Advances in Language and Literary Studies
Vol. 1 No. 2; July 2010

Materials for Assessing the Writing Skill

Vahid Nimehchisalem

doi:10.7575/aiac.alls.v.1n.2p.233

Abstract

This paper reviews the issues of concern in writing scale development in English as Second Language (ESL) settings with an intention to provide a useful guide for researchers or writing teachers who wish to develop or adapt valid, reliable and efficient writing scales considering their present assessment situations. With a brief discussion on the rationale behind writing scales, the author considers the process of scale development by breaking it into three phases of design, operationalization and administration. The issues discussed in the first phase include analyzing the samples, deciding on the type of scale and ensuring the validity of its design. Phase two encompasses setting the scale criteria, operationalization of definitions, setting a numerical value, assigning an appropriate weight for each trait, accounting for validity and reliability. The final phase comprises recommendations on how a writing scale should be used.

Introduction

Materials developed for assessing writing have undergone a good deal of change. During the dominance of product-based approach timed compositions were common. These direct writing tests were often scored impressionistically which brought about readers' idiosyncratic evaluations thereby lowering reliability of the

resulting scores (Cooper & Odell, 1999). Giving the primary emphasis to speaking, Audio-lingual method, on the other hand, generated indirect multiple-choice items which were more reliable and objective (Crusan, 2002), and yet neglected global writing skills (Attali & Powers, 2008), leading to validity and authenticity problems (Hamp-Lyons, 2001; Bachman & Palmer, 1996; White, 1994) and efficiency problems (Williamson, 1994). Therefore, once again interest revived in a type of essay tests in which the problem of subjective assessment of the written samples could to be solved. These materials with their explicit evaluative criteria, also known as rubrics or range-finders, would allow the rater to avoid implicit or impressionistic rating methods which often lack consistency and lead to unfair results. Writing scales, however, should not be designed and applied in an ad hoc manner; otherwise, they will lack reliability, validity and/or efficiency.

This paper considers the areas of concern in writing scale development. Following the framework of test development (Bachman & Palmer, 1996), it divides the process into three phases of design, operationalization and administration. Then, it discusses the issues that may rise in each of these phases and the important points that developers should regard in developing writing scales.

Phase One: Design

The area of writing evaluation is heavily researched making it easy to experience a good deal of confusion after one has reviewed the literature. Therefore, it is necessary that one avoid certain starting points that can turn to tar pits. For instance, setting off to design a scale by mapping the dimensions of the writing

construct through a review of the literature and the available scales may result in a scale that attempts too much. The final scale may irrelevant for one's present assessment situation even though its criteria that are considered essential parts of writing construct but that may be. For example, consider how variations in the age group or level of proficiency of the student writers can lead to very different written products that without any doubt will call for scales that ought to emphasize or eliminate certain aspects of writing. This section discusses the preliminary issues in scale development.

Analyzing the target samples

The primary issue about which any developer ought to be concerned is what sort of written works of what group of learners are going to be assessed using the designed scale. As it is the common practice among scale designers, the best starting point would be the target population's written samples. Odell (1981, p. 119) suggests categorizing "the errors in our students' work and limit our evaluation to those types of errors … that seem most important for the students we are concerned with."

Developing a scale by an analysis of the target samples may result in a list of descriptions that eventually can be used as the descriptors of the writing scale. This method of scale development is called a "databased approach" (Fulcher & Davidson, 2007, p. 98). Scales that are derived from the analysis of samples are more valuable since they are "empirically derived" instruments based on boundary definitions (Fulcher, 2003, p. 104). By analyzing the written works,

developers can formulate and classify the qualities that can allow the reader to differentiate between the successful and less successful essays.

Determining the type of scale

It is possible to classify writing scales into generic, task-specific and genre-specific scales. Any of these scale types may be either holistic or analytic. Holistic scales follow an almost general impression scoring procedure. The rater reads a script and grades it based on a set of descriptors that evaluate the writing performance. Analytic scales examine a written piece in terms of separate dimensions of the writing construct, like language control, content and organization. Therefore, while the result of a holistic scale is a single grade, an analytic scale provides several scores depending on the number of its subscales.

Generic scales are designed with the presupposition that all sorts of writing are equal. Generic scales may be either holistic or analytic. The Test of English as a Foreign Language (TOEFL) Writing Scoring Guide used for marking the essays of the candidates in the Tests of Written English (TWE) section can be mentioned as an example of a generic holistic instrument (Weigle, 2002, p. 113). The Tests of English for Educational Purposes (TEEP) attribute writing scale (Weir, 1993, p. 160), on the other hand, is an example of a generic analytic instrument.

Task-specific scales are designed with the primary focus on the task. The primary concern of a scale of this type is to answer the question, "Did the writer successfully accomplish the purpose of this task?" (Applebee, 2000, p. 4).

Therefore, any variation in the task will urge the rater to adapt or completely modify a scoring guide.

Finally, a genre-specific scale is developed based on the distinctive features of the genres it examines. For example, the asTTle Writing Scoring Rubrics (Glasswell, Parr, & Aikman, 2001) are analytic genre-specific scales that include a set of six analytic genre-specific scales each developed to help school teachers in New Zealand to assess their student writers' ability to explain, argue, instruct, classify, inform and recount along with a seventh scale designed specifically for conventions, like grammar, spelling and punctuation. It is also possible to have genre-specific scales that are primary trait or multi-trait. Connor and Lauer (1988, p. 145), for instance, designed a scale to assess the argumentative quality of written pieces by examining the quality of their elements of argument based on Toulmin's (1958) model of argumentative writing.

Apart from the technical reasons for choosing the type of the scale, the final decision will depend upon practical issues like the available time, budget and experts to design the instrument, train the raters and rate the scripts, the degree to which the resulting scores are going to be important for the stake holders, the number of the scripts that need to be scored, the time limit and the like. Therefore, there may be situations in which one may find oneself compelled to go for a scale that is only the second best due to problems of practicality. Appendix (A) summarizes the most appropriate types of writing scales for various testing situations.

Validity (a priori)

Messick (1989, p. 13) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". To ensure the validity or relevance of any instrument, its developers should consider the construct domain being addressed and specify the criteria. If the criteria of the scale are informed by a large number of samples collected from the target population, the problem of assigning unfairly low scores to learners who respond taking an unusual perspective may be reduced (Odell, 1981).

At this primary stage, it is important that the scale be based on a current theory and practice of ESL writing that depends on a comprehensive review of the related literature and the existing instruments used for assessing writing. As Weir (2005) states, in order to validate a scale as designers we should heed, "The more fully we are able to describe the construct we are attempting to measure at the *a priori* stage the more meaningful might be the statistical procedures contributing to construct validation" (p.18).

Phase Two: Operationalization

Having made decisions on the type, levels and the rating criteria of the scale, the developers have to operationalize these criteria. While the focus of the first phase was on the skeleton and foundation of the scale, at this stage, the focus is on its elevation.

Evaluative criteria

The evaluative criteria show the sub-traits of the writing construct that the scale is going to consider. They should be explicitly and clearly stated to avoid ambiguities. Weir's (1983) scale, TEEP, for example includes "relevance and adequacy of content, organization, cohesion, adequacy of vocabulary for purpose, grammar, punctuation, spelling and appropriateness of language to context, function and intention and appropriateness of layout" (Weir, 1993, p. 136).

As Raimes (1983) puts it, to come up with a successful written work a writer should take several dimensions of the writing construct into consideration, including syntax (sentence structure, stylistic choices, etc.), grammar (rules for verbs, agreement, etc.), mechanics (handwriting, punctuation, etc.), organization (paragraphs, cohesion and unity), word choice (vocabulary, idiom, tone), purpose (the reason for writing), audience (the reader/s) and content (relevance, clarity, etc.). The choice of the evaluative criteria will depend upon the purpose and the specifications of the writing test. Regarding their assessment situation and informed by a relevant theory, developers decide on the evaluative criteria.

Operationalizing the definitions

Once the developers have decided which aspects of the writing skill their scale is going to cover, these aspects should be operationalized for the reason that it is roughly understandable what an evaluative criterion means. That is, the end users of the scale may know what the criterion means, but the designer should operationally define the criteria so that the prospective users are clear about the

concrete meaning of these criteria which will enable them to evaluate the papers based on the criteria.

Therefore, based on the information gained from the previous phase of design, the descriptors and the related criteria are determined, classified and defined. Descriptors may include terms such as fluent, relevant, substantive, flippant and the like that are used in writing scales to describe the quality of the learners' writing ability as indicated in their scripts. Each descriptor may be further broken down into more understandable pieces, or the criteria. Such detailed features of varying dimensions of writing skill help the rater decide on the best descriptor matching the quality of a script.

To offer an example, 'elaboration' may be conceptually defined as an in-depth and detailed expression and clarification of one's reasons behind a certain view. However, in order to be used in a scale this concept needs to be operationalized by defining the quantity and/or quality of the reasons the student writers provide to support a position and the degree to which they clarify them. In a three-point scale, for instance, one may define the term in the following way:

0) Irrelevant reasons given to support a position

1) One clarified plus some unelaborated reasons given to support a position

2) Two or more clear and detailed reasons supporting a position

Setting a numerical value

When opting for an appropriate numerical value, it is essential to note two important points. First, according to evidence available in the literature, raters have a tendency to go for the middle scores; for example, in the case of five-point scales, raters usually go for 3; therefore, we are advised to avoid odd-numbered scales (McColloy & Remsted, 1965; Sager, 1972; Wong, 1989). The second point to be considered is the number of behavioral levels. According to McColloy and Remsted (1965), who compared a four-point and a six-point scale in terms of their reliability, sensitivity and applicability, no significant difference was observed between the two scales regarding their reliability and sensitivity. The four-point scale was, however, reported to be more practical due to its ease of use.

A primary factor that determines the range of the points is the different levels of performance observed through the analysis of the target scripts. In the case of placement tests, scale developers may decide on the grading system based on the number of the courses available in the curriculum. In this respect, Bachman and Palmer (1996) state that it is advisable to include one or two levels more than the present levels observed through a study of the sample scripts. Such practice can increase the reliability of the instrument.

Assigning anchor papers

Anchor papers are the sample benchmark papers representing the varying levels ranging from the most basic to the most competent student writers. Sometimes, particularly in the case of holistic scales, the rubrics *per se* may not help the rater to make a solid decision on the level of a script. Scoring guides, therefore, should

be supplemented by anchors at each level so that the raters are able to properly interpret the guides (Attali & Powers, 2008). In such situations, anchor papers can provide raters with additional guidance ensuring them of the reliability of the score they assign to the script. In Tests of Written English (TWE), or the written section of the Tests of English as a Foreign Language (TOEFL), useful anchor papers can be found in most of the preparation course books (e.g. Phillips, 2003, pp. 339-343). These benchmarks are usually chosen from among the scripts before the rater training. Once they have been briefed on the scale, the raters are given these scripts to score following the scale and when the group has reached consensus on the level of these papers, they are photocopied and given to each rater.

Typically, scale designers determine three example essays exemplifying each score level specified by the scale, yet this number is often influenced by factors like "the size of the reading, the complexity of the scale, the number of the readers," (Weigle, 2002, p. 130) plus the importance and sensitivity of the test results. Additionally, if the raters are supposed to score scripts with varying topics, it is recommended to have separate sets of anchor papers for each topic. Need also may be felt to include some samples to illustrate problematic cases like those in which the instructions in the prompt have been copied straight away (Weigle, 2002).

When the objective is to design a scale for a specific genre, it is crucial that these sample scripts consistently and clearly represent the features of that genre.

Otherwise, there will be a mismatch between the qualities of these papers and the demands determined by the task or the scale (Beck & Jeffery, 2007).

Assigning weights

The developer may decide to assign varying weights to different traits. Wong (1989) designed a scale to assess ESL narratives. She assigned the component of language twice the weight of the other three components; that is, overall effectiveness, content and vocabulary. The reason behind such a choice was "the importance of language proficiency in a proficiency test [since] most adults know *what* to say but not *how* to say it" (Wong, 1989, p. 26). Such a justification may sound appropriate when developers are aware that a certain trait will contribute more to the student writers' success in their present situation and when they have a clear picture of and intend to account for these students' needs. In ESL Composition Profile (Jacobs et al, 1981), where different weights are assigned to different components of the scale, content is given the highest weight (30% of the total score), language use, organization and vocabulary have moderate weights (25%, 20% and 20% of the total mark respectively), while mechanics receives the lowest amount (only 5% of the total mark).

In a different project, based on a survey of university-level academic staff in the UK, Weir (1983) observed relevance and adequacy together with compositional organisation to be highly important, cohesion, referential adequacy and grammatical adequacy to be moderately important, and spelling and punctuation to be the least crucial (and probably negligible) aspects determining the quality of written works. By contrast, in a similar survey, in Malaysia, in their attempt to

determine the most significant traits in the genre-specific analytic scale they were developing, Nimehchisalem and Mukundan (2009) observed their samples of 89 ESL writing experts rated content and language use as the most significant, audience awareness, vocabulary and style as moderately significant, and finally mechanics and essay length as the least (but still important) features of argumentative writing. This last finding stands in contrast with what Attali and Burstein (2006) observed in the development and evaluation process of an automated essay scoring program called *e-rater* Version 2 where they reported essay length as the most objectively predicting factor of human holistic rating.

Further support could also be provided from the related literature and theory if a certain dimension of writing is to be weighted more heavily than others. Wong, for example, cites Morris (1954) who in Wong's words contends, "While credit must always be given for the matter of the composition, the primary aim of productive writing in a foreign language is surely to make the pupils proficient to the use of the new medium as a vehicle of expression" (Wong, 1989, p. 26). Therefore, another factor that may affect weighting the criteria of a rating scale is the degree of their importance stated in the literature. Furthermore, Tedick (2002) notes factors like the task, purpose and learners' level can determine whether designers weight certain criteria or not.

Finally, scale designers may justify their weighting based on statistical evidence. One possible way is using factor analysis to see which traits account for the highest variance in the scores and then assigning them a higher weight:

> In factor analysis, only the shared variance between variables (features in our case) is analyzed, and an attempt is made to estimate and eliminate variance due to error or variance that is unique to each variable. The derived factor weights … estimate the relative contribution of each variable to the common variance among all variables. Thus, the scale scores in this study reflect the relative importance of each feature to the underlying common factor among them. (Attali & Powers, 2008, p.6)

According to Hamp-Lyons (1991), however, an equal-weight scheme is more preferable. She recommends a holistic scale focusing on a given aspect of writing more than others as a better alternative than assigning varying weights to different components. Likewise, Attali and Powers (2008) in their attempt to evaluate their automated scale to score TOEFL essays compare the reliability of a previous writing scale (Attali & Powers, 2007) which followed an equal-weights scheme with that of their present scale that adheres to an optimal-weights scheme, in which different weights are given to different traits. They report, "the equal-weights scheme was as reliable and showed several advantages over the optimal-weights scheme, such as lower correlations with essay length … and better alignment with the factor-analysis structure of the data" (Attali & Powers, 2008, p. 18).

If the developers decide to adhere to an optimal-weights scheme, a guide should be added to clearly show the raters how to assign the score for each subscale and how to calculate the total score. If such a calculation turns out to be of a complicated procedure, it may make the scale complex and confuse the raters consequently.

Validity (a posteriori)

Validity should be regarded continuously in all steps of developing an instrument. To ensure validity at this stage, the descriptors should examine for their relevance. If they cover more concepts than they should, the instrument will suffer from construct-irrelevant variances. If they cover less than what they should, it will result in the construct underpresentation variances. These are the two common threats to construct validity (Messick, 1989). In a writing scale, therefore, underpresentation variances will occur if for example the scale designed for assessing a certain group of learners' argumentative writing ability neglects an important element like the 'content'. Therefore, in order to have valid descriptors, a scale should present a relevant picture of the writing construct (McNamara, 1996).

To avoid these risks scales and their criteria are often moderated at this stage. As Weir (1993, p. 19) also points out, "The discussion of tasks and criteria of assessment is in fact a key contribution to achieving valid and reliable testing procedures." It is also common to consult with the experts in the area to gain an understanding of certain details that have probably been neglected. One systematic way to do this is through a focus group study.

The dimensions of writing skill are so varied that they cannot be defined as a unitary construct (Purves, 1992). A scale developer should have a close and comprehensive understanding of the target students' ability levels to ensure they will all benefit from the scale. Such a goal cannot be achieved unless the developer has an insider's understanding and information of these learners'

strengths and weaknesses as writers. A picture of such a context can be achieved through focus group meetings. The group may include some teachers who are native to the present assessment situation. It may focus on the issues raised and recorded by the group leader who elicits the group's interactive responses. The duration of focus group meeting and its size rely on the scope of the assessment and the available resources:

> Focus groups typically consist of ten to twelve people. The group should be small enough that everyone can take part in the discussion, but large enough to provide diversity in perspective. Focus-group discussions usually need to last at least one hour and possibly two hours. (Ary, Jacobs & Razavieh, 2002, p.435)

In an attempt to design a genre-specific writing scale, Mukundan and Nimehchisalem (2009) held a focus group meeting of over two hours with four experienced writing lecturers. Their interactive discussion concerning the most relevant features of argumentative writing as well as weighting the scale provided insightful ideas helping them make their scale more user-friendly.

To ensure the validity of a scale depending on its importance, it may be evaluated through scoring experiments and a longitudinal study. Further validation experiments may follow through exploratory and confirmatory factor analysis of the internal structure of writing performance (Attali & Powers, 2008). Factor analysis will indicate which descriptors account for the highest variance in scores and can therefore help us to eliminate some of the items in the scale. This may help us collapse some of the related aspects of writing that compose separate sections of an analytic scale which makes it more economical and efficient.

At this stage a scale may also be tested for its criterion-related validity, which is a "quantitative and *a posteriori* concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance" (Weir, 2005, p. 35). For instance, if developers of a new generic analytic scale wish to ensure its criterion-related validity, they may decide to score the same scripts once using the new scale and again using another previously validated scale. If the two sets of resulting scores have a significant correlation, the new instrument is claimed to be a valid instrument.

Finally, the end users of the scale may be surveyed on their attitudes toward it. This will help the developers account for the consequential validity (Messick, 1989). In order for a writing scale to be consequentially valid its stakeholders should indicate their satisfaction of it and the inferences made by its scores. As an example, Nimehchisalem and Mukundan (2009) used a questionnaire to test their developed scale for its consequential validity (Appendix B).

Reliability

If a scale is reliable and has "scoring validity" (Weir, 2005, p. 22), it will help us reach almost consistent scores when we keep rating the same sample. When different raters score the same script inconsistently, the scale will lack inter-rater reliability. Furthermore, it will indicate weak intra-rater reliability once the same rater assigns two discrepant scores to the same script with a time interval. Scholars have varying ways of interpreting reliability coefficients, yet generally speaking a reliability coefficient of below .50 is regarded as low, .50 to .75 as moderate and .75 to .90 as high (Farhadi, Jafarpur & Birjandi, 2001).

In testing analytic scales, usually each subscale is evaluated for its reliability. The reason is that the reliability tests may indicate a high reliability coefficient for the total scores while one of the subscales may represent a low coefficient. In such cases, the descriptors of the subscale are reworded or (if possible) deleted to improve reliability.

Different methods can be proposed to increase the reliability of a scale. A commonly practiced and viable way is to train the raters on the scale before having them use it. This involves preparing raters through formal meetings and guidelines designed for writing scales usually in a short course of 6-10 hours. The training period may vary depending on the complexity of the scale and the degree of importance of the scores it will produce. Anchor papers are also employed as benchmarks to which raters can refer in order to avoid disagreements that may result in significantly different scores. Still another way is having a third rater score the essay on whose grade the raters do not seem to reach a consensus (Hamp-Lyons, 1990; Odell, 1981)

Using clear statements to describe the determined criteria and eliminating or rewording the ambiguous terms can also contribute to the reliability of the instrument. Another factor that raises the reliability of a scale is a higher number of subscales (Brown & Bailey, 1984). This may, however, negatively affect the economy of the scale. As the topics vary so will the responses elicited from the students be measurably different (Reid, 1990). This suggests if the scale is not going to be all-purpose, it is important to test it on a number of different topics. There is evidence that inter-rater reliability is likely to decrease if the readers at a

rating session are given scripts that have a variety of topics (Weir, 1993). Testing the scale on different topics can help developers avoid this problem.

Phase Three: Administration

It is recommended to have two raters score the same group of scripts while a third more experienced reader leads the group. She cross-checks the two sets of scores for any significant discrepancies. Breland, Bridgeman and Fowles (1999, p. 8) describe scores of "4" and "2" from two different readers on a 6-point scale as discrepant scores. This varies from one assessment situation to another, however. In some testing programs where the assigned marks are not of high importance "essay scores must be at least three points apart before they are considered discrepant" (Breland, et al, 1999, p. 24).

Thus, for instance, when in a six-point scale the first rater assigns a score of 5 while the second scores the same script 3, a third rater marks the script again. The three scores can be treated in two different ways, "All three may be averaged or only the two closest scores may be averaged" (Hamp-Lyons, 1990, p. 80). If the first and second raters score 3 and 5 and the leader scores 4, the average of the three marks; that is 4, is reported as the final score. However, if the leader scores 6, the average of the two closer scores; that is 5 and 6 (i.e., 5.5), is regarded as the final mark.

The method of scoring the written work by putting the raters together in a conference setting is referred to as "conference approach" which is different from "remote scoring" approach in which raters score the scripts individually and

independently at home or office (Breland et al., 1999, p. 8). In remote scoring raters are given the scoring guidelines and the anchor papers but independently. This makes consulting with the rating leader and other raters relatively challenging. Research findings suggest that conference approach commonly is of a higher reliability and validity (Breland & Jones, 1988). When the objective is to test a scale designed for high-stake writing tests, five or even more raters may be asked to mark the same set of scripts.

Conclusion

Research shows that in writing courses most students are left confused with covert criteria on how they can achieve the highest possible mark at the end of the course (Mukundan & Ahour, 2009). A writing scale is, of course, a mere instrument which *per se* will not be able to support learners requiring the teacher to create a connection between instruction and evaluation of writing. Scales can be converted into checklists whose wording is void of complex jargon and therefore easy for learners to grasp. Such checklists can be valuable tools for self-evaluation or self-reflection activities as well as guidelines for peer critique purposes (Ferris & Hedgecock, 2005). There is empirical evidence available on the positive effect of using checklists of this sort on the improvement of students' written works (Hillocks, 1984).

The present paper sought to review some of the issues of concern among ESL writing scale developers. Following Bachman and Palmer's (1996) model, the author offered a three-phase procedure to design a writing scale. It is of paramount importance, however, to note that the task of creating a scale should

follow a reiterative and rather heuristic process of design, operationalization and administration that may have to be repeated if its developers wish to come up with a valid and reliable instrument. One may take it as composing a poem which is by no means a fixed process. Therefore, the author's intention has been far from offering a fixed mould that can be used for any scale development purposes. Rather this has been an attempt to share certain points with English language teachers and researchers who need to select, adapt or design writing scales for their own research purposes and situations.

References

Applebee, A.N. (2000). Alternative models of writing development. [Online] Available: http://cela.albany.edu/publication/article/writing.htm. (October 7, 2009)

Ary, D., Jacobs, L.C., & Razavieh, A. (2002). *Introduction to Research in Education* (6th ed.). Belmont, CA: Wadsworth/Thomson learning.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with *e-rater* V.2. *Journal of Technology, Learning, and Assessment, 4*(3). [Online] Available: http://escholarship.bc.edu/jtla/vol4/3/ (October 3, 2008)

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

Attali, Y. & Powers, D. (2008). A developmental writing scale. ETS RR-08-19, Princeton, NJ. [Online] Available: http://www.ets.org/Media/Research/pdf/RR-08-19.pdf (August 6, 2009)

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

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

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework* (College Board Report No. 99-3; GRE Board Research Report No. 96-12R). New York: College Entrance Examination Board. [Online] Available: http://www.nocheating.org/Media/Research/pdf/RR-99-03-Breland.pdf (September 16, 2009)

Breland, H. M., & Jones, R. J. (1988). *Remote scoring of essays*. College Board Report No. 88-3 (ETS RR No. 88-s4). New York: College Entrance Examination Board.

Brown, J. D. & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21-42.

Connor, U. & Lauer, J. (1988). Cross-cultural variation in persuasive student writing. In *Writing Across Languages and Cultures: Issues in Contrastive Rhetoric*, Purves, A.C. ed., Newbury Park, CA: Sage.

Cooper, C. & Odell, L. (1999). Introduction: evaluating student writing, what can we do, and what should we do? In Cooper, C. & Odell, L. (eds.), *Evaluating Writing: The Role of Teacher's Knowledge about Text, Learning, and Culture.* Urbana, Illinois: National Council of Teachers of English (NCTE).

Crusan, D., (2002). An assessment of ESL writing placement assessment, *Assessing Writing 8*, 17–30.

Farhadi, H., Jafarpur, A. & Birjandi, P. (2001). *Testing Language Skills: From Theory to Practice*. Tehran, Iran: SAMT.

Ferris, D.R. & Hedgecock, J.S. (2005). *Teaching ESL Composition.* (2nd ed.). London: Lawrence Erlbaum Associates Publishers.

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.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In: B. Kroll (ed.), *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Pre-text: Task-related influences on the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

Hamp-Lyons, L. (2001). Fourth generation writing assessment. In: T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 117–127). Mahwah, NJ: Lawrence Erlbaum.

Hillocks, G. (1984). What works in teaching composition: a meta-analysis of experimental treatment studies. *American Journal of Education 93* (1), 133–171.

Jacobs, H.; Zingraf, S. ; Wormuth, D.; Hartfiel, V.F., & Hughey, J. (1981). ESL Composition Profile. Newbury House Publishers.

McColloy, W. & Remsted, R. (1965). Composition Rating Scale for General Merit: An Experimental Evaluation, *Journal of Educational Research, LIX* , 55-57.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement,* New York: Macmillan.

Morris, L.A. (1954). *The Art of Teaching English as a Living language*. London: MacMillan & Co. Ltd

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

Mukundan, J. & Nimehchisalem, V. (2009). Development and Evaluation of a Holistic Argumentative Writing Evaluation Instrument. Unpublished Research Report. University Putra Malaysia: Serdang.

Nimehchisalem, V. & Mukundan, J. (2009). *Development and Evaluation of an Analytic Argumentative Writing Scale*. Unpublished Research Report. University Putra Malaysia: Serdang.

Odell, L. (1981). Defining and Assessing Competence in Writing. In Cooper, C. (ed.), *The Nature and Measurement of Competency in English* (pp. 95-138). Urbana, Illinois: National Council of Teachers of English (NCTE).

Phillips, D. (2003). *Longman Preparation Course for the TOEFL Test*. New York: Pearson Education.

Purves, A. (1992). Reflection on research and assessment in written composition. *Research in the Teaching of English, 26*(1), 108–122.

Reid, M. J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In: B. Kroll (ed.), *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press.

Raimes, A. (1983). *Techniques in Teaching Writing*. Oxford: Oxford University Press.

Sager, C. (1972). Improving the quality of written composition through pupil use of rating scale. PhD Dissertation. Boston: Boston University.

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. [Online] Available: http://www.carla.umn.edu/articulation/polia/pdf_files/standards.pdf (October 20, 2009)

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

White, E.M. (1994). *Teaching and Assessing Writing, 2nd ed*. San Francisco: Jossey-Bass.

Weigle, S. C. (2002). *Assessing Writing*. 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.

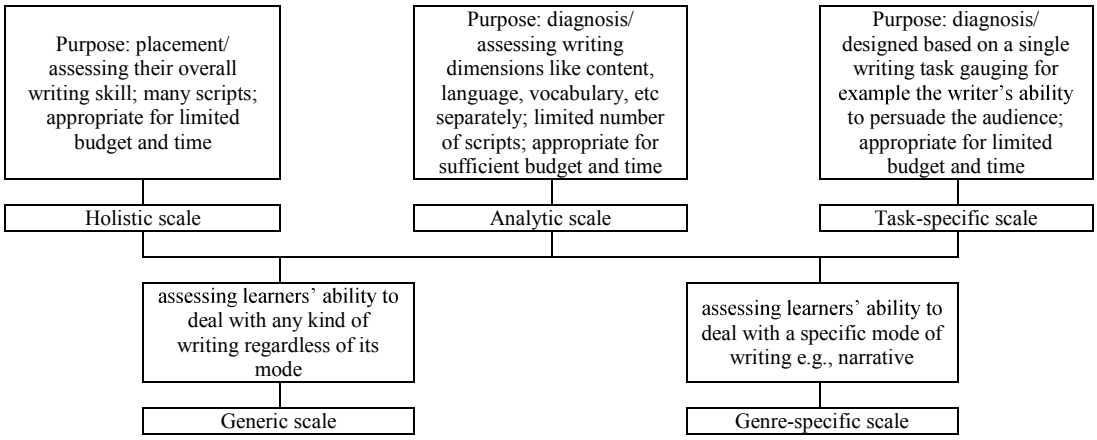
Weir, C.J. (1993). *Understanding and Developing Language Tests*. Hampshire, UK: Prentice Hall International.

Weir, C.J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Hampshire, UK: Palgrave MacMillan.

Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing*, *1*, 147–193.

Wong, H. (1989). *The Development of a Qualitative Writing Scale.* UKM: University Kebangsaan Malaysia.

Appendix A: Test purpose and scale type



Appendix B: Analytic Argumentative Writing Scale Evaluation Questionnaire

(Nimehchisalem & Mukundan, 2009)

This questionnaire seeks to evaluate the Analytic Argumentative Writing Scales based on your judgment of its quality. You, as a rater who used the scale, are kindly requested to mark the spaces in front of the statements below that best describe your evaluation of it according to the key provided below and answer questions 14-16:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Statement	1	2	3	4	5	Comments
1. I found it easy and not tiring to work with the scale.						
2. I will use this scale to correct my own students' written works.						
3. I recommend using this scale to my colleague.						
4. The scale fully covers the aspects of argumentative writing skill.						
5. The scale assesses an adequate scope of writing construct.						
6. The scores produced by the scale distinguish learners' levels.						
7. The scale helped me draw a clear line between the scripts that seemed to be of different levels.						
8. All the terms in the scale are clear and easy to understand.						
9. The sample scripts helped me get a grip of the different levels of performance.						
10. The scoring guideline is clear and leaves no concept vague.						
11. Overall the scale sounds a reliable instrument.						
12. Weighting of different aspects is fair.						
13. Overall, I am satisfied with this scale.						
Total: …… /65						
Key: 13-24 (low satisfaction), 25-42 (moderate satisfaction), 43-54 (high satisfaction), 55-65 (very high satisfaction)						