Research Article doi: 10.12973/eu-jer.9.2.523

European Journal of Educational Research
Volume 9, Issue 2, 523 - 535.
ISSN: 2165-8714
http://www.eu-jer.com/

Summative Assessment, Test Scores and Text Quality: A Study of Cohesion as an Unspecified Descriptor in the Assessment Scale

Zulfiqar Ahmad*
University of Jeddah, SAUDI ARABIA

Received: December 29, 2019 ▪ Revised: February 4, 2020 ▪ Accepted: February 29, 2020

Abstract: Summative assessment of students' writing predicts not only the extent to which the course learning objectives have been achieved but also reveals the relevance of the assessment design with the construct of writing being assessed. Any dichotomy between the assessment criteria and the construct of writing or between the assessment criteria and test scoring procedures can produce unreliable and invalid interpretations of the students' writing proficiency. Assuming cohesion as a measure of writing quality, the present study chose samples of academic writing which did not specify cohesion as a descriptor in the assessment scale. A cohesion index was, therefore, developed to investigate how cohesive devices created texture in the sample texts and correlated with the test scores. Descriptive statistics and non-parametric correlation analysis revealed that cohesive devices were positively associated with appropriate use to create texture; however, they only weakly correlated with the test scores. The findings imply that there is the need for developing assessment criteria which consistently measures the text-forming resources to reliably ascertain the writing proficiency of the students. The study recommends a research initiative based on an analytical assessment criteria to ensure a more accurate analysis of the role of cohesion in text-formation and writing quality.

Keywords: *Academic writing, cohesion, English as a Foreign Language, language testing, summative assessment, texture, writing quality.*

To cite this article: Ahmad, Z. (2020). Summative assessment, test scores and text quality: A study of cohesion as an unspecified descriptor in the assessment scale. *European Journal of Educational Research*, *9*(2), 523-535.https://doi.org/10.12973/eu-jer.9.2.523

Introduction

"The teachers and their activities are the most crucial variables in the scheme of teaching and learning at all levels of the educational system" (Saha & Dworkin as cited in Ahmad, 2020 p.1). Writing teachers in the academia are mainly responsible for assessing the writing tasks (Crusan et al., 2016; Weigle, 2007), and may engage in developing assessment rubrics, test specifications, test items/tasks, and scoring for both formative and summative assessment. White (2009, p.6) argues that writing teachers, in most cases, conduct assessment oblivious of "the principles of sound assessment", and therefore, writing scores obtained from the summative assessment are likely to predict an invalid and unreliable report on students' writing proficiency, achievement of the course learning objectives, and evaluation of the writing course itself. Issues either in the assessment design such as the test specification, test content, scoring rubrics and criteria, and inconsistent scoring of the writing task by the teachers may be vulnerable to different interpretations of both the rubric descriptors and the test scores. One solution to this problem could be to study empirically the qualities of texts from the multiplicity of perspectives including the micro and the macro level text-forming features such as the use of cohesion or coherence in text creation. Following Weigle (2007) that single draft timed-essays reveal students' actual writing competence, it is viable to analyze a sample of single draft timed-essays produced in an examination setting and put to summative assessment to find out how the assessment rubrics, scoring criteria and teachers' scoring practices correlate to produce a valid and reliable interpretation of the students' writing competence.

Summative assessment of L2 writing

Summative assessment (SA) being the Assessment of Learning (AoL) aims to measure and quantify the learning achievement of students at a particular time (Stiggins, 2001). The measurement and quantification of the learning outcomes is based on preset criteria or standards which once adhered to produce numerical data in the form of test

* **Correspondence:**
Zulfiqar Ahmad, English Language Institute (ELI), University of Jeddah, Asfan 21589, Jeddah, Saudi Arabia. ✉ zulfiqar16c@hotmail.com

scores. The test scores become the source of making inferences about students' achievement. The interpretations of the test scores, if valid and reliable, unfold teaching effectiveness and learners' progression as student writers, guide course evaluation processes, meet discourse community expectations, and facilitate realization of the institutional and national objectives. Black and William (2018, p.563) mention two other uses of SA: instruments used in SA "should be so designed that they are supportive of learning" and that "teachers' should take responsibility for serving the summative purposes". The former refers to the use of test results for what scholars call "assessment as learning" (ibid) while the latter has implications for the training needs of the teachers in language assessment literacy (LAL).

Weigle (2007) proposes four essential components of an assessment procedure for any classroom based test: a) setting measureable objectives, b) deciding on how to assess objectives (formally and informally), c) setting tasks, and d) scoring. This implies that these benchmarks are valid, reliable and practical as well as the teachers are adequately trained to not only design these benchmarks but also administer and score in compliance with the standards. Summative assessment, from this perspective, is expected to generate "the metrics to know what's working and what's not" (States et al., 2018, p.1).

These metrics can be obtained by assessing writing either holistically or analytically (Hughes, 2002). In holistic measurement, the teacher assigns a single score to the writing product. The analytic marking, on the other hand, segregates writing into different measurable components and marks or letter grade are given for each descriptor such as the structure, content, rhetorical features, language use etc. Weigle (2007) finds out that though holistic measurement can be done quickly, it is less reliable than analytic scales as "scores on different aspects of writing can tell students where their respective strengths and weaknesses are" (p. 203). This entails that holistic marking is not programmed to give a systematic appraisal of the individual text-forming resources, cohesion or coherence for instance, which have been incorporated into the writing construct being assessed.

Issues with assessing writing

Benzehaf (2017, p.2) observes that "the need for increased use of test results to improve educational outcomes is urgent". However, this projected use of test scores for understanding the learning outcomes can be seriously impeded if the issues with assessing writing are not adequately addressed. Studies (Kalajahi & Abdullah, 2016; Sultana, 2019) report inconsistencies in teachers' assessment performance which may be either due to teachers' lack of assessment literacy to accurately adhere to the assessment rubrics or disparities between course objectives and assessment criteria.

The foremost challenge in assessing writing is to ensure that the assessment interventions are reliable and valid. There are imminent gaps between what the students are taught and what they actually learn, and therefore, "we need to develop processes of eliciting and interpreting evidence so that we can draw conclusions about what students have in fact learned" (Black & William, 2018, p.570). If the assessment interventions fail to test what the assessment intends to, there are issues with validity, and if the test scores fail to produce evidence that can be replicated, there are issues of reliability (States et al., 2018).

Allocating numerical value or letter grade to students' writing is a complicated task for the teachers. If an assessment task with a prescribed analytic scoring scale is graded holistically, both reliability of the test scores and validity of the assessment criteria will be affected. Consequently, accurate interpretation of the data obtained from the test scores cannot be made. Iliya (2014, p.115) argues that "the interpretation necessarily reduces the richness of the actual performance to a score, category or mark that represents it; thus a great deal of information is lost". It is, therefore, crucial that teachers must operationalize a writing construct for the test they propose to design, and base their expectations of the students' writing and the scoring rubrics on the construct. Similarly, they should design systematic scoring criteria which can be consistently practiced by the teachers.

Lynne (2004), on the other hand, is critical of judging a piece of writing from the narrower perspective of validity and reliability for he believes that these two concepts are at odds with the tenets of the social constructivists which define modern day writing theory. Hout (2002) has pointed to the different understanding of validity by the institution and education department. The former takes a traditional view of the concept that validity refers to the fact that a test measures what it is supposed to measure, while the latter interprets it differently even including the washback effects. Typical high-stake tests require an essay to be written within a strict time limit (30 minutes to an hour is typical) in response to a given prompt, and consequently, there could be several problems with the validity of such a task, both theoretical and practical.

Assessment of writing is vulnerable to distracted focus to the extent that the teachers target only the "easily quantifiable traits of essays such as error counts" (Weigle, 2007, p.198). Studies by Bouzidias as cited in Benzehaf (2017) and Lee (2010) corroborate this view, and reveal issues with the assessment descriptors which focused scoring of the micro level features of writing such as the spelling, punctuation, and verb forms. Scoring foci of a higher priority which can unveil students' discourse competence, for instance, organization of ideas, thesis statement etc., do not figure prominent in the scoring criteria. As a result, effective feedback on the quality of writing cannot be given.

Assessment benchmarks are by default prone to ambiguity and assessor subjectivity. For instance, IELTS Task 2 has four band descriptors with a hierarchy of grading scale for each descriptor to guide the rater (Ahmad, 2019). There is no provision for how to assess cohesion and on whose framework. If, for example, assessed on Halliday and Hasan's (1976) framework, the rater has to identify and account for 18 categories which in itself is quite a challenging task. Then there is the issue of genre specificity which has preference for certain type of linguistic entity for different text types. For instance, ellipsis and substitution do not feature prominently in academic writing (McCarthy, 1991). Hence, there is ample chance of rater bias to intervene with both the construct of writing and assessment benchmarks, and so can be true of other descriptors. Following process approach to writing, the contemporary assessment practices do not account for the cognitive processes involved in the production of a text, and rely only on the text as the final product (Breland et al., 1999). As a result, the components of the assessment criteria reveal an unreliable report on students' writing competence. Moreover, erroneous writing can mislead assessors as it can very likely distract them from the assessment benchmarks and focus solely on errors which are in most part mechanical and grammatical. Other features of text formation are likely to be overshadowed by an explicit focus on students' errors. Hence, only a partial evaluation of the students' actual discourse competence could be the result.

Assessment rubrics despite their limitations are, nevertheless, an integral part of the assessment process (Hamp-Lyons, 2003). They may also regulate students' test anxiety which according to Aydin (2019, p.21) "have significant influences on essential academic outcomes". Assessment of academic writing whether performance-based (PBA) such as done for the examination purposes or classroom-based (CBA) typically assigns a score "which is assumed to reflect the underlying construct or ability to be measured, relative to descriptors included in scoring rubrics" (Becker, 2011 p.113). Following contemporary perspectives on academic writing which situate it as premised on a configuration of social, cultural, cognitive, and linguistic variables (Hyland, 2006), it becomes crucial to revisit assessment practices in order to align them with empirically founded text forming resources. One such resource is cohesion which through its repertoire of lexical and grammatical ties establishes semantic relationships in and between clauses to create texture which Halliday and Hasan (1976) consider a non-structural resource of text formation.

Cohesion as a variable of text quality in EFL contexts

Writing aims at generating sentences that are "correct, complete and logical" (Solikhahas as cited in Demir & Erdogan, 2018, p.88). This traditional view of writing instruction in EFL/ESL contexts has focused on micro-level linguistic features (Lee, 1998), and consequently teachers correlate the presence of cohesive devices in a text with the writing quality (Wahby, 2014). According to Halliday and Hasan (1976, p.4) "the concept of a tie makes it possible to analyze a text in terms of its cohesive properties, and give a systematic account of its patterns of texture". Studies on cohesion devices as a variable of text quality have, however, yielded opposing results. For instance, researchers (Johnson, 1992; Toddet et al., 2007; Zhang, 2000) contend that writing quality is not impacted by cohesion. Witte and Faigley (1981) observe that it is the writer's invention skills and not the quantitative presence of cohesive links that account for the writing quality. Cooper (1986) studied 400 persuasive texts and found no correlation between cohesion and writing quality. Similarly, a study of 38 college essays by Jafarpur (1991) found no significant correlations between holistic scores and cohesion devices in terms of frequency and category.

On the other hand, studies (Chiang, 1999; Liu & Braine, 2005; Song & Xia, 2002) provide evidence of strong correlation between cohesion and writing quality. Cameron et al. (1995) report cohesion as being responsible for 15% of the significant differences in the quality of writing among children. In another study, Chiang's (1999, 2003) analysis of cohesion revealed that non-native speakers base their notion of quality of writing in EFL on the use of discourse features like cohesion and coherence. Guiju (2005) analyzed writing samples of 85 students to test the correlation between knowledge of cohesion and the quality of writing of college students. His results indicated that high score essays had effective use of cohesive devices as compared to the low grade which did not show statistically significant use of the cohesive devices. In a comprehensive cross-cultural study involving 898 academic scripts of 145 native-speaking (NS) American, and non-native-speaking (NNS) Japanese, Korean, Indonesian, and Arab students, Hinkel (2001) made comparisons in the use of cohesive devices. She found that the Arab students used more coordinators than the NS Americans. Rahman (2013) conducted a comparative study involving NS student writers and NNS Omani students. He found significant variations in the use of cohesive devices specifically from the measure of frequency, variety, and control by the two groups of writers. The EFL Omani student writers failed to use a range of cohesive devices and were restricted to the overuse of repetition and reference. The NS writers, on the other hand, had shown variety and control in the use of a range of cohesive devices which made their text read more fluid than their counterpart Omani students. Darweesh and Kadhimi (2016) investigated Iraqi students' use of conjunctive cohesion and found that the misuse far outnumbered the appropriate use which clearly indicated that the students were unable to create what Ting (2003) calls organic text connectivity. Another latest research by Al-Khatib (2017) reveals that students' writing show inappropriate use of cataphoric and anaphoric reference, ellipsis, substitution, and other grammatical cohesive ties. He observes that "the challenge that students face while writing is increased by the fact that the rhetorical conventions of the English texts such as the structure, organization and grammar differ from those in Arabic" (Al-Khatib, 2017, p.81).

Despite sufficient research initiatives in the domain of cohesion analysis in the Arab academic context, there is relative scarcity of research that investigates the relationship of assessment criteria, test scores, and text quality with cohesion to find out the extent to which a text reflects students' ability to use text-forming resources. The study primarily aimed at bridging this research gap to find out how cohesion as an unspecified descriptor in the assessment scale can be analyzed to ascertain its association with the text scores and text quality. The findings are expected to provide some useful insights to teachers, assessment experts and raters of academic writing in assessing an ignored component of text-formation in particular, and reviewing their pedagogic and assessment literacy practices in general for the benefit of the student writers and the academia.

Research questions

To find out the extent to which the student writers' use cohesion (an unspecified assessment descriptor) as a variable of text quality or text-forming resource, and how the presence of cohesive devices relates with test scores, the following research questions were generated:

1. How do students use cohesion as a text-forming resource?
2. What is the nature of relationship between test scores and text cohesion in regard to assessment criteria and text quality?

Method

The flowchart (FC-1) illustrates the analytical procedures adopted for this study.

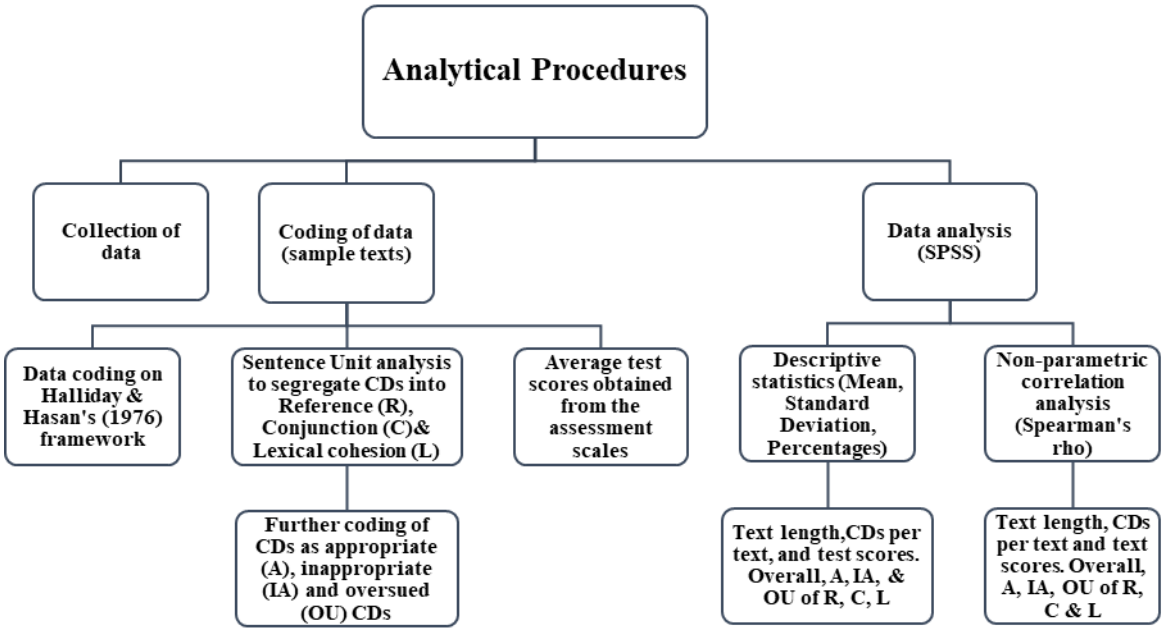

Figure 1. Research design

The sub-sections below detail the analytical procedures that were used for the analysis of the data:

Participants and sample characteristics

The study was conducted at Yanbu English Language Center (YELI), Yanbu Al Sinaiyah, Saudi Arabia. The participants (n = 30) were all male Saudi undergraduate students enrolled in different engineering specialism. Before this academic writing course, they had already completed the Foundation Year and Associate Degree English language courses. This two modules/semesters ENG 301 and ENG 302 academic writing course was delivered based on *Writing Academic English* (4th ed.) by Oshima and Hogue (2006). Chapter two "Unity and Coherence" (p. 18 -37) did not provide for cohesion as a distinct text feature. The subsections under "Coherence" included "repetition of key nouns", "key noun substitutes", "consistent pronouns", "transition signals", and "logical order" implying cohesion as 'unity' was built into coherence. The course assessed mode was summative which included In-class Assignments, a Mid and a Final Term examination resulting in a score which was then converted to a letter grade.

The data included writing samples (n = 30 of the argumentative essays) of the repeated subjects (the data was collected at the mid and final points of the semester) with different allocation for total marks for the Mid and the Final test. These texts had already been assessed by three raters on a three-point criteria(Appendix 1): *'structure and organization'* on a

scoring scale from 0.5 to 4, , *'grammar and mechanics'* on a scoring scale from 1 to 3, and *'content'* on a scoring scale from 1 to 8 respectively. However, the assessment rubrics did not provide for any explicit assessment of cohesion, and the marked scripts, therefore, did not have any scores allocated for cohesion analysis. The rubrics did not follow the prescribed content from Chapter 2 of the course book referred to earlier, and included only 'transitional signals' and that too in both *'text organization'* and '*content*' (Appendix 1). Even the allocation for the three-point assessment scale as shown in Tables T-1 and T-2 was different.

Table 1. Assessment scale for the Mid-term examination

Mid-term **Text organization**	**Grammar & mechanics**	**Content**	**Total**
5	3	12	20
25% of 20	15% of 20	60% of 20	

Table 2. Assessment scale for the Final-term examination

Final-term **Text organization**	**Grammar & mechanics**	**Content**	**Total**
4	3	8	15
26.6% of 15	20% of 15	53.33% of 15	

Therefore, in view of the uneven scales, the researcher decided to average the overall score from both the tests and then convert them to be out of ten to establish consistency. The primary reason for using this assessment scale was to investigate how students' use of cohesion devices and the test scores were related in terms of text quality. It was also anticipated that the findings would be useful to the research site as a course evaluation intervention. For this study, these sample texts were first typed in word document with all the errors whatsoever intact to maintain originality and transparency. The typed texts were then colour-coded to mark the use of referential, conjunctive, and lexical cohesion based on Halliday and Hasan's (1976) framework. Although there were other frameworks for cohesion analysis available such as that of Hoey (1991a), the revised version of Halliday and Matheissen (2004), and of Martin (2001), Halliday and Hasan's (1976) framework was chosen not only because it had been the most frequently adopted analytical taxonomy which provided a huge reference for collation or disagreement but also because the taxonomy most closely associated itself with common descriptions of cohesive elements used in the textbooks and classroom instructions such as the pronouns or reference, conjunctions, synonyms, collocations etc. This would help the stakeholders such as the teachers, course designers, test developers, and researchers to associate the results with what they actually practice.

Analysis of cohesion as a variable of text quality

Halliday and Hasan (1976) argue that cohesive devices (CD/s) appear to be critical in determining the clarity, appropriateness, and comprehensibility in writing. In other words, they play significant role in the creation of texture which is crucial to the existence of a piece of writing as a text. In order to measure the role of cohesion in creating texture in the sample texts, the researcher used measures of accuracy of cohesive ties which focused on ties that were complete (i.e., the referent was found within the text) or ambiguous (i.e., the referent must be inferred or was unclear (Cox et al., 1990; McCutchen & Perfetti, 1982). Following Halliday and Hasan (1976) and Tanskanen (2006, p. 84), Sentence Unit (SU) analysis was used for the purpose of developing an index of appropriate and inappropriate or ambiguous cohesive ties. To determine the types of cohesive relations present in students' texts, each SU within the texts was coded adapting Halliday and Hasan's (1976) coding scheme to determine instances of the following factors: (1) type of cohesive relations - reference (pronominal, demonstrative, comparative), conjunction (additive, adversative, causal, temporal), and lexical cohesion - reiteration (repetition, synonym/near synonym, superordinate, general word) and collocation; (2) number of ties per SU; (3) cohesive items within the text; and (4) the presupposed item. The coding was also extended to include the appropriate, inappropriate and overuse of cohesive devices. Substitution and Ellipsis were excluded from the analysis because of their low probability of use in academic texts (McCarthy, 1991). Table (T-3) shows the coding scheme for the present study:

Table 3. Coding scheme (adapted from Halliday and Hasan, 1976)

Types of cohesion		Coding			
		General	**Appropriate**	**Inappropriate**	**Overused**
1	Reference	R	AR	IAR	OUR
	Personal reference	R1	AR1	IAR1	OUR1
	Demonstrative reference	R2	AR2	IAR2	OUR2
	Comparative reference	R3	AR3	IAR3	OUR3
4	Conjunction	C	AC	IAC	OUC
	Additive	C1	AC1	IAC1	OUC1
	Adversative	C2	AC2	IAC2	OUC2
	Causal	C3	AC3	IAC3	OUC3
	Temporal	C4	AC4	IAC4	OUC4
5	Lexical cohesion	L	AL	IAL	OUL
	Repetition	L1	AL1	IAL1	OUL1
	Synonymy	L2	AL2	IAL2	OUL2
	Superordinates	L3	AL3	IAL3	OUL3
	General nouns/words	L4	AL4	IAL4	OUL4
	Collocation	L5	AL5	IAL5	OUL5

Next, the texts were examined for appropriate, inappropriate use and overuse of cohesive devices. Appropriate devices were identified as clearly establishing a cohesive relationship with the presupposed item to the extent that recovery of meaning was not challenging. Inappropriate items were identified as either ambiguous for which meaning was difficult to retrieve (Cox et al., 1991) or too distant to be retrieved easily or grammatically inaccurate to distort meaning relationship between the referring and the referent or existed only in the situation of composition or the writer's own private knowledge rather than being stated explicitly in the text. Following Gilquin et al. (2007, p. 322), the researcher operationalized the overuse of cohesive devices to be those instances of the more than three times repeated use of the same item for which an alternative linguistic item could be used. The role of cohesion in establishing text quality was assumed to be the presence of appropriately used cohesive devices in the creation of texture versus those devices which disrupted cohesion either through misuse or overuse.

Validity and reliability

"Reliability is the degree to which a test consistently measures whatever it measures" and "it is expressed numerically, usually as a coefficient; a high coefficient indicates high reliability" (Gay, 1997, p.145). In simple terms, research results have high reliability if they can be replicated in other contexts. As for as the present study is concerned, caution was taken to adhere to the established research procedures in terms of item construction, implementation, data collection, and analysis. But since the present study was conducted in a certain teaching context for collection of the writing samples, the results might not be as highly generalizable to other contexts as they would be to a similar Arab EFL context because of the social, cultural, and pedagogic factors that affect students' writing proficiency and performance. However, the researcher conducted a reliability test (Cronbach's Alpha) which is reported in the results section to ascertain consistency of the data being used for text analysis.

Following Best and Kahn (2003 p.297) that a typical valid research must provide validity evidence based on "three broad sources: content, relations to other variables, and construct", the researcher took care that this research study fulfills these conditions. The study was conducted in an English Language Institute which was accredited by Commission for English Language Program Accreditation (CEA) for its course designs. The student participants shared commonalities in terms of the social, cultural, and linguistic background, English language preparation, and learning objectives. The teacher participants (assessors) were all qualified and trained EFL teachers who had considerable experience of teaching in the Arab EFL settings.

The mainstay of validity is to justify the extent of data interpretation. First, the researcher operationalized the key concepts and constructs in regard to the participants and the data before finalizing the research design (Bachman and Palmer, 1996). For evidence of the test content, data from student writing which had been produced in an examination setting was chosen. Before the examination, the students had received formal instruction in writing argumentative essays. The data was carefully drawn following set criteria. Validity evidence in relation to other variables was based on what are referred to as predicative validity and concurrent validity. The data based on samples of students' writing was used to make predictions about how cohesion manifested itself in academic writing as well as in relationships with other variables and measures such as the test and cohesion scores. Validity evidence in regard to internal structure also known as construct validity is about the extent to which test item/s and test structure can be "accounted for by the explanatory constructs of a sound theory" (Best & Kahn, 2003 p.298). The construct of cohesion was modeled after the

SFL theory, and more specifically after Halliday and Hasan (1976) which is by far the most commonly framework used for analysis of cohesion.

For analysis of the data, descriptive statistics on SPSS was run to obtain sum, percentage, mean (M), and standard deviation (SD) scores for the referential, conjunctive and lexical cohesion. In addition, correlation analyses were conducted to find out significant associations between the variables of the corpus, cohesion categories, appropriate and inappropriate cohesion devices, overused cohesion devices, and the test scores. The results were then used to ascertain the role of cohesion in creating texture i.e. text quality.

Results

A reliability test was conducted on SPSS to measure the internal consistency of the data collected for the corpus of the main study. The three variables of the corpus: Words per Text (WPT), Sentence Units per Text (SUPT), and Cohesive Devices per Text (CDPT) were set to a five-point scale for the reliability analysis. The Cronbach's Alpha (α = .799) indicated that the data for the study was sufficiently reliable to be used for analysis (Sekaran, 2006, p. 311). Moreover, following Lincoln & Guba's (1985) suggestion of involving other researchers at a more general level to increase credibility in case a substitute for the inter-rater reliability is desired, the coding scheme for the textual analysis, which was adapted from Halliday and Hasan's (1976) coding for cohesion analysis, was verified by a colleague to check for consistency. The verified coding scheme was then applied to the data set for textual analysis of cohesion.

The researcher also decided to perform a data normality check before choosing the appropriate statistical tests for data analysis. The researcher used the Shapiro-Wilk test (p >.05) (Shapiro & Wilk, 1965; Hanusz & Tarasinska, 2015), which is recommended for a sample size of n > 3 and n < 2000, to get an estimate of the normal distribution of the variables i.e. WPT, SUPT, CDPT, and Test Scores (TS). The results (Table T-4) indicated that apart from SUPT with a skewness of*.472* (*SE =.427*) and a kurtosis of *-.137* (*SE =.833*), and CDPT with a skewness of*.236* (*SE =.427*) and a kurtosis of *-.914* (*SE =.833*) which were approximately normally distributed, the other variables had non-normal distribution. Therefore, following Dornyei (2007, p.227) that "if we have less precise, ordinal data, or categorical (i.e. nominal) data or if the data is not normally distributed, parametric tests are not appropriate", the researcher used the non-parametric test Spearman's rho (r_s) to find out any statistically significant correlations among the variables.

Table 4. Normality test results for the corpus

	Tests of normality				
	Skewness (SE)	**Kurtosis (SE)**	**Shapiro-Wilk**		
			Statistics	**df**	**Sig.**
WPT	.221 (.427)	-1.231 (.833)	.929	30	.045
SUPT	. 472(.427)	-.137 (.833)	.957	30	.256
CDPT	.243(.427)	-.873 (.833)	.961	30	.324
TS	-.148(.427)	-1.381 (.833)	.889	30	.004

SPSS was run to obtain descriptive statistics and non-parametric correlation (Spearman's rho) results for the sample texts. Descriptive statistics reported in Table (T-5) show that the sample texts (n=30) comprised of 11436 words with 1924 CDs which constituted 16.82% of the overall word length per text. 91.89% (n = 1768) of the CDs were appropriately used; however, 20.53% (n = 395) of the overall CDs were overused. 50.15% of the cohesion was formed of the Lexical items followed by 37% of Referential and 12.83% of the conjunctive devices. The test scores for the sample texts had *M = 7.342; SD =.972*.

Table 5. Descriptive Statistics for the corpus

	N	**Sum**	**Mean**	**Std. Deviation**
WPT	30	11436	381.2	84.076
SUPT	30	628	20.93	3.999
CDPT	30	1924	64.13	16.848
Reference	30	712	23.73	10.13
Conjunction	30	247	8.23	3.928
Lex Cohesion	30	965	32.17	10.048
ACDs	30	1768	58.93	15.472
IACDs	30	126	4.2	3.881
OUCDs	30	395	13.17	10.336
TS	30		7.3427	0.97253
Valid N (listwise)	30			

Spearman's rho (Table T-6) revealed statistically strong positive correlation between WPT and CDPT, $r_s = .759$, $p < .01$; WPT and R, $r_s = .705$, $p > .01$; and WPT and ACD, $r_s = .759$, $p > .01$;CDPT and R, $r_s = .800$, $p > .01$; CDPT and L, $r_s = .823$, $p > .01$; and CDPT and ACD, $r_s = .977$, $p > .01$. The correlation results were, however, statistically moderately significant between WPT and L, $r_s = .623$, $p > .01$ and CDPT and OUCD, $r_s = .592$, $p > .01$. The statistical associations between WPT and OUCD, $r_s = .407$, $p > .05$; CDPT and TS, $r_s = .384$, $p > .05$; TS and ACD, $r_s = .402$, $p > .05$; and TS and R, $r_s = .410$, $p > .05$ were found to be weak but positive.

Table 6. Correlation analysis of the corpus (N=30)

	WPT	CDPT	TS	R	L	ACD	OUCD
WPT	1.000	.759**	.224	.705**	.623**	.759**	.407*
		.000	.234	.000	.000	.000	·026
CDPT	.759**	1.000	.384*	.800**	.823**	.977**	.592**
	.000		.036	.000	.000	.000	.001
TS	.224	.384*	1.000	.410*	.333	.402*	.124
	.234	.036		.024	.072	.027	.513
R	.705**	.800**	.410*	1.000	.441*	.742**	.445*
	.000	.000	.024		.015	.000	.014
L	.623**	.823**	.333	.441*	1.000	..863**	.582**
	.000	.000	.072	·015		.000	.001
ACD	.759**	.977**	.402*	.742**	..863**	1.000	.652**
	.000	.000	.027	.000	.000		.000
OUCD	.407*	.592**	124	.445*	582**	.652**	1.000
	·026	.001	.513	.014	.001	.000	

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

The results unfolded that the number of words per texts (WPT) was not very consistent across the collected data and there was considerable variation among the number of words used by individual student writers in their essays. Similarly, there was considerable degree of variance in the CD use in these essays. However, there was strong probability of a linear increase in the number of CDs with an increase in the text length. Statistically strong positive association between the text length and appropriate CD use indicated that the students were able to create texture through the use of coreferential elements in the text which was unaffected by either the inappropriate or the overuse. The absence of positive relationship between the text length and the test scores showed that there were factors other than these which accounted for students' grades. Test scores proved to be a positive but weak determinant of the presence of cohesive element in the texts.

The examples from the sample texts cited below substantiate the use of cohesion as being appropriate, inappropriate and overused:

i. *by culture we can make the difference between nations. in my country Saudi Arabia keeping <u>the</u> culture is very important becuse some of the culture is from our relgios Islam.*

ii. *Because <u>the more</u> tests they perform <u>the more</u> they will get tired of it in the middle of the semester and <u>the more</u> it'll affect their grades.*

iii. *People with money know the importance of giving some of the responsibilities to others so "<u>they"</u> can help.*

iv. *Saudi people is known by eating rice by hands but now day <u>they</u> start to eat by using spon and forke and also it changes the way <u>they</u> speak to people. when <u>they</u> are traviling a lot <u>they</u> will start to hear a lot of people talking and <u>they</u> may take their accent or the way <u>they</u> speak.*

v. *Parents promise their kids if their behaiv good <u>they</u> will get the games <u>they</u> want or if <u>they</u> do good in the school <u>they</u> will get the games <u>they</u> want or the game device <u>they</u> want.*

In example (i) the student writer uses three different CDs to create semantic relationships in the text through coreferentiality. There is part-whole relationship between *nations - Saudi Arabia - the culture - religion Islam*; *the* specifies *culture* which is a lexical repetition of the same item in the previous sentence; and the causal conjunctive *because* provides a rationale for the claim in the independent clause. Example (ii) is a typical instance of how comparative reference is used to create cohesion in text. Though grammatically inappropriate, "*it*" ties with the dependent clause "... *the more tests they perform*". Moreover, the pattern of collocation use evident in *test-semester-grades* enhances lexical cohesion. The use of *"they"* in (iii) is ambiguous since the pronoun can be understood to refer to both *"people"* and *"others"*. This feature is typical of the impact of Arab culture which requires the readers to resolve ambiguity (Mohammad & Omer, 2000). Suffice it to say that this type of use is unlike the native English use where the text itself helps in the retrieval of the antecedent, and therefore, needs some sort of contextual intervention to get back

to the presupposed item. The pronominal *"they"* has been excessively used in two examples (iv & v) while referring back to *"people"* and *"parents".* This may be because the students are unable to conceive of an alternative syntactic structure where they may avoid such an overuse. Secondly, they appear context bound to use "*they*" repeatedly as they use repetition of lexical items which is a highly prominent aspect of these students' rhetorical strategy.

Discussion

The data analysis results reveal that the text length does not correlate with the test scores. This finding contradicts most studies on the relationship between writing quality and the text length (e.g. Chodorow & Burnstein, 2004; de Haan & van Esch, 2008). The test scores are only weakly correlated with cohesive devices per text. More specifically, there is significant association of the test scores with the overall appropriate CDs. Consistent with most other research findings (e.g. Llach & Catalan, 2007), referential cohesion is found statistically significant in those texts which are significant for their effect on test scores. Despite being the most frequently used cohesive element in the sample texts, lexical devices do not show significant association with the test scores. Similarly, conjunctions stated in the assessment scale as "transition signals" fail to establish positive association with the test scores. Both the lexical and the conjunctive devices which are an integral cohesive component of academic writing point to probable flaws in the assessment criteria. It may be that they have not been stated explicitly in the assessment scale as measurement descriptors as is the case with this study, especially lexical cohesion. As such, the meaning making potential of a text through the lexical and conjunctive devices cannot be reliably predicted to account for students' writing competence.

These results, moreover, do not indicate any significant association of the test scores with inappropriate, and overused CDs - a finding supported by Mohamed (2016). Keeping the results of the appropriate use and their significant relationship with the test scores, it can be argued that the students used cohesive devices appropriately which successfully aided in the creation of texture in their writing.

The analysis also revealed some visible gaps in the pedagogic, curricular, and assessment system being practiced at the research site. For instance, despite the exclusion of explicit provision for cohesion in the instructional and the assessment design, the students were able to successfully employ cohesive devices to create texture in their writing. Nevertheless, correlation between writing and cohesion has been a fluid topic unfolding results which are both supporting and contradictory. A number of variables such as the research context, the student writers' language proficiency profile, the pedagogic preferences, the assessment rubrics, the raters etc. determine the outcomes of results which often lead to significant variations in the conclusions drawn for the topic.

However, the assessment criteria ignore measurement of cohesion as a text-forming resource. Ahmad (2019, p.22) argues that cohesion being a crucial text-forming resource must feature explicitly both in the course design and assessment criteria because "it is part of writing. So, if in general, the elements that are specific to aspects of writing are not being taken care of, writing as an academic and language skill is also not being taken care of".

This seems to be a major limitation of the assessment design because a text is expected to be analyzed on it text-forming properties such as is done in the case of IELTS and TOEFL examinations. Both these exams provide for the measurement of cohesion and coherence assuming students' ability to create texts through the use of semantic associations which give the text its unity cannot be decoded otherwise. The three-point assessment scale used by the teacher assessors does not fully account for the appraisals of the textual resources. Even the measurement of transition signals is inconsistent since the teachers were expected to grade them from two assessment descriptors - organization and content - which might have produced an unreliable score. In addition, following the course book referred to in the "Method" section of this paper, transitions which are conjunctives and statistically insignificant in the present study only give a partial account of the use of cohesion. Other cohesive devices which were part of the course design such as the pronouns and the lexical devices have not been assessed. This is likely to render the scores invalid and unreliable because the learning outcomes as revealed in the test results may challenge the course learning objectives.

Limitations, Implications and Recommendations

This analysis of associations and comparisons of cohesion and test scores with writing quality has, however, limited generalizability for both the research context and beyond. First, the small sample size (n=30) collected from one research site may not produce results which can be generalized to other academic contexts. The researcher used the test scores which had been awarded by the raters at the research site. The assessment scale did not provide for any explicit provision for the assessment of cohesion, and therefore, distinct scores for cohesion could not be obtained. A scale to assess the 18 categories of cohesion (Halliday & Hasan, 1976) individually, a viable focus for a new research study, would give a more accurate measure of relationship between cohesion and writing quality. The focus of the present study on cohesion only may not fully reveal students' use of textual resources in the creation of texture. A study which investigates two other variables of texture i.e. the intra-sentence structure and the macro structure of discourse (Halliday & Hasan, 1976) is likely to produce a more comprehensive analysis of the text-forming features. Although the study included repeated subjects from the same course, it cannot predict the extent to which the student writers progressed in their writing proficiency, especially the use of cohesion, from the beginning of the course to the Mid-term

exam, and from the Mid-term to the Final exam. Therefore, a study which can compare students' performance between the different stages of a course can produce results which can help better interpret not only students' learning but also the quality of teaching, course design, and assessment practices.

Conclusion

One of the most important aims of academic writing pedagogy is to help student writers acquire discourse competence through awareness raising and practical tasks in the dynamics of text-forming resources. Such an approach to writing instruction will enable students produce texts appropriate to their respective discourse community. This cannot be achieved unless assessment of academic writing is explicitly aligned with the features that make up for genre-specific text. Cohesion not only ensures textual unity through its inherent properties of co-classification and co-referentiality but also supplements coherence which is also central to the existence of a piece of writing as a text. Importantly though, cohesion also operates at the intra-sentential structure especially in the Theme-Rheme, and marks register choices which in turn configure to create the macro-structure of the genre. Hence, absence of cohesion as an assessment descriptor in the assessment rubrics is likely to render the assessment of academic writing as invalid and unreliable.

References

Ahmad, Z. (2020). Peer observation as a professional development intervention in EFL pedagogy: A case of a reading lesson on developing the top-down processing skills of the preparatory year students. *International Linguistics Research*, *3*(1), 1-15. https://doi.org/10.30560/ilr.v3n1p1

Ahmad, Z. (2019). Teacher beliefs about students' use of cohesion in writing: What does the textual evidence reveal? *Journal of Language and Education, 5*(4), 9-24.https://doi.org/10.17323/jle.2019.9708

Ahmad, Z. (2019). Analyzing argumentative essay as an academic genre on assessment frameworks of IELTS and TOEFL. In S. Hidri (Ed.), *English language teaching research in the Middle East and North Africa: Multiple perspectives* (pp. 279-299). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-98533-6

Al-Khatib, H. (2017). The five tier model for teaching English academic writing in EFL contexts. *Arab World English Journal, 8*(2), 74-86*.* https://doi.org/10.24093/awej/vol8no2.5

Aydin, U. (2019). Test anxiety: Gender differences in elementary school students. *European Journal of Educational Research*, *8*(1), 21-30. https://doi.org/10.12973/eu-jer.8.1.21

Bachman, L. F., &Palmer, A.S. (1996).*Language testing in practice*. Oxford University Press.

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

Benzehaf, B. (2017). Exploring teachers' assessment practices and skills. *International Journal of Assessment Tools in Education, 4*(1), 1-18*.* https://doi.org/10.21449/ijate.254581

Best, J. W., & Kahn, J. V. (2003).*Research in education* (9th ed.). Allyn and Bacon.

Black, P., & Wiliam, D. (2018).Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551-575. https://doi.org/10.1080/0969594X.2018.1441807

Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999).*The performance assessment study in writing: Analysis of the SAT II: Writing subject test* (College Board Report No. 99-4). College Entrance Examination Board.

Cameron, C. A., Lee, K., Webster, S., Munro, K., Hunt, A. K., & Linton, M. J. (1995). Text cohesion in children's narrative writing. *Applied Psycholinguistics*, *16*, 257-269.

Chiang, S. Y. (1999). Assessing grammatical and textual features in L2 writing samples: The case of French as a foreign language. *The Modern Language Journal*, *83*(2), 219-232.

Chiang, S. Y. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, *31*(4), 471-484.

Chodorow, M., & Burnstein, J. (2004).*Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (ETS Research Report No.RR-04-04). ETS.

Cox, B. E., Shanahan, T., & Sulzby, E. (1990). Good and poor elementary readers' use of cohesion in writing. *Reading Research Quarterly*, *25*, 47-65.

Cox, B. E., Shanahan, T., & Tinzmann, M. B. (1991).Children's knowledge of organization, cohesion, and voice in written exposition. *Research in the teaching of English, 25*(2), 179-218.

Crusan, D., Plakans, L., & Gebril, A. (2016).Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, *28*, 43-56. https://doi.org/10.1016/j.asw.2016.03.001

Darweesh, A. D., & Kadhimi, S. A. H. (2016). Iraqi EFL learners' problems in using conjunctions as cohesive devices. *Journal of Education and Practice*, *7*(11), 169-180.

deHaan, P., & van Esch, K. (2008). Measuring and assessing the development of foreign language writing competence. *Porta Linguiarium, 9,* 7-21.

Demir, S., & Erdogan, A. (2018). The role of teaching grammar in first language education. *European Journal of Educational Research*, *7*(1), 87-101. https://doi.org/10.12973/eu-jer.7.1.87

Dornyei, Z. (2007). *Research methods in applied linguistics.* Oxford University Press.

Gay, L.R. (2005). *Educational research* (5th ed.). National Book Foundation.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, *6,* 319-335.

Guiju, Z. (2005). The cohesive knowledge and English writing quality of college student. *CELEA Journal*, *28*(3), 24-30.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* Longman.

Halliday, M. A. K., & Mathiessen, C. (2004). *An introduction to functional grammar*. Hodder Arnold.

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 University Press.

Hanusz, Z., & Tarasinska, J. (2015). Normalization of the Kolmogorov–Smirnov and Shapiro–Wilk tests of normality. *Biometrical Letters, 52*(2), 85-93. https://doi.org/10.1515/bile-2015-0008

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning, 12*(2), 111-132.

Hoey, M. (1991a). *Patterns of lexis in text.* Oxford University Press.

Hughes, A. (2002). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Huot, B. (2002). *(Re) articulating writing assessment for teaching and learning*. Logan.

Hyland, K. (2006). *English for academic purposes: an advanced resource book*. Routledge.

Iliya, A. (2014).Formative and summative assessment in educational enterprise, *Journal of Education and Practice*, *5*(20), 111-117.

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System*, *19*(4), 459-465.

Johnson, P. (1992). Cohesion and coherence in compositions in Malay and English. *RELC Journal*, *23*(2), 1-17.

Kalajahi, S. A. R., & Abdullah, A.N. (2016). Assessing assessment literacy and practices among lecturers, *Pedagogy*, *124*(4), 232–248.

Lee, I. (1998). Writing in the Hong Kong secondary classroom: Teachers' beliefs and practice. *Hong Journal of Applied Linguistics*, *3*, 61-76.

Lee, I. (2010). Writing teacher education and teacher learning: Testimonies of four EFL teachers. *Journal of Second Language Writing*, *19*(3), 143–157.

Lincoln, Y. S., & Gaba, E. G. (1985). *Naturalistic inquiry*. Sage Publications.

Liu, M., & Braine G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System, 33,* 623-636.

Llach, M. P. A., & Catalan, R. M. J. (2007). Lexical reiteration in EFL young learners' essays: Does it relate to the type of instruction? *International Journal of English Studies, 7*(2), 85-103. https://doi.org/10.6018/ijes.7.2.49001

Lynne, P. (2004). *Coming to terms: A theory of writing assessment*. Utah State University Press.

Martin, J. R. (2001). Cohesion and texture. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 35-53). Blackwell.

McCarthy, M. (1991). *Discourse Analysis for Language Teachers.* Cambridge University Press.

McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *Text-Interdisciplinary Journal for the Study of Discourse, 2*(1–3), 113–140.

Mohamed, A. H., & Omer, M. R. (2000). Texture and culture: Cohesion as a marker in rhetorical organization in Arabic and English narrative texts. *RELC Journal*, *31*(2), 45-75.

Mohamed, N. (2016). Use of conjunctions in argumentative essay by ESL undergraduates. *e-Academia Journal UiTMT*, *5*(1), 1-13.

Oshima, A., & Hogue, A. (2006).*Writing academic English* (4th ed.). Longman.

Rahman, Z. A. A. (2013). The use of cohesive devices writing by Omani student-teachers. *SAGE Open*, *3*, 1-10.

Sekaran, U. (2006). *Research methods for business: A skill building approach*. John Wiley & Sons.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

Song, M., & Xia, W. (2002). Combination of textual cohesive ties and textual teaching for the teaching of English writing: a statistical analysis of the good and poor compositions written by non-English major freshmen. *Foreign Language World, 6,* 40-44.

States, J., Detrich, R., & Keyworth, R. (2018). *Overview of summative assessment*. The Wing Institute.

Stiggins, R. (2001). *Student involved classroom assessment*. Merrill Publishing.

Sultana, N. (2019). Language assessment literacy: An uncharted area for the English language teachers in Bangladesh, *Language Testing in Asia*, *9*(1), 2-14. https://doi.org/10.1186/s40468-019-0077-8

Tanskanen, S. K. (2006). *Collaborating towards coherence*. John Benjamins. https://doi.org/10.1075/pbns.146

Ting, F. (2003). *An Investigation of cohesive errors in the writing of PRC tertiary EFL students* (Unpublished master's thesis). National University of Singapore.

Todd, W., Khongput, R., & Darasawang, P. (2007).Coherence, cohesion and comments on students' academic essays. *Assessing Writing*, *12*(1), 10-25. https://doi.org/10.1016/j.asw.2007.02.002

Wahby, M. (2014). The effect of implementing cohesive ties by Saudi prep-year pre intermediate students on their written texts. *European Scientific Journal*, *10*(4), 220-232.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), 194–209. https://doi.org/10.1016/j.jslw.2007.07.004

White, E. (2009). Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. *OnCUE Journal*, *3*(1), 3–25.

Witte, S., & Faigley, L. (1981). Coherence, cohesion, and writing ability. *College Composition and Communication, 32,* 189-204.

Zhang, M. (2000). Cohesive features in exploratory writing of undergraduates in two Chinese universities. *RELC Journal, 31*, 61–93.

Appendix

Assessment rubrics for the original assessment

Structure & Organization			
3	**2**	**1**	**0.5**
The introduction contains a clearly stated thesis statement; the body fully and clearly explains the thesis statement; the conclusion effectively ends with author's final thought; effective and varied transitions are used throughout the essay.	The introduction contains Thesis statement; the body explains the thesis statement; the conclusion presents the last step or another logical ending; transitions are used throughout the essay	The introduction contains thesis statement but it may be unclear, imprecise, or undeveloped; the body explains only some of the sub ideas mentioned in the thesis statement; the conclusion does not present the last step or any other logical reflection on the process; more or better transitions are needed throughout the essay.	The introduction lacks Thesis statement; the body does not address the points mentioned in thesis statement; the conclusion is missing or repetitive; the writing lacks transitions.
Contents			
7-8	**5-6**	**3-4**	**1-2**
The overall thesis is clear; he arguments are presented in the effective, precise and clear way in which they are mentioned in the thesis statement; transitional words and phrases that show forceful ideas are used effectively; word choice is consistently precise; there is sense of completeness.	The overall thesis is generally clear; most of the arguments are presented and explained in order of their sequence; transitional words and phrases are used; most word choices are precise; some details are missing.	The overall arguments are unclear; ideas may be presented in a sketchy way; the writing lacks effective words and phrases to describe the arguments; most word choices are imprecise, redundant, or confusing; many details are either confusing or missing; transitional words and phrases are sparingly used.	No Process is apparent in the writing; the writing does not address the intended process; word choices are imprecise, redundant, or confusing; writing is sketchy
Grammar & Mechanics			
4	**3**	**2**	**1**
There are few or no errors in mechanics, usage, grammar, or spelling.	There are some errors in mechanics, usage, grammar, or spelling	Errors in mechanics, usage, grammar, or spelling interfere with the audience's understanding of the process	Serious and numerous errors in mechanics, usage, grammar, or spelling block the audience's understanding of the process.