CHAPTER 3

PORTFOLIO ASSESSMENT: A FRAMEWORK PROPOSED

In this chapter, a framework for the current research study is presented. The assessment of portfolios as it is done in different second language study contexts and the theory related to the assessment of portfolios is discussed in this chapter. The criteria used for portfolio assessment and the approaches followed for designing the rating scales have been explained.

3.1 Portfolio-based assessment of writing

Reckase (1995: 13) indicates that more work has been done to apply portfolio assessment methodology in the writing area than in any other content area. This is because portfolio-based assessment of writing seems to have more advantages than a single essay test or an indirect assessment such as a traditional multiple choice test. As pointed out by Weigle (2002: 197), assessing writing by collecting and evaluating individual writing pieces as single performances, and making inferences from these performances about the writing ability of the test takers has several limitations. Two of the most serious limitations are: (1) writing done under timed conditions on an unfamiliar topic does not accurately reflect the conditions under which most writing is done in non-testing situations or writing as it is taught and practised in the classroom, and (2) it is difficult to generalize from a single writing sample to a much broader universe of writing in different genres and for different purposes and audiences. Weigle (ibid) observes that portfolio assessment, as seen by many, is an alternative approach to writing assessment that can allow broader inferences about writing ability than are possible with single-shot approaches to evaluating writing, both in the individual classroom and on a larger scale.

Elbow and Belanoff (1986 cited in Sommers, 2003: 378) argue that portfolio-based assessment of writing reflects the complexities of the writing process. Herman et al. (1993:
support this and point out that portfolios ‘have the potential to provide a more equitable and a more sensitive portrait of students’ strengths and weaknesses’. Ruetten’s (1994 cited in Song and August, 2002: 63) study and research showed that ESL students found holistically scored timed impromptu essay particularly difficult. The study showed that ESL students assessed on the basis of portfolios achieved better results than the students assessed by non-portfolio measures. Studies done by Hamp-Lyons and Condon (2000) and Song and August (2002) confirmed this finding and found portfolio-based assessment of writing to be a more appropriate assessment alternative for the ESL students. Hamp-Lyons and Condon (2000) present a convincing argument in favour of introducing portfolios in ESL contexts. According to them, portfolio provides an accurate description of the writer’s abilities and it does so ‘far more extensively than could a sample of the same writer’s performance on a one-shot essay test’ (p. 4).

Weigle (2007: 199) demonstrates that portfolio assessment is a potentially more valid way of assessing many aspects of writing than a single test. She points out that many individual teachers and writing programmes have adopted portfolio assessment as a potentially more valid approach to writing assessment. The reason, according to her, is that portfolios provide opportunities for a writer to compose texts in a variety of genres that are appropriate for their audience and purpose. Hence, portfolio-based assessment of writing seems more appropriate than a single test in an ESL context.

However, the studies on portfolio that have been done so far in teacher education contexts primarily examine teachers’ teaching competencies (e.g. Darling, 2001; Van der Schaaf et al., 2005). There are few examples of portfolios being used to assess teachers’ own language competence, especially in the area of writing. The study of Van der Schaaf et al. (2005), for example, looked at teachers’ abilities to design tasks needed to develop students’ research skills. Accordingly, the portfolio included a series of research assignments given to students,
the assessment of students’ work by the teacher, student evaluations of the teacher, and
teachers’ reflections on their strengths and weaknesses and on how to improve their teaching.

Furthermore, in a study conducted by Tillema and Smith (2007) in a pre-service teaching
education context, the portfolio consisted of teaching observations, accounts of lessons,
personal reflections, and lesson materials prepared by student teachers. However, they found
that the grade which the student received was very subjective, as there were no open and
specified criteria to assess the portfolios. Additionally, the students and assessors differed in
their perception of what was required of a portfolio primarily because of the lack of explicit
and shared assessment criteria. Tillema and Smith (2007: 453) argue that it is possible to
overcome these problems by having an open and shared criteria to assess portfolios.

3.2 Assessment of the portfolio

The process of designing rubrics/criteria for assessing portfolios is important in order to
develop a shared understanding among teachers and assessors. Underwood and Murphy
(1998: 203) argue that a shared understanding of the criteria among teachers and assessors is
essential, especially to achieve inter-rater reliability in portfolio assessment. They cite a
study conducted by LeMahieu et al. (1995 cited in Underwood and Murphy, 1998: 202) on
writing portfolios to support their argument. LeMahieu et al. (ibid) were able to achieve inter-
rater reliability, expressed as correlation coefficients, that varied from .74 to .87, for different
dimensions and grade levels examined. The success was attributed to ‘the development of a
shared understanding and interpretive framework, that the teachers in Pittsburgh had
developed as a result of extensive professional development activities and a rubric
development process’ (p. 203).

Several researchers argue that students should be involved in the process of developing the
criteria and standards by which portfolios are judged. Wade and Yarbrough (1996: 65)
indicate that students might ‘give input as to what parts of the portfolio are evaluated and which criteria are used for judging merit.’ According to Lynch and Shaw (2005: 265), one of the essential features of portfolios is that ‘[t]he students participate in deciding the criteria for evaluating the portfolios’. They (ibid) conducted a longitudinal study in an MA TESOL programme where they had developed criteria for assessing portfolios from a process of student - faculty consultation and decided that both the process and product of assessment tasks should be evaluated. Arguing along the same line, Darling (2001: 118) states that negotiating evaluative criteria helps students ‘become clearer about the broader purposes for constructing portfolios and the goods associated with them...’

Hamp-Lyons and Condon (2000: 5) draw our attention to the two factors that may have a great influence on the components of the portfolio and the process of assessing the portfolio. The two factors are the present needs of the writers producing portfolios and the concerns of teachers teaching the students and evaluating the portfolios. Hence, they argue that it is essential to involve all the stakeholders, especially the writer and the teacher, in the evaluation process. The portfolio system has this potential. Hamp-Lyons and Condon (2000: 6) point out that the strength of portfolio-based writing assessment, vis-a-vis traditional psychometric test, is that portfolio has the potential for open, shared assessment. Above all, as Paulson et al. (1991: 63) argue, a portfolio becomes a portfolio ‘when the student is a participant in, rather than the object of, assessment’.

However, in many contexts, the criteria for assessing portfolios appear to have been designed by the programme staff and evaluators. For example, in a study conducted by Song and August (2002), the evaluation criteria for portfolios were determined by the department based on discussions with instructors and a survey of literature. Hence, for them, presenting portfolio work – labelling the drafts, stapling the final draft on top of the others, typing or writing neatly, including the writer’s name, date, professor’s name, etc. – becomes one of the
important criteria. Adopting a top-down approach to portfolio assessment, as indicated by several researchers, goes against the basic tenants of portfolio theory.

As portfolios are used in many different contexts, there is a wide variation in terms of how portfolios are assembled, evaluated and used (Weigle, 2002: 198). Furthermore, as Tillema and Smith (2007: 443) report, studies done by Darling-Hammond and Snyder (2000) and Zeichner and Wray (2001) suggest that teacher educators are not open to the idea of maintaining strict criteria. This may be because of the fear that maintaining strict criteria disrupts the process of competence development in their students (Tillema and Smith, 2007: 443).

However, establishing standardised criteria is necessary for assessing portfolios. Without this, Williams (2000 cited in Song and August, 2002: 52) argues that portfolio may become unfair because it ‘increases the subjectivity teachers bring to evaluation’. Hamp-Lyons and Condon (2000: 54) reiterate this view and argue that we need to have a clearly defined assessment criteria for judging outcomes from the portfolio to measure progress along specific parameters. In addition, it is important to establish the criteria while designing curriculum. The requirements specified in the criteria will allow both learners and assessors to track the learning that has occurred and to measure the extent of progress the writer has made (ibid). Portfolio, by including the drafts of the finished products, allows us to trace the development of each piece and also to perceive ways the writer has improved over time.

The criteria used for assessing portfolios, as stated earlier, vary considerably. LeMahieu et al. (1995 cited in Johnson et al., 2000: 69), for example, used the following evaluative dimensions for assessing the writing portfolios in the Pittsburgh public school district.

- accomplishment in writing
• use of processes and strategies for writing and
growth, development and engagement as a writer

On the other hand, the dimensions used by Willard-Traub et al. (1999: 60) at the University of Michigan included:

i. attributes related to the reflective piece;

ii. attributes related to the range of tasks represented in the portfolio;

iii. attributes related to the writer’s engagement with the subject matter; and

iv. attributes related to control of grammar, mechanics, and style.

In this set, the first two categories were specific to portfolio assessment whereas the latter two were important for assessing single samples of writing. These attributes were evaluated from ‘consistently absent or low’ to ‘consistently present or high’.

Song and August (2002: 69) used a holistic scale for assessing the portfolios of ESL students. The evaluation criteria included finding and organising ideas, using the writing process, editing and presenting work. Some of the specific items included in the criteria were as follows:

• writing in depth

• establishing a focus

• writing clearly

• using drafting process effectively
• basic mechanical competence (correct verb tenses, subject-verb agreement, correct punctuation, etc.)

It is worth noting that the presentation of portfolio work was one criterion for portfolio evaluation.

The different criteria discussed, thus far, may be used in two ways: (i) to assess the portfolio as a whole, putting all the pieces together to provide a single report of the results (Black et al., 1992; cited in Reckase, 1995) or (ii) to look at the individual pieces, score them separately and sometimes in a number of ways, and then either combine the scores into a single total score (Nystrand et al., 1993 cited in Reckase, 1995) or report a profile (Koretz, et al., 1992 cited in Reckase, 1995). Birenbaum (1996 cited in Lynch and Shaw, 2005: 265) suggests that a set of criteria with descriptors and clearly distinguished levels is useful in reporting the assessment results. According to him, it is important to report the assessment results as a qualitative profile rather than a single score or other quantification.

3.3 Approaches to rating scale design

It may be useful to state the fundamental characteristics of a scale. Hudson (2005: 207) lists three important features of a rubric or scale. According to him, a scale

1. Is based on a continuum of performance quality, with a scale of varying potential score points to be assigned

2. Identifies the significant traits or dimensions to be examined and assessed

3. Provides key criteria of performance for each level of scoring, in “descriptors,” which reflect whether and to what extent the key requirements of the performance have been demonstrated.
Further, as Alderson (1991: 73) states, language scales can be created for many different functions: (a) for describing levels of performance; (b) for providing guidance for assessors who are rating the performance; and (c) for guiding test constructors with a set of specifications. The purpose of constructing a rating scale, in this study, is two-fold: to describe levels of performance and to provide guidance for assessors who are rating the portfolios.

3.3.1 Analytic and holistic scales

Different types of rating scales have been used in different contexts for scoring portfolios. Two of the more common types of rubrics used for the scoring of the portfolios are analytic and holistic scales (Johnson et al., 2000: 71). An analytic scale is used to provide separate scores on specified features of the text produced e.g. accuracy of language, relevance and adequacy of content (Shaw and Weir, 2007: 150). A holistic scale, on the other hand, is used to award one global score according to the overall properties of the text. Weigle (2007: 203) states that an analytic scale tends to be somewhat more reliable than a holistic scale. This is because of two reasons: (i) an analytic scale has the effect of focusing the raters’ judgements on the most salient features of the text; and (ii) as each candidate is awarded a number of scores, a reliable award can be derived from a set of summed, aggregated or ‘averaged’ multiple ratings (Shaw and Weir, 2007: 151).

Furthermore, analytic scales provide more useful feedback to students. The scores on different aspects of writing reveal to the students their respective strengths and weaknesses. Shaw and Weir (2007: 153) indicate that an analytic scale may be especially important in tests involving only one marker. Holistic scoring, on the other hand, is appropriate for ranking candidates and is useful for discriminating across a narrow range of assessment
bands (ibid). It is also suitable for arriving at a rapid overall rating and for large-scale assessments – multiple markings.

However, Weigle (2002: 114) argues that holistic scoring is ‘especially problematic for second language writers, since different aspects of writing ability develop at different rates for different writers’. White (2005: 581) also points to this and emphasises the need to develop a new scale to assess portfolios. He observes that most portfolio evaluation (e.g. Spalding and Cummins, 1998 and Wolcott, 1998 cited in Weigle, 2002: 222-4) currently uses some adaptation of holistic scoring. He maintains that the system developed for essay testing by the Educational Testing Service in the 1960s has been imported without much question and used to assess portfolios holistically. The problems with scoring portfolios holistically are many and they are not readily resolvable (ibid). Furthermore, many aspects of holistic scoring work against the principles behind portfolio assessment. White (2005: 583) states that ‘[w]e have from the start needed a scoring methodology that responds to and reflects the nature of portfolios, not merely an adaptation of essay scoring’. There are, however, instances of analytic scales being used to evaluate portfolios which will be discussed later in this work. There is, thus, a need to design a new scale to evaluate portfolios.

3.3.2 Intuitive and empirical methods

The literature on rating scale development reveals that there are two approaches to designing a rating scale (Fulcher, 2003: 91 - 92). They are intuitive and empirical methods. In intuitive methods, Fulcher (ibid) states that there are three sub-categories. The first one is called ‘expert judgement’ in which an experienced teacher or language tester writes a rating scale in relation to existing rating scales or a teaching syllabus. The second category is called ‘committee’ where a small group of experts discuss and agree on the wording of the descriptors and the levels of the scale. Another way, within the intuitive method, is to evolve
a rating scale by starting with expert judgment or committee design and the scale is refined by those who use it ‘so that over a period of time the users intuitively understand the meaning of the levels in relation to sample performances’ (Fulcher, 2003: 92). According to Fulcher (ibid), the third category (experiential rating scale) is by far the most common intuitive method of scale development. The Canadian Language Benchmarks Assessment (CLBA; Norton and Stewart, 1999; Pawlikowska-Smith, 2002) is an example of scales based on an intuitive developmental approach.

A different approach to developing a rating scale is to base the design upon some kind of empirical data. Fulcher (2003: 92) discusses three categories under empirical methods: (i) data-based or data-driven scale development; (ii) empirically derived, binary-choice, boundary definition scales; and (iii) scaling descriptors. The first method requires the analysis of actual learner performance on tasks, and the description of key features of performance. The number of occurrences of a particular feature (for example, the number of cohesive devices used) is used as a guide to scale construction. Fulcher (1996) designed a fluency rating scale for speaking based on the data-driven method.

In the second method, which is developed by Upshur and Turner (1995), expert judges divide the samples of learner performance into better or poorer categories. ‘The reason for the categorisation is recorded, and used to write a sequence of yes/no questions that lead the rater to the score’ (Fulcher, 1996: 92). In ‘scaling descriptors’ method, many band descriptors are collected in isolation from a scale, and experts are asked to rank them in order of difficulty. The descriptors are then sequenced to create the scale. Scales such as the Common European Framework of Reference for Languages developed by the Council of Europe (North and Schneider, 1998 cited in Fulcher, 2003: 107) and the related scales of the ALTE (Association of Language Testers in Europe, 1998), and the Dialang Project (Alderson, 2005) are examples of scales following empirical methods.
However, there seem to be few instances of studies done on the development of scales to assess portfolios. As pointed out in the previous section, in many contexts, top-down approaches, involving the programme staff and raters, have been adopted to design criteria for assessing portfolios. Hamp-Lyons and Condon (2000: 5) claim that portfolios have the potential for employing a bottom-up structure, in which the roles of all the major stakeholders are taken into account.

This study is an attempt to design a rating scale based on a bottom-up approach in which a combination of intuitive and empirical methods have been employed. Following a purely intuitive method may not be desirable in the case of portfolios as it will be top-down in its approach. On the other hand, adopting a data-driven method to scale development for portfolios may be a mammoth exercise and not feasible in many ESL contexts. Additionally, a data-based scale for assessing portfolios may lack meaningfulness and utility which, as pointed out by Henning (1992: 369), are essential requirements of any rating scale. This is primarily because of two reasons:

(i) a rating scale designed for the portfolio is not a proficiency scale as in the case of assessing speaking or assessing writing. A portfolio-based assessment of writing examines not only the characteristics of writing as well as the characteristics of portfolio.

(ii) the portfolio consists of a variety of tasks, and thus, a scale developed for the rating of one kind of task may not be appropriate for a different task (Upshur and Turner, 1995: 6).

Therefore, a different approach to designing a scale for assessing portfolios may be desirable and this study is an attempt in that direction.
3.4 Summary

In sum, portfolios vary from context to context, likewise their assessment. Portfolios are frequently used to develop as well as to assess writing skills in ESL contexts. Two types of scales may be used to assess writing portfolios: an analytic scale and a holistic scale. These scales may be designed by following different approaches such as intuitive, empirical or a combination of both. While designing rating scales for portfolios, it may be desirable and useful to involve various stakeholders so that full justice is done to portfolios.

This study involves teachers, trainers and raters in constructing a rating scale for portfolios. In the following chapter, I describe the study carried out in three phases, report views of teachers on assessing portfolios, the criteria used by expert raters to assess portfolios and the perceptions of expert raters about a common rating scale. Before addressing these issues, it is important to describe the context of the study. The next chapter gives an overview of the background in which this study is undertaken and the field work done in three different phases.