Determining the Evaluative Criteria of an Argumentative Writing Scale

Vahid Nimehchisalem
Resource Center, Department of Language and Humanities Education, Faculty of Educational Studies
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
E-mail: nimechie22@yahoo.com
Jayakaran Mukundan (Corresponding author)
Department of Language and Humanities Education, Faculty of Educational Studies
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
Tel: 60-389-468172 E-mail: jaya@educ.upm.edu.my

Abstract

Even though many writing scales have been developed, instructors, educational administrators or researchers may have to develop new scales to fit their specific testing situation. In so-doing, one of the initial steps to be taken is to determine the evaluative criteria on which the scale is supposed to be based. As a part of a project that was proposed to develop a genre-based writing scale, a survey was carried out to investigate Malaysian lecturers' views on the evaluative criteria to be considered in evaluating argumentative essays. For this purpose, a group of English as a Second Language (ESL) lecturers (n=88) were administered a questionnaire. The subscales of organization, content and language skills were recommended by factor analysis. A fourth subscale, task fulfillment, was added as a result of the qualitative analysis of the data. The findings can be useful for language teaching or assessing purposes.

Keywords: Assessing writing, Scale development, Evaluative criteria, Argumentative writing

1. Introduction

Generally, argumentation is the art and science of civil debate, dialogue and persuasion (Glenn, Miller, Webb, Gary & Hodge, 2004). More specifically, argumentation involves statement of an issue, discussion of its pros and/or cons, and justification of support for one with the primary focus on the reader (Kinneavy, 1971). In order to write a successful argumentative essay, the writer should use an appropriate style to invent relevant and rational ideas that are linked and arranged logically with the help of the writer's language, world and strategic competencies (Bachman, 1990). In order to learn to write and revise effectively, student writers should know how to differentiate successful from unsuccessful pieces, which suggests that assessing is an indispensable part of teaching writing (Huot, 2002).

A valid and reliable assessment of learners' written works is facilitated through writing scales that provide the evaluator with a set of descriptors for each level of writing performance. Writing scales should present the evaluative criteria explicitly and clearly, based on which the results of performance assessments are determined and important decisions are made (AERA/APA/NCME, 1999). Any *ad hoc* decision in this matter may undermine the construct validity of the scale. Construct validity is "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (Bachman & Palmer, 1996:21). This would necessitate the scale developers' awareness of the construct domain that is being addressed if they wish to account for the construct validity of their instrument (Huot, 1996). The criteria emphasized by the scale depend on the specifications determined for a particular test. Thus, it is possible to find two writing scales that aim at quite different aspects of the writing ability. This could be the reason why in the literature, different criteria have been identified as the sub-traits of the writing skill, which in turn, raises the need for further investigation as new test specifications are encountered.

1.1 Evaluative Criteria

A wealth of literature is available on studies on the evaluative criteria in evaluating the written pieces. Jacobs, Zingraf, Wormuth, Hartfiel and Hughey (1981) developed the ESL Composition Profile, a generic writing scale that evaluates compositions in terms of five dimensions of the writing skill namely as content, organization, vocabulary, language and mechanics. In a different project, based on a survey of university-level academic staff in the Great Britain, Weir (1983) developed another generic scale. Weir specifies seven dimensions in his scale, including relevance and adequacy of content, compositional organization, cohesion, adequacy of vocabulary for purpose, grammar, punctuation and spelling.

For Reid (1993), content, purpose and audience, rhetorical matters (organization, cohesion, unity), and mechanics (sentence structure, grammar, vocabulary) include the main sub-traits of the writing skill in ESL situations. Cohen (1994) regards content, organization, register (appropriateness of level of formality), style (sense of control and grace), economy, accuracy (correct selection and use of vocabulary), appropriateness of language conventions (correct grammar, spelling and punctuation), reader's acceptance (soliciting reader's agreement), and finally, reader's understanding (intelligibility of the text) as the major dimensions of the writing construct.

Attali and Burstein (2006) recognize grammar, usage, mechanics, style, organization, development, vocabulary and word length as the eight important features to be assessed in an automated writing scale, called e-rater V.2. In another study, Attali and Powers (2008) focus on the same features, but substitute organization and development with essay length, which is easier to measure using computers. They found a very high correlation between the two features and essay length (around .95) and reported essay length as "the most important objective predictor of human holistic essay scores" (Attali & Powers, 2008:6). That is to say, the longer the essay, the higher the score assigned by the rater.

Besides such common sub-traits like content, organization and language use, other under-investigated aspects of the writing construct have also been identified by scholars like Lee Odell. In a very interesting review and application of Pike's (1964) tagmemics, Odell (1977) classifies intellectual processes and linguistic cues to differentiate mature written works from basic written pieces. He mentions focus, contrast, classification, change, physical context and sequence as the six features that count in the maturity of a piece of writing. The linguistic cue to recognize focus is the grammatical subject. Mature writers are capable of shifting the focus more often than the basic writers. Contrast is the second feature and shows the writer's ability to discuss what an item/issue is not, or how it differs from other items. Connectors like 'although' or 'but', and words such as 'not' can indicate contrast. Classification, as the third feature, shows the writer's skill to highlight similarities between two entities, label people, actions, feelings or ideas compared with others. A mature writer may classify with the help of relevant examples or witty metaphors. The next feature is change, a part of the writer's experience that is crucial for understanding it. Verbs like 'become' or 'turn' are the linguistic cues that show change. Physical context, or the writer's precise description of a given setting, is the fifth feature that can help distinguish mature writers. Finally, skilful writers highlight time sequences, using cues like 'subsequently' and logical sequences with the help of linguistic cues such as 'consequently'.

Useful models are also available on argumentative writing. Among others Toulmin's Model of Argument stands out for its practicality and accuracy. According to Toulmin (1958), a good piece of argument commonly consists of six elements:

i. Claim [C]: the statement of the thesis
ii. Data [D]: the evidence providing proof for C
iii. Warrant [W]: the principle that bridges D to C implicitly/explicitly, proving the legitimacy of D
iv. Qualifiers [Q]: the linguistic cues that show the strength of the C, D or W
v. Backing [B]: further support for W
vi. Rebuttal [R]: response to the anticipated objections against the arguments

The following example illustrates the elements of argument discussed above:

Narcotics are quite [Q] harmful [C] because they are addictive [D]. Anything addictive is dangerous [W] since it stops the mind from thinking [B]. These drugs are dangerous [C] unless they are used for medical reasons [R].

Besides the writer's mature development and linking of the arguments, her awareness of the audience has also been emphasized in the literature (Ryder, Lei & Roen, 1999). The audience can determine the style. A change in the audience may result in an entirely different paper. The writer's awareness of the audience will account for grounding; that is, her written piece will cognitively, linguistically and socially be appreciated by her reader, (Mäkitalo, 2006). It sounds particularly essential to consider audience awareness in the evaluation of argumentative pieces since it deals with the socio-cultural aspects of the pieces that may finally influence the reader's acceptance or rejection of the argument (Clark & Brennan, 1991). Ryder *et al.* (1999) mention four ways to account for the audience:

i. Naming moves: addressing the reader using pronouns like 'you' or 'we' or placing them in certain groups like democrats
ii. Context moves: sharing the background information based on the audience's prior knowledge
iii. Strategy moves: connecting to the audience by appealing to their interests, circumstances, emotions to ensure they will keep reading
iv. Response moves: anticipating the reader's probable responses and objections

Because of the importance of audience awareness in argumentative writing, it seems necessary to include these moves in the evaluative criteria in addition to the preceding dimensions of the writing construct.

1.2 Weighting

In developing writing scales, besides determining the aspects of the writing construct, another important step is to make decisions on the weight of each sub-construct. Two choices are available, equal-weight or optimal-weight schemes. In other words, either equal or varying weights are assigned to each dimension of the writing skill. In the ESL Composition Profile (Jacobs *et al.*, 1981), different weights are assigned to each subscale. Content has the highest weight (30% of the total score). Moderate weights are given to language use, organization and vocabulary (25%, 20% and 20% of the total mark, respectively), while mechanics receives the lowest (only 5% of the total mark). In developing his scale, called Tests of English for Educational Purposes (TEEP), Weir (1983) observes relevance and adequacy together with compositional organisation to be highly important; cohesion, referential adequacy and grammatical adequacy to be moderately important; and spelling as well as punctuation to be the least important aspects of the writing skill. Unlike the ESL Composition Profile, TEEP follows an equal-weight scheme. However, its subscales have been sequenced in the order of their importance. For instance, relevance appears first and is thus emphasized over punctuation and spelling, as the last criteria.

The weighting scheme of any scale would depend on factors like the task, purpose and learners' level (Tedick, 2002). The related literature and previous scales should also be considered before deciding on the weight of each criterion. Finally, scale developers may intend to justify the weightage through statistical evidence. Factor analysis is commonly used to recognize the components that account for the highest variance in the scores. Subsequently, a higher weight is assigned to the component that explains a higher variance (Attali & Powers, 2008).

1.3 Rationale behind Genre-based Scales

A new approach to teaching of writing emerged in the 1980s, known as the genre-based approach (Jones, 1996; Matsuda, 2003). It seeks to show the learner how writers make certain linguistic and rhetorical choices as the message, purpose and audience shift (Hyland, 2003). Accordingly, in assessing writing interest grew in "genre-specific" as opposed to "all-purpose" evaluation (Cooper, 1999:30). While all-purpose scales are generic and do not consider the type of the genre of the essays, genre-specific scale are sensitive to the changes that any shift in the genre of the text may bring about (Tedick, 2002).

Research shows that a variation in the type of the genre; that is, whether it is argumentative, descriptive or narrative, can affect the schematic structure of the text (Lock & Lockhart, 1999; Strong, 1990). Beck and Jeffry (2007) observe that reports first present an overview of the topic, then describe the information in a logical sequence and finally may or may not include a conclusion. In contrast, argumentative essays typically begin with the statement of a thesis followed by the supporting evidence and end with conclusion where the thesis is reiterated. As a result of these and similar findings, genre-specific scales have been developed over the last few decades to account for these variations (Connor & Lauer, 1988; Glasswell, Parr & Aikman, 2001). As an example, Glasswell *et al.* (2001) developed the Assessment Tools for Teaching and Learning (asTTle) writing scoring rubrics for grades (2-4) of school students in New Zealand. These scoring rubrics consist of six genre-specific scales developed to assess students' ability to explain, argue, instruct, classify, inform and recount. Each scale has four subscales namely as 'audience awareness and purpose', 'content inclusion', 'coherence' and 'language resources'.

2. Research objectives and questions

The study followed the objective of determining the evaluative criteria that ESL lecturers regard as important in evaluating argumentative essays. More specifically, it aimed at investigating ESL lecturers' views on the importance, wording, and inclusiveness of the evaluative criteria. In order to meet these objectives the following research questions were posed:

i. What evaluative criteria are regarded as important in evaluating argumentative essays?
ii. What weightage can be assigned to the determined evaluative criteria?
iii. How can these criteria be grouped?

The following section discusses the way in which the researchers sought to answer the research questions.

3. Method

The sequential explanatory model, a type of mixed-method design, was used (Creswell, 2003). In this method, quantitative data collection precedes collecting qualitative data. The priority is placed on the quantitative results, while the qualitative findings shed light on the quantitative data, thus deepening the understanding of the results (Creswell, 2007). This section discusses the instrument, respondents and data analysis method.

3.1 Instrument

Based on the literature, the Evaluative Criteria Checklist for ESL Argumentative Writing (Appendix), an eleven-item, six-point scale Likert style instrument, was developed. The respondent would indicate how significant each criterion was by assigning it a score from zero to five, or the least to the most important. The first five items, including syntax, usage, mechanics, style and essay length as well as their sub-categories were taken from other similar studies like Attali and Powers' (2008). While these items focused on the form, the criteria in items 6-11 emphasized the meaning domain of the writing ability. The item on intellectual maturity came from Odell (1977). A review of Harmer (2004) and similar literature resulted in the next two criteria, cohesion and coherence. The next item, effective argumentation, represented Toulmin's (1958) model. The last two items concerned audience awareness and invocation (Ryder *et al.*, 1999). Once it was ready, three experts were consulted to determine its adequacy.

The respondents were ESL writing lecturers that are typically very busy people, so the checklist had to look brief. To this end, the criteria were limited to the main sub-constructs of the ESL writing ability. The items in the checklist represented the sub-constructs while their subcategories appeared together in brackets next to each sub-construct. The respondents were free to reword the criteria if they found them ambiguous and to comment on any one of the criteria. The final row of the checklist was left open. If the experts found an important criterion was missing in the list, they could add it in this part.

3.2 Respondents

A group of ESL writing lecturers was enquired about their views on the evaluative criteria that should be regarded in assessing university students' argumentative essays. The statistical method that was employed to analyze the data collected from the ESL lecturers was factor analysis. For this method, Nunnally (1978) suggests a sample size of 10:1 ratio of subjects to items or a minimum of 5:1 ratio. To offer an example, providing the instrument consists of 11 items (as in the present study), the appropriate size will range between 55 and 111.

In factor analysis the data do not have to meet the assumption of random selection, so non-probability sampling method was used to select the respondents. Purposive sampling method, in which "elements are selected based on the researcher's judgment that they will provide access to the desired information" (Dattalo, 2008:6) was followed. Thus, a copy of the instrument was sent to a group of 110 ESL lecturers that had an experience of two years or above in rating in Malaysia. However, since after the second follow-up the number of the respondents reached only 69, the researchers snowballed to gain access to a few more samples. For this purpose, a group of respondents were sent several copies of the instrument and were requested to forward them to their colleagues, who were experienced raters. As a result, 88 checklists were finally collected. The number was far greater than the minimum required size (55) and therefore adequate for factor analysis.

3.3 Statistical Analysis Method

The statistical analysis was carried out using SPSS version 14. Exploratory Factor Analysis was used to analyze the data. The method can indicate how much variance is explained by each factor and can recommend the researcher how to classify several items in the instrument under a limited number of categories (Hair, Black, Babin, Anderson & Tatham, 2006). This latter application can be highly helpful in scale development as it contributes to the economy and practicality of the scale by aiding the developer to collapse certain components.

4. Results

The checklist was administered to the experts. The survey resulted in both quantitative and qualitative findings that are discussed in this section.

4.1 Quantitative findings

The researchers collected, tabulated and analyzed the data using descriptive statistics and factor analysis. This section presents the descriptive statistics results followed by the factor analysis output. Table (1) shows the descriptive statistics results while Figure (1) illustrates the importance of each criterion rated by the respondents. As the table and figure show, the criteria can be divided into three groups in terms of their importance:

i. Important/very important (4-5)
ii. Fairly important/important (3-4)
iii. Almost important/fairly important (2-3)

The results are expressed as mean ± SD (n = 88). The respondents rated coherence (4.4±.7), cohesion (4.34±.77), effective argumentation (4.22±.88) and syntax (4.15±.85) as the important/very important criteria. This suggests that 10 percent of the total score had to be assigned to each criterion in this category. Usage (3.82±.99), audience

awareness (3.61±1.09), audience invocation (3.57±1), style (3.54±1.02), mechanics (3.36±1.14) and intellectual maturity (3.14±.73) included the criteria rated as fairly important/important. The sub-traits in this category would account for about 8-9 percent of the total score. The only criterion that was regarded as the least important in the checklist was essay length (2.8±1.21). These findings will later be discussed in comparison with the results of qualitative analysis of data after the factor analysis results.

Prior to factor analysis, the data were tested for sampling adequacy (Coakes & Steed, 2007). For this purpose, Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity were used. Table (2) indicates the results of these analyses. The value of the Bartlett's test of sphericity (p=.000) was less than alpha (α=.05), and therefore significant. Moreover, the Kaiser-Meyer-Olkin measure of sampling adequacy (.654) was more than the threshold of (.6). Thus, the data set was appropriate for factor analysis.

Next, Kaiser's Criterion, or the eigenvalue rule, was used to identify the number of the factors. Table (3) shows the results of this analysis. As it can be observed from the table, there are three components with eigenvalues of more than one. According to the results of Rotation sums of squared loadings, about 22, 21 and 15 percent variance will be explained by the first, second and third components, respectively, with the cumulative percentage of about 57 percent variance.

Besides showing the components which account for the highest variance, factor analysis can also group the similar components together. The Varimax Rotation Technique was used to determine the factor loading of each factor. This technique is commonly employed in factor analysis because it "reduces the number of complex variables and improves interpretation" (Coakes & Steed, 2007:131). Table (4) demonstrates the Rotated Component Matrix, according to which the eleven items have been divided into three distinct sub-constructs of argumentative writing skill.

The first group of components comprises audience invocation, audience awareness, usage and intellectual maturity with factor loadings ranging between .291 and .876. The researchers labeled the group as 'content'. The second column indicates the factor loadings of the components in the second group. These components are composed of cohesion, coherence and effective argumentation with factor loadings ranging from .759 to .87. The category was labeled as 'organization'. The final group included mechanics, syntax, essay length and style with factor loadings of between.426 and .766. The category was called 'language skills'.

By right, the researcher's decision on the number of the factors to be extracted should rely on the literature, and factor analysis can only make recommendations in this respect (Hair *et al.*, 2006). The results of the Rotated Component Matrix were quite consistent with the literature. However, a closer look at the three groups and their subcategories revealed that two of the sub-categories did not fit in their groups. Usage did not sound appropriate under 'content'. The second group, 'language skills', was more suitable for this component. Effective argumentation would also fit better under 'content' than under 'organization'. Excluding usage and effective argumentation, the remaining items fit their groups.

Table (4) presents the three groups of factor loadings. Factor loadings indicate the correlation between the original variables and the factors (Coakes & Steed, 2007). According to Hair *et al.* (2006), factor loadings are commonly interpreted following this rule of thumb:

i. >.30: unacceptable factor
ii. .30-.40: minimally acceptable factor
iii. .40-.50: acceptable factor
iv. .50<: significant factor

Based on this rule of thumb, intellectual maturity with a low factor loading of (.291) is the only unacceptable factor. This implies that the dimension is not suitable to evaluate argumentative essays. There may have been two reasons behind the respondents' poor rating of this dimension. They must have believed that the elements of intellectual maturity are rather inappropriate for evaluating argumentative essays. Indeed, some of them (like physical context) sound more suitable for the narrative genre. The other component that also has a relatively low factor loading is style, but according to the rule of thumb mentioned above, style is an acceptable factor since its factor loading (.426) is between (.40-.50). The other components indicate practically significant values of more than .50. The results are almost in line with those of Weir (1983). As discussed previously in the introduction, he reported compositional organisation as highly important; grammatical adequacy as moderately important; and mechanics as the least important traits of the writing ability. Essay length is also identified as a significant factor, a finding that is consistent with that of Attali and Power's (2008), who regard essay length as an important factor in determining the score of a written piece. These findings will further be discussed following the discussion of the qualitative results in the next section.

4.2 Qualitative findings

As mentioned previously, the checklist enabled the respondents to reword any criterion that they considered ambiguous. Furthermore, they could add comments for the criteria or their sub-categories. They were also informed that the list was not exhaustive and that more criteria could be added. A number of the respondents (about 10 percent) took time to provide such qualitative data, which helped the researchers improve the quality of the checklist considerably. A summary of the comments made by the respondents is presented in this section.

A few respondents warned about an overlap between syntax and usage since 'articles', 'prepositions' and 'faulty comparisons' could be regarded as the elements of syntax. As a respondent noted:

In some places, usage is considered the same as grammar (morphology & syntax), but sometimes, it has assumed a broader meaning comprising a number of competences such as vocabulary, morphology, syntax, and phonology/graphology. In the latter case, it is very important.

As this excerpt suggests, some raters may be unable to differentiate between usage, syntax and morphology. This will create confusion and influence the validity of the scale that is developed based on these criteria. The respondent also points out vocabulary should also be added to the list. There were three other respondents who changed usage to vocabulary. According to two of these experts, usage involved sub-categories like 'word forms' and 'confusable words'. In order to avoid confusion usage was reworded as vocabulary. 'Articles', 'prepositions' and 'faulty comparisons' were also moved under syntax. The new classification of the dimensions of the writing construct was identical with most of the available writing scales. Most of them have a separate subscale for vocabulary. In addition, they do not divide language use into syntax and usage that may sound ambiguous for the raters.

In addition to syntax and usage, style was another item that received a few comments. According to the descriptive results, style was not rated as a highly important item. This was not what the researchers had been expecting. Factor analysis recommended it as an acceptable item in the checklist; however, its factor loading was lower than the values of nine other items. Qualitative findings shed light on the reason behind the respondents' low support for this criterion. "If by style, you mean rhetorical organization included in part 3," an ESL lecturer with 15 years of experience commented, "it is, in my opinion, important." As this comment demonstrates, style had not been defined properly and had confused some of the respondents. This along with other similar comments drew the researchers' attention to the fact that the 'appropriate use of words, phrases and passive voice, avoiding repetitious words and unnaturally long/short sentences' was too narrow a definition for style. The criterion, therefore, was defined from the perspective of classical rhetoric. As a result, following Crowley and Hawhee (2004), style was defined more broadly as the 'correct, clear, appropriate and ornate use of language'.

Quite a few comments had also been made on the evaluative criterion of essay length. It had been rated almost/fairly important with the lowest mean of (2.84), according to the descriptive statistics results. Even though its factor loading was over .50, it was the item with the third lowest value. The finding was in contrast with Attali and Burstein's (2006) as well as Attali and Power's (2008). Essay length was reportedly the most important factor in these studies. Qualitative data revealed what some respondents really thought about the criterion: "To me, essay length is not very important," a respondent maintained, "But, if it was part of task fulfillment, it would be a totally different kettle of fish. Then, I'd give it 5". Task fulfilment had also been mentioned by two other experts. A respondent added task fulfilment to the checklist and defined its sub-categories as 'shows understanding and mature treatment of the topic' and 'shows development of ideas and view points'. Following these comments, the researchers changed essay length to a more accurate term, task fulfillment. Furthermore, they defined it as the 'understanding of topic' and 'development of ideas'.

Intellectual maturity had also been given a few comments. For some respondents, it was 'out of place'. For others, its sub-category, sequence, was more appropriate for coherence, the eighth item in the checklist. As it was also pointed out, intellectual maturity and its subcategories such as 'change' and 'physical context', in particular, would be more suitable for the narrative than argumentative genre. These comments encouraged the researchers to examine the item more critically. At a closer look, its first three sub-categories, 'focus', 'contrast' and 'classification' looked more appropriate under style with its broader definition. Shifting the focus and using contrast help writers avoid repetitions and thus create a more engaging style. What is more, 'classification'; that is, using examples, metaphors and other figures of speech can make the style ornate. Style seemed to cover most of the components of intellectual maturity. Those that it did not either fell under other criteria (like sequence that went under coherence) or sounded more appropriate for narrative writing (change and physical context). Therefore, after moving sequence under coherence, the researchers removed intellectual maturity to avoid redundancy. The findings of qualitative analysis of data were consistent with those of factor analysis according to which intellectual maturity was recognized as an unacceptable factor due to its low factor loading.

Additionally, according to the comments, effective argumentation, audience awareness and audience invocation (items 9-11) were recommended to be collapsed under one category. As one of the lecturers explained, the three criteria had several overlapping features. Audience awareness had a cause and effect relationship with audience invocation. Obviously, the writer's awareness of her audience is not ensured unless she invokes the reader. Further, the fourth technique of audience invocation, 'response moves', was in fact the same as 'accounting for counter-arguments' under effective argumentation. Thus, the three criteria were collapsed in order to make the list more economical.

A final comment made by a respondent concerned the theoretical framework on which the checklist had been based. As he stated, establishing the checklist on a sound taxonomy of language ability and then adding it an argumentative flavor would contribute to the construct validity of the final list of criteria. The list would be more inclusive as compared with when different models and frameworks from a variety of sources in the literature were patched together. The researchers found Bachman's (1990) tree diagram of language competence useful for this purpose. A description of this model would be beyond the scope of the paper, but for the purpose of the present discussion it should be stated that it helped the researchers improve the theoretical framework of the scale in the end. These findings emphasize the importance of qualitative data that can enrich and improve the quality of quantitative findings.

Table (5) presents a summary of the revised list of the evaluative criteria based on both quantitative and qualitative findings. The table shows the list of the evaluative criteria, their importance and the recommended groups to which they belong. The criteria were categorized in four sub-constructs of the writing ability, namely as task fulfillment, organization, content and language skills, in order of importance. Task fulfillment was added as a result of the qualitative analysis of the findings. Effective argumentation, audience awareness and audience invocation were collapsed as a result of qualitative analysis of the data while factor analysis grouped the three together under content. Coherence and cohesion were categorized under organization. Syntax, vocabulary, style and mechanics were grouped as the components of language skills. With regard to their importance, all of the components were practically significant with an exception of style. Nevertheless, the qualitative analysis of data revealed this was due to the narrow definition of style in the checklist.

5. Conclusion

The paper began with a review of the evaluative criteria to be considered in developing an argumentative writing scale. Based on this review, a checklist was developed. Mixed method was followed to investigate a group of ESL lecturers' views on the wording, inclusiveness and importance of the criteria. The findings of factor analysis showed which factors were practically significant. The analysis also helped the researchers group the components in three categories. The qualitative analysis of data revealed further useful points that urged the researchers to reconsider the list of the criteria. The insight shared by the respondents in the survey aided the researchers to reform and improve the list.

As it was observed, an integration of qualitative and quantitative methods to collect data helped the researchers gain a better understanding of the respondents' views. In the absence of qualitative findings, quantitative data would have led the researchers to a misconceived interpretation of the components in the checklist. However, qualitative results illuminated the reasons behind the choices that the respondents had made. An appreciation of these points made the criteria less ambiguous, more relevant and more economical. Neglecting these important data could have significantly lowered the validity of the scale that was to be developed based on the criteria.

The findings of this study can be useful for test and scale developers. They can follow the same procedure to find what counts in the evaluation of a particular area of language ability from the viewpoint of the practitioners in their testing situation. Similarly, writing instructors may apply the list of the criteria to develop checklists to assess their learners' argumentative writing. By so-doing they can systematically diagnose the particular problem areas of their student writers. They may also introduce the criteria to their learners and help them use the checklist as a guide for a peer feedback activities or self-assessment purposes. Research shows that most Malaysian students are unaware of the criteria according to which their written pieces are scored (Mukundan & Ahour, 2009). There is empirical evidence that such an implicit method of evaluation can increase learners' test anxiety and lower their motivation (Brennan, Kim, Wenz-Gross & Siperstein, 2001). Checklists of this type can, therefore, improve the quality of teaching-testing ESL writing (Campbell, 1998).

When the writing instructors do not have access to evaluative criteria checklists or writing scales, they commonly evaluate their learners' written works impressionistically. This method can be highly subjective and brings about the challenge of evaluators' idiosyncratic judgment (Cooper & Odell, 1999). In other words, two different raters may

assign quite discrepant scores to the same piece. Scale-based assessment of writing can aid evaluators to score more reliably (Cooper & Odell, 1977; Crusan & Cornett, 2002).

Employing checklists and scales to evaluate written samples can help the reader guard against rater bias, or the rater's idiosyncratic beliefs about successful writing (Tedick, 2002). Research has indicated that when raters are untrained, they show a tendency to emphasize sentence level accuracy and language skills over other sub-traits of the writing skill, like content and organization (Sweedler-Brown, 1993). This suggests that scales and checklists can increase the validity of the evaluator's judgment by controlling the problem of rater bias.

Further research is required to gain a deeper understanding of the experts' views about the evaluative criteria through focus group studies. Analysis of a batch of argumentative samples can also empirically indicate which criteria should be considered in evaluating argumentative pieces. Then a comparison of the findings of all these data with those of the present study would unveil an interesting account of the criteria that should be considered in assessing argumentative writing.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME) (1999). *Standards for educational & psychological testing*. Washington, DC: American Educational Research Association.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with *e-rater* V.2. *Journal of Technology, Learning, and Assessment, 4*(3).

Attali, Y. & Powers, D. (2007). *Construct validity of* e-rater *in scoring TOEFL essays* (ETS Research Rep. RR-07-21). Princeton, NJ: ETS.

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

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

Beck, S. W. & Jeffry, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing, 12* (1), 60-79.

Brennan, R., Kim, J., Wenz-Gross, M. & Siperstein, G. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review, 71*(2): 173-216.

Campbell, C. (1998). *Teaching second language writing: Interacting with text*. Boston: Heinle & Heinle Publishers.

Clark, H. H. & Brennan, S. E. (1991). Grounding in communication. In Lauren B. Resnick, John M. Levine, & Stephanie D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Coakes, S. J. & Steed, L. (2007). *SPSS version 14.0 for Windows: Analysis without anguish.* Melbourne: Wiley.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston: Heinle & Heinle Publishers.

Connor, U. & Lauer, J. (1988). Cross-cultural variation in persuasive student writing. In A.C. Purves (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 206-227). Newbury Park, CA: Sage.

Cooper, C. (1999). What we know about genres, and how it can help us assign and evaluate writing. In C. Cooper, and L. Odell, (Eds.). *Evaluating writing: The role of teacher's knowledge about text, learning, and culture* (pp. 23-52). Urbana, Illinois: National Council of Teachers of English (NCTE).

Cooper, C. R & Odell, L. (1999). Introduction: evaluating student writing, what can we do, and what should we do? In C. Cooper, and L. Odell, (Eds.). *Evaluating writing: The role of teacher's knowledge about text, learning, and culture* (pp. vii-xiii). Urbana, Illinois: National Council of Teachers of English (NCTE).

Cooper, C. R. & Odell, L., (Eds.), (1977). *Evaluating writing: Describing, measuring, judging*. Ill: National Council of Teachers of English.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed method approaches.* California: Sage Publications, Inc.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks: Sage.

Crowley, S. & Hawhee, D. (2004). *Ancient rhetorics for contemporary students*. New York: Pearson Education, Inc.

Crusan, D. & Cornett, C. (2002). The cart before the horse: Teaching assessment criteria before writing. *The International Journal for Teachers of English Writing Skills, 9*, 20–33.

Dattalo, P. (2008). *Determining sample size: Balancing power, precision, and practicality*. Oxford: OUP.

Glasswell, K., Parr, J. & Aikman, M. (2001). Development of the asTTle writing assessment rubrics for scoring extended writing tasks. Technical Report 6, *Project asTTle*, University of Auckland.

Glenn, C., Miller, R. K., Webb, S. S., Gary, L. & Hodge, J. C. (2004). *Hodges' harbrace handbook* (15th ed.), Boston: Thompson Heinle.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). NJ: Pearson.

Harmer, J. (2004). *How to teach writing?* Petaling Jaya, Malaysia: Longman Pearson Education Limited.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Logan, Utah: Utah State University Press.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*, 549-566.

Hyland, K. (2003). Genre-based pedagogies: A social response to process. *Journal of Second Language Writing*, *12* (1), 17-29.

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V. F. & Hughey, J. (1981). *Testing ESL composition: A practical approach*. MA: Newbury House Publishers.

Jones, A. M. (1996) Dialogue: Genre and pedagogical purposes. *Journal of Second Language Writing, 4*(2), 181-190.

Kinneavy, J. A. (1971). *A Theory of discourse: The aims of discourse*. Englewood Cliffs, NJ: Prentice Hall.

Lock, G. & Lockhart, C. (1999). Genres in an academic writing class. *Hong Kong Journal of Applied Linguistics, 3*(2): 47-64.

Matsuda, P. K. (2003). Process and post-process: A discursive history. *Journal of Second Language Writing, 12* (1), 65-83.

Mäkitalo, K. (2006). *Interaction in online learning environments: How to support collaborative activities in higher education settings*. Research Reports. Jyväskylä, Finland: Institute for Educational Research. University of Jyväskylä.

Mukundan, J. & Ahour, T. (2009). Perceptions of Malaysian school and university ESL instructors on writing assessment. *Journal Sastra Inggris, 9*(1), 1-21.

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

Odell, L. (1977). Measuring changes in intelligence processes as one dimension of growth in writing. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 107-132). Ill: National Council of Teachers of English.

Pike, K. L. (1964). A linguistic contribution to composition, *College Composition and Communication 15:* 82-88.

Reid, M. J. (1993). *Teaching ESL writing*, New Jersey: Regents/ Prentice Hall.

Ryder, P. M., Lei, E. V. & Roen, D. H. (1999). Audience considerations for evaluating writing. In C. Cooper, and L. Odell (Eds.), *Evaluating writing: The role of teacher's knowledge about text, learning, and culture.* (pp. 53-71). Urbana, Illinois: National Council of Teachers of English (NCTE).

Strong, W. (1999). Coaching writing development: Syntax revisited, options explored. In C. Cooper and L. Odell, (Eds.), *Evaluating writing: The role of teacher's knowledge about text, learning, and culture* (pp. 72-92). Urbana, Illinois: National Council of Teachers of English (NCTE).

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

Tedick, D. J. (2002). Proficiency-oriented language instruction and assessment: Standards, philosophies, and considerations for assessment. In Minnesota Articulation Project, D. J. Tedick (Ed.), *Proficiency-oriented language instruction and assessment: A curriculum handbook for teachers (Rev Ed.)*. CARLA Working Paper Series. Minneapolis, MN: University of Minnesota, The Center for Advanced Research on Language Acquisition. Retrieved October 20, 2009 from http://www.carla.umn.edu/articulation/polia/pdf_files/standards.pdf

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Weir, C. J. (1983). Identifying the language needs of overseas students in tertiary education in the United Kingdom. Unpublished PhD Thesis, University of London Institute of Education.

Table 1. Descriptive statistics

No	Criteria	n	Sum	Mean	Std.*	%
1	Coherence	88	387	4.40	.70	10.7
2	Cohesion	88	382	4.34	.77	10.6
3	effective argumentation	88	371	4.22	.88	10.3
4	Syntax	88	365	4.15	.85	10.1
5	Usage	88	336	3.82	.99	9.3
6	Audience awareness	88	318	3.61	1.09	8.8
7	Audience invocation	88	314	3.57	1.00	8.7
8	Style	88	312	3.54	1.02	8.6
9	Mechanics	88	296	3.36	1.14	8.2
10	Intellectual maturity	88	276	3.14	.73	7.6
11	Essay length	88	250	2.84	1.21	6.9

*Std.: Standard deviation

Table 2. Results of KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.654
Bartlett's Test of Sphericity	Approx. Chi-Square	253.385
	Df	55
	Sig.	.000

Table 3. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.951	26.828	26.828	2.951	26.828	26.828	2.387	21.704	**21.704**
2	2.017	18.334	45.162	2.017	18.334	45.162	2.262	20.560	**42.264**
3	1.346	12.240	57.402	1.346	12.240	57.402	1.665	15.138	**57.402**
4	.962	8.744	66.146						
5	.813	7.391	73.537						
6	.764	6.944	80.482						
7	.678	6.163	86.644						
8	.531	4.825	91.470						
9	.445	4.044	95.513						
10	.266	2.422	97.935						
11	.227	2.065	100.000						

Extraction Method: Principal Component Analysis.

Table 4. Rotated Component Matrix(a)

	Component		
	1	2	3
Audience invocation	.876	.047	-.080
Audience awareness	.852	.051	.055
Usage	.680	-.026	.288
Intellectual maturity	.291	.231	.133
Cohesion	-.019	.870	.159
Coherence	-.069	.787	.059
Effective argumentation	.311	.759	-.034
Mechanics	-.057	.017	.766
Syntax	.097	.256	.683
Essay length	.349	-.313	.539
Style	.331	.292	.426

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Table 5. List of criteria according to quantitative and qualitative results of the survey

No	Criteria	Importance (mean)	Recommended sub-constructs
1	Task fulfillment	5	Task fulfillment
2	Coherence	4.40	Organization
3	Cohesion	4.34	Organization
4	effective argumentation	4.22	Content
5	Syntax	4.15	Language skills
6	Vocabulary	3.82	Language skills
7	Style	3.54	Language skills
8	Mechanics	3.36	Language skills

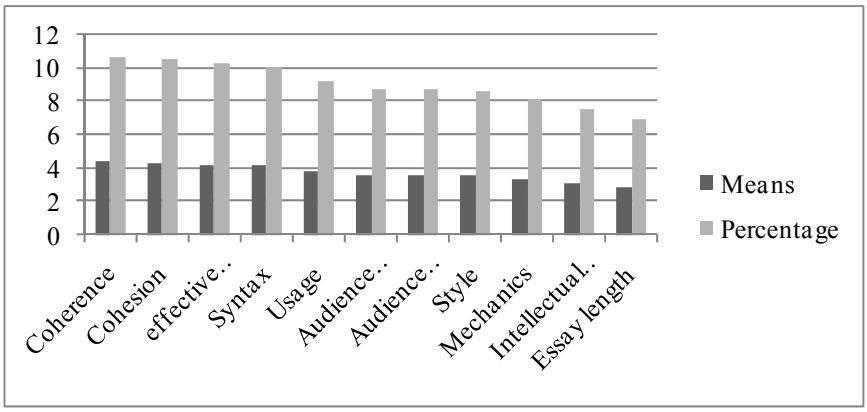

Figure 1. Degree of importance of the criteria

Appendix: Evaluative Criteria Checklist for ESL Argumentative Writing

Below you will kindly find a short list of criteria that will be used to develop the rubrics of an argumentative writing scale. You are requested to type (0-5) to indicate the level of importance of each criterion in scoring a successful argumentative essay written by university students in an ESL setting, according to this key:

0: Unimportant **3: Fairly important**

1: Not very important **4: Important**

2: Almost important **5. Very important**

If you think a criterion is missing, you may add it to the end of the list and indicate its level of importance. In addition, if there is a term that, according to your experience, would be hard for novice raters to understand, you may add the term that you recommend in the column, *Reword the Term*. Finally, if you have any further comments about each criterion, kindly mention it in the *Comment* column.

Personal Information

Kindly underline the item that applies to you:

Rating experience: - None - Below 1 year - 2 to 3 years - Over 3 years

Criteria	Importance (0-5)	Reword the Term	Comment
1. syntax (pronouns, verb forms, possessives, plural/singular noun, subject-verb agreement, avoiding fragments, run-on sentences (two or more sentences connected together only with commas) and garbled sentences (sentences with confusing meanings due to their disorganized forms), using complex structures)			
2. usage (articles, prepositions, word length, avoiding wrong word forms (e.g. 'Her father is a cook.'), confusable words (e.g. 'advise' and 'advice') and faulty comparisons)			
3. mechanics (spelling, capitalization, punctuation)			
4. style (appropriate use of words, phrases and passive voice, avoiding repetitious words and unnaturally long/short sentences)			
5. essay length			
6. intellectual maturity (frequent shift of grammatical subject or focus, focusing on what something is not or how different it is from other things or contrast, labeling people, actions, feelings or ideas compared with other entities, or classification; showing how the course of action changes, or change, describing the physical context and describing the order in which events occur, or sequence)			
7. cohesion (lexical set chains (words in the same topic area) grammatical cohesion(pronoun and possessive reference, article reference, tense agreement, linkers, substitution and ellipsis) and repetition of words)			
8. coherence (transition, organization)			
9. effective argumentation (making a claim, providing data to support the position, providing warrants; i.e., bridging claim to data to show the connection between them, backing to show the logic used in the warrants is good, accounting for counter-arguments)			
10. audience awareness (basing the argument on readers' values and perceptions, attitudes and background knowledge)			
11. audience invocation (by addressing readers with pronouns like 'you' or 'we', or naming moves; by limiting the background information based on the audience's prior knowledge, or context moves; by connecting to the audience by appealing to their interests and emotions, or strategy moves, and by accounting for the reader's probable objections, or response moves)			
12. ……			