International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:9, No:1, 2015

Development of a Rating Scale for Elementary EFL Writing

Mohammed S. Assiri

***Abstract*—**In EFL programs, rating scales used in writing assessment are often constructed by intuition. Intuition-based scales tend to provide inaccurate and divisive ratings of learners' writing performance. Hence, following an empirical approach, this study attempted to develop a rating scale for elementary-level writing at an EFL program in Saudi Arabia. Towards this goal, 98 students' essays were scored and then coded using comprehensive taxonomy of writing constructs and their measures. An automatic linear modeling was run to find out which measures would best predict essay scores. A nonparametric ANOVA, the Kruskal-Wallis test, was then used to determine which measures could best differentiate among scoring levels. Findings indicated that there were certain measures that could serve as either good predictors of essay scores or differentiators among scoring levels, or both. The main conclusion was that a rating scale can be empirically developed using predictive and discriminative statistical tests.

***Keywords*—**Analytic scoring, rating scales, writing assessment, writing performance.

I. Introduction

SINCE the inception of writing assessment as a practice in the late 1970s, language testers have looked for ways to design suitable and consistent measures. In direct assessment of writing, students are asked to write a brief essay. Each essay is then scored holistically or analytically by means of a set of criteria. Such criteria comprise what is referred to as a rating scale. More specifically, a rating scale can be defined as an assessment tool that incorporates skills or constructs in graded levels (or bands) alongside descriptions of mastery requirements for each given level.

Although rating scales are widely used in writing assessment, they may have inherent problems that threaten the assessment validity and reliability. First, rating scales are usually developed with the view that they exactly reproduce writing abilities [1]. Thus, a tendency among raters is to oversimplify writing constructs by designing intuition-based rating scales [2]. Such rating scales are usually constructed by means of personal judgments, which are often based on existing rating scales or teaching syllabi. The formulation of rating criteria in these scales is largely influenced by persons' theories and experiences of how writing should be assessed [3]. Therefore, considerably diverse opinions are often imposed on the development of a rating scale [4]. Nevertheless, the end result of these efforts is a rating scale whose use is likely to furnish scores that may not adequately represent writing abilities [5].

In EFL programs in Saudi Arabia, rating scales are not often used; and, when used, they tend to be intuitively constructed. In the English Language Center (henceforth, ELC), where the researcher works as an EFL instructor, most of the writing teachers score their students' essays, including those intended for assessment purposes, by determining the extent to which a student has met the requirements of the essay directions or instructions. A few teachers however use simple, intuition-based rating scales. Nevertheless, there is a concern among these teachers that they do not follow similar measures in assessing their students across their groups of students. Writing in this context is by far the most contentious skill that often diverts from other skills in terms of assessment and scores. Presumably, if an empirically-developed or data-based rating scale is used, it would serve as a yardstick on the basis of which students' writing abilities are assessed across various groups in a fair and consistent manner.

II. Literature Review

A number of theories or models have been proposed in order to capture the distinct aspects of L2 writing. The most influential model has been that of [6] who proposed a framework for a general theory of writing based on an extensive review of literature. This framework has seven major components: syntactic structures, semantic functions, cohesion devices, coherence structures, lexical relations, stylistic and register dimensions, and non-linguistic knowledge. Syntax and semantics play their roles at the local or sentential level of a text in that both result in the production of a meaningful sequence of forms. Cohesion links sentences together and coherence endorses the unity of the text as a whole, both in a manner that conveys the author's intent. The interpersonal style of communication between the author and the reader is shaped by posture and stance. And, lexicon assumes a median position since the use of lexis feeds into, and thus endorses the functions of all other components.

On the basis of their model, [6] developed a taxonomy of writing skills and contexts, which can be summarized in Fig. 1. Reference [6] taxonomy is clearly quite inclusive of important aspects of L2 writing. Although [6] model cannot be directly applied in the area of writing assessment [7], it has informed decisions taken in the development of rating scales. This has paved the way for the construction of empirically-developed rating scales that draw on the taxonomy of the model.

Mohammed Assiri is with the Institute of Public Administration, Riyadh 11141, Saudi Arabia (96611-474-5661; fax: 96611-476-5620; e-mail: assirim@ipa.edu.sa).

1) Academic writing settings: topics, texts, tasks …
2) Writer background: intents, attributes, attitudes …
3) Linguistic knowledge: code, morphology, vocabulary, syntax, typological differences …
4) Discourse knowledge: inter- and intra-sentential relations, informational structuring, semantic relations, genre structures and constraints, organizing schemes, inferences …
5) Sociolinguistic knowledge: functional uses, Gricean maxims, situation and register parameters …
6) Audience: considerations pertinent to audience
7) World knowledge: non-linguistic knowledge resources
8) Writing skills and strategies: planning, elaborating, revising …

Fig. 1 Reference [6] taxonomy of writing skills and contexts

In speaking assessment, the most obvious merit of empirically-developed rating scales is that the observable aspects of the speaker behavior and task characteristics are brought together [8]. One empirical approach to developing a rating scale was proposed by [9] who requested from experienced raters to classify writing scripts into different levels. The researchers could then identify salient aspects that distinguish among the various levels. A set of yes/no questions was formulated to assign scores to writing scripts at a given level.

Another empirical approach was followed in [10] research on Cambridge ESOL writing examinations at various levels of proficiency. The researchers used previous ratings of writing scripts to sort them into different performance levels. Then, they analyzed the writing scripts at each level to find out what aspects could set the performance levels apart. Such aspects were then re-examined to decide which ones were most suitable for a rating scale.

Scaling descriptors represent the third choice for designing a rating scale. Reference [11] used scaling descriptors in his work on the common European framework. He drew descriptors from a sample of thirty rating scales. These descriptors were then categorized into several tasks that are communicatively tailored. In the second phase, a group of teachers were provided with the descriptors and requested to classify them into categories. They were also asked to indicate the extent to which they thought each descriptor was useful and applicable to their own assessment. The teachers were then requested to rank the descriptors within each category into three levels of competency. The descriptors with consistent rankings were used to construct questionnaires that shared the same anchor items. In the third phase, the teachers rated a sample of their students' essays using the questionnaires. Multi-faceted Rasch was then run to use fit descriptors in the design of a rating scale with cut-off points determined by estimates of difficulty and grouping. Nevertheless, reference [12] critiqued this method, suggesting that it does not stem from well-grounded assumptions of language proficiency. Therefore, each of the approaches to the development of empirically-based rating scales has its merits and weaknesses, which highlights the need for more sound approaches that lay aside hunches.

The design of a rating scale is a systematic process. Reference [3] provides a step-by-step framework for the development of a rating scale. First, the developer decides as to what type of rating scale is needed: analytic, holistic, primary- or multiple-trait. An analytic scale deals with writing as separate skills, whereas a holistic scale reflects a whole, integrated assessment of a written product. A primary-trait scale focuses on one skill of writing, whereas a multiple-trait scale measures writing performance in the form of a number of skills. Second, the developer identifies who will use the results furnished by the rating scale. In this regard, a rating scale can be one of three categories: first, a user-oriented scale which informs the user about the test-takers' abilities; second, an assessor-oriented scale which aids decisions about a test-taker for academic or professional purposes; and third, a constructor-oriented scale which enables test makers to include tasks that match learners' abilities [13]. Language testers typically use assessor-oriented scales [2]. Third, the developer determines what skills and sub-skills of writing the rating scale should include. Fourth, the developer specifies the number of bands (levels) on the rating scale as well as the number of descriptors, and how they should look like. The number of bands should allow the rater to make fine and fair distinctions among writers. This has proven to be the case with scales that have bands ranging from five to nine [14]. Fifth, the developer decides as to how scores are assigned. Sixth, the developer chooses a style for reporting scores depending on the purpose of the writing test and the interests of the score users. Reference [7] adds a seventh step according to which the developer should decide about how the rating scale can be validated.

Reference [7] also describes validity facets of a rating scale on the basis of [15] model of language ability and use. First, the rating scale should possess adequate representation of the underlying constructs of writing. Second, it should have a reasonable discriminatory power to distinguish adjacent levels of writing performance. Third, it should give similar results every time it is used to rate the same writing products (or reliability). Fourth, it should exhibit aspects of writing typical of L2 texts that are understandable by its readers. Fifth, writers ought to be provided with full and accurate feedback on their performances. Sixth, the scores furnished to stakeholders should fulfill their demands. Seventh and last, the development of the rating scale and its use should not be in conflict with practicality.

The method of assigning scores on the rating scale can be one of four. First, the rater subtracts points of a total score for each deficiency. Second, the rater selects the point, value, or a label along a continuum, which best matches the writers' ability. Third, the rater selects the value along a vertical scale, which best matches the writer's ability as determined by a detailed description next to the value. As for phrasing the descriptors, the developer of a rating scale may use abstract forms including the use of qualifiers or quantifiers (e.g., a lot, a little, some … etc.). Another option is to use concrete terms to denote the main features of the written task the existence of which can be signaled with 'yes' or 'no'. Otherwise, the rater can employ more objective forms by using quantifiable features (e.g., the number of error-free clauses). Noticeably, using concrete and objective descriptors, in particular,

precludes chances of subjective or biased rating.

A number of researchers have criticized the use of rating scales on several grounds. Rating scales are often developed in the absence of considerations of the writing task or writer characteristics [16]. Band levels are usually distinguished by relative adjectives or adverbs [17] that do not offer sophisticated and defined differences between levels. Scaling descriptors are often formulated in a style that causes raters to perform obscurely and less objectively [18]. Besides, certain elements of descriptors may not necessarily co-exist at a given band level [16]. Research exploring rater's reactions and attitudes towards rating scales pointed out certain problematic issues, including disagreement over criteria [19], judging writing features by means of adjectives such as *appropriately* and *well* [20], inconsistency in assigning ratings using relativistic terms [21], passing personal judgment when criteria did not capture a writing feature [22], and lack of elements such as length and vocabulary range [23].

In brief, the process of developing a rating scale should take into account considerations underlying models of both writing and language learning. A rating scale can be developed empirically when emphasis is placed on aspects of writing that are important and relevant to the learning setting. The design of a rating scales proceeds through orderly steps that begin with type identification and end with score reporting. Facets of validity of a rating scale include content, discriminability, reliability, face, consequentiality, and practicality. There are certain issues that can deter the validity of a rating scale such as prior nature, unspecific descriptors, and relativistic and vague terms.

III. Method

The current study aimed to develop a rating scale for use by teachers with elementary-level writers at the ELC. Such a rating scale would preferably have as many quantitative measures (descriptors) of writing constructs as determined practical at this level. It was hoped to encourage all writing instructors at this level to use one rating scale when scoring their students' essays. This would ensure a unified criterion for assessing composition skills across elementary-level groups. Towards the goals of this study, a systematic procedure similar to that proposed by [24] was pursued.

At the outset, reference [7] taxonomy of writing constructs and their measures was employed. Such taxonomy was developed in light of various models of writing and language learning such that it included all writing constructs and measures and their operationalizations (see Table I). It was postulated that the use of such general, all-encompassing taxonomy in the development of the rating scale sought by this study would boost its validity and reliability [see 25].

In Table I, measures from 1A to 1F are of accuracy, 2A to 2C of fluency, 3A to 3C of grammatical complexity, 3D to 3F of lexical complexity, 4A to 4D of mechanics, 5A to 5C of cohesion, 6A to 6D of coherence, 7A to 7G of writer-reader interaction, and 8A to 8D of content. The measures of content were adapted to match the research goals.

TABLE I
List of Construct Measures and Their Operationalizations

1A. error-free t-units: # error-free independent clauses + any other clauses
1B. error-free clauses: # error-free independent & dependent clauses
1C. error-free t-unit ratio: #error-free t-units /# t-units
1D. error-free clause ratio: #error-free clauses /# clauses
1E. errors per t-unit: #errors/# t-units
1F. errors per clause: #errors/# clauses
2A. number of words: # words in essay
2B. number of self-corrections: #insertions, deletions, or modifications
2C. average length of self-corrections: #letters in self-corrections/#self-corrections
3A. clauses per t-unit: #clauses/#t-units
3B. dependent clauses per t-unit: #dependent clauses/#t-units
3C. dependent clauses per clause: #dependent clauses/# clauses
3D. average word length: #characters/#spaces between words
3E. content words: #words with semantic function
3F. sophisticated words: #words from academic word list
4A. punctuation: #errors in punctuation
4B. spelling: #errors in spelling
4C. capitalization: #errors in capitalization
4D. main parts of a paragraph(introduction, body, conclusion): # missing parts/3
5A. Anaphoric pronominals: #reference pronouns
5B. Linking devices: #conjunctions
5C. Lexical chains: #lexically-related forms
6A. parallel progression: #instances where topics of successive sentences are the same
6B. direct progression: #instances where the exact comment of a sentence is the topic of the next one
6C. indirect progression: # instances where the topic/comment of a sentence is the topic of the next one by inference or exemplification
6D. extended progression: #instances where the topic/comment of an earlier sentence is the topic of a new sentence
7A. hedges: #instances where claims are moderated or softened
7B. boosters: #instances where claims are emphasized or asserted
7C. attributors: #instances where arguments are supported
7D. attitude markers: #instances of writer's personal feelings and attitudes
7E. markers of writer identity: #instances where the writer positions himself in relation to the world, reader, or text
7F. passive voice: # instances of passive voice
7G. commentaries: # instances of writer's involving the reader into discourse
8A. topic sentence: (1 if exists or 0 if not) /1
8B. supporting sentences: #supporting sentences /3
8C. supporting details #supporting details /3
8D. concluding sentence: (1 if exists or 0 if not) /1

#: number of, /: divided by.

Next, in order to determine which measures of writing constructs are important, a regression analysis was performed. Then, to identify measures that discriminate well between performance levels, a nonparametric analysis, equivalent of one-way ANOVA, was conducted. Finally, the results of the regression and differentiation statistics were consolidated to decide which measures should be in included in the rating scale.

A. Research Questions

With the research goals in mind, the research questions the current study attempted to answer were as follows:

1) What taxonomy constructs and measures are most relevant to writing assessment?
2) Which construct measures best distinguish different levels of writing performance?
3) Based on the answers to questions 1 and 2, what taxonomy constructs and measures can be included in the rating scale?

B. Setting and Participants

The research took place in an intensive English program that is part of a government institute of public administration in Saudi Arabia. This program prepares students with English skills required for degree programs in a variety of specialties. It is a year-long, four-level program with a focus on reading, grammar, writing, listening and oral skills. Each level offers an eight-week course of study.

The participants were 89 students, aged from 20 to 25. They were at the elementary or second level of the program. Generally, students at this level are expected to have attained a command of English writing skills that accord with the ACTFL proficiency guidelines for the Intermediate Mid-level. That is, on average, participants "can write short, simple communications, compositions, and requests for information in loosely connected texts about personal preferences, daily routines, common events, and other personal topics … [They] show evidence of control of basic sentence structures and verb forms" [26].

C. Data Collection

The data of this study comprised essays that students wrote in response to their final exam of writing. One main section of the exam prompted students to write a composition about one's favorite city, which includes a topic sentence, three main supporting sentences, three detailed supporting sentences, and a concluding sentence. Students had levels of both background knowledge and topic familiarity that precluded any possibility of a prompt effect [see 27]. They were allowed one hour to complete the whole exam.

Students' essays were scored using the rating scale in Fig. 2. Each essay was assigned a total score out of 30, with due consideration to content structure and components as well as grammar and writing mechanics.

Table II shows descriptive statistics of essay scores. Each essay was then coded using [7] taxonomy of writing constructs and measures. The coding process relied exclusively on the operationalizations of measures in the taxonomy (refer to Table I).

TABLE II
DESCRIPTIVE STATISTICS FOR ESSAY SCORES

Range	Minimum	Maximum	Mean	Standard Deviation	Variance
20.77	6.92	27.69	22.08	3.53	12.45

Component	# instances	Weight Sub-score
Topic sentence	1	x 5
Main supporting sentences	1/2/3	x 4
Detailed supporting sentence	1/2/3	x 3
Concluding sentence	1	x 4

Component	Weight	#mistakes	Sub-score
Grammar	Every3 mistakes = -1		
Mechanics (spelling, capitalization, punctuation)	Every5 mistakes = -1		
			Total score = (___ / 30)

Fig. 2 Rating scale used to score students' essays

IV. RESULTS

To answer the first research question, "*what taxonomy constructs and measures are most relevant to writing assessment?*", automatic linear modeling (ALM) was run in order to determine which measures of writing constructs had important effects on essay scores. ALM is a robust test of predictability of an outcome on the basis of its linear relationship(s) with one or more predictors. It demonstrates how importantly each predictor contributes to the existence of an outcome through a model, at high levels of accuracy and stability [28]. A standard model was used because its results can be directly and readily interpreted. Also, no automatic data preparation was employed in order not to transform the predictor values. Forward stepwise was used to select the model such that effects were added and removed during regression according to their importance in the model. The decision as to whether add or remove an effect was determined by adjusted R-squared (.92), which is considered a reasonable estimate. The model fit was indicated by the small estimate of information criterion (=27.823), and its accuracy which amounted to 92%. Strong correlations existed between the observed scores and the predicted ones. Outliers did not show to have substantial effects on the model accuracy.

Table III shows effects of measures on essay scores. Measures are identified by both a number referring to a construct and a serial letter, where 1 is for content, 2 for mechanics, 3 for accuracy, 4 for complexity, 5 for coherence, 6 for cohesion, 7 for fluency, and 8 for writer-reader interaction. Because mean square values in this model were identical to those of sum of squares, sum of squares are reported here. Out of the 32 measures that entered the ALM, 18 measures were important in predicting essay scores. The ALM yielded estimates of importance for each measure. Estimates of importance for each measure as well as each construct are reported here in a descending order. Importance estimates for constructs were computed by averaging out those estimates of their respective measures. The number of these measures varied across eight constructs. Thus, whereas content was represented by all of its four measures, other constructs like fluency and writer-reader interaction each had only one measure. Each of accuracy, cohesion, and coherence had a representation of two measures. As for constructs like complexity and mechanics, each had three measures.

TABLE III
EFFECTS OF MEASURES ON ESSAY SCORES

Measure	Sum of squares	Importance	
		Measures	Respective constructs
1A. supporting details	196.85	0.161	0.0885
1B. supporting sentences	56.29	0.076	
1C. concluding sentence	44.74	0.069	
1D. topic sentence	8.97	0.048	
2A. capitalization	15.83	0.052	0.0500
2B. spelling	15.76	0.052	
2C. punctuation	5.63	0.046	
3A. error-free t-unit ratio	17.01	0.053	0.0480
3B. error-free clause ratio	1.14	0.043	
4A. dependent clauses per t-unit	10.37	0.049	0.0457
4B. dependent clauses per clause	4.24	0.045	
4C. clauses per t-unit	1.34	0.043	
5A. parallel progression	7.50	0.047	0.0455
5B. indirect progression	3.21	0.044	
6A. linking devices	2.63	0.044	0.0435
6B. anaphoric pronominals	1.01	0.043	
7A. number of self-corrections	1.72	0.043	0.0430
8A. attitude markers	1.17	0.043	0.0430

df = 1; p =0.01-0.05

Evidently, each construct is represented by one, two, three, or four measures. Measures of each given construct differed in their importance. Measures of content ranked first with supporting details being the most important one. Measures of mechanics ranked second with a roughly shared level of importance of spelling and capitalization. Error-free t-unit ratio and error-free clause ratio were both important measures of accuracy. Complexity, with three measures, had the fourth rank of constructs. Measures of coherence occupied the fifth rank with two measures representing parallel and indirect modes of progression. The other measures represent constructs in close proximity, ranging from 0.0430 to 0.0435. Such measures included linking devices and anaphoric pronominals of cohesion, number of self-corrections for fluency, and attitude markers for writer-reader interaction.

Based on the ALM analysis, 18 measures predicted essay scores to an important extent. Nevertheless, such measures had different levels of importance in this regard. Consequently, these measures gave diverse weights to their respective constructs. Measures of content, mechanics, accuracy, complexity, and coherence were by far most predictive of essay scores. This finding suggests that these constructs are worth high emphasis in writing assessment at this level. The next are the three measures of cohesion, fluency, and writer-reader interaction. It might be the case that writers were considerably acquainted with the first five constructs than with the latter ones. This draws attention to the nature of skills that instructors should focus on in their writing classes and tests. These findings point out the vital roles the constructs represented by these skills play in writing at this level of language learning. They also suggest that the measures of these constructs are indispensable for the rating scale being developed to serve writing assessment at this level.

In the answer to the second research question, "*which construct measures best distinguish different levels of writing performance?*", the 32 measures that resulted from the coding process were examined to determine which measures could differentiate between five scoring levels (graded as A, B, C, D, & F). Because the measures were not all assumed to be normally distributed, the Kruskal-Wallis test was used. This test is the nonparametric equivalent of one-way ANOVA. It measures the extent to which two or more groups are different in terms of one or more variables. The results of this test indicated that only 17 measures significantly differentiated among the five scoring levels. Table IV shows values of the test statistic (H) (or index of differentiation to reflect the extent to which a measure differentiated among scoring levels) for each of the 17 measures. Estimates of differentiation for each measure as well as each construct are reported here. Tabulated in a descending order, differentiation estimates for constructs were computed by averaging out those estimates of individual measures.

TABLE IV
ESTIMATES OF DIFFERENTIATION

Measure	H	Average H of measures per construct
1A. error-free t-units	27.178	24.907
1C. error-free t-unit ratio	27.061	
1E. errors per t-unit	24.870	
1B. error-free clauses	24.478	
1F. errors per clause	23.530	
1D. error-free clause ratio	22.326	
8C. supporting details	47.457	22.713
8B. supporting sentences	33.899	
8D. concluding sentence	24.116	
8A. topic sentence	10.125	
4D. main parts of paragraph	22.033	15.295
4B. spelling	15.295	
6B. direct progression	15.103	15.103
2C. average length of self-corrections	11.690	11.262
2A. number of words	10.834	
7E. markers of writer identity	9.519	9.518
7G. commentaries	9.516	

df= 4; p: 0.00-0.05

It is obvious that all of the six measures of accuracy served as high differentiators. This is especially the case with the measures that applied to the level of t-units. All four measures of content were very good differentiators. These measures can be listed in a descending order—*supporting details*, *supporting sentences*, *concluding sentence*, and *topic sentence*. Also, two measures of mechanics, *main parts of paragraph* and *spelling*, had fairly high values as differentiators. The *main parts of paragraph* differentiated among scoring levels even better than *spelling*. *Direct progression*, as a measure of coherence, was also a good differentiator. Two measures of fluency, *average length of self-corrections* and *number of words*, differentiated well. Both measures had adjacent values along the differentiation index. Two measures of writer-reader interaction, *markers of writer identity* and *commentaries*, were reasonable differentiators.

On the basis of the Kruskal -Wallis test results, the five

constructs of accuracy, content, mechanics, coherence, fluency, and writer-reader interaction appeared to have the measures that adequately differentiated among scoring levels. All measures of accuracy set scoring levels apart, which highlights the role of grammar in writing performance at this level. As for content measures, it seems that the more a measure was detailed the better it served as a differentiator. In other words, content measures that required students to write detailed information tended to differentiate very well. This indicates that both measures of accuracy and content are essential to be incorporated in writing assessment at this level. *Main parts of paragraph* as a measure of mechanics is linked to content in that content measures themselves could make up the main parts of a paragraph. *Spelling* as a differentiator outperforms other measures of mechanics (i.e., *capitalization* and *punctuation*).

Interestingly, *direct progression* as a measure of coherence can determine the level of scoring on writing tests at this level. It involves the use of a comment in the previous sentence as a starting point to develop another sentence. This clearly correlates positively with writing ability. Fluency also determines the level of scoring on an essay, which can be measured by means of the *average length of self-corrections* or the *number of words*. Accordingly, elementary writers who paraphrase their thoughts and elaborate on them are likely to score higher than those who do not. Last, through the use of *markers of writer identity* and *commentaries*, an elementary writer can interact with his reader in a manner that reflects his writing skill.

In the answer to the third research question, "*based on the answers to Research Questions 1 and 2, what taxonomy constructs and measures can be included in the rating scale*?", the results of the ALM and the Kruskal-Wallis analyses were converged. This approach was adopted in order to produce a rating scale with measures that can predict and differentiate between levels of writing performance. Therefore, 18 measures importantly predicted essay scores. These measures had the highest regression coefficients among the 32 measures that entered the ALM analysis. The results of the Kruskal-Wallis analysis, on the other hand, indicated that 17 out of the 32 measures significantly differentiated between five levels of scoring. Therefore, a set of criteria were devised so as to make systematic decisions about which measures to include in the rating scale.

First, *error-free t-unit ratio* was chosen as a measure of accuracy. This was because it had the highest values on the two indexes of prediction and differentiation among all six measures of accuracy. Because of its sheer similarity to *error-free clause ratio*, it was decided that *error-free t-unit ratio* would more ideally serve as the only measure of accuracy in the rating scale. Second, as a measure of fluency, *number of words* was a high discriminator. Although it shared this feature with *average length of self-corrections*, *number of words* as a measure is more doable and practicable. *Number of words* was found to distinguish between writers at lower levels [7]. *Number of self-corrections*, on the other hand, appeared to be a strong predictor of scores, but lacking in terms of practicality. This is because self-corrections may not always be clear to the rater, especially with essays written in pencil. Three measures of complexity were high predictors; therefore, *number of clauses per t-unit* was selected because of its higher simplicity and popularity when compared to *number of dependent clauses per t-unit* or *clause*. *Number of clauses per t-unit* was shown to correlate with the level of proficiency [29].

Spelling was the only measure of mechanics that showed to be an important predictor and discriminator. *Spelling* was found to be a successful differentiator among lower-level writers [7]. Each one of the other measures of mechanics had a high value on either index of prediction or differentiation, but not on both. As for measures of cohesion, *linking devices* predicted essay scores more importantly than did *anaphoric pronominals*; besides, *linking devices* as a measure is more doable and practicable. Based on research evidence suggesting that notable variability was observed among writing levels in the use of linking devices [30], linking devices can also serve as a differentiator. *Parallel progression* and *indirect progression* were the only two measures of coherence that predicted essay scores with the former being more important than the latter in this respect.

Direct progression was another measure of coherence with a high value as a discriminator. The use of *direct progression* has been shown to correlate with proficiency level in previous research [e.g., 31]. Because of the evident effect of coherence on the writers' performances, a decision was made to come up with a measure of coherence that combines the predictive and discriminative features of these three measures (i.e., *parallel progression*, *indirect progression,* and *direct progression*). Thus, *logical progression* was proposed for inclusion in the rating scale, which can be operationalized as the number of instances in which the topic or comment of a sentence is the topic of the next one.

Attitude markers was the only measure of writer-reader interaction with a high value as a predictor. Two other measures of writer-reader interaction, *markers of writer identity* and *commentaries*, showed to be adequate discriminators. Once again, because writer-reader interaction evidently affected the writers' performances, a decision was made to come up with a measure of writer-reader interaction that combines the predictive and discriminative features of these three measures (i.e., *attitude markers*, *markers of writer identity*, and *commentaries*). Thus, *personal makers* is proposed for inclusion in the rating scale, which can be operationalized as the number of instances in which the writer expresses himself or involves the reader into discourse. Last, all four measures of content (i.e., *topic sentence, supporting sentences, supporting details,* and *concluding sentence*) were important predictors and significant differentiators.

In the design of a rating scale, measures of various relevant constructs should possess qualities of high prediction of writing performance and differentiation among performance levels. In elementary writing, the importance of the measures of content, accuracy, and mechanics as predictors correlates with their roles as differentiators. This suggests that these

measures are essential components of a rating scale. Measures of other constructs may exhibit themselves as either efficient predictors or discriminators, but not both. This fact applies to measures of constructs such as fluency, complexity, cohesion, coherence, and writer-reader interaction. Therefore, the choice of which measures of these constructs would fit the design and use of a rating scale can be inspired by factors including simplicity and doability. In fact, these factors besides others impose certain requirements on the choice of which measures can be included in a rating scale. Such conditions have to do with the extent to which a given measure can be used with minimal amount of time and effort on the part of the rater(s).

The answer to the third research question led to the design and formulation of a new rating scale for writing assessment at level two of the ELC program (see Fig. 3). The rating scale is composed of eight constructs. Each construct has its own measures that are operationalized and quantified such that they are bounded by their values in the data of this study. The possibility of a writer's earning a rate beyond these boundaries was catered for at both ends of the scale for each.

			Levels				
	Constructs	Measures	1(poor)	2(fair)	3(average)	4(good)	5(super)
1)	content	*8 sentences:* (#1 topic), (#3 main), (#3 detailed), (#1 concluding)	0-1 sentence	2-3 sentences	4-5 sentences	6-7 sentences	All 8 sentences
2)	accuracy	*error-free t-unit ratio:* # error-free t-units / # t-units	0.00-0.20	0.21-0.40	0.41-0.60	0.61-0.80	0.81-1.00
3)	fluency	# words	0-40	41-80	81-120	121-160	≥161
4)	mechanics	*spelling:* # mistakes	≥21	16-20	11-15	6-10	0-5
5)	complexity	# clauses / # t-units	1.00-1.25	1.26-1.50	1.51-1.75	1.76-2.00	2.01-3.00
6)	coherence	*logical progression*: # instances the topic or comment of a sentence is the topic of the next	≤2	3-4	5-6	7-8	≥9
7)	cohesion	# linking devices	≤2	3-4	5-6	7-8	≥9
8)	writer-reader interaction	*personal markers*: #instances the writer expresses himself or involves the reader	≤2	3-4	5-6	7-8	≥9

Fig. 3 The new rating scale

V. Conclusions

The scale development can make use of an existing taxonomy of constructs, and measures and their operationalizations. The taxonomy is used to code essays composed by the target population of writers. Decisions about what constructs and measures to include in a rating scale can be informed by means of predictive and discriminative statistics applied to essay data. The statistical results are considerably revealing as to what writing constructs and measures are predictive of varying levels of writing performance and/or discriminative among these levels. Certain constructs would manifest high prediction of and/or discrimination among performance levels through certain measures. This is obviously useful when determining which constructs are applicable at a given level of proficiency. The more measures linked to a particular construct are, the higher is the importance of including this construct in a rating scale. This approach can also benefit validation of writing tasks by checking the extent to which a given task is construct-relevant.

In content-based writing assessment, measures of content are definitely highly predictive of and discriminative among scoring levels. This suggests that raters should allow content a heavier weight than other constructs in their ratings of elementary-level writers. Well-composed essays, as far as content is concerned, have embedded in them advanced aspects of other writing constructs including accuracy and fluency. This in turn implies that measures of accuracy and fluency can themselves predict and discriminate among writing levels very well. However, such measures may vary among themselves in this regard, which makes the choice of which measures to use a matter to be decided in light of the writing course objectives.

When deciding which measures to select for a rating scale, there are chances that two measures of a given construct are high predictors and differentiators and both measures represent the construct in similar ways. In such cases, the measure with the highest values on both indexes of prediction and differentiation can be selected. In other cases, one measure with a high value both as a predictor and differentiator can serve as the only measure of a given construct. This is especially the case if the other measures of the same construct have lower values on either index of prediction or differentiation, or both. There are also chances that two measures of a given construct show high prediction and differentiation; however, in the actual rating process, one measure is more practical than the other. In other words, applying one measure is more economical in terms of time and effort than the other; thus, the former should be selected.

Generally, in the context of language testing, examiners and raters should strike a balance between valid and reliable assessment on one hand and practical and efficient rating on the other. The choice of measures for a rating scale can also benefit from previous research. The previous findings pointed to correlations between certain measures and levels of proficiency and writing performance. Other findings related to how certain measures can be useful on particular forms of writing assessment. Construct measures on taxonomies of writing features are usually stated in general terms; nevertheless, such measures can be made more specific to a given context of writing assessment.

Certain construct measures (e.g., spelling as a measure of mechanics) predict and differentiate well among performance levels better than other measures of the same construct. This can make intriguing topics for future research that may

consider the use of qualitative approaches to data analysis besides the quantitative ones. Thus, future research may attempt to explore which constructs and measures are applicable at different proficiency levels. Also, a possible research inquiry may seek to explain why certain measures do well as predictors of writing scores whereas others are high differentiators among scoring levels; and, whether or not this has a link to proficiency level. Perhaps, this can be explained in terms of the discourse features of the given writing tasks. Therefore, descriptive writing calls for use of more linking devices than anaphoric pronominals that may be characteristic of narrative writing.

Accordingly, when formulating prompts for essay writing, care ought to be taken as to the nature of the writing product expected (be it descriptive, narrative, expressive … etc). This also suggests that if writers are expected to employ certain aspects of a given construct, they should be provided with writing prompts that demand the use of such features. This implication is especially useful in the context of proficiency-oriented assessment where writers can demonstrate their abilities in response to a variety of writing purposes. The inclusion of as-many-as-possible construct measures in a rating scale is necessary in order to tap into, and so assess different writing abilities. Writers even at the same level of language learning exhibit different abilities. Therefore, the more versatile the rating scale is the more it allows for accurate and fair assessment of writing.

The taxonomy features excluded from the design of the current rating scale may fit writing assessments at the upper levels of the program. However, this can be examined in another study that may replicate the current study with student writers at the Intermediate and Advanced levels. The practical suggestions this study offers aim to have a rating scale with measures that are both reasonable in number and justifiable in quality. Scoring students' essays is a demanding task that may result in exhaustion, and using a lengthy rating scale may threaten the assignment of accurate and fair scores.

There are certain limiting factors in this study. First, the five scoring levels in the new rating scale were determined by means of an intuitive rating scale. These levels may not necessarily conform to the levels that can be discerned using the new rating scale. Second, the opinions of elementary-writing instructors were not taken into account when developing the new rating scale. Nonetheless, their evaluative remarks would certainly be indispensable in the validation of the new rating scale. Due to the extensive nature of rating-scale validation, no attempts were made to check the extent to which the new rating scale is valid or reliable. Therefore, this is a matter worth considering in subsequent research.

On the other hand, there are a number of features of this study that counterbalance its weaknesses. First, the rating scale that resulted from this study is empirically developed, which makes it highly reliable. This is besides the fact that it is a 5-point scale [see 32]. It also has substantial construct validity since it was developed on the basis of actual writing performance. Furthermore, it caters for the fact that because writing abilities differ in terms of their rates of development, each ability should be assessed independently. Second, the current study sets the stage for similar efforts to design rating scales for writing classes using powerful statistical tests (such as automatic linear modeling and the Kruskal-Wallis test). The results of such tests provide accurate estimates of the extent to which construct measures can predict scores and differentiate among scoring levels.

References

[1] T. McNamara, "Discourse and assessment," *Annual Review of Applied Linguistics*, vol. 22, pp. 221-242, 2002.
[2] B. North, *Scales for Rating Language Performance: Descriptive Models, Formulation Styles, and Presentation Formats*. TOEFL Monograph 24. Princeton: Educational Testing Service, 2003.
[3] S. Weigle, *Assessing Writing*. Cambridge: Cambridge University Press, 2002.
[4] T. Stewart, S. Rehorick, and B. Perry, "Adapting the Canadian language benchmarks for writing assessment," *TESL Canada Journal*, vol. 18, no. 2, pp. 48-64, 2001.
[5] U. Knoch, "Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?" *Assessing Writing*, vol. 16, no. 2, pp. 81-96, 2011.
[6] W. Grabe and R. Kaplan, *Theory and Practice of Writing.* New York: Longman, 1996.
[7] U. Knoch, "The Development and Validation of an Empirically-developed Rating Scale for Academic Writing", University of Auckland, Unpublished PhD dissertation, 2007.
[8] G. Fultcher, *Testing Second Language Speaking.* London: Pearson Longman, 2003.
[9] J. Upshur and C. Turner, "Constructing rating scales for second language tests," *ELT Journal*, vol. 49, no. 1, pp. 3-12, 1995.
[10] R. Hawkey and F. Barker, "Developing a common scale for the assessment of writing," *Assessing Writing*, vol. 9, no. 2, pp. 122-159, 2004.
[11] B. North, "The development of a common framework scale of descriptors of language proficiency based on a theory of measurement," *System*, vol. 23, no. 4, pp. 445-465, 1995.
[12] B. North and G. Schneider, "Scaling descriptors for language proficiency scales," *Language Testing*, vol. 15, no. 2, pp. 217-263, 1998.
[13] C. Alderson, "Bands and scores", in *Language Testing in the 1990s: The Communicative Legacy*, C. Alderson and B. North, Eds. London: Modern English Publications/British Council/Macmillan, 1991, pp. 71-86.
[14] C. Myford, "Investigating design features of descriptive graphic rating scales," *Applied Measurement in Education*, vol. 15, no. 2, pp. 187-215, 2002.
[15] L. Bachman and A. Palmer, *Language Testing in Practice.* Oxford: Oxford University Press, 1996.
[16] C. Turner and J. Upshur, "Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores," *TESOL Quarterly*, vol. 36, no. 1, pp. 49-70, 2002.
[17] P. Mickan, *'What's Your Score?' An Investigation into Language Descriptors for Rating Written Performance*. Canberra: IELTS Australia, 2003.
[18] R. Todd, P. Thienpermpool, and S. Keyuravong, "Measuring the coherence of writing using topic-based analysis," *Assessing Writing*, vol. 9, no. 2, pp. 85-104, 2004.
[19] S. Shaw, "IELTS writing: Revising assessment criteria and scales (Phase 1)," *Cambridge Research Notes*, vol. 9, pp. 16-18, 2002.
[20] S. Claire, "Assessment and moderation in the CSWE: Processes, performances, and tasks", in *Studies in Immigrant English Language Assessment*,2nd ed. vol., G. Brindley and C. Burrows, Eds. Sydney: National Center for English Language Teaching and Research, 2001, pp. 15-57.
[21] D. Smith, "Rater judgments in the direct assessment of competency-based second language writing ability," in *Studies in Immigrant English Language Assessment*, vol. 1, G. Brindley, Ed. Sydney: National Centre for English Language Teaching and Research, 2000, pp. 159-189.
[22] T. Lumley, *Assessing Second Language Writing: The Rater's Perspective*. Frankfurt: Peter Lang, 2005.
[23] J. Yi, "The use of diaries as a qualitative research method to investigate

teachers' perception and use of rating schemes," *Journal of Pan-Pacific Association of Applied Linguistics*, vol. 12, no. 1, pp. 1-10, 2008.
[24] T. McNamara, *Measuring Second Language Performance*. Harlow, Essex: Pearson Education, 1996.
[25] S. Luoma, *Assessing Speaking*. Cambridge: Cambridge University Press, 2004.
[26] ACTFL, "ACTFL Proficiency Guidelines for Writing", 2012, at the following link: http://actflproficiencyguidelines2012.org/writing, accessed on 17 Mar., 2013.
[27] G. Lim, "Investigating prompt effects in writing performance assessment," *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, vol. 8, pp. 95-116, 2010.
[28] IBM, *SPSS Statistics* (Version 19). (Software). 2012. Available from http://www-01.ibm.com/software/analytics/spss/
[29] K. Wolfe-Quintero, S. Inagaki, and H-Y Kim, *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*. Technical Report No. 17. Honolulu, HI: University of Hawai'i Press, 1998.
[30] C. Kennedy and D. Thorp, *A Corpus-based Investigation of Linguistic Responses to an IELTS Academic Writing Task.* Birmingham, University of Birmingham, 2002.
[31] J. Wu, "Topical Structure Analysis of English as a Second Language (ESL) Texts Written by College South-east Asian Refugee Students", ProQuest Dissertations and Theses database, 1997.
[32] Y. Sugita, "The development and implementation of task-based writing performance assessment for Japanese learners of English," *Journal of Pan-Pacific Association of Applied Linguistics,* vol. 13, no. 2, pp. 77-103, 2009.