Pertanika J. Soc. Sci. & Hum. 21 (1): 85 - 104 (2013)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Developing a Content Subscale to Assess University Students' Argumentative Essays

Vahid Nimehchisalem[#] and Jayakaran Mukundan*

Department of Educational Studies, Faculty of Education, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

ABSTRACT

Genre-specific scales are available to evaluate students' writing in English as a Second Language (ESL) situations, but instructors may still feel a need to develop new scales to match their specific testing situations. In order to develop a valid instrument for their testing situation, the researchers reviewed the literature and carried out a survey as well as a focus group study. These led them to a number of subscales, namely, content, organization, vocabulary, language conventions and overall effectiveness. The paper reviews how the band descriptors for the content subscale of the Analytic Scale of Argumentative Writing (ASAW) were determined. Toulmin's (1958/2003) model was used to analyze the patterns of argument in 20 purposely selected argumentative essays written by a group of Malaysian students. The results of the analysis provided the researchers with descriptors for five levels of writing ability. The subscale was tested for inter- and intra-rater reliability as well as concurrent validity. Positive results were observed. ESL writing instructors and evaluators may find the subscales useful for formative assessment purposes. In addition, the samples can be useful models for ESL students to differentiate the successful from unsuccessful argumentative content in writing courses.

Keywords: Assessing English as a second language writing, writing scale development

INTRODUCTION

In order to assess their bachelor degree students' argumentative essays in their English writing courses, the present researchers required a rating scale. These students were mainly freshmen in a public Malaysian university, Universiti Putra Malaysia (UPM), who are generally of

ARTICLE INFO

Article history:
Received: 5 January 2011
Accepted: 20 October 2011

E-mail addresses:
nimechie22@yahoo.com (Vahid Nimehchisalem),
jaya@educ.upm.edu.my (Jayakaran Mukundan)
* Corresponding author
[#] *Current address:*
Department of English Language, Faculty of Languages and Linguistics, Universiti of Malaya, 50603 Kuala Lumpur, Malaysia

ISSN: 0128-7702

a lower intermediate level of English proficiency. This university follows a grading system with a numerical value ranging from 0 to 100 categorized into the five grades of A, B, C, D and F. Therefore, this study followed the objective of developing a writing scale that could differentiate five levels of performance from 'excellent' (A) to 'very limited' (F) level.

A writing scale is an instrument that provides writing teachers, researchers or raters with a scoring guide to help them reach more reliable and valid measures of students' writing performance. Writing scales are of different types. They may be either all-purpose or genre-specific (Cooper, 1999). All-purpose scales are generic in nature and are developed to score scripts regardless of the genre in which they were written. A generic scale does not take the genre of the written works into account, whereas a genre-specific scale is sensitive to the type of the text written, that is, whether it is narrative, descriptive or argumentative. As the genre of a text shifts, so does its schematic structure (Lock & Lockhart, 1999). For example, argumentative essays commonly start with the statement of a position, continue with supporting evidence and end with a reiteration of the position. By contrast, descriptions in the form of scientific reports begin with an overview of the classification of the topic under discussion, followed by a presentation of certain information in a logical and thematic order and end with and sometimes without a conclusion (Beck & Jeffry, 2007). This suggests "when we ask students to explain or argue in writing, we are implicitly asking for certain kinds of sentences" (Strong, 1999, p. 83). In assessing writing, such variations should be considered by including certain criteria that are particularly devised to gauge student writers' ability to handle a specific genre (Cooper & Odell, 1999).

Generic or genre-specific scales may be either holistic or analytic. Holistic scales help the rater assign a grade for a certain script considering the writer's overall writing skill. They are useful for large-scale tests and placement purposes (Cohen, 1994). Analytic scales, on the other hand, divide the writing construct into its various dimensions such as content, language use, organization and the like. They are appropriate for classroom use and diagnostic purposes because they can indicate the particular weaknesses or strengths of student-writers (Weigle, 2002).

Most widely used writing scales that are available in the literature are generic in nature and are not sensitive to the genre of essays. However, the available few genre-specific scales have their own drawbacks. Some of them cover only a few traits of writing and not the whole construct. Connor and Lauer (1988, p. 145), for instance, developed a scale to assess the argumentative quality of written pieces in terms of their 'claim', 'data' and 'warrant'. This scale focuses on student writers' ability in argumentative writing. Moreover, it ignores other traits like language use among other important dimensions of writing in ESL situations.

Admittedly, there are genre-specific scales that cover all important traits of writing ability; however, they may have been developed in the light of test specifications different from those in another assessment situation. For example, the Writing Scoring Rubrics (Glasswell *et al.,* 2001) were developed for school students in New Zealand. Thus, these rubrics are appropriate for the performance levels of school children's written pieces in New Zealand, and they cannot be used for assessing Malaysian university student's essays.

To best of our knowledge, no analytic argumentative scale has been developed to match with this grading system. Meanwhile, adjusting the existing scales to their situation would complicate the development procedure since their descriptors differentiated only three levels of performance. For this reason a project was proposed that involved developing the content subscale of the Analytic Scale of Argumentative Writing (ASAW). This paper presents one of the phases of development of this scale. Before a discussion of this particular phase, an overview of the project will follow.

In its development, the ASAW went through the four phases of design, operationalization, trial and validation. In the first phase, design, one of the primary issues of concern, was to specify the evaluative criteria. These criteria indicate the features of writing construct that should be considered in scoring the scripts. Fig.1 presents the procedure in which the evaluative criteria were determined.

As the figure shows, the band descriptors were determined in three different ways. On the one hand, a review of the available scales and the related literature resulted in a list of evaluative criteria. This list was converted into a checklist which was refined quantitatively (a survey of ESL writing

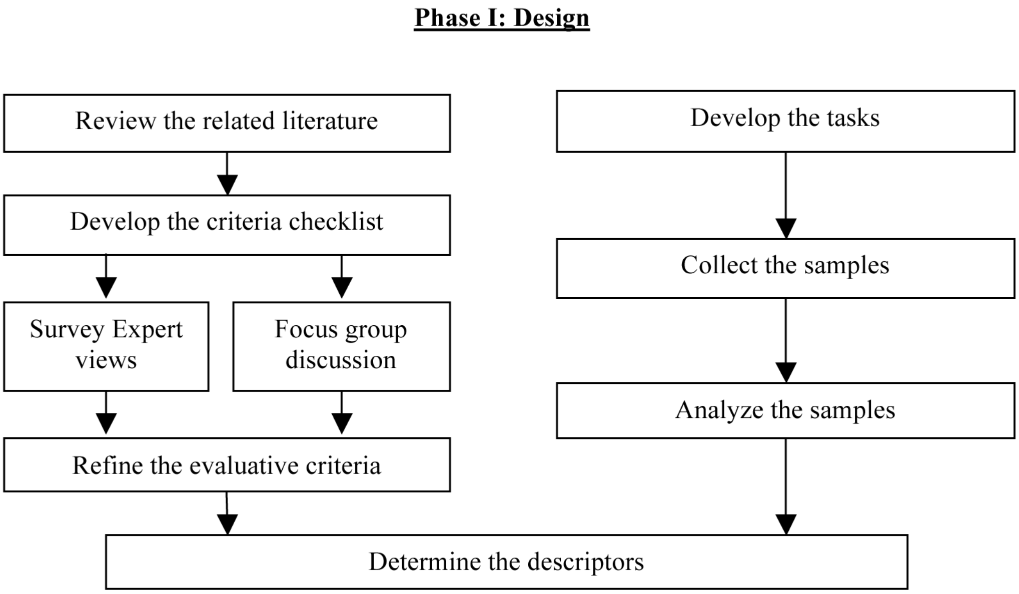

Phase I: Design
Review the related literature
Develop the criteria checklist
Survey Expert views
Focus group discussion
Refine the evaluative criteria
Develop the tasks
Collect the samples
Analyze the samples
Determine the descriptors

Fig.1: Research Procedure (Design Phase)

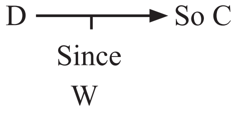
instructors' views) and qualitatively (a focus group discussion). A review of the methods and findings of these studies would be beyond the scope of the present paper. However, for an extensive discussion on the development of the checklist based on the literature review as well as the procedure and findings of the survey the reader may refer to Nimehchisalem and Mukundan (2011). Additionally, Nimehchisalem (2010, pp. 167-175) provides the procedure of the focus group study and the modifications that it caused in the final scale.

On the other hand, four argumentative tasks (Appendix A) were designed to collect samples of argumentative essays. The analysis of a number of these samples enables the researchers to describe the varying levels of writing performance in terms of the dimensions of writing construct. This paper involves the analysis of one of these dimensions, i.e., the 'content' that was carried out using Toulmin's model of argument. Other models are also available to analyze arguments, and these include Beardsley's (1950) diagrammatic approach and Scriven's (1976) tree diagram. While Beardsley's theory deals with the argumentative relationships, Scriven's method indicates the argumentative roles that each statement plays in a given text. Even though these theories have been widely used for describing arguments, they are not practical (Johnson, 2000) and also lack the ease and precision of Toulmin's model (Yeh, 1998).

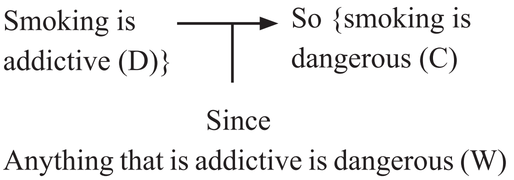
Toulmin (1958/2003) describes claim, data, warrant, qualifier, backing and rebuttal as the elements of a good argument. A claim (C) is the thesis that is being argued. It demonstrates the arguer's standpoint. The data (D) are the facts and pieces of evidence that support the claim. Example (1) shows the relationship between a claim and a datum:

Example (1)
Smoking is dangerous (C) because it is addictive (D).

In this example, the claim that is made on the danger of smoking is supported by a datum that argues smoking is addictive. A warrant (W) is a bridge between a claim and a datum. In Toulmin's words, warrants are "general, hypothetical statements, which can act as bridges, and authorise the sort of step to which our particular argument commits us" (2003, p. 91). Warrants are often implicit propositions like rules and principles that prove the legitimacy of a datum. In Example (1), the reader is convinced that smoking is dangerous because it is addictive since there is an implicit bridge between the claim and the datum that holds anything that is addictive is dangerous. Toulmin (2003, p. 92) demonstrates the relationship between the C, D and W, as follows:

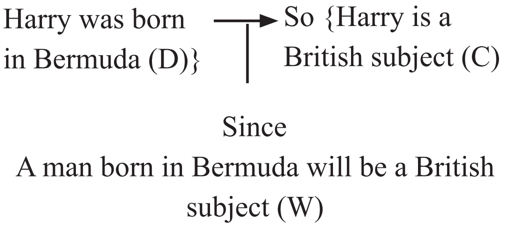


Therefore, the relationship between the argumentative elements of Example (1) can be shown in this way:


Smoking is addictive (D)}
So {smoking is dangerous (C)
Since
Anything that is addictive is dangerous (W)

When the reader has little background knowledge on the relationship between the claim and the data, the writer must explicitly state the warrant. Otherwise, the argument may sound unclear and ambiguous. In Example (2), from Toulmin (2003, p. 92), if the reader is unaware of certain rules, she will find it hard to link the datum to the claim:

Example (2)

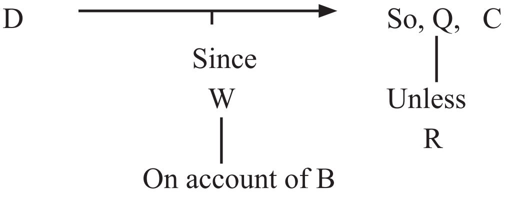

Harry was born in Bermuda (D)}
So {Harry is a British subject (C)
Since
A man born in Bermuda will be a British subject (W)

In addition to the three major elements discussed above, an argument may have three optional elements as well, including the qualifier, backing and rebuttal. Qualifiers (Q) are the cues that indicate the strength of an argument. On Toulmin's words, a qualifier shows "the degree of force which our data confer on our claim in virtue of our warrant" (2003, p. 93). As an example, definitely, in Smoking is definitely dangerous, is a qualifier. As the next element of argument, backing (B) provides further support for the warrant. One may state, "Addiction disables one's thinking" to back the warrant Anything that is addictive can be dangerous. A final element of a good argument is rebuttal (R) that shows the arguer's awareness of certain conditions in which his/her claim may not hold true. Example (3) provides a rebuttal for the claim in the first example:

Example (3)

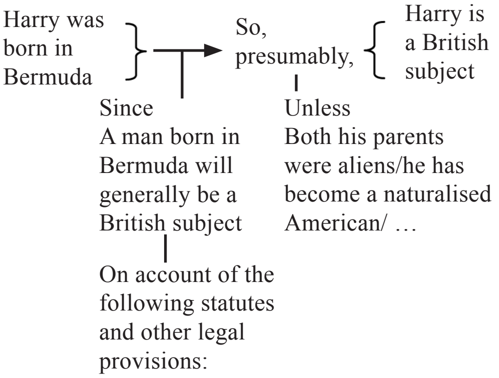
Cigarettes are dangerous (C) unless they are used for medical reasons (R).

The rebuttal "unless they are used for medical reasons" defines the exceptional cases in which cigarettes may not prove harmful. Toulmin (2003, p. 97) illustrates the distinction between the six elements, thus:



Additionally, he provides the following example to further clarify the elements in his model (Toulmin, 2003, p. 97):

Example (4)


Harry was born in Bermuda
So, presumably,
Harry is a British subject
Since
A man born in Bermuda will generally be a British subject
Unless
Both his parents were aliens/he has become a naturalised American/ …
On account of the following statutes and other legal provisions:

This brief review of Toulmin's model is followed by the method of the study described in the next section.

METHOD

The method that was followed to determine the descriptors of the subscale on content is referred to as the "databased approach" (Flucher & Davidson, 2007, p. 98). In this method, the description of samples may be carried out through a direct analysis of some written works with the help of discourse analysis (Fulcher, 1996). Wong (1989) followed the same method to develop a scale for Malaysian learners' narratives.

Samples

Since the ASAW was supposed to be used to measure argumentative writing performance of the students in Universiti Putra Malaysia (UPM), a state university in Selangor, Malaysia, the samples of the study were chosen from this university. The researchers collected samples from a variety of faculties to have access to an appropriate representation of target population's argumentative writing. They selected the participants from among male and female students from six faculties of Economy and Management, Health and Medicine, Design, Communication, Agriculture and Ecology in order to achieve a collection of samples from students with diverse levels of writing ability.

Tasks

Four different tasks (Appendix A) were developed to collect the samples. As the topics vary, so will the quality of responses elicited from the students be measurably different (Reid, 1993). There is evidence showing that inter-rater reliability may decrease if the raters have to score scripts with different topics (Weir, 1993). This suggests that in scale development establishing the descriptors on the samples that have been collected based on a variety of topics can contribute to the reliability of the scale.

In addition, four different tasks (Appendix A) were used to ensure that all the four types of argument, identified by Reid (2000), had been taken into account. The topic of the first task concerned an argument of a policy. The second topic involved a combination of an argument of fact and an argument of solution while the third was an argument of value. Finally, the last topic would prompt an argument about cause and effect.

Procedure

The researchers read all the collected samples and then impressionistically categorized them into five different levels of performance regarding their content. Next, four different samples were purposely selected for each level of performance. This resulted in a total of 20 samples which were analyzed for their argumentative elements. The sample size was equal to that of a similar study, in which Wong (1989) aimed at developing an analytic narrative scale. The content of the selected samples was analyzed using Toulmin's (2003) model. The analysts were two postgraduate students of applied linguistics in UPM. They were familiar with the model, but they were still briefed on it and given examples. The two analysts were expected to describe the patterns of use of each element of argument in the samples. Following the briefing session, they analyzed the same batch of samples independently. After the results had been collected, they were cross-checked for inconsistencies. Whenever the descriptions presented by the two analysts did not match, the elements were analyzed once more by the researchers to ensure reliability. The next section shares the results of the analysis of a number of samples based on the model that was described above.

RESULTS

The results of the analysis indicated five levels of performance, namely, <u>very limited</u>, <u>basic</u>, <u>modest</u>, <u>competent</u> and <u>excellent</u>. This is consistent with the researchers' impressionistic categorization of the samples and the five categories in the grading system of UPM. This section presents the results of the analysis of the samples which were accompanied by some examples for each level of performance. What should be stated at this point as a word of caution is that by the term 'mature' arguments, frequently used in this section, the authors simply mean arguments that are 'well-developed' or 'well-elaborated'.

Very limited samples

The first sample, which was reviewed as an example for a 'very limited' or 'F' paper, was written in response to task three that prompted comparing the advantages and disadvantages of three mass media and stating which one can be considered the most effective.

Example (5)

> Communication is very important for us even we are at school, university or at work **(C)**. their are many type we can communicate like picture, music, grafic and animation **(C)**. Right now many media are use for communicate like Internet, television, radio, book or newspaper **(C)**. The most effective media for communication is television, radio and film **(C)**. (112E232)

As both analysts agreed, the writer of this piece merely makes a number of claims. The sample lacks any kind of supporting data provided for these claims. At times, irrelevant claims are evident. As an example, in the last sentence, the writer mentions television, radio and films as the most effective media. However, according to the task, test takers were supposed to mention only one medium as the most effective. The next sample whose content will be reviewed here was written in response to the fourth task that was in relation with the most suitable age for children to start school:

Example (6)

> I'm not experience about that topic but I have more own knowledge

> about the advantages of attending school from a young age **(C)**. For me, I believe that it is important for young children to go to school as soon as possible **(C)** because that young children can be know ability since from a young age and at the same time that young children more easy become something done incorrectly as first behavior and new knowledge that they can be put in themselves it easy for future. Main purpose why, are see young children… (143E241)

The sample began with two claims. Then, it continued with a run-on sentence that was in fact a random collection of words with no meaning. Words like 'because' and 'since' in lines 3 and 4 signal the presence of some data whose meaning is entirely blurred due to the writer's lack of language knowledge. The results of analysis of a few samples at this level led the researchers to this description: 'A very limited sample only makes a number of claims, some of which may be irrelevant.'

Basic samples

The example that had been selected to present the patterns of argumentative elements at a basic level was written in response to the first task that concerned allowing an equal number of boys and girls to have higher education. The sample is presented below, along with its elements of argumentation:

Example (7)

> Nowadays, the number of females is more than male **(C)**. About the questions, me as a student totally agree with that **(C)**. In my oppinient, when the numbers of male and female students in every subject is equal, relation among the student will improve **(D)**. It is because, when they get the assignment, they can discuss and make a mix group **(B)**. Other than that, we knew that our country has so many races which is Malay, Chinese, Indian, Iban and others **(D)**. With this environment we as a student have to make friends with student from any kind of races **(B)**. This will improve the spirit of Malaysian **(W)**. (21C712)

The writer of this sample began with a claim that could be linked to the topic; however, it was left isolated from the rest of the argument. Following this irrelevant claim, the standpoint has clearly been stated in the second sentence. Next, a datum was suggested with a rather far-fetched implicit warrant, which holds, 'If there is an even distribution of males and females in a group, the chances for a good relationship between them will rise.' The next sentence has been written to back this warrant, but it cannot be accepted as a strong piece of support. The fifth sentence is an attempt to present an additional foundation in support of the claim made in the second sentence. However, this reason is quite irrelevant to the claim.

This divergence from the topic makes both the backing in the sixth sentence and the warrant in the last sentence irrelevant. The first claim as well as the datum, backing and warrant at the end of the paragraph have been underlined in order to illustrate the irrelevant arguments of this sample. Observing the results of the analysis of the argumentative content of the samples at this level of performance, the researchers defined it in this way: 'A basic sample presents claims, data, warrants and backings, some of which may be irrelevant.'

Modest samples

The sample that has been selected to be presented as a model of a modest paper was written in response to the third task. The sample is presented below. The elements of argumentation have been indicated in brackets in front of each element:

Example (8)

> [Paragraph 1]
>
> Media play an important role in our daily life **(C)**. We can know many information through the media **(D)**. Nowadays types of media increase because of technology **(C)**. We can gain information through many ways such as comics, books, radio, television, film, theatre and so on **(C)**.
>
> [Paragraph 2]
>
> Books are one type of media that very useful for us **(C)**. We can improve our knowledge through buy books from bookstore **(D)**. Books are quite easy to get **(D)** and book's price also cheaper than other types of media **(B)**. Through the book we can learn many things such as knowledge about science, economy, accounting and so forth **(D)**. Books also suitable for people whose not going out and can also gain many knowledge through read books at home **(D)**. The disadvantages of books are many people will feel boring when reading books **(C)**. I also didn't like to read book which contain many words or uninteresting topic **(D)**. Moreover, some people maybe cannot understand the meanings of word in book and this lead communication become ineffective **(D)**.
>
> [Paragraph 5]
>
> In these three type of media, I feel that television is type of media that is most effective **(C)**. It is because we can watch the picture, and listen to the voice of people **(D)**. People whose are maybe not study also see the picture and know a bit what happen shows on television **(B)**. Furthermore, people whose blind can use their's ears to listen **(B)**.

> [Paragraph 6, Final]
> In conclusion, media play important roles in our lives and we cannot without them **(C)**. (105E433)

This sample made a claim about the media in general, followed by a datum that supported it. Then, two more claims made are linked to the topic. In the second paragraph, the writer listed four data in support of a claim on the advantages of books. Next, another claim is made on the disadvantages of books followed by two data in its support. The implied warrant of the second datum is backed.

Apart from backing an implicit warrant, the writer makes no other effort to elaborate on the data about the advantages of books. Additionally, as Connor and Lauer (1988, p.145) note, "everyone-knows" kind of data are evident in this sample, like the third datum that states books enable individuals to acquire knowledge. This can be the advantage of any other media, in fact. Likewise, the fourth datum lacks maturity since people can also learn from television or most other types of media without having to leave home. Therefore, even though the data are relevant to the topic, they do not sound mature. In addition, the writer only touches upon the advantages by listing a few data without elaborating on them.

In the fifth paragraph, a claim was made presenting television as the most effective medium. Two reasons were mentioned and a warrant bridged the first datum to the claim. The second reason, however, sounded out of place. It seemed more relevant for radio that sounds a more appropriate medium for the blind. Finally, the essay ended with a claim on the importance of media in the last paragraph.

As this sample demonstrates, the argumentative content of the samples at this level is relevant but superficial and unelaborated. In other words, the data are outlined rather than being developed. Thus, with regard to its content, a modest paper may be defined in this way: 'A modest sample presents relevant claims and data, but the data sound immature and are not well-elaborated.'

Competent samples

The samples at this level showed a relatively more mature use of elements of argumentation. This can be observed by a review of the following paragraphs. They have been selected from a sample written in response to Task 3:

Example (9)

> [Paragraph 1]
> Since the world progresses fast especially in information technology **(D)**, people now have more and more options in getting the information **(C)**. Selecting suitable media to communicate information now also becomes very important in business and many other fields **(C)** because number of media increase everytime **(D)**. Each media has its own advantages and disadvantages **(D)**, so we should choose the suitable one carefully **(C)**.

[Paragraph 2]

The most famous media, in my opinion **(Q)**, is television **(C)**. Each family should have at least a television today **(D)**, and now the price of a television has dropped sharply **(D)**. We can easily communicate our companies information, for example, to public and audience through television **(D)**. The effect is good **(C)** because television not only provides the sound **(D)**, it also provides the visual that we can see and observe **(D)**. So we can remember the information in longer time **(W)**. However, television will also transfer or send the wrong information, or information that is opposite with our cultural values to society **(C)**, causing the teenagers to learn the wrong messages like smoking **(D)**.

[Paragraph 6, Final]

In a nutshell, there are too many media for us to choose to communicate the information **(C)**. We should choose the most suitable media to transfer the information **(C)** and we also must accept the message and information carefully **(C)**. (107E434)

The sample started with three data and claimed one after the other in the first paragraph. In the second paragraph, the first claim came with a qualifier and was followed by three data supporting it. The claim was repeated before another datum was added on the advantage of television. This datum is bridged to the claim with a warrant. The paragraph ended with a claim on the disadvantage of TV that is backed by an unelaborated datum. The final paragraph is a conclusion of claims that have been put forth rather hastily with no data to support them.

In comparison with the modest sample reviewed in the previous example, this essay presents a more mature development of ideas. Each claim has been supported by a minimum of one datum. Warrants have also been employed to tie the claims to the data that are relevant to the topic. The paper presents a competent model of argumentative content. However, it lacks the maturity of an excellent model. Some of the data sound disconnected from their respective claims. For example, the third datum of the second paragraph takes it for granted that the reader will understand how it is possible to "easily communicate our companies information, for example, to public and audience through television." The reader here expects to know how television may make this happen. Based on the results of the analysis of this and similar samples, a 'competent' or 'B' level sample may be defined thus: 'A competent sample presents a reasonably mature and extensive account of relevant claims and data, but at times lacks adequate backing.'

Excellent samples

'Excellent', or 'A' level, samples demonstrated the most mature and elaborated arguments. A few paragraphs of two excellent samples are presented in this section. The first sample addresses the questions whether children these days have too much free time, whether they should be given more school work and how they should spend their free time (Task 2):

Example (10)

> [1st Paragraph]
>
> The word, *children*, reminds me of innocence, happy moments and big dreams **(D)**. Childhood was by all means **(Q)**, the best time in my life and I believe, in many people's lives **(B)**. And childhood memories are sweet and filled with laughter and fun without worries **(B)**. Thus, I feel that children should not just use their free time to do only school work **(C)**. Their time should be filled by more meaningful and memorable activities **(C)**.
>
> [2nd Paragraph]
>
> Of course, unlike adults, children do not need to worry about their career, money and means to support themselves **(D)**. Hence, they have abundant time **(C)**. So, the question is how should they spend their free time? As all of us know, knowledge does not only come from school work **(D)**. In fact, reading other books like encyclopedias and story books could enhance the children's general knowledge and creativity **(B)**. For this part, the parents play an important role **(C)**. They have to make these reading materials available to the children **(D)**.
>
> [Final Paragraph]
>
> I think **(Q)**, it's best for children not to spend their time only on school work **(C)**. As it will only limit their creativity and options **(D)**. Thus, children should be free to explore the world **(C)**, but of course, under the watchful eyes of their guiding parents **(R)**. After all, we cannot ever be children again **(D)**. (54M625)

The first paragraph began with a datum that made reference to an implicit warrant, that is, 'Any experience filled with innocence, joy and big dreams deserves to be cherished.' This warrant was backed by the next two ideas that preceded two claims. The two claims addressed the topic. In addition, they clearly indicate the position that the writer has chosen to support.

In the second paragraph, a datum preceded the claim that was followed by another datum. This datum conveyed a warrant that argued, 'Children can acquire knowledge from sources other than school.' The next sentence backs this implied warrant. This backing is skilfully linked to another claim that is related to the main claim and is followed by another datum.

Finally, in the conclusion, the two claims were restated. Each of these claims was accompanied by a relevant datum. In addition, the writer had effectively responded to the probable objections by including a rebuttal in the last paragraph. In this sample, the qualifiers had also been used skilfully, in the first paragraph in order to emphasize the warrant and in the last paragraph to mitigate the claim. As it can be noticed, this sample presents an in-depth and effective account of all elements of argumentation in Toulmin's model. Setting off with the data and backings and only then pointing out the claims have contributed to the lucid flow of the argument in the sample. Another sample is reviewed below that compares the advantages and disadvantages of three mass media, stating which one was the most effective:

Example (11)

> [Paragraph 2]
>
> The first media that I would pick for comparison of advantages and disadvantages is comic. As we know, comic are very popular among teenagers **(C)**, thus the information that we want to send to teenagers can somewhat reaches teenagers faster than other channels **(D)**. In addition, the availability of comic is very high **(D)** because we can see comics are being sold in many places such as book stores, roadside newspaper stalls, convenience stores and so on **(B)**. The disadvantages of comics include the exaggerat of information and the price of comics **(C)**. Many publishers do exaggerate the information in comics in order to get people to buy the comics they publish **(D)**. The price of comics is getting higher **(C)**. People might abandon this media after realizing the price is getting higher **(D)**.
>
> [Paragraph 4]
>
> Last but not least, the media that I would choose to compare is television **(C)**. I think television is the most effective tool to communicate information **(C)** because nowadays almost every family has a television in their home **(D)**. After working whole day long most people would spend their leisure time watching television comfortably in their home to relax their tired brain **(D)**. The rich variety of channels and popularity of satellite television also increase the effectiveness of television for being a information commuting tool **(D)**. Some more, television stimulates both the hearing pleasure and is able to visualize the information conveyed **(D)**. The only disadvantage I could think off for television being the effective communication tool **(C)** is the price of sending information via television **(D)**.

> [Paragraph 5, Final]
>
> All in all, I think **(Q)** television is the best media to communicate information **(C)** because of its high popularity **(D)**, availability **(D)** and attractiveness **(D)**. Although people usually label television as an idiot box **(R)**, when it comes to the aspect of communicating message **(D)**, it is no longer an idiot box **(C)**, but a mighty box **(C)**. Thus, I would consider television as the best media to communicate information **(C)**. (82E435)

The first claim in the second paragraph was followed by two data that were bridged to the claim with the implicit warrants 'Availability is a merit.' The writer chose to provide further backing for the warrant. Next, the disadvantages were discussed by two claims, each of which was followed by a separate datum. Like the second paragraph, paragraph 4 also started by taking a position by putting forth the claims that were supported extensively by four data. However, this has left little time for the writer to elaborate more on the disadvantage of television. Indeed, the claim for the disadvantage was made in a hurry. The final paragraph restated the claim with a qualifier, followed by a summary of the data. A rebuttal preceded a datum and the three claims that concluded the essay.

This sample presents a rather extensive account of merits and demerits of television with effective reasoning. Despite its rather immature claim and its datum in the discussion on the disadvantage of TV, it sets an example of another excellent paper. However, unlike example 10, the writer of this sample chooses to present a good deal of data rather than elaborating on and backing each premise. An excellent sample may be defined as, 'An excellent sample effectively introduces the claim(s), maturely provides an in-depth or extensive account of relevant data in support of the claim(s), backs the

TABLE 1
Content subscale of ASAW

Level	Description	Grade
Excellent	An excellent sample effectively introduces the claim(s), maturely provides an in-depth or extensive account of relevant data in support of the claim(s), backs the warrants, accounts for rebuttals and may employ qualifiers.	A
Competent	A competent sample presents a reasonably mature and extensive account of relevant claims and data, but at times lacks adequate backing.	B
Modest	A modest sample presents relevant claims and data, but the data sound immature and are not well-elaborated.	C
Basic	A basic sample presents claims, data, warrants and backings, some of which may be irrelevant.	D
Very limited	A very limited sample only makes a number of claims, some of which may be irrelevant.	F

warrants, accounts for rebuttals and may employ qualifiers.' Table 1 summarizes the results discussed so far.

Once all the descriptions were ready, they were compared with the findings of the focus group discussion (Fig.1). The participants of the focus group and respondents in the survey had unanimously defined 'content' in terms of 'relevance', 'development of ideas', 'maturity of ideas' and 'consistency of stance'. The results of the analysis of the samples based on Toulmin's model covered all but one of these features, that is, 'consistency of stance'. As it was observed in the samples analyzed for their content, all the writers (either of very limited or of excellent samples) invariably took a consistent position. The sub-trait would not, therefore, differentiate between various levels of performance among the samples of this study. Therefore, 'consistency of stance' was discarded from the final version of the descriptors in the content subscale. Thus, the content subscale of ASAW was developed. Appendix (B) indicates the descriptors that differentiate between the five various levels of argumentative writing content from 'Excellent' to 'Very limited'. The descriptors of this subscale draw the rater's attention to the way the student writer employed the argumentative components in Toulmin's model. The descriptors indicate as students become less competent, they employ fewer elements of argument.

Reliability and Validity Test Results

A group of university lecturers (n =5) were trained on the subscale to score a batch of argumentative essays (n =110). They had a minimum experience of 12 years of rating. Meanwhile, SPSS (version 14) was used to analyze the data. Pearson correlation was used to test the inter-rater reliability between the scores, the results of which are presented in Table (2).

According to the table, the reliability coefficients ranged between .71 and .82. A correlation coefficient of below .50 is generally regarded as low, .50 to .75 as moderate and .75 to .90 as high (Farhadi, Jafarpur, & Birjandi, 2001). Based on these criteria, the scores assigned by the raters using the content subscale indicated moderate or high reliability coefficients.

Additionally, the subscale was tested for its intra-rater reliability. For this purpose, 50 samples from among the same batch of 110 samples were scored by the first rater after a time interval of six weeks. The scores that the rater had assigned were tested for correlation with the scores she had previously given to the same samples and a

TABLE 2
Inter-rater reliability results

Raters	1 & 2	1 & 3	1 & 4	1 &5	2 & 3	2 &4	2 & 5	3 &4	3 & 5	4 & 5
Pearson Correlation Coefficients	.786	.706	.802	.710	.811	.783	.818	.767	.797	.733

TABLE 3
Concurrent validity test results

	Correlation coefficient	Significant value
Content scores and MUET bands	$\rho_S = .79$	$r_S = .000$
Content and Argumentative Quality Scale scores	$\rho = .62$	sig-r =.000
Content and ESL Composition Profile content subscale scores	$\rho = .74$	sig-r =.000

high intra-rater reliability coefficient of .85 was achieved.

Finally, the scores that had been assigned using the content subscale were tested for their concurrent validity. For this purpose, the scores of the same samples, which had been marked using the content of the ASAW, were tested for correlation with the same students' Malaysian University English Test (MUET) band scores. MUET is recognized as a well-established high-stakes testing system in Malaysia and its bands indicate students' general proficiency in English. The scores were also tested for any correlation with the scores assigned to them using two other well-established writing scales (Argumentative Quality Scale by Conner & Lauer, 1988, as well as English as a Second Language (ESL) Composition Profile by Jacobs *et al.*, 1981). Table (3) summarizes the results of these correlation tests.

A value of "sixty or above provides strong empirical support for the concurrent validity" (Jacobs *et al.*, 1981, pp. 74-75). Therefore, the students' MUET bands as well as the scores assigned to their samples using the two other scales strongly support the validity of the scores assigned using the developed subscale. As indicated in the table, the significant values for all three tests of significance are (r_S/sig-r = .000), which are smaller than the level of significance at (α =.05); therefore, there is a significant relationship between the scores produced by the content subscale of ASAW and the other instruments.

DISCUSSION AND CONCLUSION

This paper started with a brief overview of the first phase, or design, of a writing scale called ASAW. The focus was on the part of the phase that dealt with the analysis of a number of samples. They were selected to cover a variety of levels of writing performance of the target population. The results of the analysis based on Toulmin's model led to descriptions that differentiated between the argumentative content in five levels of performance. The analysis of the samples was an essential step in developing the content subscale. Establishing the descriptors on such an analysis would contribute to the empirical value of the scale (Fulcher, 2003). It helped the researchers to formulate and classify the distinguishing qualities of the successful and unsuccessful essays. Furthermore, it helped them make the aspects of writing skill, which were emphasized in the subscale, relevant to their testing situation. It also offered a way to detect the range of writing ability levels of the target population.

It may be argued that it would be perfectly possible for a student writer to follow Toulmin's model but produce a dull and unconvincing argument. Such a narrow view toward content in this scale would, however, be accounted for with the descriptors of the final subscale of ASAW, that is overall effectiveness, which considers broader and crucial features like the following:

- How is the argument presented and justified?
- Is the style engaging, correct, clear, appropriate and/or ornate?
- Is the task fulfilled?
- Is the word limit considered?

The findings of an analysis of this kind can show ESL writing teachers the areas of argumentative writing skill that can be emphasized in writing courses to improve learners' writing performance. Likewise, L2 student writers can also benefit from such findings. In addition, presenting the examples of this study to students and having them analyze the elements of their own argumentative essays can also provide them with invaluable benchmarks of successful writing in this genre (Campbell, 1998). This issue is of primary importance since research shows that L2 learners are frequently unaware of the criteria according to which their written works are scored (Mukundan & Ahour, 2009). There is evidence when learners are unaware of the evaluative criteria in writing tests, their test anxiety rises, which in turn lowers their motivation, and in extreme cases, can discourage some learners from completing or continuing their studies (Brennan *et al.*, 2001).

A related discussion in this respect is that today criterion-referenced tests are preferable to norm-referenced tests (Weir, 2005). In criterion-referenced approach, students' performance is assessed using a well-defined set of criteria and test objectives whereas in norm-referenced approach their performance is measured in comparison to other students' (Brown, 1996). A writing scale like ASAW promotes criterion-referenced approach to testing language.

A final point worth mentioning is that although in this study the descriptors were developed with a picture of Malaysian university students in mind, they can be used for evaluating argumentative essays in other testing situations. The findings shared in this paper are by no means local and can prove helpful for writing instructors whose learners are cognitively ready to analyze the elements of argument. Most language learners do not regard assessment as an educational tool; rather, they perceive it as a tricky guessing game of their teachers' expectations (McLaughlin & Simpson, 2004). The descriptors in Table 1 offer a useful tool that can unveil what writing instructors want from their students in ESL writing courses. Formatively, they can aid writing instructors, anywhere in the world, diagnose their learners' areas of strength and weakness in developing the content of argumentative writing.

REFERENCES

Beardsley, M. C. (1950). *Practical logic*. Englewood Cliffs. NJ: Prentice-Hall.

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.

Brown, J. D. (1996). *Testing in language programs*. NJ: Prentice Hall Regents.

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

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 Purves, A. C. (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 206-227). Newbury Park, CA: Sage.

Cooper, C. R. (1999). What we know about genres, and how it can help us assign and evaluate writing. In Cooper, C. R., & Odell, L. (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 Cooper, C. R., & Odell, L. (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).

Farhadi, H., Jafarpur, A., & Birjandi, P. (2001). *Testing language skills: From theory to practice.* Tehran, Iran: SAMT.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (2003). *Testing second language speaking.* London: Longman/Pearson Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* London: Rutledge.

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.

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

Johnson, R. H. (2000).*Manifest rationality: A pragmatic theory of argument.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

McLaughlin, P., & Simpson, N. (2004). Peer assessment in first year university: How the students feel. *Studies in Educational Evaluation, 30*(2), 135-49.

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

Nimehchisalem, V. (2010). *Developing an analytic scale for evaluating argumentative writing of students in a Malaysian public university.* (Unpublished PhD Thesis dissertation). Universiti Putra Malaysia, Serdang, Malaysia.

Nimehchisalem, V., & Mukundan, J. (2011). Determining the evaluative criteria of an argumentative writing scale. *English Language Teaching, 4*(1), 58-69.

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

Reid, S. (2000). *The Prentice Hall guide for college writers.* New Jersey: Prentice Hall.

Scriven, M. (1976). *Reasoning*. New York: McGraw-Hill.

Strong, W. (1999). Coaching writing development: Syntax revisited, options explored. In Cooper, C. R., & Odell, L. (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).

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

Toulmin, S. (2003). *The uses of argument.* Cambridge: Cambridge University Press.

Weigle, S. C. (2006). Investing in assessment: Designing tests to promote positive washback. In Matsuda, P. K., Ortmeier-Hooper, C., & You, X. (Eds.), *The politics of second language writing: In search of the promised land* (pp. 222-244). Indiana: Parlor Press.

Weir, C. J. (1993). *Understanding and developing language tests*. Hampshire, UK: Prentice Hall International.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Hampshire, UK: Palgrave MacMillan.

Wong, H. (1989). *The development of a qualitative writing scale*. UKM: University Kebangsaan Malaysia.

Yeh, S. S. (1998). Validation of a scheme for assessing argumentative writing of middle school students. *Assessing Writing, 5*(1), 123-150.

APPENDIX

Tasks

Task 1
The editor of an entertainment magazine asked you to write an article for the next month's issue about the most suitable ratio of boy to girl students in Malaysian universities.

Do you think it is fair to admit only boys to universities? Or, do you think girls should also have an equal chance of higher education? Why?

The readers of this magazine are young people and adults. Since you are a busy person, you decide to spend only one hour to write a paper that is about 300 words or more.

Task 2
The editor of an entertainment magazine asked you to write an article for the next month's issue about school children's free time.

Do you think children should only play in their free time after school? What is best for them to do in their free time? Why?

The readers of this magazine are young people and adults. Since you are a busy person, you decide to spend only one hour to write a paper that is about 300 words or more.

Task 3
The editor of an entertainment magazine asked you to write an article for the next month's issue about mass media like TV, magazines, books, … .

You decide to choose three mass media and tell your readers what is good or bad about each. Then, you conclude which is the best means of sharing information.

The readers of this magazine are young people and adults. Since you are a busy person, you decide to spend only one hour to write a paper that is about 300 words or more.

Task 4
The editor of an entertainment magazine asked you to write an article for the next month's issue. He asked you to write about the best age for kids to start school. Do you think children should start school only after they are 7 years old or when they are younger? Why?

The readers of this magazine are young people and adults. Since you are a busy person, you decide to spend only one hour to write a paper that is about 300 words or more.