Iranian Journal of Applied Linguistics (IJAL), Vol. 16, No. 1, March 2013, 25-57

Using an Analytic Dichotomous Evaluation Checklist to Increase Inter- and Intra-rater Reliability of EFL Writing Evaluation

Masoumeh Ahmadi Shirazi [a]
Assistant Professor of TEFL, University of Tehran, Tehran, Iran

Received 24 October 2012; revised 19 January 2013; accepted 2 February 2013

Abstract

The present study reports the processes of development and use of an Analytic Dichotomous Evaluation Checklist (ADEC) which aims at enhancing both inter- and intra-rater reliability of writing evaluation. The ADEC consists of a total of 68 items that comprises five subscales of content, organization, grammar, vocabulary, and mechanics. Eight raters assessed the writing performance of 20 Iranian EFL students using the ADEC. Also, the raters were asked to rate the same sample of essays holistically based on Test of Written English (TWE) scale. To examine the inter-rater and intra-rater reliability of the ADEC, multiple approaches were employed including correlation coefficient, the dichotomous Rasch Model, and many-faceted Rasch measurement (MFRM). The findings of the study confirmed that the ADEC introduces higher reliability into scoring procedure compared with holistic scoring. Future research with greater number of raters and examinees may provide robust evidence to use analytic scale rather than holistic one.

[a] *Email address:* ahmadim@ut.ac.ir
Corresponding address: Department of English Language and Literature, Faculty of Foreign Languages and Literature, University of Tehran, between 15th St. and 16th St. - Karegar-e Shomali St., Tehran, Iran. P.O. Box: 14155-6553

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Keywords: Analytic scoring; Inter-rater reliability; Intra-rater reliability; EFL writing evaluation

Introduction

Consistency of measurement has been among the major concerns of language testers. To Weigle (2002), reliability is the consistency of measurement across different characteristics or facets of a testing situation. A number of factors are likely to distort reliability including the choice of raters, rating scales, descriptors, rubrics and scoring methods. To bring more consistency between raters, a number of scholars suggested rater training as a remedy (Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994, 1998). Rating scales may bring about more agreement among raters. Connor-Linton (1995) indicates how rating scales increase inter-rater reliability arguing that "rating scales (holistic or analytic) with relatively few proficiency levels promote inter-rater reliability by compression and shaping the possible space in which individual raters may express their responses to compositions" (p. 763). Knoch (2007) contends that detailed rating scales, when empirically based, result in higher rater reliability. She concludes that descriptors of rating scales when developed on an empirical basis are of great value. Controversy over which method of scoring, especially holistic or analytic ones, is likely to alter the degree of agreement among raters, still continues. O'Loughlin (1994) found that holistic ratings could result in higher levels of inter-rater agreement across raters; Song and Caruso (1996) found significant differences among raters when holistic scoring was utilized and this was not true for analytic rating; Bacha (2001) found high levels of inter- and intra-rater reliability for both holistic and analytic rating scales; Barkaoui (2007) examined holistic vs. multiple-trait scoring and documented higher inter-rater agreement when holistic scoring method was employed. However, in another study, Barkaoui (2008) found that raters tend to be more self-consistent while utilizing the analytic scale. This study is an endeavor to develop a method of scoring, considering empirically driven rating scales and more detailed descriptors (Knoch, 2009), which may result in higher inter- and intra-rater reliability.

Literature review

In recent years the significant role of raters in assessing the writing performance of learners has proved to be a determining factor because different ratings of raters can introduce some degrees of subjectivity as a potential source of error into scoring and, hence, pose a threat to reliability of scoring results. As pointed out

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

earlier, various sources of errors deemed influential in scoring written essays include the choice of method of scoring, raters, rating scales, descriptors and rubrics.

The controversy on the priority of holistic and analytic methods of scoring is a prevailing issue in writing assessment literature. On the one hand, the holistic method of scoring is prioritized because it can be less time consuming and cost-effective (e.g. Hughes, 1989; Nakamura, 2004; White, 1998; Wolcott and Legg, 1988 as cited in Blatner, 1999). Moreover, it is thought to be an authentic rating method (Nakamura, 2004; White, 1998) and a number of scholars associate reliability to holistic scoring (Barkaoui, 2007; Charney, 1984; Cooper, 1977; Harris, 1968; O'Laughlin, 1994; White, 1998). On the other hand, this method has been criticized for several reasons; for example, Huot (1990) and White (1998) doubt the validity of this method. Elbow (1993) casts doubts on the reliability of holistic scoring. Similarly, Song and Caruso (1996) found significant differences among raters when holistic scoring was used. Some scholars argue that holistic scoring would direct the attention of raters to certain features of the text at the cost of not paying due attention to other important features (e.g. Blatner, 1999; Francis, 1977 as cited in Weir, 1990; Hamp-Lyons and Kroll, 1997; Nakamura, 2004; Sakyi, 2001; Wolcott and Legg, 1998 as cited in Blatner, 1999); and, finally, Hamp-Lyons and Kroll (1997) raise some questions on the issue of test accountability.

Having considered the demerits associated with holistic scoring, we will examine why some scholars prefer analytic method of scoring (e.g. Bacha, 2001; Barkaoui, 2008; Charney, 1984; Connor-Linton, 1995; Elbow, 1993; Hamp-Lyons, 1995; Heaton, 1975; Nakamura, 2004; Raimes, 1990; Sasaki and Hirose, 1999; Wolcott and Legg, 1998 as cited in Blatner, 1999; Weigle, 2002; White, 1998). It is argued that analytic scoring may sit on the following vantage points: higher reliability (Bauer, 1981; Hartog, Rhodes, & Burt, 1936; and Cast, 1939 as cited in Weir, 1990; Hughes, 1989). Maybe, the reason lies in the multiple scores given to each separate part which may, in turn, cause variations in measurement; hence, increasing the reliability. There are some studies, however which prioritize neither holistic nor analytic method of scoring. Analytic scoring can be pushed aside since rating dimensions are highly correlated not only among themselves but also with holistic scores (e.g. Bacha, 2001; Huot, 1990; Veal & Hudson, 1983; Wiseman, 2006). Bacha (2001) found high levels of inter- and intra-rater reliability for both

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

holistic and analytic rating scales. Lee, Gentile, and Kantor (2006) found that the analytic and holistic scores highly correlated with each other resulting in similar reliability indexes. If this is the case, then other factors may affect the reliability of writing assessment. Lee, Gentile, and Kantor (2006) considered essay length affecting reliability; a number of scholars investigated the effect of the type, number, and wording of prompt on reliability (e.g. Breland, Lee, and Muraki, 2005; Chalhoub-Deville, 1995; Hamp-Lyons and Kroll, 1997). Among these factors, the choice of raters seems to be the most crucial one since it serves as the heart of scoring process.

Different issues related to raters have been the focus of many studies in writing assessment literature. Raters with different language backgrounds were found to be inconsistent in scoring. ESL/EFL raters scored students' writings differently compared with their counterpart native English raters (e.g. Brown, 1991; Cumming, Kantor, Powers, 2001; Johnson and Lim, 2009; O'Loughlin, 1994; Shi, 2001). ESL and non-ESL raters scored the papers written by native or non-native English test takers differently; even if the scores they reached were quite similar , the components they considered were different (e.g. Carlson, Bridgeman, Camp, & Waanders, 1985; O'Loughlin, 1993; Sweedler-Brown, 1993; Vann, Meyer, & Lorenz, 1984). Also an extensive body of research has addressed writing assessment by raters' being lay or professional (Cumming, 1990; Shohamy, Gordon, & Kraemer, 1992; Schoonen, Vergeer, and Eiting, 1997; Wolfe & Ranney, 1996).

In addition to rater variables affecting the reliability of scoring, rating scales can either contribute or reduce reliability of scoring. Connor-Linton (1995) asserts that both holistic and analytic rating scales when limited to few proficiency levels would increase inter-rater reliability because the scope of scores is not too broad to confuse raters. However, Bachman (1990) refers to the problem associated with the number of level descriptors; he states that the points on a rating scale are "typically defined in terms of either the types of language performance or the levels of abilities that are considered distinctive at different scale points" (p. 36). Knoch (2007) suggests that rating scales and their descriptors should be developed empirically. She states that "rating scale developers should consider this method of scale development as a viable alternative to intuitive development methods which are commonly used around the world" (p. 122). She further contends that the more

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

detailed this empirically developed rating scale is, the higher the rater reliability would be.

Descriptors also affect the rating processes. A descriptor as Davies, Brown, Elder, Hill, Lumley, and McNamara (1999) depict, is a "statement which describes the level of performance required of candidates at each point on a proficiency scale" (p. 43). Pollitt and Murray (1996) think of scores being affected not only by testees' ability but the way a rater interprets the descriptors as well. North and Schneider (1998) cast doubts on the validity and reliability of the scale descriptors and maintain that "there is no guarantee that the description of proficiency offered in a scale is accurate, valid or balanced; raters may actually be trained to think the same" (p. 220). Shaw (2002: 13) holds that "the shared interpretation of rating scale descriptors cannot be assumed and unless the rating scale points define clearly-differentiated levels or bands, precise interpretation by different audiences will vary." Considering three approaches to formulating descriptors as suggested by North (2003), Knoch (2010) suggests that the descriptors of each band, when concrete formulation is attempted, can be converted to a checklist of 'yes' or 'no' questions. She stresses that such descriptors usually result in greater inter-rater reliability.

Rubrics, the wordings or statements differentiating each level descriptor, may lead to different interpretation by raters, less consensus among raters, and lower reliability. Matthews (1990) contends that there are many problems associated with categories and subcategories in the assessment criteria:

> … they are not clearly defined; they are not always appropriate for the particular task assigned; or they straddle too obviously the linguistic/non-linguistic divide. The same descriptions make reference to abilities which were not tapped by the task set. Bare statements such as 'may pause to prepare next utterance' are of little assistance to the assessor, because they describe behavior which is ambiguous. (p. 119)

DeRemer (1998) believes that "scoring rubrics which identify criteria for assigning scores are relied upon for the achievement of reliable scoring" (p. 8). However, she raises the following question about rubrics: "do the rubric guidelines adequately characterize lexical, syntactical and semantic characteristics of a text's organization or do the guidelines offer highly-abstracted and not widely-understood concepts" (p. 26)? Along the same line, Marby (1999) asserts that

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

"writing rubrics can fail to predict the actual features of a student's writing, thereby creating a mismatch between scoring criteria and actual performance" (675). Reviewing 75 studies, Jonsson and Svingby (2007) reported that reliability may be enhanced by the use of rubrics which are analytic, topic-specific, and complemented with exemplars and/or rater training.

As was previously mentioned, a number of problems are associated with both holistic and analytic scoring. This study attempts to develop an analytic dichotomous evaluation checklist (ADEC) assuming that detailed rating criteria can lead to more consistency (higher reliability thereof). The rationale for choosing the above name for the instrument developed in this study is threefold. First, it is assumed that analytic scoring would result in higher reliability. Second, the rating criteria are evaluated dichotomously, that is ADEC requires raters to check mark the presence or absence of a trait in a piece of writing, therefore, the scores are within a limit which, in turn, may lead to further agreement among raters. Finally, the term *checklist* implies two specifications: (1) to present a number of criteria to evaluate and (2) to check for the presence or absence of these criteria. This checklist, if proved to be reliable, can suggest an alternative scoring procedure which may result in higher inter- and intra-rater reliability. The study addresses the following research questions:

1. Does the use of the analytic dichotomous evaluation checklist result in higher inter-rater reliability?
2. Does the use of the analytic dichotomous evaluation checklist result in higher intra-rater reliability?

Methodology

Participants

The raters of the study were four English native speakers and four Persian speakers of English with TEFL education background in Iran. Table 1 presents a quick profile of the raters' characteristics.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Table 1
Raters' Characteristics

	Raters	Education	Major	Gender	Age	Teaching Experience	Assessment Experience
Native	A	MA	Applied Linguistics	Male	>50	28	18
	B	PhD	F/SL Education	Female	31-40	17	7-10
	C	MA	TESOL	Female	31-40	15	15
	D	MA	TEFL	Female	<30	7	5
Non-native	E	MA	TEFL	Female	31-40	5	2
	F	MA	TEFL	Female	<30	4	3
	G	MA	TEFL	Female	>50	15	10
	H	PhD	TEFL	Female	31-40	7	4

Instruments

ADEC as the main instrument. The ADEC was developed in several stages. During the first stage, the theoretical basis of the ADEC was established. The rubrics found in the literature were gathered, classified, and divided into writing features. Then their frequency of occurrence in the literature helped the researcher to determine priorities in the ADEC. In the next stage, the researcher worked for qualitative support for the newly-developed checklist. This stage began with collecting think-aloud protocols from raters. The transcripts of the protocols were analyzed carefully in order to find clues for formulating ADEC items. Next the reliability and validity of the ADEC were checked.

A taxonomy of writing features was developed just after they were specified; then, they were put into the macro categories of content, organization, grammatical, lexical and mechanical features. Some features were called *miscellaneous* since they either bore little or no relation to the main categories or could be classified under almost all of the macro categories.

After reducing the names to codes, the basis of taxonomy was set up. The number of writing features came to 288. The frequency of occurrence of macro categories/features is provided in Table 2.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Table 2
The Frequency of Occurrence of Macro-categories of Writing

Macro-categories	Frequency of occurrence
Content	~68
Organization	~65
Grammar	~66
Vocabulary	~50
Mechanics	~33
Miscellaneous	~1-8

In a preliminary try, the number of the most frequent items came to 80 (see Appendix A for these writing components). However, the items went under some modifications. The modifications included: (*1) Wording of the Items, (2) Providing Examples, and (3) Reducing Items*

After these modifications, the ADEC incorporated 68 items (see Appendix B) whose inclusion in the final instrument would depend on (1) the qualitative support of protocol analysis, (2) the proof for its validity and reliability.

Qualitative Support for the ADEC

Verbal protocol analysis (VPA). The type of VPA used for data collection in this study was non-mediated concurrent Think-Aloud (TA). Table 3 presents the TA components ranging in number between *3, 141* and *10, 749* words.

Table 3
The Approximate Number of Words Counted through TA

	Raters	Approximate number of words
Non-native	RA	3,377
	RB	3,141
	RC	6,105
	RD	3,234
Native	RE	10,042
	RF	6,055
	RG	5,541
	RH	10,749

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Results of protocol analysis. The results of protocol analysis provided support for the inclusion of macro- and micro-categories of the ADEC. The raters commented on five macro-categories: content, organization, grammar, vocabulary, and mechanics; also the number of micro-categories reached 126 (see Appendix C for the full classification).

Validity of ADEC

Rasch model was used to probe the construct validity of the newly-developed instrument. To test unidimensionality of the data, Winsteps reports fit indices. The main indices include infit and outfit. According to Bond and Fox (2007), an acceptable infit range should be within 0.75 - 1.3. The results of Rasch analysis confirm that almost all items have acceptable fit indices, hence, supporting unidimensionality of the ADEC. Item map (Figure 1) of the ADEC items shows that the instrument (i.e., ADEC) is easy. Also, items are mostly between +2 and -2 logits bearing witness to the fact that they measure the same construct.

Reliability of the ADEC

Cronbach alpha, a measure of internal consistency, was chosen since it is appropriate for the continuous data gathered in this study. In order to examine the reliability of the checklist, alpha analysis was run the result of which (.70) confirmed an acceptable reliability index. Moreover, Rasch analysis also showed high separation reliability for persons (.87) as Table 4 illustrates.

Table 4

Summary Statistics for Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	53.3	68.0	1.87	0.38	0.99	0.1	1.04	0.1
S.D.	9.5	0.0	1.11	0.12	0.14	0.9	0.57	1.0
MAX.	67.0	68.0	4.81	1.02	1.53	4.0	3.14	3.8
MIN.	24.0	68.0	-0.77	0.28	0.63	-3.2	0.18	-2.3

REAL RMSE .41 ADJ.SD 1.03 SEPARATION 2.54 PERSON RELIABILITY .87
MODEL RSME .40 ADJ.SD. 1.03 SEPARATION 2.60 PERSON RELIABILITY .87
S.E. OF PERSON MEAN = .09

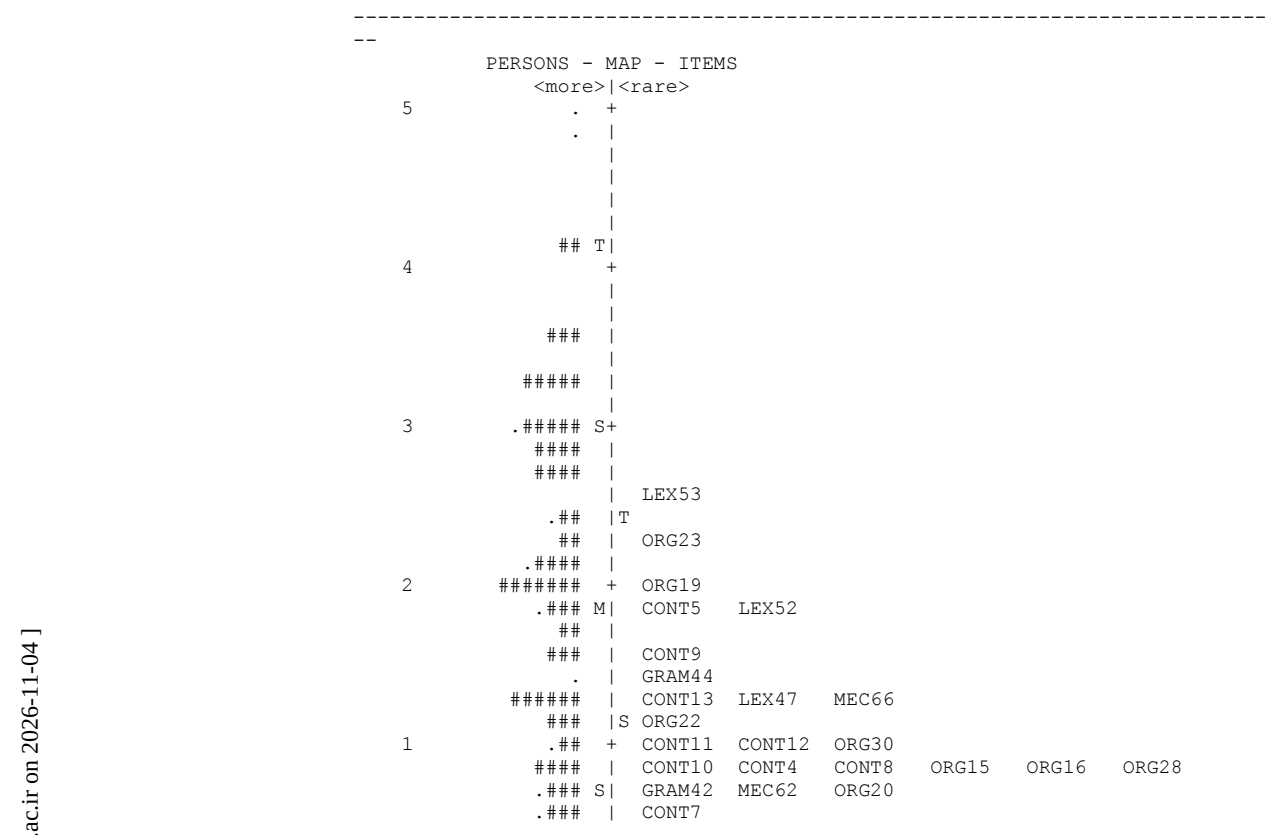
[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Quantitative Data Collection

At first the raters scored 20 essays holistically and then they attempted analytic scoring neither of which required TA. Each holistic scoring session lasted not more than 30 minutes whereas each analytic scoring took 45 minutes.

Quantitative and Qualitative Data Collection

The quantitative and qualitative data collection involved the raters in both scoring and Thinking Aloud. TA accompanied just holistic scoring and not the analytic one. Each rating session according to holistic scoring and involving TA took a maximum time of three hours.



[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

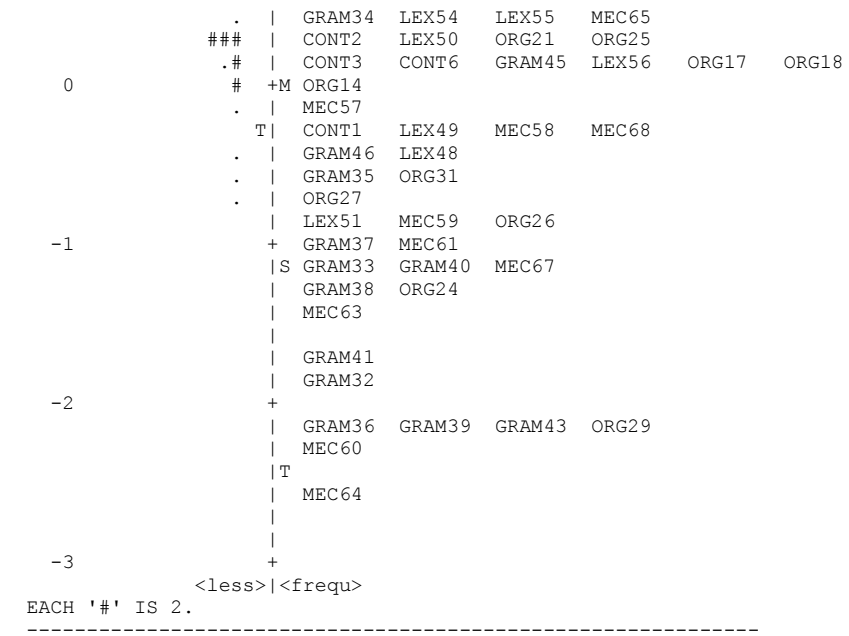


Figure 1: Item Map of the ADEC items

Data Collection Duration

As indicated in Table 5, data collection took two months and a half.

Table 5
Rating Time line

Week 1	2 Weeks Interval	Week 4	1 Week Interval	Week 6	2 Weeks Interval	Week 9
Holistic Scoring + TA		Holistic Scoring		ADEC		ADEC

Raters' Orientation

The ADEC scoring guide familiarized the raters with the newly-developed instrument. The main categories and their subcategories were explicated so that the raters could grasp the rating process.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Raters' Scoring

The raters were briefed on how to score the essays. Holistic scores followed a band scale including six bands. The ADEC required the raters to check mark the presence or absence of the given traits. Positive answers earned one while the negative scored zero. The aggregate of positive answers made the total score for the essays.

Data Analysis

In order to investigate if the use of the ADEC results in higher inter-rater reliability, correlational analyses and many-faceted Rasch measurement model were used. Table 6 provides inter-rater reliability coefficients for both types of rating. As can be observed, correlation coefficient of analytic scorings almost remained the same showing that the raters were more consistent whereas holistic scorings changed highlighting less consistently among the raters.

Table 6
Inter-rater Correlation Coefficients among Raters

	Analytic 1	Holistic 1	Analytic 2	Holistic 2
Inter-rater Coefficient	.71	.66	.72	.70
Average (Using Z Transformation)	.88	.78	.89	.87

FACETS Analysis

In this study, three facets were taken into account: judge severity, item difficulty, and examinee ability. The FACETS Software, Version 3. 6. 0, (Linacre, 2008) was used for the analysis in order to provide information on the facets including judges' consistency.

Inter-rater reliability (IRR) reported under MFRM can be interpreted with regard to the purpose for which the ratings are collected. If we would like our raters to behave like rating machines (i.e., exact agreement with criteria determined by rating scales), then MFRM reports higher inter-rater exact agreement than inter-rater expected agreement. In contrast, if raters act like independent raters, inter-rater exact agreement should be close to expected agreement or inter-rater expected agreement should be higher than inter-rater exact agreement. As Linacre (2008)

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

observes, "If Obs% ≈ Exp% then the raters may be behaving like independent experts; If Obs% ≫ Exp% then the raters may be behaving like rating machines" (p. 200).

As Tables 7 and 8 illustrate, in analytic scoring exact agreement is higher than expected one, whereas the reverse is true for holistic scoring. In fact, the raters acted like rating machines in the case of analytic scoring which contributed to higher reliability, but they behaved like independent raters when they used holistic method of scoring, thus, introducing less consistency into the scoring results.

Table 7
Raters Measurement Report (ADEC)

Raters	Difficulty estimate	Error estimate	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Exact Obs%
C	-.83	.06	.99	-.3	.96	-.8	1.03	72.9
D	-88	.06	1.01	0.1	1.01	0.2	.99	72.2
F	-1.37	.07	1.07	2.1	1.25	3.2	.85	74.5
G	-1.69	.07	1.02	.4	1.05	.6	.97	77.5
E	-2.07	.08	.94	-1.2	.92	-.7	1.06	78.4
B	-2.18	.08	.99	-.1	1.01	.1	1.03	70.7
A	-2.26	.08	1.03	.5	1.25	1.9	.94	81.9
H	-2.76	.10	.88	-1.8	.82	-1.2	1.10	81.9

Model Populn: RMSE .08 Adj (True) S. D. .64 Separation 8.26 Reliability (not Inter-rater) .99
Model Populn: RMSE .08 Adj (True) S. D. .68 Separation 8.84 Reliability (not Inter-rater) .99
Model, Fixed (all same) chi-square: 569.9 d. f. 7 Significance (probability) : .00
Model, Random (normal) chi-square: 6.9 d. f. 6 Significance (probability): .33
Inter-rater agreement opportunities: 12444 **Exact agreement: 9463 = 76.0%**
Expected: 8938.5 = 71.8%

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Table 8
Rater Measurement Report (Holistic Scoring)

Raters	Difficulty estimate	Error estimate	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Exact Obs %
G	.16	.26	.97	.0	.96	.0	1.04	25.7
E	.03	.26	.75	-.8	.73	-.9	1.37	22.9
H	-.38	.26	.76	-.7	.77	-.7	1.20	30.7
A	-.65	.27	.61	-1.3	.62	-1.3	1.33	25.7
F	-.93	.27	1.10	.4	1.13	.4	.93	27.1
B	-1.38	.28	.68	-1.0	.70	-.9	1.34	27.9
D	-2.12	.30	1.88	2.3	1.85	2.2	-.09	25.7
C	-2.21	.30	.80	-.5	.95	-.0	1.07	24.3

Model Populn: RMSE .28 Adj (True) S. D. .80 Separation 2.91 Reliability (not Inter-rater) .89
Model Populn: RMSE .28 Adj (True) S. D. .86 Separation 3.13 Reliability (not Inter-rater) .91
Model, Fixed (all same) chi-square: 71.8 d. f. 7 Significance (probability) : .00
Model, Random (normal) chi-square: 6.4 d. f. 6 Significance (probability): .33
Inter-rater agreement opportunities: 560 **Exact agreement: 147 = 26.2%**
Expected: 153.8 = 27.5%

To check for the higher intra-rater reliability through using ADEC, we carried out three analyses: *Correlational analyses, coefficient alpha, self-consistency through FACETS.*

Correlational Analyses

One way to investigate the intra-rater reliability is to compute correlation coefficients. The results of correlations within a single rater for both analytic and holistic scoring are summed up in Tables 9 and 10.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Table 9
Intra-rater Correlations of ADEC for Native and Non-native Raters

	Raters	RA	RB	RC	RD
	RA	.69**			
Native	RB		.27		
	RC			.80**	
	RD				.78**
		RE	RF	RG	RH
	RE	.90**			
Non-native	RF		.77**		
	RG			.92**	
	RH				.80**

**Correlation is significant at the 0.01 level

Table 10
Intra-rater Correlations of Holistic Scoring for Native and Non-native Raters

	Raters	RA	RB	RC	RD
	RA	.79**			
Native	RB		.30		
	RC			.20	
	RD				.32
		RE	RF	RG	RH
	RE	.89**			
Non-native	RF		.42		
	RG			.90**	
	RH				.87**

** Correlation is significant at the 0.01 level

When compared, raters show more consistency in analytic scoring. Rater A and Rater H came more consistent in holistic scoring. Rater B simply was consistent in neither of scorings. Rater C, Rater D, and Rater F showed much more consistency in analytic scoring.

Besides, common variance r^2 was calculated to show the amount of overlap. As Tables 11 and 12 clearly show, raters are more self-consistent when ADEC is used; while low overlap can be seen when raters score the essays holistically.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Table 11
Intra-rater Coefficient of Determination of ADEC for Native and Non-native Raters

	Raters	RA	RB	RC	RD
	RA	.47			
Native	RB		.07		
	RC			.64	
	RD				.60
		RE	RF	RG	RH
	RE	.81			
Non-native	RF		.59		
	RG			.84	
	RH				.64

Table 12
Intra-rater Coefficient of Determination of Holistic Scoring for Native and Non-native Raters

	Raters	RA	RB	RC	RD
	RA	.62			
Native	RB		.09		
	RC			.04	
	RD				.10
		RE	RF	RG	RH
	RE	.79			
Non-native	RF		.17		
	RG			.81	
	RH				.75

Coefficient Alpha

To compute coefficient alpha, two ratings for each individual rater are added, then two variances should be computed: (1) the variance of the ratings for a given rater and (2) the sum of the variances of different raters' ratings (Bachman, 1990, p. 181). Table 13 provides the results of these computations for both analytic and holistic scoring.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Table 13
Reliability within Raters for the Analytic and Holistic Scoring

	Raters	Analytic	Holistic
Native	A	.82	.88
	B	.38	.48
	C	.88	.34
	D	.87	.48
Non-native	E	.94	.96
	F	.87	.60
	G	.96	.96
	H	.87	.94

As is shown, Rater A and Rater H scored more consistently in holistic scoring which supports the results of correlational analyses. Rater E and Rater G performed equally well in the two scoring procedures. Rater B failed to be consistent whereas Raters C, D, and F showed much more consistency in analytic scoring.

Self-consistency through FACETS

FACETS analysis provides fit statistics for each facet specified in the study. As pointed out previously, raters, items, and examinees comprised the main facets of this study. As for the raters, fit statistics show rater consistency. Wright and Linacre (1994) suggest the following reasonable mean square ranges for infit and outfit is between upper and lower limits of 1.3 and 0.7 respectively. As Table 14 depicts, the infit values fall within an acceptable range.

Table 14
Rater Measurement Report for the ADEC

Raters	Infit Mnsq	ZStd	Outfit Mnsq	ZStd
RA	1.03	.5	1.25	1.9
RB	.99	-.1	1.01	.1
RC	.99	-.3	.96	-.8
RD	1.01	.1	1.01	.2
RE	.94	-1.2	.92	-.7
RF	1.07	2.1	1.25	3.2
RG	1.02	.4	1.05	.6
RH	.88	-1.8	.82	-1.2

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Table 15 provides raters' measurement report for holistic scoring. Rater D apart, the other raters showed consistency in scoring.

Table 15
Rater Measurement Report for Holistic Scoring

Raters	Infit Mnsq	ZStd	Outfit Mnsq	ZStd
RA	.61	-1.3	.62	-1.3
RB	.68	-1.0	.70	-.9
RC	.80	-.5	.95	.0
RD	1.88	2.3	1.85	2.2
RE	.75	-.8	.73	-.9
RF	1.10	.4	1.13	.4
RG	.97	.0	.96	.0
RH	.76	-.7	.77	-.2

Discussion

This study was primarily designed to probe the contribution of ADEC to raters' consistency in scoring writing. Between-group consistency can suggest that raters can obtain similar results by using ADEC. Our findings indicated that the raters reached further agreement when scored essays analytically rather than holistically. The reason for lower inter-rater reliability of holistic scoring can be due to raters' training effect. The ADEC was applied by the raters without training sessions; they were not forced to agree and they did not experience scoring under imposed conditions. As indicated previously, holistic scoring forces raters to aggregate a set of objective hypotheses imposed by the criteria, determined by rubrics, and mixed with raters' own value systems, whereas analytic scoring encourages raters to sum up the objective quantities of clearly-stated features to come up with a right score for a piece of writing. As is clear, the lack of consistency among raters can be due to subjective nature of holistic scoring. Elbow (1993) casts doubt on the reliability of holistic scoring when he asserts that "reliability in holistic scoring is not a measure of how texts are valued by real readers in natural settings, but only of how they are valued in artificial settings with imposed agreements" (p. 189).

Perhaps more important than inter-rater reliability is the question of how internally consistent the raters are, i.e. intra-rater reliability. To date, a few studies (e.g. Cho, 1999) have addressed the concept of intra-rater reliability. This study

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

indicated conflicting results about intra-rater reliability of holistic vis-à-vis analytic scoring. Correlational analyses showed that the majority (five of eight) of raters were consistent in their analytic scoring. Rater A and Rater H were more consistent in holistic scoring. Rater B showed inconsistency in both scoring systems. However, due to the fact that correlations can sometimes result in false impressions, coefficient alpha was also computed to see whether considering mean differences in two sets of scores can change the result of simple correlations. Rater A and Rater H were again more consistent in holistic scoring; Rater B failed to be consistent in both holistic and analytic scoring; Rater E and Rater G did not show any difference in their analytic and holistic scores; in fact, they performed equally well in the two scoring procedures. But Rater C, Rater D, and Rater F showed more consistency in holistic scoring. The results of FACETS analysis provided more support for analytic scoring. Using the ADEC, all raters showed self-consistency whereas in holistic scoring all raters except for one were internally consistent. Although the analytic scoring stood higher than the holistic scoring in terms of causing self-consistency among the raters, it showed a trend, that is, they both induced the same consistency level within individual raters. The differences between raters' self-consistency may emanate from their rater types (Eckes, 2008, 2012), rater training context (Wolfe & McVay, 2010; Sugita, 2011), rater experience (Lim, 2011) and other factors affecting the scoring process. The present researcher suggests deeper investigations of writing assessment based on the findings of this study and the challenges raised.

Conclusion

The findings of the present study show that inter-rater reliability was higher while scoring analytically rather than holistically; in fact, holistic scoring turned out to be much more challenging for the raters than analytic scoring. Second, intra-rater reliability is so significant a concept as inter-rater reliability; the current study attempted to thrust intra-rater reliability into limelight by appreciating the great import of this concept. The hunch is that if raters happen to be inconsistent in their scoring, they are liable to show inconsistency with other raters. Finally, we assume that the ADEC developed in this study may take a turn to relieve raters' difficulty in judgment by itemizing significant features of an essay and, hence, facilitating the process of mapping these criteria onto evaluation of writing tasks. Further investigations are required to show the efficiency of the ADEC considering such factors as rating context, rater types, motivational style, educational background, teaching experience and other features that may alter the consistency of scoring.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Notes on Contributors:

Masoomeh Ahmadi Shirazi *received Ph.D. in TEFL at the University of Tehran in December 2008. She received her MA in TEFL from the same university in May 2003. She is now an assistant professor at the faculty of foreign languages and Literatures in University of Tehran. She takes interest in academic writing, writing assessment, SLA, research methodology, and statistics.*

References

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

Bacha, N. (2001) Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29* (3), 371-383.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12* (2), 86-107.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Unpublished PhD dissertation). University of Toronto, Canada.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7* (1), 54-74.

Bauer, B. A. (1981). A study of the reliabilities and cost-effectiveness of three methods of assessment for writing ability. (*ERIC Document Reproduction Service No. ED216357*).

Blatner, N. (1999). Demystifying writing assessment: Empowering teachers and students. [Review of the book: *An overview of the writing assessment: Theory, research, and practice*]. *Assessing Writing, 6* (2), 229-237.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences. (2nd ed.).* Mahwah NJ: Lawrence Erlbaum Associates.

Breland, H., Lee, Y.-W., & Muraki, E. (2005). Comparability of TOEFL CBT essay prompts: response-mode analysis. *Educational and Psychological Measurement, 65* (4), 577-595.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12* (1), 1-15.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(4), 587-603.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. (*TOEFL Research Report No. RR-19*). Princeton NJ: Educational Testing Service.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12* (1), 16-33.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18* (1), 65-81.

Cho, D. (1999). A study on ESL writing assessment: Intra-rater reliability of ESL compositions. *Melbourne Papers in Language Testing, 8* (1), 1-24.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29* (4), 762-765.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: National Council of Teachers of English.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7* (1), 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype written tasks: An investigation into raters' decision making and development of a preliminary analytic framework. (*TOEFL Monograph Series No. 22*). Princeton NJ: Educational Testing Service.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5* (1), 7-29.

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

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9* (3), 270-292.

Elbow, P. (1993). Ranking, evaluating, and liking: Sorting out three forms of judgment. *College English, 55* (2), 187-205.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Elder, C. (1993). How do subject specialists construe language proficiency? *Language Testing, 10* (3), 235-54.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29* (4), 759-762.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12* (2), 1-9.

Hamp-Lyons, L., & Kroll, B. (1997). Issues in ESL writing assessment: An overview. *College ESL, 6* (1), 52-72.

Harris, D. P. (1968). *Testing English as a second language*. New York: McGraw Hill.

Heaton, J. B. (1975). *Writing English language tests*. London: Longman Group, Ltd.

Henning, G. H. (in press). *Language test development.* Rowley Massachusetts: Newbury House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60* (2), 237-263.

Johnson, S. J., & Lim, S. G. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26* (4), 485-505.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2* (2), 130-144.

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12* (2), 108-128.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26* (2), 276-304.

Lee, Y-W., Gentile, C., & Kantor, P. (2006). *Analytic essay scoring: Validity evidence and diagnostic potential for automated scores*. Paper presented at the annual meeting of National Council on Measurement in Education (NCME) San Francisco: California.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28* (4), 543-560.

Linacre, J. M. (2008). *Facets Rasch measurement computer program*. Chicago: Winsteps.com

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Marby, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappa, 80* (9), 673-680.

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT, 44* (2), 117-121.

Nakamura, Y. (2004). A comparison of holistic and analytic scoring methods in the assessment of writing. *JALT Pan SIG Proceedings,* 45-52.

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

O'Loughlin, K. (1994). The assessment of writing by English and ESL teachers. *ARAL, 17* (1), 23-44.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic and N. Saville (Eds.), *Performance testing, cognition and assessment* (74-91). Selected papers from the 15th Language Testing Research Colloquium Cambridge and Arnhern. Cambridge: Cambridge University Press.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill NJ: Hampton Press.

Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly, 24* (3), 427- 442.

Sakyi, A. (2001). Validation of holistic scoring for ESL writing assessment: A study of how raters evaluate ESL compositions on a holistic scale. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 130-153). Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16* (4), 457- 478.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: expert readers vs. lay readers. *Language Testing, 14* (2), 157-184.

Shaw, S. (2002). IELTS writing: Revising assessment criteria and scales (Phase 2). (*Cambridge ESOL Research Notes No. 10, pp. 10-13*). Cambridge: University of Cambridge.

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing*. Cambridge: Cambridge University Press.

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18* (3), 303-325.

Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal. 76* (1), 27-33.

Song, C. 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* (2), 163-182.

Sugita, Y. (2011). *Differences in raters' severity, consistency, and biased interactions between trained and untrained raters in the context of a task-based writing performance test*. Proceedings of the 16th Conference of Pan-Pacific Association of Applied Linguistics. Hong Kong, pp.171-176.

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing, 2* (1), 3-17.

Vann, R., Meyer, D., & Lorenz, F. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18* (3), 427-440.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English, 17* (3), 285-296.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11* (2), 197-223.

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.

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

White, E. M. (1998). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance (2nd ed.).* Maine: Calendar Islands Publishers.

Wiseman, C. S. (2006). *The validation and comparison of a holistic and an analytic scoring rubric in the assessment of second language writing*. Poster session presented at the six annual meeting of the East Coast Organization of Language Testers Maryland: Washington D.C.

Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training context*. US: Pearson.

Wolfe, E. W., & Ranney, M. (1996). Expertise in essay scoring. In D. C. Edelson and E. A. Domeshek (Eds.), *Proceedings of ICLS 96* (pp. 545-550).

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Charlottesville VA: Association for the Advancement of Computing in Education.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8* (3), 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm.

Appendices

Appendix A

Preliminary Design of the ADEC

Items	Y	N
1. Is there a thesis statement?		
2. Is the thesis statement related to the prompt?		
3. Is the thesis statement the copy of the prompt?		
4. Is the topic sentence an exemplification of the prompt?		
5. Does the writer interpret the prompt correctly?		
6. Does the writer use a part of the prompt to write about?		
7. Does the writer change the topic?		
8. Does the writer rephrase the prompt?		
9. Does the writer repeat ideas throughout the text?		
10. Is there a need for the reader to re-read the text to understand it?		
11. Does each paragraph contain a topic sentence?		
12. Is the topic sentence of each paragraph followed by examples and reasons?		
13. Do supporting sentences develop the main topic?		
14. Does the text have an introduction?		
15. Does the text have a conclusion?		
16. Do paragraphs have chronological order?		
17. Do paragraphs have logical order?		
18. Does the writer use enumerators to show paragraphing?		
19. Does the writer use cohesive devices to join the paragraphs?		
20. Does the writer underuse cohesive devices?		
21. Does the writer overuse cohesive devices?		
22. Does the writer stick to topic?		
23. Does the writer make use of signaling phrases to make paragraphs coherent?		

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

24. Are the topic sentences of all paragraphs understandable?		
25. Does the logic of ideas govern the whole text?		
26. Does the writer make use of cohesive devices in a mechanical way?		
27. Is it clear what words like "it", "that", and "they" refer to?		
28. Is the piece formal in style?		
29. Is the piece informal in style?		
30. Is the piece neutral in style?		
31. Is the tone of the text informative?		
32. Does the writer give accurate and logical information?		
33. Does the writer give detailed but off-topic information?		
34. Do cohesive devices include conjunctions?		
35. Do cohesive devices include lexical sets?		
36. Do cohesive devices include articles?		
37. Does the writer use possessive adjectives for the sake of cohesion?		
38. Does the writer use compounding with the word "and"?		
39. Does the writer use simple sentences?		
40. Does the writer use adverbial, adjective or gerund phrases?		
41. Does the writer use embedded sentences?		
42. Does the writer use correct tenses?		
43. Are modals used correctly?		
44. Is there subject-verb agreement?		
45. Does the writer use quantifiers?		
46. Is the manipulation of quantifiers correct?		
47. Is there a repetitious use of a single grammatical pattern?		
48. Is there a correct word order pattern throughout the text?		
49. Does the writer copy words from the title?		
50. Does the writer use foreign language vocabularies when in need?		
51. Does the writer use collocations?		
52. Does the writer make use of a varied word choice?		
53. Does the writer repeat some words?		
54. Is the choice of words in harmony with the prompt?		
55. Does the writer make use of figurative speech?		
56. Does the writer exploit simple and common words?		

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

57. Does the writer choose complex and rarely-used words?		
58. Does the writer use synonyms to avoid repetition?		
59. Does the writer use antonyms to avoid repetition?		
60. Is word length a matter of writer's concern?		
61. Does the writer use memorized sentences?		
62. Does the writer make use of adjectives and adverbs correctly?		
63. Is there the correct use of contractions?		
64. Does the writer make use of hypothetical structures correctly?		
65. Does the use of lengthy sentences far outweigh the short sentences?		
66. Is there a correct, consistent use of punctuation in the text?		
67. Does the writer use punctuation mechanically?		
68. Is the spelling of the words correct?		
69. Does the incorrect spelling cause misunderstanding?		
70. Does the absence of punctuation make the text difficult to understand?		
71. Has punctuation led to easy communication of ideas to the reader?		
72. Is capitalization practiced by the writer?		
73. Are proper nouns, if exist, capitalized?		
74. Is there right spacing between words?		
75. Is there right spacing between paragraphs?		
76. Do paragraphs break in the right place?		
77. Is the writer aware of the standard number of paragraphs in the essay?		
78. Does the writer make use of indentation to indicate beginning of successive paragraphs?		
79. Is the piece of writing legible?		
80. Does the writer practice neat handwriting?		

Appendix B
Final Design of the ADEC

Content	Yes	No
1. 1. Is there a thesis statement/topic sentence?		

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

2. Is the thesis statement/topic sentence related to the prompt?		
3. Is the thesis statement/topic sentence copied verbatim from the prompt?		
4. Does the writer rephrase the prompt?		
5. Does the writer use a part of the prompt to write about?		
6. Does the writer change the topic?		
7. Does the writer interpret the prompt correctly?		
8. Is there redundancy throughout the script?		
9. Is there a need for the rater to re-read the text to understand it?		
10. Does each paragraph contain a topic sentence?		
11. Do supporting sentences of each paragraph develop the main topic?		
12. Are the topic sentences of all paragraphs understandable?		
13. Does the writer well develop the topic in the body of the essay?		
14. Does the same logic of ideas govern the whole text?		
15. Does the writer stick to topic?		
16. Is the text informative enough to exhaust the topic?		
17. Does the writer give relevant information?		
18. Does the writer give off-topic information?		
Organization	Yes	No
19. Does the text have an introduction?		
20. Does the text have a conclusion?		
21. Are the paragraphs arranged in a logical order?		
22. Does the writer use cohesive devices to join the paragraphs?		
23. Does the writer use enumerators (e.g. First, Second, Next, Finally …) or signaling phrases (e.g. I'd now like to discuss the advantages...; my second argument against this statement is ...; finally I would like to) to show paragraphing?		
24. Are conjunctions used correctly?		
25. Does the writer make use of cohesive devices in a mechanical way (overuse without understanding)?		
26. Is it clear what referent words like "it", "that", and "they" refer to?		
27. Are articles used correctly?		
28. Does the writer correctly use possessive adjectives for the sake of cohesion?		
29. Are lexical sets (words that are used from a set of lexis e.g. car, engine, steering wheel, driver, exhaust, etc) used correctly?		
30. Does the writer use lengthy sentences to get the meaning across (Circumlocution)?		
31. Is the piece formal in style?		
Grammar	Yes	No
32. Does the writer use sentences according to right grammatical order i.e. SVO pattern?		
33. Does the writer use compounding with the use of coordinators (such		

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

as and, but, so, or, for, etc)?		
34. Does the writer use adverbial, adjective or gerund phrases (in an attempt for making complex sentences?		
35. Does the writer use embedded sentences (that clauses, relative clauses, etc)?		
36. Does the writer use correct tenses?		
37. Are modals used correctly?		
38. Is there subject-verb agreement?		
39. Does the writer use correct word forms (parts of speech)?		
40. Is the use of active/passive voice appropriate?		
41. Is the use of prepositions appropriate?		
42. Does the writer use quantifiers (such as many, much, few, little, very, etc) correctly?		
43. Does the writer make use of adjectives and adverbs correctly?		
44 Does the writer make use of hypothetical structures (if clauses, for example) correctly?		
45. Is there a repetitious use of a single grammatical pattern?		
46. Does the writer use memorized (clichés/set expressions/formulaic sentences) sentences?		
Vocabulary	Yes	No
47. Does the writer use the words that are taken from the topic repeatedly throughout the text?		
48. Does the writer use L1 and/or foreign vocabularies due to the lack of knowledge in L2?		
49. Is the use of collocations appropriate?		
50. Does the writer avoid repetition by a varied word choice?		
51. Is the choice of words in harmony with the topic (prompt)?		
52. Does the writer use figurative speech?		
53. Does the writer use idioms?		
54. Does the writer enhance the clarity of the text by using simple and common words?		
55. Does the writer use complex and rarely-used words?		
56. Does the writer use synonyms/antonyms to avoid repetition?		
Mechanics	Yes	No
57. Is there a correct, consistent use of punctuation in the text?		
58. Does the writer use punctuation mechanically?		
59. Is the spelling of the words correct?		
60. Does incorrect spelling cause misunderstanding?		
61. Does the absence of punctuation make the text difficult to understand?		
62. Has punctuation led to easy communication of ideas?		
63. Is capitalization practiced by the writer?		
64. Is there right spacing between words?		

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

65. Do paragraphs break in the right place (through indentation or spacing to indicate the beginning of successive paragraphs)?		
66. Is the writer aware of the standard number of paragraphs (usually five) in the essay?		
67. Is the piece of writing easy to read (legible)?		
68. Does the writer have neat handwriting?		

Appendix C
Frequencies of Categories and Subcategories of Writing Obtained through Think Aloud

Categories / **Subcategories**	**Frequency**
Content	549
Clarity of ideas	22
Rhetorical function	3
Task development	1
Relevance	34
Adherence to the main topic	10
Appropriacy of ideas	2
Necessity to reread the essay	27
Inadequate development of ideas	1
Well-stated ideas	16
Control ideas through examples	37
Control ideas through details	19
Incomplete/complete essay	15
Well-developed ideas	46
Adequate addressing of the topic	19
Comprehensibility	58
Fluency	8
Repetition of ideas	4
Adequate amount of information	24
Task response	40
Focus	8
Redundancy	8
Explicit thesis statement	36
Implied thesis statement	1
Copying materials	1

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Thesis statement in each paragraph	1
Well-developed paragraphs	6
A good response to the topic	5
Paraphrase and/or rephrase of the title	2
Supporting sentences	17
Unclear arguments	1
Interpretation on the part of the reader	1
Correct interpretation of the prompt/task/topic	1
Clarity of details	1
Communication of the message	5
Message diversion	1
Addressing part of the task	1
Flow	9
Off topic essay	1
Organization	747
The length of the essay	85
The length of the paragraph	10
Paragraphing	203
Introduction	152
Body	41
Conclusion	96
Referencing	7
Cohesion	4
Coherence	31
Style	37
Cohesive ties/Conjunctions/Enumerators	38
Blueprint	40
Transitions/Markers	6
Logical relations	8
Logical order of ideas	1
Grammar	462
Prepositions	24
SVO order	4
Articles	15
Word form	19

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Participles	8
Repetition of the same pattern	12
Verb forms/ tense	20
Conditionals	12
Pronouns	11
Subject-verb agreement	12
Passive/active voice	8
Simple sentences	4
Complex/compound sentences	10
Syntactic variety	9
Appropriacy	7
Accuracy	2
Syntactic complexity	3
Conjunctions	31
Adverbs/adjectives	32
The length of the sentences	30
Structural ambiguity	2
Incomplete sentences	2
Coordinators	4
Comparatives	2
Relative clauses/embedded clauses	4
Possessive adjectives	6
Modals	11
Memorized phrases/sentences	11
L1 in L2	3
Redundancy	1
Correct use of nouns	12
Pronoun agreement	7
Numbering	5
Prepositional phrases	1
Parallelism	7
Case	3
Noun phrases	2
Negation	1
Quantifiers	3
Comprehensibility of sentences	1
Vocabulary	461

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]

Appropriacy	139
Accuracy	5
Idioms	9
Wide/restricted range of vocabulary	79
L1 in L2	11
Collocations	25
Complex words	11
Lexical complexity distorts understanding	2
Vocabulary relevant to the topic	1
Figurative speech	18
Repetition	40
Lexical variety	3
Simple common words	3
Synonyms/Antonyms	4
Phrasal verbs	7
Words copied from the topic	1
Redundancy	18
Comprehensibility	5
Missing words	4
Mechanics	564
Handwriting	232
Spelling	72
Punctuation	287
Capitalization	72
Paragraph spacing/break	3
Neatness	25
Number of paragraphs	33
Spacing between words	3
Spacing between paragraphs	14
Pagination	14
Spacing between lines	5
L1 in L2	3
Spelling mistakes distorts understanding	1
Indentation	1

[Downloaded from c4i2016.khu.ac.ir on 2026-11-04]