Journal of English Language
Teaching and Learning
Tabriz University
No. 17, 2016

Developing an Analytic Scale for Scoring EFL Descriptive Writing*

Mohammad Khatib
Associate Professor of TEFL, Allameh Tabataba'i University
Mostafa Mirzaii**
PhD Candidate of TEFL, Allameh Tabataba'i University, (Corresponding Author)

Abstract
English language practitioners have long relied on intuition-based scales for rating EFL/ESL writing. As these scales lack an empirical basis, the scores they generate tend to be unreliable, which results in invalid interpretations. Given the significance of the genre of description and the fact that the relevant literature does not introduce any data-based analytic scales for rating EFL descriptive writing, the researcher conducted a three-strand mixed study aimed at the empirical development of an analytic rating scale for scoring descriptions written by EFL learners. Composed of one quantitative and two qualitative strands, this mixed study factor-analyzed 172 ELT experts' analyses of the genre of description, and it content-analyzed 20 authentic and 30 inauthentic descriptive texts. Resulting from two meta-inferences made in the course of this study, the Analytic Rating Scale for EFL Descriptive Writing was constructed. Hopefully, employing this scale will lead to more reliable scores and more valid interpretations and decisions.

Keywords: EFL descriptive writing, analytic rating scale, empirical development, mixed methods research, factor analysis, content analysis.

---

* Received date: 2016/03/23 Accepted date: 2016/05/31

**** E-mail:** mustafa.mirzaei@yahoo.com

Introduction

Writing is widely held to be a linguistic enterprise, through the application of which a considerable proportion of communication is made possible and established. It follows that the ability to write, be it first or second language writing, requires the mastery of a number of mechanical, linguistic, and rhetorical conventions. To determine mastery of these conventions, language practitioners tend to engage learners in written English tests that are in turn scored using rating scales. Rating scales (also referred to as marking schemes and scoring rubrics) "act as a useful guide for evaluating the quality of students' written response" (Bacha, 2001, p. 113). With respect to foreign/second language learning and teaching, writing pedagogy scholars have conventionally assessed language learners' writing ability by means of one of the three commonly used scoring procedures, that is to say, primary-trait scoring, holistic scoring, or analytic scoring (East and Young, 2007; Fulcher and Davidson, 2007).

The least commonly employed scoring procedure, primary-trait rating entails making a decision about a particular aspect singled out and known to be central to the success of a writing task (Freedman, 1991). The primary-trait rating scale "is developed in regards to a single feature of writing that is determined to be essential to a particular writing task" (Becker, 2011, p. 117). The advantages of adopting primary-trait scoring are that attention is directed to one composing aspect at a time and that the scale fits well the specific tasks at hand (Cohen, 1994). On the other hand, it should be noted that the developmental process of preparing a primary-trait rating scale is labor-intensive (Fulcher & Davidson, 2007). Regarding this, Shaw and Weir (2007) hold that owing to "the lack of generalizability and the requirement to produce detailed rating protocols for each task, the primary trait approach is regarded as time-consuming and expensive to implement" (p. 149). Considering this, Becker (2011) maintains that primary-trait ratings scales are generally reserved for research situations, as well as for situations where information regarding learners' mastery of particular writing skills is required.

The second type of scoring, holistic scoring involves taking into consideration the written product in its entirety and assigning an overall score to the product. In so doing, holistic scoring, to the total exclusion

of what is deficient or missing, concentrates upon what the composition in question achieves (White, 1985). Therefore, one of the principal downsides of holistic scoring concerns its neglecting the provision of diagnostic information; meager, inadequate information as to what prospective instruction should target is supplied (Nelson and Van Meter, 2007). Moreover, as Hamp-Lyons (2003) argues, the assignation of ratings tends to be dependent on the rater rather than on the quality of the written product. As such, holistic scoring brings about a reduction in reliability of scores (Song and Caruso, 1996). In addition, according to Iwashita and Grove (2003), "the main problems in the use of holistic rating concern validity … [since] certain aspects outweigh others as assessors form an overall judgment of test-taker performance" (p. 26). This argument is also supported by Huot (1990) who while attempting to validate a holistic scale used for rating native speakers' compositions found that raters primarily focused on organization and content, and, in so doing, they mainly ignored language-related features of compositions. Unlike these contentions, however, White (1985) holds that through employing holistic rating scales validity of inferences is enhanced, for the score assigned is reflective of the authentic reaction of the rater. There are other advantages of using holistic scoring rubrics, the most salient of which is their widely agreed-upon practicality (Weigle, 2002). It is probably due to this perceived fact that "For several well-known language tests, such as the Cambridge ESOL Exams and the Internet-based Test of English as a Foreign Language (TOEFL iBT), holistic rubrics are used to score examinees' written responses" (Becker, 2011, p. 116).

Analytic scoring, the third type of scoring method, incorporates separately defined criteria, or elements, of written products. An analytic rating scale generally includes a number of writing elements—namely, organization, content, cohesion, register, coherence, mechanics of writing, and accuracy of linguistic devices (Weigle, 2002), with each element being marked independently of other components. One of the upsides of adopting analytic scoring is that thanks to its not collapsing components into a single, inflated score, raters can be trained easily (Cohen, 1994). A second merit concerns the fact that generalization to various writing tasks is made possible (Weigle, 2002). Additionally, reliability is improved (knoch, 2009). Fourth, analytic scoring helps

raters and instructors pinpoint student writers' weaknesses, as well as strengths; sufficient diagnostic information is, thus, provided (Carr, 2000). One should recall that the adoption of analytic rating is not without its shortcomings, however. Firstly, scoring the composition according to one criterion can potentially, and probably heavily, influence scoring the same composition on other criteria (Myford and Wolfe, 2004). Second, raters may fall into the trap of judging the scales holistically and thereby reach holistic impressions of compositions (Nakamura, 2004).

Resonating with Fulcher and Davidson (2007), favoring analytic rating scales over the other types of marking schemes, Knoch (2007) observes that "rating scales with detailed level descriptors are used in writing performance assessment to give raters an explicit basis on which to award scores" (p. 1). Knoch goes on to contend that these analytic rating scales tend to be developed by a team of experts who rely mainly on intuition of what they believe EFL/ESL writers produce rather than what these writers actually produce. Making reference to the criticisms made of intuition-based rating scales for lacking an empirical basis, for being too vaguely defined, and for often culminating in unreliable scores and invalid uses and interpretations, Knoch seems to be delivering a case for more empirically based methods of scale development, the results of which are data-based analytic rating scales.

Given the non-existence in the literature of a data-based analytic rating scale for scoring EFL descriptive writing, the present study aiming to empirically develop and validate such a scale undertook a mixed study consisting of three strands, i.e., a quantitative strand and two qualitative strands. The research questions were as follows:

1. What are the scoring criteria, their corresponding sub-criteria, and their respective weightings for rating EFL descriptive essays?
2. How does the empirically developed, data-based analytic scale for rating EFL descriptive essays perform in terms of reliability?

Methodology

Instruments and Materials

Existing, intuition-based scoring rubrics. These included the Descriptive Writing Rubric (suggested by the website msjohnsononline.com), the Descriptive Essay Rubric (developed by

Grey Nun Academy), and the Rubric for Descriptive Writing (constructed by the American Academy K8).

Rating criteria checklist. The checklist incorporated 18 items extracted from intuition-based scoring rubrics and thought to be of importance when rating descriptions. Each item taking the form of a phrase was followed by a Likert scale requiring the ELT expert to note their importance.

Authentic descriptive texts. The sample incorporated 20 texts gathered from ELT books and coursebook series including Writers at Work: The Short Composition (Strauch, 1994), Essay and Letter Writing (Alexander, 1965), Writing with Confidence (Meyers, 2006), First Steps in Academic Writing (Hogue, 2008), Introduction to Academic Writing (Oshima and Hogue, 2007), New Headway Intermediate Student's Book (Soars and Soars, 2009), and Total English Intermediate Student's Book (Clare and Wilson, 2005). These texts were used in Strand 2 as a basis to identify the salient elements of the genre of description.

Inauthentic descriptive essays. To help determine the number of band levels and define the rating scales analytically in the third strand, the researcher referred to essays composed by EFL learners at the elementary, intermediate, and advanced levels of language proficiency. The sample included 30 essays equally divided into three sets.

Memos. In the second strand, memos were written and used as a basis for developing templates of salient generic elements of the genre of description. Similarly, in the third strand, memos were kept in order to identify band levels and to help define the descriptors for each criterion in analytic terms.

Templates of generic elements of descriptive texts. Developed as a result of the pre-coding stage, these research instruments were used in the second strand. More specifically, these templates were used during the coding stage to provide a basis for counting the frequencies of the generic elements of descriptive texts.

Statistical package for the social sciences. IBM SPSS Statistics 21 fulfilled a variety of purposes, of which running principal components analysis on the ELT experts' analyses was the most important. Additionally, the package was employed to calculate Cronbach α as a way of investigating the internal consistency of the

resulting rating scale. Third, in the second strand, it was used to calculate the percentages, along with other statistics, of the frequencies of the salient generic elements of the descriptive texts.

MonteCarloPA.exe software. The software, relying on the number of variables being factor-analyzed, the number of participants in the sample, and the number of replications, was utilized to compute the average eigenvalues for a specific number of randomly generated samples. The researcher used this package to determine the number of factors to retain when performing principal components analysis.

Participants and Procedure

This mixed methods study was conducted in three strands. The first and only quantitative strand involved six steps. The initial step involved extracting rating criteria from existing intuition-based marking schemes. These criteria were distilled by the researcher focusing on three scoring rubrics (see Instruments and Materials). The process of extracting these scoring criteria involved analyzing these marking schemes and selecting the criteria identified by all or most of them. Next, the researcher prepared a checklist, in which these criteria, or items, were written in the form of phrases followed by a five-point Likert scale, requiring ELT experts to determine the importance of the item in question (see Appendix A).

Considering the number of extracted items (i.e., 18), the researcher, complying with the guidelines proposed by Nunnally (1978), needed to put between 90 and 180 analyses in the principal components analysis. Therefore, the checklist was distributed among 480 experts. The distribution process took place through sending the hard copy or the e-version of the checklist to the experts. Upon receiving the checklist, 182 of these experts agreed to analyze the items and return their analyses back to the researcher. Of these, 107 participants (nearly 58 %) were Iranian ELT experts and the rest (approximately 42 percent) were experts coming from other countries. Having received the analyses, the researcher decided to keep 172 analyses as they had correctly and completely analyzed the checklist.

The fourth step involved running principal components analysis, the output of which revealed the factors underlying the genre of description. As well as revealing the underlying factors, this analysis also showed how much of the total variance each of the factors accounted for. As

such, the weightings of the factors were also determined. Subsequently, the results were subjected to the test of Cronbach's alpha in order to examine the internal consistency of the factors identified. Finally, the output of the factor analytic measure was interpreted and labeled. The labels given to the factors needed to be chosen such that they represented all the items they subsumed. This was done in keeping with the literature, as well as with the theories and models of EFL writing pedagogy. The figure below clearly depicts the three strands of this study:

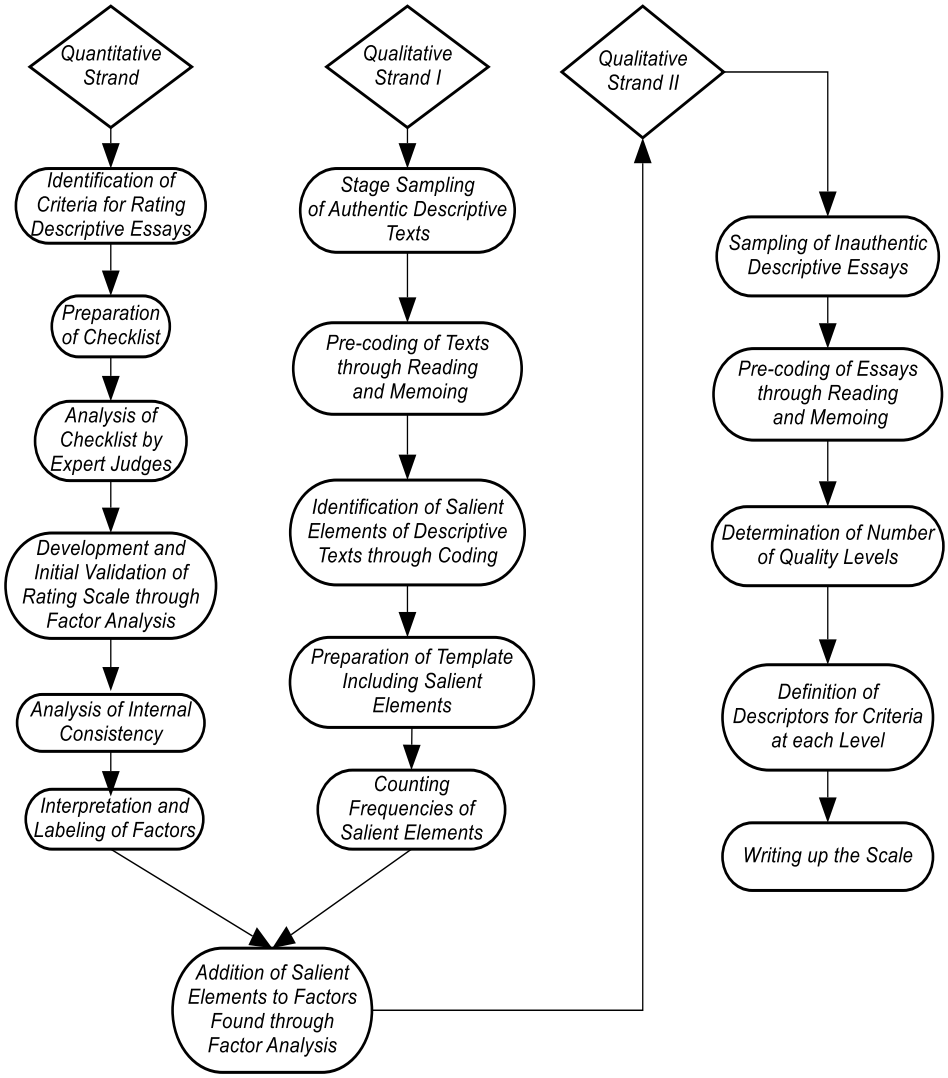

Figure 1. Graphic representation of the study's three strands

The second strand was in progress in conjunction with Strand One. This qualitative strand consisted of five steps. The purpose was to seek and identify the core and salient elements of the genre of description and to next add them to the factors derived through the first strand. The first step involved the stage sampling of authentic texts. To stage sample authentic descriptive texts, the researcher first selected a number of books and coursebook series through stratified sampling (see Instruments and Materials). The researcher then located all the descriptive texts available in these materials. Following this, an initial sample of 20 texts was selected through simple random sampling. Next, pre-coding was conducted by the researcher through reading and rereading these texts and memorizing, taking notes of all the unique features of the genre of description. This step aimed to arrive at a template of salient elements generically found in descriptive texts. The process continued until data saturation was reached. As the initial sample of 20 texts fulfilled the saturation requirement, there was no need to include more texts. In other words, by reading and taking notes of the elements found in 15 of these texts, the rest did not add any more elements. Subsequently, the researcher read the texts once more and counted the frequencies of the salient elements. To corroborate the findings, the researcher asked an ELT expert to independently code the texts. Marking the end of the first two strands, Meta-inference One involved the addition of the salient elements found in the second strand to the factors formed through factor analysis during the first strand. This was done in an attempt to render the rating scale more comprehensive. In so doing, the researcher added five elements to the rating scale.

The third strand began next. Prior to this strand, the rating scale incorporated four factors, each subsuming a number of items, or sub-criteria. However, these sub-criteria were not categorized into band levels. Put in other words, there were no descriptors defining in precise words how a low-quality descriptive essay differed from a high-quality one in terms of, say, its content and organization or its mechanics. Therefore, the purpose of the third strand was to define descriptors for various levels of performance. Additionally, this strand aimed to determine the necessary number of band levels for each criterion. To this end, the researcher, through purposive sampling, collected a sample of 30 EFL descriptive essays equally divided into elementary,

intermediate, and advanced levels. This collection was then used as a basis for the pre-coding and coding stages of content analysis. Throughout the content analysis process, the researcher read the essays over and over again, making notes of how well they had been composed. At this stage, the quality criteria the researcher took account of were those figuring in the initial meta-inference, this is, in the rating scale constructed at the end of Strands 1 and 2. Having content-analyzed the whole sample, the researcher arrived at either three or four general band levels. Next, corresponding descriptors were defined for each criterion of performance.

Results

Prior to running principal components analysis, the researcher using three pieces of information assessed the suitability of the data for the analysis. The first piece of information was provided by inspecting the correlation matrix which contained several coefficients of 0.3 and above (see Appendix B). Moreover, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, computed to be 0.827, exceeded the recommended value of 0.6. Finally, Bartlett's Test of Sphericity was highly significant ($p<0.001$). All this information indicated the suitability of the data for the analysis. Principal components analysis was, therefore, performed, the results of which revealed the presence of four factors having eigenvalues exceeding the value of one.

Table 1

Results of Principal Components Analysis

Total variance explained						
Component	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	4.363	24.240	24.240	4.363	24.240	24.240
2	3.638	20.209	44.449	3.638	20.209	44.449
3	2.617	14.541	58.989	2.617	14.541	58.989
4	2.299	12.770	71.759	2.299	12.770	71.759
5	.994	5.524	77.283			
6	.658	3.654	80.938			

Note. The rest of the cells were removed in order to save space.

As the Kaiser's criterion suggested extracting four factors, it was necessary to inspect the scree plot to check whether or not Catell's scree test supported this suggestion (see Figure 2 below).

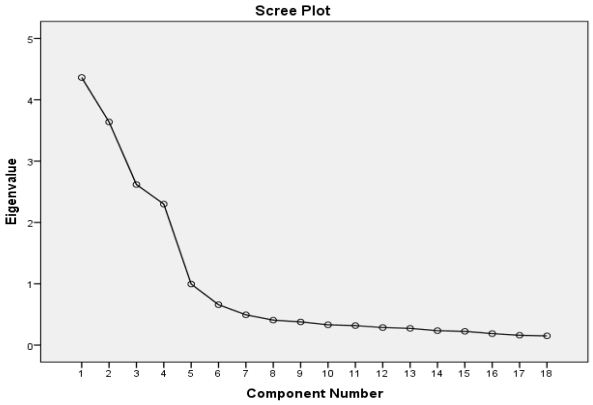

Figure 2. Scatterplot for scree test

The scree plot contained more than one break. Consequently, it was not possible to unambiguously interpret these angles and to clearly decide how many factors to retain. The researcher, thus, resorted to parallel analysis to further investigate the necessary number of factors to retain.

Table 2
Results of Parallel Analysis

Factor number	**Actual eigenvalue from principal components analysis**	**Random value from parallel analysis**	**Decision**
1	4.363	١,٦١٤٠	**accept**
2	3.638	1.4988	**accept**
3	2.617	1.3933	**accept**
4	2.299	1.3126	**accept**
5	.994	1.2400	**reject**
6	.658	1.1726	**reject**

The analysis, similar to the results of the Kaiser's criterion, revealed four factors whose eigenvalues exceeded those of their counterpart values from a randomly generated sample of data of the same size (i.e., 18 items * 172 experts). The four-factor solution accounted for a total of 71.75 percent of the variance, with the first, second, third, and fourth factors respectively explaining 24.24, 20.21, 14.54, and 12.77 percent

of the total variance. To facilitate the interpretation and labeling of these factors, the researcher rotated the factors through the Varimax method.

Table 3
Varimax Rotation of Factors

Rotated component matrix[a]				
	Component			
	1	2	3	4
c8	.914			
c16	.859			
c11	.849			
c17	.843			
c3	.816			
c1	.713			
c18		.902		
c6		.872		
c12		.835		
c14		.816		
c15		.812		
c5			.929	
c9			.923	
c4			.907	
c2				
c13				.919
c7				.901
c10				.876

The rotated solution indicated the presence of a relatively simple structure (Thurstone, 1943), with all sub-criteria loading strongly on only one factor and all the factors having a number of strong loadings. Nonetheless, as Sub-criterion 2 did not load strongly on any of the factors, it was decided to remove this criterion, repeat the analysis, and compare the results

Table 4
Results of Repeated Principal Components Analysis

Total variance explained						
Component	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of variance	Cumulative %
1	4.362	25.661	25.661	4.362	25.661	25.661
2	3.623	21.310	46.971	3.623	21.310	46.971
3	2.617	15.395	62.366	2.617	15.395	62.366
4	2.286	13.449	75.815	2.286	13.449	75.815
5	.664	3.907	79.721			
6	.497	2.922	82.643			

Note. The rest of the cells were removed in order to save space.

After removing Sub-criterion 2 from the analysis, the four-factor solution explained 75.82 percent of the total variance, which was approximately four percent more than the corresponding value of the previous solution. Moreover, most of the loadings increased, albeit marginally, once the Varimax rotation was carried out with the second sub-criterion removed.

Table 5
Repeated Varimax Rotaion of Factors

Rotated component matrix[a]				
	Component			
	1	2	3	4
c8	.914			
c16	.861			
c11	.851			
c17	.843			
c3	.815			
c1	.713			
c18		.903		
c6		.872		
c12		.835		
c14		.818		

c15	.814		
c5		.930	
c9		.924	
c4		.910	
c13			.919
c7			.901
c10			.875

The repeated analysis resulted in a 17-item scale, with 6, 5, 3, and 3 sub-criteria, loading strongly on factors 1, 2, 3, and 4, respectively. All sub-criteria loaded above 0.7 on their corresponding factors, which was a considerably high loading. Next, the internal consistency of the resulting scale was examined.

Table 6

Internal Consistency of the Scale

Reliability statistics		
Cronbach's alpha	Cronbach's alpha based on standardized items	N of items
.745	.760	18

The analysis indicated that the internal consistency of the factors, or criteria, forming the scale was acceptable, for it exceeded the critical value of 0.7 which is commonly reported as the minimum amount by many scholars (e.g., Pallant, 2013).

As stated in the previous section, Strand 2 basically aimed to complement the results of the quantitative strand. To this aim, the researcher, through analyzing authentic descriptive texts, first identified their most salient elements and prepared a template including all these elements. After that and in order to corroborate the findings, the researcher and another ELT expert analyzed the same texts, coded them for the identified elements, and counted their corresponding frequencies. The following tables present the results of the two independent coders' analyses (see Tables 5 and 6).

Table 7

Salient Generic Elements of Descriptive Texts at the Paragraph Level

No.	Feature/Element	Frequency of occurrences		Percentage of occurrences	
		Coder 1	Coder 2	Coder 1	Coder 2
1	Inclusion of telling, concrete details	17	17	100 %	100 %
2	Use of the five senses	5	4	29.41 %	23.53 %
3	Use of prepositions/prepositional phrases	17	14	100 %	82.35 %
4	Use of descriptive language (esp. adjectives and/or adverbs)	16	13	94.12 %	76.47 %
5	Use of figures of speech	6	7	32.29 %	41.18 %
6	Use of present simple	14	15	82.35 %	88.24 %
7	Use of passive voice	8	8	47.06 %	47.06 %

The two coders' analyses were strikingly similar, with the frequencies and percentages either resembling each other or being very close to one other. The inter-coder reliability coefficient was calculated to be 0.949.

Table 8

Salient Generic Elements of Descriptive Texts at the Text Level

No.	Feature/Element	Frequency of occurrences		Percentage of occurrences	
		Coder 1	Coder 2	Coder 1	Coder 2
1	Text organized in spatial order	8	7	88.89 %	77.78 %
2	Elaboration on a central idea through using details	9	8	100 %	88.89 %
3	Structural variety	7	8	77.78 %	88.89 %
4	Lexical variety	7	8	77.78 %	88.89 %

Although the results of the previous analyses were more similar, these analyses, too, revealed extremely similar findings. The inter-

coder reliability for these analyses was 0.826. Following this, the study's first meta-inference was made. This meta-inference is displayed in tabulated form below.

Table 9
Criteria, Sub-criteria, and Total Weightings

Rating criterion	**Rating sub-criteria**	**Total weighting of criterion**
Genre-related elements	creative, interest-sparking title; inviting introductory sentences; use of descriptive, vivid vocabulary; adoption of a personal, unique voice or style when describing; inclusion of concrete, precise details in body paragraphs; use of figurative language (e.g., metaphor, hyperbole, and personification; *use of the five senses*	**34 %**
Language-related elements	use of cohesive devices (i.e., substitution, conjunction, ellipsis, repetition, and lexical devices); consistent and accurate use of tenses in both active voice and passive voice; well-constructed sentences, written in correct word order and avoiding run-ons, fragments, and comma splices; accurate use of prepositional phrases, pronouns (subject and object), and determiners (articles, quantifiers, possessives, and demonstratives); precision and accuracy in word formation (morphology) and use (awareness of appropriate register); *structural variety; lexical variety*	**28 %**
Content and organization	Body content closely addressing the title and subtitles; Appropriate paragraphing of body content; descriptive language (esp, adjectives and/or adverbs); *text organized in spatial order; elaboration on a central idea through using details*	**20 %**
Mechanics	Accurate, consistent spelling; accurate use of punctuation marks; correct capitalization of letters	**18 %**

The resultant scale, reflecting Meta-inference One, encompassed the rating criteria, their sub-criteria, and their corresponding weightings, which put together make up a score of 100. These weightings were computed by taking into consideration the variances each criterion was shown by the principal components analysis to

account for. The sub-criteria included those arrived at as a result of Strand One and those arrived at as a result of Strand Two. The latter sub-criteria are italicized.

Subsequent to this and in order to determine the number of band levels, as well as define descriptors analytically, for each rating criterion, the researcher conducted Strand Three. As a result, four general band levels were identified and defined as follows:

Poor. An essay rated as Poor in relation to a particular criterion fails to fulfill any of the sub-criteria which constitute that very criterion.

Average. An essay rated as Average in relation to a particular criterion fulfills to a minimum extent some of the sub-criteria which constitute that very criterion.

Good. An essay rated as Good in relation to a particular criterion fulfills to an acceptable extent almost all the sub-criteria which constitute that very criterion.

Excellent. An essay rated as Excellent in relation to a particular criterion fulfills to the maximum extent all the sub-criteria which constitute that very criterion.

After defining these levels, the researcher reanalyzed the sample of 30 descriptive essays, coded them, and determined the number of levels necessary to score EFL descriptive essays.

Table 10

Results of Strand Three

Criterion	**Number of band levels**	**Numbers and levels of essays at each band level**			
		Poor	Average	Good	**Excellent**
Genre-related elements	٣	9 E, 3 I	1 E	7 I, 5 A	**5 A**
Language-related elements	٤	6 E, 1 I	4 E, 5 I	4 I, 4 A	**6 A**
Content and organization	٣	9 E, 2 I	-	1 E, 7 I, 1 A	**1 I, 9 A**
Mechanics	٣	8 E, 1 I	1 I	2 E, 6 I, 2 A	**2 I, 8 A**

Note. E, I, and A denote elementary, intermediate, and advanced, respectively.

Based on these results, three of the criteria, i.e., Genre-related Elements, Content and Organization, and Mechanics, were judged to

possess three band levels, that is, Poor, Good, and Excellent. This judgment was reached as the analyses showed that the majority of the essays were evaluated to be either Poor, Good, or Excellent in terms of these criteria. As a rule of thumb, for a band level to be judged necessary to be included in the scale, it had to receive at least five tallies, i.e., at least five essays had to be placed at that very level for the criterion in question. On the other hand, the criterion Language-related Elements was judged to possess four levels. Following this step, the researcher through rereading the essays wrote analytic descriptors for each criterion. This was actually the second meta-inference of the present study, leading to the finalized rating scale, i.e., the Analytic Rating Scale for EFL Descriptive Writing (see Appendix C).

Discussion

This study was conducted as the related literature did not include any studies which had empirically developed a data-based analytic scale for scoring EFL descriptive essays. The present three-strand mixed study arrived at the Analytic Rating Scale for EFL Descriptive Writing which incorporates four weighted criteria defined analytically in terms of their corresponding sub-criteria. Noteworthy is the fact that these criteria are defined in terms of varying numbers of band levels. Three of the criteria constituting this scale consist of three levels, while the remaining criterion, Language-related Elements, is defined in terms of four band level descriptors. Previously devised scales all define rating criteria using the same number of levels. As such, the end product of this study, relying on the tenets of mixed methods research and making use of factor and content analytic procedures, indicates that in some cases it might be necessary to define rating criteria in terms of varying numbers of band descriptors. The determination of the number of descriptors depends on a host of factors, among which one can name the complexity of the construct in question, degrees of precision sought, and the sample of essays on the basis of which a data-based rating scale is constructed.

A second issue meriting attention is the composition of the criteria and their importance in relation to one another. Among the four criteria, Genre-related Elements were shown to be the most important rating criterion, which, since this rating scale is directly related to the description genre, is a logical finding. Moreover, Language-related Elements, unlike a considerable number of rating scales which separate

grammar-related components from vocabulary-related components (e.g., Jacobs, Zingraf, Warmuth, Hartfiel, and Hughey, 1981; Weir, 1990), incorporate sub-criteria which relate to both grammar and vocabulary. Furthermore, Content and Organization, prioritized and deemed to be of prime significance by some scales (e.g., the Descriptive Essay Rubric developed by Grey Nun Academy and the Rubric for Descriptive Writing constructed by the American Academy K8) was identified as the third most important criterion which should be considered when scoring descriptive essays. Finally, the sub-criterion paragraphing which some marking schemes, such as the ESL Composition Profile (Jacobs et al., 1981), categorize under the rubric of Mechanics is subsumed under Content and Organization.

Another issue regards the identification of only four criteria shown to be important elements of EFL descriptive essays. Of the intuition-based rating scales previously devised, the Descriptive Essay Rubric (designed by Grey Nun Academy) and the Rubric for Descriptive Writing (prepared by the American Academy K8) are respectively composed of eight and seven criteria. One should note that the majority of the sub-criteria constituting the aforesaid rubrics are actually included in the Analytic Rating Scale for EFL Descriptive Writing. Thus, it seems that this empirically developed scale offers a more efficient, and probably less complicated, means of scoring descriptive essays.

Moreover, given the inclusion of such sub-criteria as cohesive devices, sensory imagery, and spatial order, it can be argued that, in comparison to other rating scales of its type, the Analytic Rating Scale for EFL Descriptive Writing is a more comprehensive scale which has, thus far, been constructed empirically. Last, one may argue that, given its assigning different weightings to the four criteria it incorporates, the end product of this mixed study offers a more logical means of marking descriptive essays. In this relation, it should be stated that intuition-based scales all tend to assign equal weightings to the criteria they include, which can be claimed to be a weak point. Apparently, criteria need to be prioritized and organized in a scale based upon their significance in relation to the construct which is to be assessed. Assigning equal weightings to various components, or criteria, tends to neglect this unquestionable fact.

Conclusion

Since the teaching and learning of the genre of description is of particular significance and as the literature included no study specifically aiming to construct a data-based rating scale for this genre, this three-strand mixed methods study developed the Analytic Rating Scale for EFL Descriptive Writing. The findings of this study point to a number of pedagogical implications. These implications can be categorized into those which relate to ELT instructional aspects and those which are connected with ELT assessment aspects. With respect to the former aspects, the most striking implication is for the teaching of the genre of description. As the scale encompasses the generic criteria and their sub-criteria, language teachers can base their teaching of this genre on these criteria and sub-criteria. Instruction hours should be proportionate to the weighting each criterion has. Therefore, language teachers can benefit from the scales by identifying both the content and the timing of their genre-based instructional practices.

With respect to the assessment aspects, the criteria and sub-criteria identified and included in the Analytic Rating Scale for EFL Descriptive Writing can be used to serve various purposes. Basically, these can be used for diagnostic, as well as achievement, purposes. Related to this issue, it is advised that alternative approaches to conventional assessment, i.e., peer assessment, self-assessment, and portfolio assessment, make use of this scale. As regards peer assessment, English language teachers can use the scale to inform their students of the areas on which they need to focus when assessing and providing feedback on their peers' descriptions. Likewise, when attempting to engage learners in self-assessment, EFL/ESL teachers can guide their students to self-assess their descriptive essays with reference to the criteria and sub-criteria included in this scale. Portfolio assessment, too, can be conducted based on these weighted criteria and their corresponding sub-criteria.

As illuminating as the findings and implications of this study appear to be, they should be considered in light of the limitations imposed upon it. First and foremost, due to the unavailability of native ELT experts, the second and third strands relied on the assistance of only one Iranian ELT expert helping the researcher corroborate the findings of Strand 2. Moreover, this mixed study was limited as its third strand did not

involve an external check by an ELT expert. Although this check was not essential, its presence might have enhanced the resulting analytic rating scale. Finally, owing to the non-existence of a valid analytic rating scale for EFL descriptive writing, the researcher could not examine the criterion-related validity of the Analytic Rating Scale for EFL Descriptive Writing. However, it should be noted that factor analysis definitely contributed to the construct validity of the resulting scale constructed in the present study.

References

Alexander, L. G. (1965). *Essay and letter writing*. London: Longman Publishing Group.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic scoring tell us? *System, 29,* 371-383.

Becker, A. (2011). Examining rubrics used to measure writing performance in U.S. intensive English programs. The CATESOL Journal, 22(1), 113-130.

Carr, N. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition texts. Issues in Applied Linguistics, 11, 207-241.

Clare, A., & Wilson, J. J. (2005). *Total English: Intermediate students' book*. London: Pearson Education Inc.

Cohen, A. (1994). Assessing language ability in the classroom. Boston, MA: Heinle & Heinle. *The descriptive essay rubric*. Retrieved June 2, 2012, from www.gnaedu.org. *The descriptive writing rubric*. Retrieved May 29, 2012, from msjohnsononline.weebly.com/uploads/.../1c_rubric_decriptive writin. Pdf

East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics, 13*, 1-21.

Freedman, S. W. (1991). *Evaluating writing: Linking large-scale testing and classroom assessment*. Retrieved April 21, 2012, from www.nwp.org/cs/public/download/nwp_file/50/OP27.pdf?xr=pcfile

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Taylor & Francis Group.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), Exploring the dynamics of second language writing (pp. 162-189). Cambridge: Cambridge University Press.

Hogue, A. (2008). First steps in academic writing (2nd ed.). New York: Pearson Education Inc.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. College Composition and Communication, 41, 201-213.

Iwashita, N., & Grove, E. (2003). A comparison of analytic and holistic scales in the context of a specific purpose speaking test. *Prospect, 18*(3), 25-35.

Jacobs, H. L., Zinkgraf, D. R., Wormuth, D. R., Hartfiel, V. F., and Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, Massachusetts: Newbury House.

Knoch, U. (2007). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5*, 1-36.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275-304.

Meyers, A. (2006). Writing with confidence: Writing effective sentences and paragraphs (8th ed.). London: Pearson Education Inc.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-faceted rasch measurement: Part 1. *Journal of Applied Measurement, 4*, 386-422.

Nakamura, Y. (2004). *A comparison of holistic and analytic scoring methods in the assessment of writing*. Retrieved March 10, 2012, from jalt.org/pansig/2004/HTML/Nakamura.html

Nelson, N. W., & Van Meter, A. M. (2007). Measuring written language ability in narrative samples. Reading & Writing Quarterly, 23, 287-309.

Nunnally, J. C. (1978). Psychometric theory. (2nd ed.). New York: McGraw-Hill.

Oshima, A., & Hogue, A. (2007). Introduction to academic writing. New York: Pearson Education Inc.

Pallant, J. (2013). A Step by step guide to data analysis using IBM SPSS: Survival manual. New York: The McGraw Hill Companies. *The rubric for descriptive writing*. Retrieved May 27, 2012, from www.pinterest.com/pin/285697170081236048/

Shaw, S. D., & Weir, C. J. (2007). Examining writing: Research and practice in second language writing. Cambridge: Cambridge University Press.

Soars, L., & Soars, J. (2009). *New headway: Intermediate student's book* (4th ed.). Oxford: Oxford University Press.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*, 163-182.

Strauch, A. O. (1994). *Writers at work: The short composition*. Cambridge: Cambridge University Press.

Thurstone, L. L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (1990). Communicative language testing. NJ: Prentice Hall Regents.

White, E. M. (1985). Teaching and assessing writing. San Francisco: Jossey-Bass.

Appendix A

Rating Criteria Checklist

Expert background: University degree: Major:

Please check the appropriate option below.

a. Gender: Male ☐ Female ☐

b. Age: 20-25 ☐ 26-30 ☐ 31-40 ☐ 41+ ☐

c. Years of experience in ELT: 1-2 ☐ 3-5 ☐ 6-10 ☐ 11-15 ☐ 16+ ☐

Directions: The checklist below lists the criteria thought to be of relevance when rating **descriptive essays.** You are requested to check the option which best indicates your judgment as to the importance of each criterion.

Criterion	Very important	Important	Fairly important	Not very important	Unimportant
1. Creative, interest-sparking title					
2. Relevant subtitles, revealing content of body paragraphs					
3. Inviting introductory sentences					
4. Body content closely addressing the title and subtitles					
5. Appropriate paragraphing of body content					
6. Use of cohesive devices (i.e., substitution, conjunctions, ellipsis, repetition, and lexical devices)					
7. Accurate, consistent spelling					
8. Use of descriptive, vivid vocabulary					
9. Coherence in body content (i.e., The author's track of thought can be followed without undue difficulty.)					
10. Accurate use of punctuation marks					
11. Adoption of a personal, unique voice or style when describing					
12. Consistent and accurate use of tenses in both active and passive voice					
13. Correct capitalization of letters					
14. Accurate use of prepositions, pronouns (subject and object), and determiners (articles, quantifiers, possessives, and demonstratives)					
15. Well-constructed sentences, written in correct word order and avoiding run-ons, fragments, and comma splices					
16. Inclusion of concrete, precise details in body paragraphs					
17. Use of figurative language (e.g., metaphor, hyperbole, and personification)					
18. Precision and accuracy in word formation (morphology) and use (awareness of appropriate register)					

Appendix B

Correlation Matrix

		c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18
Correlation	c1	1.000	-.010	.538	.026	.052	-.096	.053	.549	-.029	-.048	.509	.021	.070	-.048	.005	.579	.464	-.096
	c2	-.010	1.000	.078	.089	.085	.051	-.055	.027	.092	-.002	.027	.080	-.033	.100	.017	.025	.085	.086
	c3	.538	.078	1.000	.020	.020	-.120	-.024	.697	.024	-.092	.606	-.101	-.043	-.061	-.061	.612	.639	-.048
	c4	.026	.089	.020	1.000	.787	-.011	.086	.040	.756	.164	.109	.003	.128	.111	.014	.093	.096	.047
	c5	.052	.085	.020	.787	1.000	.042	.029	.031	.805	.105	.120	.057	.022	.124	.069	.086	.065	.103
	c6	-.096	.051	-.120	-.011	.042	1.000	-.006	-.067	.037	.060	-.001	.662	.050	.632	.626	-.019	-.144	.778
	c7	.053	-.055	-.024	.086	.029	-.006	1.000	-.001	.006	.669	.051	.010	.762	.053	-.080	.057	.029	.033
	c8	.549	.027	.697	.040	.031	-.067	-.001	1.000	.070	-.044	.766	-.119	.020	-.107	-.040	.768	.769	-.045
	c9	-.029	.092	.024	.756	.805	.037	.006	.070	1.000	.085	.135	.009	.014	.109	.011	.092	.110	.064
	c10	-.048	-.002	-.092	.164	.105	.060	.669	-.044	.085	1.000	.025	.109	.719	.079	-.009	.022	.006	.077
	c11	.509	.027	.606	.109	.120	-.001	.051	.766	.135	.025	1.000	.017	.020	-.069	-.019	.675	.699	-.001
	c12	.021	.080	-.101	.003	.057	.662	.010	-.119	.009	.109	.017	1.000	.099	.581	.614	-.031	-.178	.692
	c13	.070	-.033	-.043	.128	.022	.050	.762	.020	.014	.719	.020	.099	1.000	.059	-.020	.046	.041	.060
	c14	-.048	.100	-.061	.111	.124	.632	.053	-.107	.109	.079	-.069	.581	.059	1.000	.585	-.023	-.097	.725
	c15	.005	.017	-.061	.014	.069	.626	-.080	-.040	.011	-.009	-.019	.614	-.020	.585	1.000	.005	-.115	.641
	c16	.579	.025	.612	.093	.086	-.019	.057	.768	.092	.022	.675	-.031	.046	-.023	.005	1.000	.664	.003
	c17	.464	.085	.639	.096	.065	-.144	.029	.769	.110	.006	.699	-.178	.041	-.097	-.115	.664	1.000	-.081
	c18	-.096	.086	-.048	.047	.103	.778	.033	-.045	.064	.077	-.001	.692	.060	.725	.641	.003	-.081	1.000

Appendix C

Analytic Rating Scale for EFL Descriptive Writing

Criterion	Band Score	Descriptor
GENRE-RELATED ELEMENTS	25-34	creative and inviting title and introductory sentences closely addressing topic; inclusion of concrete, precise details given through descriptive, vivid vocabulary, figurative language, and sensory imagery; adoption of a personal, unique voice or style
	17-25	typical title and introductory sentences closely addressing topic; inclusion of adequate details given through descriptive vocabulary; adoption of a typical voice or style
	8-17	title and/or introductory sentences missing, or, if present, off topic or not related to body content; inclusion of inadequate details given through non-descriptive vocabulary; adoption of an inappropriate voice or style
LANGUAGE-RELATED ELEMENTS	22-28	well-constructed and varied structures; accurate and effective use of prepositional phrases, pronouns, and determiners; accurate and elaborate use of tenses in active/passive voice; sufficient and appropriate inclusion of cohesive devices (substitution, conjunction, ellipsis, reference, collocation); use of a wide range of accurately formed vocabulary of appropriate register
	17-22	well-constructed but limited structures; accurate use of prepositional phrases, pronouns, and determiners; accurate use of tenses in active/passive voice; appropriate inclusion of cohesive devices (substitution, conjunction, ellipsis, reference, collocation); use of a narrow range of accurately formed vocabulary of appropriate register
	12-17	minor problems in structural accuracy; structures of very limited variety; few problems in use of prepositional phrases, pronouns, and determiners; minor problems in use of tenses in active/passive voice; insufficient and in some cases inappropriate inclusion of cohesive devices (substitution, conjunction, ellipsis, reference, collocation); very limited lexical variety; minor problems in vocabulary formation and register appropriateness
	7-12	major problems in structural accuracy; structures of very limited variety; major problems in use of prepositional phrases, pronouns, and determiners; major problems in use of tenses in active/passive voice; insufficient and inappropriate inclusion of cohesive devices (substitution, conjunction, ellipsis, reference, collocation); very limited lexical variety; major problems in vocabulary formation and register appropriateness
CONTENT & ORGANIZATION	15-20	richly descriptive body content paragraphed and/or organized in spatial or other logical order; elaborate body content closely addressing topic and introductory sentences
	10-15	descriptive body content paragraphed and/or organized in spatial or other logical order; body content related to topic and introductory sentences
	5-10	inadequate body content not completely related to, or totally off, topic
MECHANICS	14-18	correct spelling and capitalization of words; accurate use of punctuation marks
	9-14	minor problems in spelling and capitalization of words; few problems of punctuation
	4-9	major problems in spelling and capitalization of words; major problems of punctuation