasal structure	2	Paragraph structure	2
Paragraphing	4	Pronoun	1		
Space	3	Number	1		
Face	2	Article	1		
		Contraction	1		
		Agreement	.5		
		Structure complexity	.5		
Total[a]	30	Total	22	Total	13

[a]Total weight of the scale was 100

Table 2 Rater measurement report—whole scale

Rater	Measure (logits)	InfitMnsq	OutfitMnsq
10	.93	1.65	1.61
4	.92	1.50	1.48
6	.88	.99	.89
7	.76	.89	.87
3	.73	1.10	1.00
9	.71	1.37	.97
2	.71	.93	.80
5	.61	1.21	1.12
8	.59	.62	.53
1	.43	.47	.39
Mean (count, 140)	.73	1.07	.96
S.D	.16	.37	.37
Rater point biserial, .92			
Rater exact agreement, 14.5%			
Rater separation ratio, 3.79; rater reliability, .94			

Table 3 Writer measurement report—whole scale

Writer	Measure (logits)	InfitMnsq	OutfitMnsq
1	.37	.41	.42
2	− .17	1.01	1.06
3	− .36	1.01	.89
4	.29	.70	.61
5	.33	.72	.64
6	− .48	2.39	2.03
7	.26	.98	.83
8	.12	1.01	.96
9	− .03	1.67	1.31
10	.06	.96	.93
11	.07	1.38	1.07
12	− .41	1.39	1.08
13	.33	1.10	1.04
14	.36	.80	.85
15	− .23	1.24	1.05
16	− .37	1.18	.96
17	.23	.76	.80
18	.03	.87	.77
19	− .05	.96	.97
20	− .3.4	1.11	1.01
Mean (count, 20)	.00	1.08	.96
S.D	.29	.42	.32
Writer separation ratio, 5.27; reliability, .97			

that the categories of analytic rating scale were not unidimensional for these two writers while they were for other writers in the study.

As Table 3 shows, the rating scale had a considerable discrimination. The measure of writers' separation ratio was high and it confirmed the ability of the scale to discriminate the writers into individual ability levels. The high estimate of reliability also showed that the rating scale had discriminated among the writers with a high degree of consistency.

The last facet to investigate here was the categories of the rating scale. FACETS analyzed the individual rating categories to find whether they fitted the expectations of the model. As Table 4 shows, "content" was the easiest category with the value of 1.22, while "mechanics" with − .1.77 was the most difficult. In addition, the fitness values of all the categories were appropriate. This shows that the seven rating categories are properly targeting the writer population in terms of their ability and the level of difficulty of the categories.

In addition, from the vertical ruler provided by FACETS (Fig. 3), it was found that the three facets of raters, writers, and rating categories met the predictions of the model when using the rating scale. Thus, the construct validity of the whole scale was quantitatively verified through the FACETS analysis.

Investigating the fourth research question

Following the rating task, some of the raters were asked to participate in a retrospective interview. In what follows, first the advantages of the rating scale in the raters' idea are explained, and then its weaknesses would be discussed.

Comprehensiveness of the scale

In raters' idea, the scale was detailed enough and it specifically addressed different aspects of the text. They believed that an analytic scale which includes different aspects of writing in a detailed way is more appropriate for the academic writing courses. In addition, the scale with its detailed and explicit structure of the categories and their components considerably improved the decision-making of the raters. Rater 1, for example, stated that:

> Rater 1: Look! When I started rating the essays I felt happy because I knew how much to assign for punctuation, for handwriting, etc. You know, everything was clear to me! I think this scale is inclusive enough to help the raters value all

Table 4 Rating category measurement report—whole scale

Rating category	Measure (logits)	InfitMnsq	OutfitMnsq
Mechanics	− 1.77	1.40	1.38
Grammar	− 1.25	1.07	1.11
Organization	− .05	1.18	1.25
Style	.31	.96	1.02
Cohesion and coherence	.65	.79	.80
Vocabulary	.90	.64	.66
Content	1.22	.52	.53
Mean (count, 7)	.00	.94	.96
S.D	1.12	.31	.31

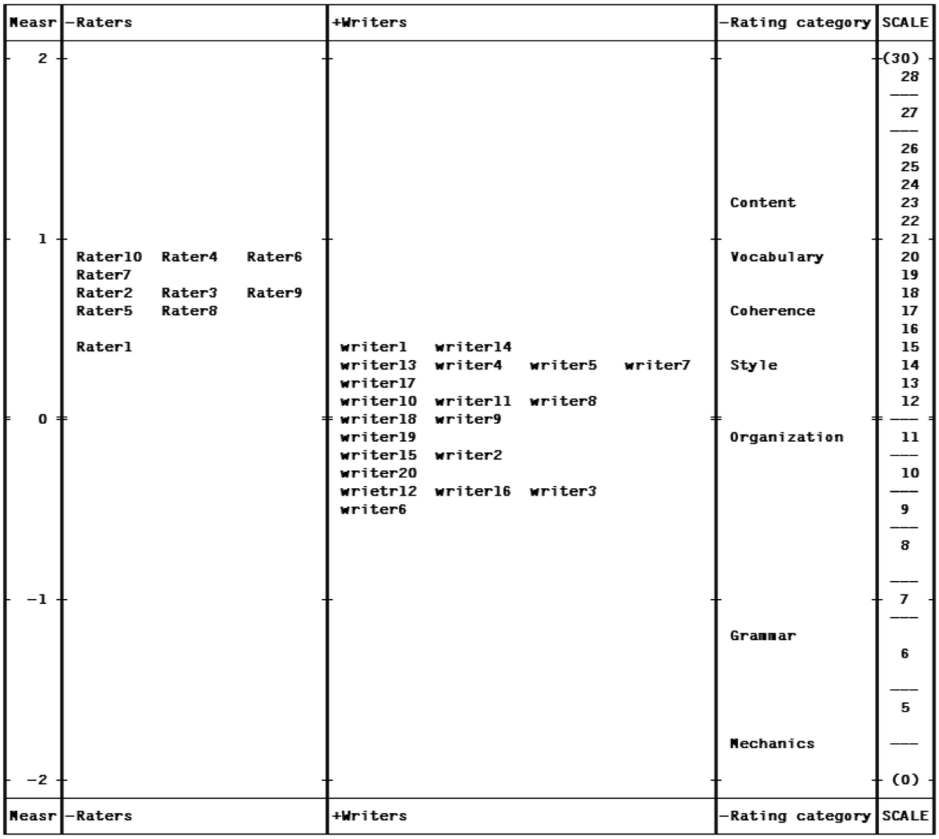

Fig. 3 All facets vertical ruler

components of writing, so it reduces the biases of the raters especially in academic contexts where students compete a lot for scores!

As another example, Rater 5 believed that writing assessment as a multi-faceted task necessitates a detailed rating scale to incorporate different components of the writing; hence, the detailed nature of the developed rating scale was a privilege for this rater.

Fair assessment of writing

The raters believed that the scale could promote a fair assessment among the raters. They claimed that by introducing the scale, the assessment of writing would be conducted with a more scientific rigor. Moreover, the raters believed that the scale can enhance accountability in the context and while awareness of the internationally well-known rating scales is necessary, having a scale that is responsive to the realities of the local context will make the assessment results more valid and hence more sound inferences will be made. Rater 2 posed the issue in this way:

> Rater 2: Mostly we do a kind of norm-referencing in assessing different writing texts. We compare the texts a lot which creates a lot of problems for a fair rating. I think through providing a uniform rating procedure the scale provides a trusted frame of reference for our writing assessment in the country.

Time-consuming

Although the raters appreciated the detailed structure of the scale, they believed that it was time-consuming. For example, rater 1 believed that when it comes to actual rating contexts, the task of rating based on the new rating scale would burn out the rater and it affects his/her scoring:

> Rater 1: I think some of the categories such as grammar, mechanics or punctuation were too detailed. Don't you think that they should be shortened in a way? You know, for example, it took me about 8 hours to rate 20 texts. Well, here I was interested to know the results of rating with this new scale but what about a final exam condition when you have about 50 essays to rate! If you want to rate with this scale, those essays which come first or last might be overestimated or underestimated. I mean when you come to final essays you may get tired and lose patience to rate each of the essays with care!

Ambiguity of the components and their weighting

Majority of the raters believed that the wording and the weighing of some categories and their components were vague. They suggested that some of the components should be deleted, some should be rephrased and some should be merged together. Some of the raters pointed to some inconsistencies in the language of the scale. For example, rater 4 commented that the inconsistency in the wording of some components would hurdle the easy implementation of the scale. She claimed that, "'contraction' in 'grammar' and 'awkward structure' in 'style' were two components that were not consistent in their language with the other components of the scale which might create problems for the implementation of the scale."

Moreover, the raters disagreed with the weights of some categories and their components. They believed that the scale in its current weighing structure is not appropriate for different levels of writing proficiency. Rater 5, for example, believed that the particular weighing structure of the scale could impose a specific view of the writing in which mechanical aspects of writing are prioritized and hence it was only suitable the lower-proficiency learners.

> Rater 5: Weighting of the categories is in a way that it ignores different proficiency levels of the students. For example, if a rater wants to put more emphasis on content, the scale does not allow him/her.

Introducing the validated rating scale

Upon the modification of the scale after the validation phase, the new scale was developed. The scale with changes in the categories, components, and their associated weights was suggested for assessing academic writing in the Iranian EFL context (Appendix).

Discussion

In this study, results of FACETS analysis showed that there was strong harmony between the ratings assigned and the expectations of the model. The measurement reports for the raters, essays, and scale categories showed that there were few instances

of misfit for each facet. This indicated that the newly developed scale was received well by the raters. The analysis of the results showed that the ratings were aligned with the structure of the scale. All of the categories were rated and the scale measurement report showed that content had the lowest score while mechanics had the highest total score. This is because the raters found their internalized criteria matched the ones in the scale. In addition to the probable effect of the training manual on their scoring, the components of the scale categories in the new scale were quite relevant to the writing courses they taught. They rated essays they had often read and were experienced with. In addition, the empirical data-driven nature of the scale that had used the raters' own criteria in developing the scale significantly affected the function of the scale among the raters.

Moreover, the raters' over-emphasis on mechanical and surface aspects of the texts along with ignoring aspects such as organization and content reflected the fact that the Iranian EFL raters had a limited conception of the writing construct which seriously calls for having consistent rater training programs in the context.

In addition to validity, FACETS analysis in this study showed high estimates of reliability for the newly developed scale. Rater separation ratio which measures the difference between the harshest and the most lenient rater was found to be appropriately low for the scale. The second estimate of reliability provided by FACETS was the rater point biserial correlation. This aspect of reliability measured the correlation between single rater and the rest of raters. It was found that almost all the raters had similarly scored the essays using the new scale. According to Knoch (2007a), a high point biserial correlation results in higher candidate discrimination and it is required as a necessary condition for a valid rating scale. The third estimate of reliability provided by FACETS in this study was the percentage of exact agreement. In this study, it was found that the percentage of exact agreement was mostly high and in many cases it exceeded the expectations of the FACETS. Finally, as the last piece of reliability evidence, the fit statistics which showed the degree of variations (high or low) in the ratings could provide evidence for the reliability of the scale and subscales. In this study, the percentage of the raters to rate with too high or too low variation which might result in high or low infit mean squares was very low. There might be two reasons for this pattern. One might be the way the raters reacted to the components of the scale. If the components be vague to them and do not offer sufficient information for the raters to make a choice, it is possible that they choose a play-it-safe method and tend to the scores in the inner of the scale and avoid the outer scores. In this situation, ratings come together and this conservative scoring would create low infit mean squares. In this study, this reason cannot be ignored. Post-hoc interviews with the raters showed that they struggled with some score values and therefore they tended towards the mean of the scale or subscales. Possibly, lack of band levels with the associated scoring for the components in the scale had created difficulties for the raters to make a decision.

The other reason for the raters' consistent ratings might be their familiarity with the categories and their components. In fact, except some of the components, the raters could understand the categories and their constituting components. Training manual also was an aid in the scoring process for them. More importantly, the special development procedure of the rating scale in this study contributed to the ease of scoring among the raters. In other words, since rating criteria were directly derived from the actual rating practice of the raters in the study, they could easily understand the nature

of the categories and their components in the scale. Therefore, the variability in the raters' performance caused by different interpretations of the rating criteria considerably reduced, and therefore, it caused the consistent performance of the raters.

Despite the appealing statistical evidence provided by FACETS, the results of the interviews indicated to some aspects that might pose a threat to the validity of the scale. Although the interviewees appreciated the scale as an attempt for more objectivity in writing assessment in the Iranian context, they believed that the scale was too long. Some of the raters even claimed that the length of the scale caused them to do a holistic rating. In other words, despite the analytic nature of the scale that had put seven categories along with the components, they rated based on their impression of texts. In a usual analytic scoring, the raters are supposed to read the text for each aspect and then assign a score at the appropriate level. Instead of having levels and band scores for each level, the scale in this study had a number of components for each category. This detailed profile of components for each category was boring for many of the raters and although they had rated all the components within their specified weighting boundaries, it increased the possibility of halo effect. Hence, their performance on some categories should be interpreted with caution. This scale-related factor was made more distinct when studying the individual categories in the scale. Overall, the raters were positive about the new scale.

Present study used concurrent verbal protocols to develop the rating scale. Although some have questioned the validity of the TAPs with regard the reactivity and veridicality threats (e.g., Barkaoui, 2011; Stratman & Hamp-Lyons, 1994), the results on the validation of the scale showed that it had homogenized the scoring of the raters. In addition, analysis of the post-rating interviews showed that the raters considered the scale to be comprehensive and they could easily understand the components of the scale. Therefore, the present study advocated the validity of TAPs to obtain the L2 writing process data. Moreover, results of the present study provided further evidence that concurrent verbal protocols caused no noticeable interference in the L2 writing processes (Yang, Hu, & Zhang, 2014).

Conclusions and implications

In sum, findings of the study re-emphasized the significant role of rating scale in the writing assessment. The study also showed that there is a pressing need for rater training courses in the country. The over-emphasis on mechanical and surface aspects of writing by the raters showed that they had a limited conception of the writing construct. The weighting structure of the first draft of the scale was in a way that it addressed the superficial aspects of writing and those criteria such as organization or style which are important at more advanced levels of writing ability were not attended enough. Therefore, in order to modify the raters' practice, having rater training courses seems necessary in the context.

The current study has also a number of theoretical and pedagogical implications for the EFL writing assessment. The first theoretical implication of this study is its contribution to EFL rating scale development literature. Studies on EFL rating scale development mostly have used questionnaire surveys (Nimehchisalem & Mukundan, 2011; Sasaki & Hirose, 1999) or reviews of available literature (Nakatsuhara, 2007). In this study, criteria for developing the rating scale were directly obtained from the raters in

the think-aloud sessions. The procedural innovation of the study caused the developed rating scale to have a robust empirical basis which in turn facilitated the rating task since the raters feel more attached with a scale developed based on their own criteria.

The next theoretical contribution of the current study is the weighting structure of the scale. In spite of the equal-weighting structure of the rating scales in the literature, the present study showed that a flexible weighting scheme gives the authority to the raters to manipulate the weighting based on the particular context of assessment. Although this requires a degree of rater ability, it would provide a fair assessment for different groups of the students in diverse assessment contexts.

Another contribution of the study is proposing a more-detailed analytic rating scale. Knoch (2007a) suggested that although holistic and analytic rating scales have been described as being distinct from each other, it was necessary to distinguish two types of less detailed, a priori developed and more detailed, empirically developed analytic scale. Therefore, the analytic scale developed in this study can be considered as an attempt for more analyticity in the analytic rating scale development procedure.

The present study also contributed to the EFL performance assessment models and models of rater decision-making processes. Since most of the performance assessment models (Fulcher, 2003; McNamara, 1996; Skehan, 1998) were conceived in the ESL oral performance context and when needed were adapted to EFL writing assessment contexts, the rating scale developed in this study indicated to some idiosyncratic aspects of EFL context. In addition, as the raters' cognition was the basis of scale development in this study, it acted as a verification check over the studies that have investigated the EFL raters' decision-making processes (e.g., Ostovar & Hajmalek, 2010).

There were also some limitations in doing the present study. Since validation was a major goal of the present study, using FACETS (Linacre, 2010) to validate the new scale was to adopt a limited definition of validity. In fact, the narrow view of validity presented here falls short of Messick's definition of validity (1989) which considers the empirical and theoretical rationales that support the adequacy and appropriateness of inferences and actions based on the test scores. Enlarging the scope of validation would be a significant improvement on the rating scale developed in the present study.

In addition, despite the country-wide scope of the study, it was conducted with a limited number of raters, i.e., 30 raters. The categories emerged out of think-aloud protocols and the following scale developed would be further refined with a larger sample of the raters in future studies. Also, due to providing rich data on the mental processes of the participants, TAPs was a major data collection instrument in the study. In addition, some of the raters were not well-oriented with the procedure. Although the researcher provided them with a training manual prior to data collection, it might be that the raters' mental processes were not completely represented in their verbalized thoughts. Having some retrospective measures following think-aloud sessions would supplement the results obtained from the verbal protocols and reduce the veridicality threat (Barkaoui, 2011).

In this study, only argumentative genre as the most common genre in the academic context (Plakan, 2009) was selected. Broadening the scope of study to include other genres would provide a more comprehensive view on the range of features involved in the final developed scale. Although the concern of the study was to develop a scale that assesses the writing performance of the writers, the salient aspects of performance such as fluency, accuracy, and complexity were missing in the developed scale. Future

studies should attempt to incorporate the three aspects in the scale. A further area of research can be the involvement of the student writers in the scale development process along with the raters. That is student writers can be asked for the rating criteria that they think they should be assessed on. The incorporation of the students' voices as important stake-holders would increase the fairness of writing assessment. Also, a detailed analysis of the rating behaviors of the raters using the scale developed in this study would show the extent the scale works in practice and also would add to the construct validity of the scale.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40468-020-00112-3.

Additional file 1:. Questionnaire used in the first phase of the study

Additional file 2:. Sample coding of TAPs in the development phase of the study

Additional file 3:. FACETS Results for the seven categories of the scale

Abbreviations

EFL: English as a foreign language; EBB: Empirically derived binary choice, boundary definition; TAPs: Think-aloud protocols; TEFL: Teaching English as a foreign language; ADEC: Analytic dichotomous evaluation checklist; MELAB: Michigan English Language Assessment Battery; MI: Maturity index; MFRM: Multi-Faceted Rasch Measurement; MC: Measure of cohesion

Acknowledgements

The authors wish to thank all thirty Iranian raters who kindly participated in the first and second phases of the study. Additionally, the authors acknowledge the help of ten raters who accepted to participate in the validation phase of the study.

Authors' contributions

Nasim Ghanbari was involved in the data collection and analysis. Dr. Hossein Barati participated in the write-up stage. He also organized different parts of the study. The authors read and approved the final manuscript.

Authors' information

Nasim Ghanbari holds a PhD in ELT (English Language Teaching). Currently, she works as an assistant professor in the English Language and Literature Department of Persian Gulf University in Bushehr, Iran. Her areas of interest include language assessment, writing assessment, and second language acquisition.

Dr. Hossein Barati is an associate professor of ELT in the English Department of University of Isfahan, Iran. His areas of interest include language assessment, second language acquisition, and EFL/ESL reading.

Funding

Not applicable: there are no sources of funding for the research study

Availability of data and materials

The data supporting this study is shared as an attachment.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]English Language and Literature Department, Faculty of Literature and Humanities, Persian Gulf University, Bushehr 75169, Iran. [2]Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran.

Received: 21 June 2020 Accepted: 15 September 2020
Published online: 11 November 2020

References

Ahmadi-Shirazi, M. (2008). *Towards more objectivity in writing assessment* Unpublished PhD dissertation. University of Tehran, Iran.

Alderson, C. (2005). *Diagnosing foreign language proficiency. The interface between learning and assessment*. London: Continuum.

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.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing*, *26*, 5–19.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Journal of Language Testing*, *28*(1), 51–75.

Bateni, M. (1998). *Computer-assisted proposition-based measure for EFL writing tasks* Unpublished MA thesis, Iran University of Science and Technology, Iran.
Boettger, R. K. (2010). Rubric use in technical communication: Exploring the process of creating valid and reliable assessment tools. *IEEE Transactions on Professional Communication*, *53*(1), 4–17.
Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman, & A. D. Cohen (Eds.), *Interfaces between SLA and language testing research*, (pp. 112–114). Cambridge: Cambridge University Press.
Campbell, A. (2005). Application of ICT and rubrics to the assessment process where professional judgement is involved: The features of an e-marking tool. *Assessment & Evaluation in Higher Education*, *30*(5), 529–537.
Chan, S., Inoue, C., & Taylor, L. (2015) Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, 26, 20-37
Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor: The University of Michigan Press.
Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing*, *26*, 1–4.
Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing*, *14*(1), 38–61. https://doi.org/10.1016/j.asw.2008.12.003.
Ducasse, A. M. (2009). Raters as scale makers for an L2 Spanish speaking test: Using paired test discourse to develop a rating scale for communicative interaction. In A. Brown, & K. Hill (Eds.), *Task and criteria in performance assessment: Proceedings of the 28th language testing research colloquium*, (pp. 1–22). New York: Peter Lang.
Dunbar, N. E., Brooks, C. F., & Kubicka-Miller, T. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, *31*(2), 115–128.
Ewert, D., Shin, S.J. (2015) Examining instructors' conceptualizations and challenges in designing a data-driven rating scale for a reading-to-write task. *Assessing Writing* 26, 38-50
Farzanehnejad, A. R. (1992). *A new objective measure for calculating EFL writing tasks* Unpublished MA thesis, University of Tehran, Iran.
Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 287–291.
Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238.
Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.
Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking test: Performance decision trees. *Language Testing*, *28*(1), 5–29.
Green, A. (1998). Verbal protocol analysis in language testing research. In M. Milanovic (Ed.), *Studies in language testing 5*. Cambridge: UCLES.
Hamp-Lyons, L. (1991). Reconstructing "academic writing proficiency". In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 127–154). Norwood: Ablex.
Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, *29*, 759–762.
Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 237–263.
Jacobs, H. L., Zinkgraf, S.A., Wormouth, D.R., Hartfiel, V. F. and Hughey, J. B. (1981). *Testing ESL composition: a practical approach*. Rowely, MA: Newbury House.
Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*, 51–66.
Jung, Y., Crossley, S., & McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *Journal of Asia TEFL*, *16*(1), 37–52.
Khatib, M., & Mirzaii, M. (2016). Developing an analytic scale for scoring EFL descriptive writing. *Journal of English Language Teaching and Learning*, *17*, 49–73.
Knoch, U. (2007a). *Diagnostic writing assessment: The development and validation of a rating scale* PhD dissertation, University of Auckland. Retrieved from: http://researchspace.auckland.ac.nzResearchSpace@Auckland.
Knoch, U. (2007b). Little coherence, considerable strain for reader: A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*, 108–128.
Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304.
Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96.
Lim, G. S. (2012). Developing and validating a mark scheme for writing. *Cambridge ESOL: Research Notes*, *49*, 6–9.
Linacre, J. M. (2010). *A user's guide to FACETS: Rasch measurement computer program* Program manual 3.67.1. Retrieved from www.winsteps.com.
Lukácsi, Z. (2020). Developing a level-specific checklist for assessing EFL writing. *Language Testing*. https://doi.org/10.1177/0265532220916703.
Maftoon, P., & Akef, K. (2010). Developing rating scale descriptors for assessing the stages of writing process: The constructs underlying students' writing performances. *Journal of Language and Translation*, *1*(1), 1–17.
McNamara, T. (1996). *Measuring second language performance*. Harlow: Longman.
McNamara, T. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, *22*, 221–242.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed.,). New York: Macmillan.
Nakatsuhara, F. (2007). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language & Linguistics*, *9*, 83–103.
Nimehchisalem, V., & Mukundan, J. (2011). Determining the evaluative criteria of an argumentative writing scale. *English Language Teaching*, *4*(1), 58–69.
North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, *23*(4), 445–465.
North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–263.
Osana, H. P., & Seymour, J. R. (2004). Critical thinking in preservice teachers: A rubric for evaluating argumentation and statistical reasoning. *Educational Research and Evaluation*, *10*(4–6), 473–498.
Ostovar, F., & Hajmalek, M. (2010). *Writing assessment: Rating rubrics as a principle of scoring validity* Paper presented at the fifth conference on Issues in English Language Teaching in Iran (IELTI-5). University of Tehran, Iran.

Plakan (2009). The role of reading strategies in L2 writing tasks. *Journal of English for Academic Purposes*, *8*, 252–266.
Rakedzon, T., & Baram-Tsabari, A. (2017). Assessing and improving graduate students' popular science and academic writing in an academic writing course. *Educational Psychology*, *37*(1), 1-19.
Reddy, Y. M., & Andrade, H. (2009). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448.
Reitmeier, C. A., Svendsen, L. K., & Vrchota, D. A. (2006). Improving oral communication skills of students in food science courses. *Journal of Food Science Education*, *3*(2), 15–20.
Sasaki, M., & Hirose, K. (1999). Development of analytic rating scale for Japanese L1 writing. *Language Testing*, *16*(4), 457–478.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
Song, K. H. (2006). A conceptual model of assessing teaching performance and intellectual development of teacher candidates: A pilot study in the US. *Teaching in Higher Education*, *11*(2), 175–190.
Stratman, J., & Hamp-Lyons, L. (1994). 2007a Reactivity in concurrent think-aloud protocols: Issues for research. In U. Knoch (Ed.), *Diagnostic writing assessment: The development and validation of a rating scale* PhD dissertation, University of Auckland. Retrieved from: http://researchspace.auckland.ac.nzResearchSpace@Auckland.
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.
Tyndall, B., & Kenyon, D. M. (1996). Validation of a new holistic rating scale using Rasch multi-faceted analysis. In A. Cumming, & R. Berwick (Eds.), *Validation in language testing*, (pp. 39–57). Clevedon: Multilingual Matters.
Upshur, J., & Turner, C. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing*, *16*(1), 82–111.
Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12.
Weigle, S. C. (1998). Using facets to model rater training effects. *Language Testing*, *15*(2), 263–287.
Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. In U. Knoch (Ed.), *(2007a). Diagnostic writing assessment: the development and validation of a rating scale* PhD dissertation, University of Auckland. Retrieved from: http://researchspace.*auckland*.ac.nz ResearchSpace@*Auckland*.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.
Yang, C., Hu, G., & Zhang, L. J. (2014). Reactivity of concurrent verbal reporting in second language writing. *Journal of Second Language Writing*, *24*, 51–70.
Young, R. (1995). Discontinuous interlanguage development and its implications for oral proficiency rating scales. *Applied Language Learning*, *6*, 13–26.

Publisher's Note