Pre-service teacher assessment practices: A gender-based analysis

By
Markus Rene Baer
B.G.S., Simon Fraser University, 1995

A thesis submitted in partial fulfillment of the requirements for the degree
of
MASTER OF ARTS
In the Department of Educational Psychology and Leadership Studies

We accept this thesis as conforming to the required standard

[redacted]

Dr. D.G. Bachor, Supervisor (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. J. O. Anderson, Departmental Member (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. L. Yore, Outside Member (Department of Curriculum and Instruction)

[redacted]

Dr. A. Preece, External Examiner (Department of Curriculum and Instruction)

Supervisor: Dr. Daniel G. Bachor

ABSTRACT

The purpose of this study was to provide further understanding of the emergent patterns of classroom assessment using fictional students' portfolios through a gender-based analysis of the practices noted among a group pre-service teachers (n=38) enrolled in the Faculty of Education at the University of Victoria. Participant diaries recorded as part of a larger investigation were analyzed across gender to determine if there were differences in reported assessment practices. Diary contents were categorized and coded for comparison using *Atlas.ti* software (Muhr, 1997). This analysis was conducted at various levels, ranging from a general examination of the diary structures to specified comparisons of the written contents. As an initial step, a t-test was conducted on the length of the diary contents from all 38 participants, with the result that the response lengths were not significantly different across gender. Following this analysis, a detailed textual comparison using *Atlas.ti* was conducted. Both male and female participants recorded evaluative comments of students or student work; however, females offered more interventions and solutions as part of their assessment discussions than did their male counterparts. A higher number of inferences classifying students as possibly having a special education need were also recorded by the female pre-service teachers. Although this is consistent with previous studies identifying complex interaction effects between gender and assessment, limitations within this investigation call attention to future research. The finding that gender does not exhibit main effects but may be influential in such areas as student intervention feedback and the types of inferences made, suggests directions for further investigation.

Examiners:

[redacted]

Dr. D.G. Bachor, Supervisor (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. J.O. Anderson, Departmental Member (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. L. Yore, Outside Member (Department of Curriculum and Instruction)

[redacted]

Dr. A. Preece, External Examiner (Department of Curriculum and Instruction)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge the assistance and support of my supervisor Dr. Dan G. Bachor for his work and dedication. In addition I would like to thank the members of my committee their contributions.

I would also like to express gratitude to my family for their unabated tolerance and support of my undertakings. Jennifer and Charlotte in particular for living with this.

DEDICATION

To the Family.

Chapter 1

INTRODUCTION TO THE PROBLEM

It has been said that the day-to-day evaluation of students is one of the most important, complex and demanding tasks that educators face (Rogers, 1999; Stiggins, Conklin & Bridgeford, 1986). Classroom assessment absorbs a substantial portion of teacher time and energy (Butterfield et al., 1999; Gullickson, 1984; Stiggins, 1990a, 1999), and has far reaching impacts for all manner of individuals who come into contact with it (Crooks, 1988). Despite its prevalence and significance, a thorough understanding of classroom assessment has not been achieved, and although advances have been made, greater knowledge regarding the details of these practices is still required (Boxall & Gilbert, 1999; Gullickson, 1985; Stiggins, 1985, 1990b; Wilson, 1989; Wilson & Martinussen, 1999). While measurement experts have traditionally espoused such recommendations as standardization, objectivity and large-scale testing (Airasian & Jones, 1993; Anderson, 1989; Gipps, 1994; Stiggins, 1985, 1986; Wilson, 1989, 1990a), these have not necessarily been intended for daily classroom assessment, nor have they typically been adopted into teacher practice -- particularly at the lower grade levels (Brookhart, 1991, 1993; Cross & Frary, 1999; Nagel, 1992; Shafer, 1989; Stiggins, 1985; Whittington, 1999). There exists what might be described as two strands of thinking about assessment; that of the measurement tradition, and what it is that teachers -- at least in the elementary classroom -- seem to do as part of their daily activities outside of formal testing (Airasian & Jones, 1993; Anderson, 1989; Brookhart, 1999; Stiggins, 1990a; Wilson, 1990a). The latter of these topics is of interest here. It is suggested that further research on classroom assessment is a necessary step towards understanding teacher practice, both in and of itself, and in relation to the measurement tradition that has historically shaped much of the discussion around assessment (Airasian & Jones, 1993; Gullickson, 1984, 1985; McIntyre, 1990; McLean, 1990; Rogers; Shulha, 1999; Stiggins & Bridgeford, 1985). In moving towards that understanding, an area that has received relatively little attention concerns possible gender differences in classroom assessment. The gender of educators is one of the variables that has been largely overlooked in the

efforts of assessment research to date (Goodwin & Stevens, 1993; Hopf & Hatzichristou, 1999).

Purpose and Definition

The purpose in conducting this study was to carry out a gender-based analysis of the assessment practices of a group of pre-service teachers (n=38) enrolled in the Faculty of Education at the University of Victoria. More specifically, these participants had maintained diaries as part of a larger study of classroom assessment (see Anderson, 2000; Bachor & Baer, 2000; Anderson, Bachor & Baer, 2001), and these data formed the basis of this investigation. The overarching question posed was as follows: Are there differences between the assessment practices of male and female pre-service teachers as indicated in the diaries these individuals had maintained? More generally, this study is seen as part of the move towards achieving a greater understanding of classroom assessment.

Due to the breadth of topic areas considered in the review for this investigation and the diversity of methodologies and practices that are included, a wide definitional berth was given to terms such as assessment, evaluation and measurement. A 'generic' use of these constructs was thought to be most applicable, and hence the terms are essentially taken to be synonymous. This is consistent with the loose definitions stipulated by others (for example Mavrommatis, 1997; Wyatt-Smith, 1999), who have described assessment and evaluation as involving such diverse processes as collecting information, making interpretations, appraising, arriving at a judgement and so forth. Although some educational theorists such as Gullickson (1994) have distinguished between the terms -- evaluation being defined as involving both description and judgement, with assessment and measurement pertaining only to description -- this was seen as too limiting for this study. While it is further recognized that the term measurement is generally reserved for more formal practices, and it is used in that sense here as well, all three constructs were taken to represent a continuum of activities from highly structured testing to classroom observation.

In addition to the interchangeable and broad nature of the definitions noted, the use of the term 'classroom assessment' is similarly intended to represent a range of activities. The distinguishing factor being that classroom assessment refers to the actions

of teachers in the classroom context. This is, however, a significant distinction, for as is argued below, classroom assessment appears to be a unique and distinct undertaking.

Finally, when describing the 'measurement tradition', the term is intended to illustrate the practices and theories that have been foundational to experts and specialists in measurement and evaluation. This refers to the notions that have historically shaped formal measurement research and the recommendations of best practice, including such things as standardization, objectivity and other matters of technical concern described in more detail later (Airasian & Jones, 1993; Anderson, 1989; Stiggins, 1985, 1986; Wilson, 1989, 1990a). The inclusion of the term 'tradition' or 'traditional' is intended to set these practices apart from other more recent research initiatives. Further, as the word tradition implies, these are facets of measurement that not only have strong historical precedence, but continue to be influential and persistent in the measurement community.

General Overview

Regardless of personal perspective, training or particular context, the majority of teachers find a substantial portion of their time taken up by classroom assessment in one form or another (Butterfield et al., 1999; Gullickson, 1984; Stiggins, 1990a, 1999). Whether it is formative or summative, designed to produce results for reporting or intended to inform curriculum, instructional and learning decisions, teachers necessarily evaluate their charges. Of interest for this thesis is the finding that despite its significance, educators tend to have very little training in assessment (Maguire, 1990; Mavrommatis, 1997; Stiggins, 1986, 1990) and often view measurement courses as irrelevant to classroom realities, especially at the elementary level (Airasian & Jones, 1993; Brookhart, 1999; Schafer, 1989; Whittington, 1999).

While teachers recognize the importance of evaluation (Gullickson, 1984), a difference appears to exists between classroom assessment as practiced daily by teachers -- at least what is understood of this -- and the measurement tradition that has shaped courses and materials on formal assessment (Airasian & Jones, 1993; Anderson, 1989; Brookhart, 1999; Stiggins, 1990a; Wilson, 1990a). The paramount issues that have traditionally concerned measurement experts, such as the large-scale testing of student achievement (Anderson, 1989, 1990; Bachor, Anderson, Walsh & Muir, 1994; Cross & Frary, 1999; Stiggins, 1986) and an emphasis on technical sophistication (Airasian &

Jones, 1993; Wilson, 1989, 1990a) are not seen as relevant by many teachers, and have not been readily incorporated into practice, at least at the lower grade levels (Brookhart, 1991, 1993; Cross & Frary, 1999; Nagel & Driscoll, 1992; Shafer, 1989; Stiggins, 1985; Whittington, 1999). It has been suggested that a greater understanding of classroom assessment begins to offer a means of addressing these concerns (Airasian & Jones, 1993; Gullickson, 1984, 1985; McIntyre, 1990; McLean, 1990; Rogers, 1999; Shulha, 1999; Stiggins & Bridgeford, 1985). It is thought that a firm knowledge base of classroom assessment -- in its multitude of guises -- is a step towards better informed practice. Although traditional measurement training may offer many useful and necessary functions for educators (for example in the design of test items and the interpretation of results), it fundamentally serves different and limited functions relative to the entire domain of classroom assessment. It is with that in mind that this research into the teacher practice is considered.

It may be argued that until the past decade or so, there had existed a relative paucity of information on assessment as it takes place in the classrooms of practicing teachers (Boxall & Gilbert, 1999; Gullickson, 1985; Lomax, 1996; Stiggins, 1985, 1990b; Wilson, 1989; Wilson & Martinussen, 1999). However, from what has been learned of classroom assessment, teachers tend to reject large-scale external testing (Anderson, 1989; Gullickson, 1985; Maguire, 1990; Stiggins et al., 1986), favoring a diversity of practices that are neither standardized nor intended to serve a single purpose (Bachor & Anderson, 1994; Cross & Frary, 1999; Mavrommatis, 1997; Wilson, 1990a). In contrast to the technical concerns of traditional measurement, educators rely on a variety of assessment procedures and practices, that are often informal, intuitive and implicit (Airasian & Jones, 1993; Bachor & Anderson, 1994; Chase, 1986; McCallum et al., 1993; Mavrommatis, 1997; Wyatt-Smith, 1999). Further, student characteristics and social interactions are known to influence assessment (Brookhart, 1991, 1993; Whittington, 1999). It must be added however, that these findings apply largely to the elementary and middle school levels, where the influence of large-scale testing and formal assessment practices are not as germane.

From this level of understanding, it appears that classroom assessment differs from practices that have been the mainstays of traditional measurement, and indeed this

would be expected considering their different intentions and objectives. Notions of standardized measures and technical sophistication, for example, are not necessarily of primary concern to the elementary teacher looking to offer constructive feedback and encouragement on a fine arts project. Recognizing these differences, there have recently been calls for further understanding of assessment and measurement through research aimed at understanding teacher practice within the classroom context (Anderson, 1989, 1999; Crooks, 1988; Gullickson, 1985; McIntyre, 1990; McLean, 1990; Shulha, 1999; Stiggins, 1990a, 1990b).

Following a review of the measurement tradition, what is known of classroom assessment and the some of the differences between the two, the move towards greater research into teacher practice is highlighted. This study is considered within this classroom assessment research agenda. As part of a larger project, termed "The Classroom Assessment Project - National" (CAP-Nat), the assessment practices of pre-service teachers have been examined (Anderson, 1999, 2000; Bachor & Baer, 2000; Shulha, 1999; Wilson & Martinussen, 1999). Described in detail below, these pre-service teacher investigations offer a novel approach to the further understanding of evaluation as carried out by these beginning teachers. Of these initiatives, the study undertaken at the University of Victoria produced a series of diaries or journals that were maintained by pre-service teacher participants as part of a research project. Essentially these diaries contained the thoughts, ideas, concerns and other notions that these individuals had regarding their assessment of a series of fictional students. This data source was originally described and examined by Bachor and Baer (2000). Through the course of their investigation, it became apparent that further examination of this large quantity of qualitative information was warranted. The second author noted that a gender-based analysis of these journals might provide a means of obtaining further information about a variable that has to date received little attention.

A review of the limited research on educator gender as a variable in classroom assessment serves as the final background for this study. It is noted that few investigations have examined the influence of educator gender on assessment and related activities (Goodwin & Stevens, 1993; Henebry & Diamond, 1998; Hopf & Hatzichristou, 1999), and of those that have been conducted, the results have typically been conflicting

and inconclusive (Brandt, Hayden & Brophy, 1975; Centra & Gaubatz, 2000; Gunderson, Tinsley & Terpstra, 1996; Stake & Katz, 1982). Following Brophy's (1985) suggestion that thick description and attention to qualitative details offers a means to further investigating this factor, this investigation was undertaken. An analysis of the pre-service teacher diaries through a comparison of entries from female and male participants was therefore conducted.

Research Questions

As noted above, this study is an attempt to parse out educator gender as a potential factor influencing assessment. The limited research and conflicting findings regarding gender and classroom assessment, along with the call for qualitative analysis on this topic made such an investigation an obvious choice. Therefore, the following nine questions were posed:

1) What are the similarities and differences in the overall contents of the participant diaries?

2) In what ways does gender influence the focus and concern these participants have regarding evaluation in general?

3) How do the means of evaluating (evaluation criteria) differ between gender?

4) How are the comments about the students and overall evaluation of these pupils different between female and male participants?

While the above questions form the general queries of interest here, several more focused questions are posed. In particular, those of interest here include the following five:

5) What are the differences between male and female participants regarding the significance of non-achievement variables including the importance of knowing the student, concerns over the fictional pupils affective state, and desiring information on the subject or classroom in order to carry out assessment?

6) How do the interventions suggested as part of these pre-service teachers' evaluations differ between gender?

7) In what ways do the decision paths these pre-service teachers appear to follow in their assessment practices differ between gender?

8) What are the discrepancies between male and female participants in regards to the assessment of inferences that classify the fictional student(s) as having a possible special learning need?

9) What are the gender differences in the assessment of "extreme" inferences concerning the influence of student(s)' quality of life on their performance products?

Chapter 2

REVIEW OF THE LITERATURE

The literature review for this study is organized into five basic sections. Although, these divisions serve organizational ends, it will be noted that they are not discrete and much of the information overlaps or otherwise connects across various portions. The initial section is devoted to a brief examination of assessment as it has traditionally been approached by measurement experts and taught in formal assessment courses. This is followed by an examination of classroom investigations, with a focus on the differences noted between the traditional measurement perspective and the contextual assessment activities that constitute teachers' daily activities. The third section is focused on research aimed at examining exactly what it is that educators do in the classroom and how measurement and evaluation take place within this context. Following this, the immediate predecessors to this study -- which constitute part of a larger investigation -- are described. In the fifth and final section, the research on educator gender and differences in classroom assessment practices are discussed.

The Measurement Tradition

A discussion of the measurement tradition is included here in order to illustrate the differing intentions, aims and objectives of this body of research in relation to the classroom practices of teachers, and in particular teachers at the lower grade levels. The cornerstones of traditional measurement (discussed below) have a strong historical precedence in assessment and continue to exert a powerful influence in the educational arena, and therefore cannot be ignored in setting the context for this topic. Large-scale mandated testing and some aspects of special education assessment (e.g., Bachor, 1990; Bachor & Crealock, 1986) are examples. However, it is argued that this represents a strand of assessment that does not encompass the realm of classroom practices; nor is it intended to reflect it. Appeals to traditional measurement as the sole means of improving teacher practice are therefore likely to fall short, despite their relevance to certain aspects of assessment. Through an elucidation of the differences between the measurement tradition and classroom assessment as practiced by educators, it is hoped that the justification for a more thorough understanding of the latter is highlighted. Moreover, it

is suggested that the application of traditional measurement recommendations cannot in and of themselves address the needs of classroom assessment.

As noted in the introductory chapter of this thesis, the terms evaluation, assessment and measurement are used interchangeably. The diversity of topic areas considered necessitates that a wide definition be afforded these constructs and a broad description such as that suggested by Mavrommatis (1997) or Wyatt-Smith (1999) serves this function. However, in the measurement tradition the sorts of practices that are said to constitute measurement and evaluation are much more narrowly defined. Assessment activities that teachers may employ on a regular basis (described in the section to follow) have not generally been included here, while issues that often take a back seat for educators in the primary and intermediate grades are of prime concern to measurement experts. A brief examination of the measurement communities traditional views regarding the focus of assessment, the function of educational measurement and the content of assessment courses is used to highlight this point. It must be pointed out that this section is concerned with the traditional and pervasive themes that have historically been the center of measurement research. There are many in the field who have been critical of these views in terms of their application in education (for example, Stiggins, 1990b; Whittington, 1999; Wilson, 1999).

The assessment focus. Central to educational measurement is what Wilson (1989, 1990a) has referred to as the "technical sophistication" or "technical characteristics" of a measure. Airasian and Jones (1993) have similarly used the term "technical rationality" for this view. The emphasis is on matters of standardization, generalizability and objectivity (Anderson, 1989; Stiggins, 1985, 1986). Practices that are structured and objective are promoted (Shulha, Wilson & Anderson, 1999). As Gipps (1994) has written, "With the psychometric model comes an emphasis on standardization and reliability... Standardization, which in part gives rise to reliability... is a key feature of the psychometric model" (p. 284). This strive for technical mastery of measurement has honed in on matters such as those described, and in turn has led to a particular manner of assessment.

The desire for objectivity, standardization and other 'technical' concerns is noted in tandem with an emphasis on the ubiquitous 'paper-and-pencil' tests as the evidence

gathering techniques of choice (Anderson, 1989; Stiggins, 1986). Further, these measures have often taken the form large-scale testing initiatives (Anderson, 1989, 1990; Bachor et al., 1994), typically employing multiple-choice items (Stiggins, 1990a). As Anderson (1989) has stated, "... the focus of educational measurement is on large-scale testing projects and standardized tests..." (p. 124). Large-scale testing of this nature may be described as a 'best-fit' with the desire for objectivity and technical sophistication. An additional facet of this approach is that these measures are seen as falling within a paradigm that attempts to document student achievement (Anderson, 1999; Cross & Frary, 1999; Stiggins, 1986)

Student achievement is thought to be the fundamental characteristic for assessment by testing. As Stiggins (1986) put it, "Available evidence suggests that the dominant view regards measurement in education as a means of documenting student achievement by using collections of standardized paper and pencil test items..." (p. 5). Measures are therefore developed with the specific intent of assessing student achievement and nothing else. When an individual sits an exam, for example, it is desirable that the results of the exam reflect only the achievement level of the examinee. This is in line with what Anderson (1999) has cited as the predominant view in educational measurement, whereby testing is unidimensional and considered to assess only a single underlying trait. Potentially confounding variables are controlled so that outcomes reflect only the achievement results. In turn these results are considered as the basis for the marks and grades student are awarded. As noted by Cross and Frary (1999), "... there is widespread agreement among measurement specialists that grades, at least in academic subjects, should be based exclusively on measures of current achievement and that growth, ability, effort, conduct and other nonachievement factors should not be considered" (p. 53). To summarize, it may be said that measurement experts in education have traditionally been concerned with matters of technical adequacy and the development of assessment tools that have generally taken the form of standardized paper-and-pencil tests in order to assess current student achievement levels. The assessment of achievement is then transformed into the letter grades or other reporting methods with which students are familiar. In turn, this has had an influence on the contents of assessment courses.

Assessment course content. In 1964, the inaugural issue of the *Journal of Educational Measurement* carried an article titled "What Experts Think Teachers Ought to Know About Educational Measurement" by Samuel T. Mayo. This foundational article set a tone for measurement courses and materials -- at least in North America -- that is still reflected in textbooks on the subject (Gullickson, 1986). Mayo's (1964) survey, which purported to rank what beginning teachers should know about measurement, focused primarily on matters of testing (standardized testing in particular), as well as the uses of measurement/evaluation and statistical concepts; essentially, the mainstays of the measurement tradition. A more recent survey by Gullickson concluded that professors' emphases in undergraduate educational measurement courses had remained consistent with the opinions reported in Mayo's earlier investigation.

Several authors have underscored the notion that the content of measurement courses and corresponding materials have essentially continued to reflect the traditions described above. For example, Lomax (1996) has written that assessment courses and textbooks devote a great deal of time to standardized testing and statistics, while Wilson (1990b) noted that texts in educational measurement generally involve adaptations of psychometrics. Brookhart (1999) echoed these sentiments when she wrote that most measurement professionals have received an education and training which "...emphasized psychometrics for large-scale assessment" (p. 5). Airasian and Jones (1993) have similarly noted that instruction in classroom measurement is guided primarily by a view towards technical rationality. Although there have been calls for a shift away from these traditional perspectives (to be described shortly), topics such as reliability and standardized testing have had a profound impact on measurement courses. As stated by Shulha (1999), "To be adequate for the task of assessing students, teachers are encouraged to learn what measurement specialists know about reliable and valid measures and to structure their work such that this knowledge can be put into practice" (p. 289). Thus, the measurement tradition can be described as having substantial roots in educational assessment, as well as a continuing influence in the field.

Teacher Practices and their Discrepancy with Traditional Measurement

An examination of teachers' classroom assessment practices reveals that in many senses they are different from the recommendations that have come out of the

measurement tradition. In fact, some of the activities that elementary and intermediate educators have come to rely on might be described as the antitheses of best psychometric practices. By examining teacher perspectives on measurement, what they see as the purpose of assessment, how they view tests, what sort of assessments are frequently employed and the implicit nature of many of these activities, the discrepancies between classroom assessment and the measurement tradition come to the fore.

Prior to examining classroom assessment from the perspective of the practicing teacher, it must be added that until the past decade, relatively little was understood about this process. Although student evaluation is something that has occupied educators for as long as teachers have been teaching, there had been a general lack of research in this area. For instance, Lomax (1996) has noted that few studies have directly examine teacher assessment literacy, while Wilson and Martinussen (1999) have likewise claimed that not much is understood about the process of how teachers arrive at judgements regarding student achievement. Others (see for example, Crooks, 1988; Gullickson, 1985; Stiggins, 1985, 1990b) have similarly stated that knowledge of teacher evaluation practices is lacking. Nonetheless, the available evidence points to discrepancies between what it is that teachers do and what the measurement tradition has espoused.

Teachers and measurement. The 'relationship' between educators and assessment provides a foundation from which perspectives and attitudes begin to shape and influence the application and utilization of various measures. The measurement training teachers receive and how they incorporate it into their teaching activities is one example. Further, the position educators see themselves as occupying relative to student evaluation and their view of assessment practices underpins some of the more obvious questions concerning the purpose of assessment, instruments used by teachers and so forth.

Having briefly discussed the impact of the measurement tradition on educational assessment courses and related materials, it might well be assumed that teachers are versed in at least the basics of the discipline. However, it has been found that most teachers have very little or no formal course-work in assessment (Stiggins, 1986, 1990a). Others (for example Mavrommatis, 1997) have described a situation in which assessment is all but overlooked in teacher training programs. Within the Canadian context for example, Maguire (1990) has written, "In many institutions, education students are not

required to take a course in classroom assessment so it is not surprising that they know very little about concepts like grade equivalence, validity and error of measure" (p. 87). It appears that the tenets of educational measurement, as recommended for instance in Mayo's (1964) seminal paper, have not translated into teacher training. The concepts that have often been advanced as fundamental for educators -- those that they "ought to know" -- are taught to relatively few. While courses and materials may reflect these topics, it seems that they are not required of many teacher candidates. Most teachers enter the classroom with assessment strategies (assuming they have any strategies) that are uninformed by knowledge of reliability, standardized testing and so forth. Yet, if the measurement tradition has had little influence on teacher training and teacher assessment knowledge, this begs the question about where teachers learn such practices and how they deal with this topic.

The measurement practices that teachers use are thought to be influenced by personal experiences and exposure to the methods employed by colleagues (Anderson, 1989; Gullickson, 1984). Regarding testing, for example, Gullickson has stated that most teachers believe they learned how to test students through experience acquired on-the-job. Even in cases where teachers had been taught something about assessment, accepted school practices were found to exert considerable sway (Nagel & Driscoll, 1992). For example, Nagel and Driscoll discovered that when it came to various teacher practices (including assessment practices), "...students were willing to abandon what they had learned at the university in order to align themselves with the school" (p. 8.). These school practices in turn may be influenced by mandated policies, or what Wilson (1990b) has described as levels 'above' in the administrative hierarchy, and thus channel evaluation into particular areas. Whatever the direct and indirect influences, even educators that come into the classroom with some assessment background are not immune to the impacts of 'learning on- the-job'. In the context of this thesis, the notion that gender may be one of the potential influences is considered.

With many teachers having relatively little formal training in classroom assessment, and others noting the influence of colleagues, accepted and expected school practices and so forth, it is understandable that measurement courses tend to be viewed as somewhat mismatched with classroom realities (Airasian & Jones, 1993; Schafer, 1989;

Whittington, 1999). The apparent differences between traditional measurement courses and the approaches learned in the classroom creates a dilemma whereby practices recommended as part of teacher training may contrast with those that are generally valued within schools, at least at the lower grade levels where testing is not as prevalent (Nagel & Driscoll, 1992). In a survey of teachers and professors, Gullickson (1986) found strong disagreement between the two groups in regards to the importance of several key assessment issues, while Cross and Frary (1999) noted that even among those educators trained in the measurement, there is considerable resistance to embracing these practices. Airasian and Jones (1993) summarized the situation when they wrote "Currently, a mismatch exists between the content of many classroom measurement and assessment courses and the classroom realities these courses are intended to inform" (p. 241). When traditional evaluation is included as part of teacher training, it is not necessarily seen as addressing the gambit of classroom experience. Therefore, it can be argued that the practices promoted within the measurement tradition are not always included as part of teacher education and when they are taught, they may be considered as contrasting with on-the-job practices.

Teachers recognize the significance of evaluation (Gullickson, 1984) and it has been noted that they are not averse to examining their assessment methods or changing practices (Bachor & Anderson, 1990, 1994). However, part of the classroom reality (or classroom perception is perhaps more appropriate here) deserves particular mention due to its perceived constraint on practice; namely, time. Educators are not only concerned about their assessment methods, they express worry over the length of time evaluation activities take and the demanding nature of the tasks (Stiggins, 1990a; Stiggins & Bridgeford, 1985). As Stiggins and Bridgeford concluded from the results of their survey, "...teachers frequently reported concern about their ability to effectively integrate assessment given the time constraints imposed by the classroom. Overall, teachers' responses in this study indicated concern about assessment quality and frustration at the lack of time available to deal more adequately with the problem" (p. 282). Thus, while teachers are concerned over inadequate understanding and they may be open to new assessment methods, they also feel pressed for time. The demands of the classroom environment do not invite assessment practices that are seen as taking away from

instructional time. In fact, Butterfield et al. (1999) reported that many teachers view assessment as a chore and an interruption to teaching. In a context where formal education in measurement is limited, the influence of school and classroom 'realities' is great, assessment courses may be seen as narrow and teachers feel particularly pressed by time constraints, it is not surprising that the recommendations traditionally advanced by measurement specialists have not necessarily translated into daily teacher practice. This is further noted when considering teachers' perspectives on the purposes of assessment.

The purpose of assessment. The aforementioned concern with the unidimensionality of testing in the measurement tradition -- expressed as achievement outcomes -- and the adopted function of these assessments as means to accountability measures contrasts with the multiplicity of purposes that are cited for classroom assessment (Wilson, 1989). As Wilson has put it, "A most striking feature of the data concerning purpose is that evaluation instruments invariably serve at least two purposes, and frequently more than two.... Thus the professional testmaker's concern with clarity and singleness of purpose, a unitary index of reliability, and a view of validity based on the administration of a single instrument may not have much applicability in the complexity represented by this [*the high school classroom*] milieu" (p. 140-141). This is particularly true of the alternative assessments employed by many teachers, which fit neither a single set of procedures nor a uniform purpose (Bachor & Anderson, 1994). Classroom assessment is both multipurpose and multimeasure, with what at times appear to be disparate intentions to conducting evaluations.

Providing an all-inclusive description of the purposes of classroom assessment is beyond the scope of this review, however, a brief list may serve to illustrate some of the diversity of purposes educators ascribed to these classroom practices, as opposed to the more restrictive intentions of the measurement tradition. These include checking students' progress, diagnosing weaknesses, enabling students to monitor their own progress, assisting with decisions regarding what to teach, checking how well material has been taught, generating marks for reporting and so forth (Wilson, 1990b). Butterfield et al. (1999) have stated that assessment is also seen by some educators as serving the purpose of meeting mandated accountability requirements, as well as maintaining control and discipline over the class. The multiplicity of purposes is perhaps best considered in

relation to the comment made by Nuttall (1987) that "Assessment (like learning) is highly context specific..." (p. 115). The purpose it serves may be driven as much by the situation at hand as it is by notions of anything else, and certainly not by uniform intent. One consideration is however particularly salient and worthy of further mention, and that is the contextual situation created by the relationship educators' maintain with their students.

Teachers do not operate in a social vacuum. In fact, their position may be described as one steeped with social interaction and personal rapport. The teacher-student relationship involves persons engaged in the active exchange of information and human contact. Despite the inherent nature of authority and power structure between educators and pupils, it is a relationship nonetheless. In turn, the influence of this relationship has been noted at the level of classroom evaluation (Airasian & Jones, 1993; Brookhart, 1991, 1993; Whittington, 1999). Brookhart (1991, 1993) has suggested that teachers naturally consider the consequences of assigned grades and grading practices, and that any consideration of the validity of a grade should include an exploration of these social consequences. She (1993) has further stated that, "When considering what consequences a grade will have for a student, the teacher functions as an advocate for the student. Concern about this function does not differ for those who have had measurement instruction" (p. 140). As Airasian and Jones (1993) have argued, it is therefore difficult for teachers to separate their knowledge and perception of pupils from their grading judgements. This, however, places educators in the rather awkward position of serving dual roles when it comes to assessment practices.

The position of teachers has been described as one of being caught in the conflicting roles of 'judge and advocate' (Whittington, 1999). There is an expectation that educators ought to promote student learning, encourage growth and otherwise provide support, while at the same time act as judges when it comes to assessment for grading. This point has similarly been made in describing the classroom as a context whereby formative assessment for student growth is encouraged alongside the demands of summative accountability measures (Butterfield et al., 1999; Boxall & Gilbert, 1999). Whittington has likewise stated that teachers often perceive that adapting to the characteristics and needs of the individual student is preferable to upholding a common

standard; the latter of which is paramount to traditional measurement practices. In other words, there is a conflict between individualization and adaptation versus standardization and reliability. Teachers feel a need to know and understand students, and are pulled to individualize marks by including such factors as effort, participation, attitude and so forth (Wilson & Martinussen, 1999). These, however, are precisely the sort of variables that measurement specialists have typically attempted to exclude. Nonetheless, in the classroom context, it appears that the nature of the social interactions teachers have with pupils are particularly influential on assessment practices, even when teachers have had some formal measurement training (Brookhart, 1993). This makes sense in light of the different intentions and roles that traditional measurement has relative to the aims of many aspects of classroom assessment.

The varying purposes of assessment illustrate that educators (particularly at the elementary and intermediate grades) and traditional measurement specialists may be described as occupying different spectrums when it comes to evaluation. The singularity of purpose in the measurement tradition contrasts with the multiplicity of intentions in classroom assessment. It is the relationship teachers maintain with their students that is an important factor in the classroom context. A closer examination of a particular form of assessment, testing, offers a means to explore further such discrepancies.

Tests as used by teachers. The use of large-scale paper-and-pencil tests -- most often based on multiple choice items -- was described as the evaluation tool of choice within the measurement tradition (Anderson, 1989; Stiggins, 1986, 1990a). Large-scale testing initiatives have also been adopted and put into practice as part of the growth in accountability assessment, particularly in North America (Bacon, 2000; Butterfield et al., 1999; Mavrommatis, 1997; Rogers, 1990). The significance of tests, not only as they exist within the measurement community, but also as part of accountability assessment warrants particular attention in regards to the role they play in teacher practice. That is, questions concerning the use of tests by teachers, the role of tests and the reaction of educators to large-scale testing programs serve to illustrate differences that may exist in regards to this fundamental assessment tool.

Although there had until fairly recently been limited field research on educators' use of tests (Gullickson, 1984), it seems to be the case that tests are a key assessment tool

for many classroom teachers (Bateson, 1990; Gullickson, 1985; Stiggins & Bridgeford, 1985). Bateson, for example, found that teachers of science in British Columbia depended primarily on their own "objective-type" tests to evaluate their students (p. 45). Similarly, Stiggins and Bridgeford reported the prominence of teacher developed tests in their investigation of classroom evaluation practices, while Anderson (1989) noted that teacher-made tests received the most emphasis in determining student achievement among the educators in his study. The use of tests, however, is only one in a number of assessment strategies found in the classroom (some of which are discussed in the section below). Moreover, the form of these tests at the lower grade levels appears to differ from large-scale measures.

Unlike the paper-and-pencil tests of the measurement tradition, teacher developed tests tend to take the form of smaller more routinely administered assessments of performance that are used along with a variety of other measures (Anderson, 1989, 1990; Bachor & Anderson, 1994; Gullickson, 1985). Concern over standardization gives way to tests that are context specific measures of performance on relatively narrow topics. Weekly spelling tests, end of unit quizzes and so forth are examples of teacher developed tests. They are considered part of ongoing student evaluations, rather than one-time measures of achievement. Again, the emphasis is on individualization as opposed to standardization. While results from teacher-made tests are linked directly to establishing scores for grading (Bachor & Anderson, 1994), test results tend not to be the sole judgement of those grades (Gullickson, 1984). Investigators (Gullickson, 1984; Stiggins & Bridgeford, 1985) have additionally noted that educators express concern both over their own abilities at testing and about the evaluative merits of these measures. Yet, self-made tests continue to hold a prominent place in the classroom, and this despite the wide availability of large-scale externally developed assessments.

Teachers' reactions to external testing, offers another avenue to understanding where classroom educators stand relative to the sorts of practices that have come out of the measurement tradition. In general it has been found that teachers tend to reject large-scale external testing initiatives and commercially prepared tests (Anderson, 1989; Gullickson, 1985; Maguire, 1990; Stiggins et al., 1986). For example, Gullickson has concluded that while the use of teacher-made tests increased from the elementary to

secondary grades, commercially developed tests were not emphasized at any grade level. Among science teachers, Anderson likewise found that teachers would not like to base assessment on external exams, nor do they express interest in the data generated from large-scale testing. These sentiments have been reflected in Bateson's (1990) description of the noticeable lack of Ministry of Education developed achievement tests in science classrooms and Bachor et al's (1994) finding that formal tests are "rare events" that tended to be removed from classroom activities (p. 248). Thus, while large-scale testing may find its way into schools via mandated assessments, these tend to be handed down from above rather than voluntarily adopted by teachers. The use of individualized tests developed by classroom teachers for their particular contexts and specified needs are the preferred measures here. The recommendations that have shaped large-scale testing within the measurement tradition do not appear to have fully addressed the concerns of classroom teachers, who use self-made and small scale tests as part of their ongoing student assessment activities. These tests then become part of a larger repertoire of measures that teachers use on a regular basis.

A complex variety of assessment practices. As noted above, classroom assessment as practiced daily outside of formal testing initiatives has neither a single set of procedures nor a single purpose (Bachor & Anderson, 1994). This point is revisited here not to provide an exhaustive list of assessment varieties, but rather, to highlight further the fact that what is recommended by measurement experts does not necessarily reflect the breadth of teacher practice. Classroom assessment has been described in various investigations as a hodgepodge of activities (Cross & Frary, 1999), a complex procedure (Mavrommatis, 1997), involving an eclectic array of measures (Anderson, 1990), using a diversity of evaluation techniques (Gullickson, 1985) and lacking the tidiness of aim that measurement experts seek (Wilson, 1990a). Teachers do not appear to share the concerns over standardization, uniformity of procedures and so forth that occupy their measurement colleagues. Rather, practices vary tremendously both within and between classrooms.

Beyond a diversity of practices, notions of objectivity and control of 'extraneous variables' may also be set aside. It has been pointed out for example, that the personal characteristics and individual philosophies of educators exert an influence over

assessment (McCallum, McAlister, Brown & Gipps, 1993; Wyatt-Smith, 1999). For example, Wyatt-Smith noted that reading and evaluating student written work involved at least four major elements: individual teacher philosophy, attitudes towards and purpose of reading, available teacher knowledge, and the teacher's attempt at reconstituting conceptions of quality. Cross and Frary (1999) argued that educators do indeed consider such factors as student effort and growth, as well as the social consequences of evaluation. These however are precisely the sorts of variables that are seen as confounding measures of achievement.

Looking more closely at classroom assessment practices, Bachor and Anderson (1994) found that teachers preferred an expanded repertoire of tools. This repertoire was noted to include portfolios, tests, student self-evaluations, reviews of student work samples as well as various other methods. The authors continued, stating that "These assessment procedures are not discrete, specific activities; rather, they constitute broad categorizations of assessment practice and vary considerably in application from one time to another, and from one teacher to the next" (p. 71).

Within this categorization of practices, the use of observation comes up time and again as one of the most commonly cited assessment procedures (Bachor & Anderson, 1994; Stiggins & Bridgeford, 1985; Stiggins et al., 1986; Wilson & Martinussen, 1999). The prominence of observation is particularly noteworthy because it is not easily amenable to standardization or objectivity, nor are results from such observations generally recorded; although checklists, rubrics and other procedures are employed by some teachers (Bachor & Anderson, 1994). Student observation perhaps best epitomizes one of the varieties of commonly used classroom assessment practices that defy the recommendations of traditional measurement specialists. Not only do teachers employ multimethod approaches, many of these practices truly lack the tidiness of aim that measurement specialists have proposed. This point becomes particularly salient considering that many classroom evaluation methods are not articulated, let alone recorded.

Implicit aspect of classroom evaluations. Classroom assessment practices, including teacher-made tests, work sample evaluations and observations are described as complex and diverse relative to large-scale paper-and-pencil testing initiatives (Stiggins,

1990b). Teachers carry out their evaluations in situ, as part of an ongoing information gathering process. While these tests, portfolios and other measures provide tangible evidence of this process, they are generally not the standardized techniques and protocols of the measurement specialist. However, many of the influences on classroom assessment are far less visible.

It was noted above that student observations, although one of the prominent methods of evaluation, often go unrecorded (Bachor & Anderson, 1994). Stiggins and Bridgeford (1985) likewise found that in 40 percent of performance evaluations they examined, teachers relied on "mental record-keeping" to store and retrieve information (p. 283). At least a certain amount of assessment therefore may be considered dependent on educator memory, recall and mental interpretation. Carrying investigations of this further, several authors have described teacher practices that are informal and hidden (Airasian & Jones, 1993), existentially generated (Wyatt-Smith, 1999), intuitive (Bachor & Anderson, 1994; Mavrommatis, 1997), implicit (McCallum et al., 1993) and otherwise influenced by variables unrelated to content (Chase, 1986). As Mavrommatis has put it, "...in practice, classroom assessment is often intuitive. Indeed, the teacher may not even be aware that it is taking place" (p. 382). More formal assessment practices are tempered by the informal assessment of pupils, the social context and other ill-defined factors (Airasian & Jones, 1993). These influences noted among classroom practices contrast with the measurement tradition, where efforts to control such 'confounding' variables are paramount. At least at the lower grade levels, factors that have traditionally been deemed undesirable among measurement experts play a role in classroom assessment. The inclusion of educator gender as a possibly intuitive or otherwise influential variable is examined in this thesis.

Classroom assessment is a complex and demanding undertaking that is often implicit, time consuming and difficult for teachers to explain (see for example, Bachor & Anderson, 1994; Stiggins 1990a). Although some evaluation methods display a closer affinity towards objective measures than others do, teachers have been described as bringing in a range of "...existentially generated considerations" (Wyatt-Smith, 1999, p. 220). It is these factors interacting in multiple ways that are not easily quantified or explicitly defined. For example, in the flurry of classroom life, teachers have been found

to form quick impressions and judgements of student that in turn have a substantial impact on subsequent evaluations (Stiggins et al., 1986; Wilson, 1990a). Stiggins et al. described this process as tremendously complex, with teachers including scholastic, social and personal student characteristics as part of their evaluative judgements. Largely, it is the inclusion of such implicit factors in classroom assessment that led the authors to conclude, " There are fundamental far-reaching differences between the science of testing and the assessment demands of the classroom" (p. 14).

Summary of discrepancies between classroom practices and the measurement tradition. Classroom assessment is a substantial and ubiquitous facet of education. Teachers spend a significant portion of their working lives attending to the demands of assessment and a larger community is concerned with its outcomes. It is arguably one of the pivotal and most important activities to take place in the classroom. While traditional measurement has informed many aspects of assessment and it continues to be influential in several areas, there are essential differences between these practices and what is known about classroom assessment as conducted by teachers.

Some of the overarching principles to come out of the measurement tradition have been the centrality of technical sophistication and the focus on large-scale standardized testing. These traditions were subsequently emphasized in courses and related materials on educational measurement. Exploration of teacher practices reveals that their perspectives and classroom realities -- at least at the lower grades -- differ from those of the measurement tradition. Four areas are relevant in highlighting these discrepancies. First, teachers tend to receive relatively little formal training in assessment, learning on the job and being influenced by a variety of factors. Second, in contrast to the unidimensional nature of standardized tests, teachers use various assessment measures for a diversity of purposes and intentions. Likewise, teachers do not tend to depend on large-scale single tests, but instead prefer individual and context specific tests along with a collection of other measures. Finally, the technical concerns of traditional measurement contrast with teacher practices that are informal, intuitive and complex. These differences in turn suggest that further investigation into classroom assessment and the variables that play a role in these practices is warranted.

At the outset, teachers appear to come into the classroom with relatively little assessment training, many having limited to no formal course-work on the topic (Maguire, 1990; Mavrommatis, 1997; Stiggins, 1986, 1990a). More importantly, those having some exposure to traditional measurement courses report finding them limited when it comes to many of the aspects and activities taking place in the classroom (Brookhart, 1999; Stiggens & Bridgeford, 1985). Consequentially, teachers may learn their assessment 'skills' on the job, rely on intuition, adapting to the methods promoted by colleagues or have their practices influenced by other variables -- educator gender being one of the possible factors examined here (Anderson, 1989; Gullickson, 1984; Nagel & Driscoll, 1992). Continued research into classroom assessment has therefore been recommended as a means to better understanding the 'messiness' of the classroom (Airasian & Jones, 1993), the perception of teacher testing behaviour (Gullickson, 1984) and the differences between traditional measurement and the full range of teachers' practical measurement needs (Stiggins & Bridgeford, 1985).

The apparent differences between traditional measurement and classroom assessment are noted in the purposes which these practices are seen as serving. For example, the unidimensional nature of standardized achievement tests contrasts with the multiplicity of purposes ascribed to evaluation by elementary and intermediate level educators (Bachor & Anderson, 1994; Wilson, 1989). Teachers often turn to assessment with an array of intentions (Butterfield et al., 1999; Wilson, 1990b), and various contextual and social factors may enter into the milieu, including consideration of the students they are evaluating (Airasian & Jones, 1993; Brookhart, 1991, 1993; Whittington, 1999). The inclusion of this diversity of factors is naturally considered undesirable when they are seen as conflicting variables to the accurate measurement of achievement outcomes.

Even when it comes to testing, teachers do not depend solely upon large-scale measures (Anderson, 1989; Gullickson, 1985; Maguire, 1990; Stiggins et al., 1986), preferring to rely on their own individual tests that form part of a variety of assessment activities (Anderson, 1989; Bachor & Anderson, 1994; Gullickson, 1984, 1985). The standardization and other technical concerns of the measurement specialist give way to a need for context specificity and tailor-made tests that are a part of a larger repertoire of

measures. However, it must be added that this is much more the case at the lower grade levels and within particular disciplines, as the impacts and influence of formal testing tend to increase in high school and within non-elective subject areas. Nonetheless, at least a substantial portion of the teacher repertoire can be said to consist of a complex and varied collection of procedures and practices (Cross & Frary, 1999; Gullickson, 1985; Mavrommais, 1997; Wilson, 1990b) that is further influenced by personal characteristics, philosophies and so on (McCallum et al., 1993; Wyatt-Smith, 1999). Of particular interest is whether gender plays a role within this 'hodgepodge' of methods and multiplicity of interloping variables (Cross & Frary, 1999).

Finally, in contrast to the technical concerns of traditional measurement, teachers may confound classroom assessment with practices that have been described as informal, intuitive and implicit (Airasian & Jones, 1993; Bachor & Anderson, 1994; Chase, 1986; McCallum et al., 1993; Mavrommatis, 1997; Wyatt-Smith, 1999). The complex variety of activities in the classroom, coupled with contextual and social factors, impacts the evaluative judgements teachers make (Stiggins et al., 1986). The desire for cleanliness and rigor of measurement differs from these often messy and loose practices (Mavrommatis, 1997). It is with the goal of better understanding assessment in the classroom context that there have been increasing calls over the past decade for further research into these teacher practices (see for example Brookhart, 1999; McLean, 1990; Rogers; 1999 and others discussed below).

Classroom Assessment Research

At this juncture, it may be prudent to restate that the intent behind examining some of the differences between traditional measurement and classroom assessment has not been to evaluate or judge either of these undertakings. In fact, the manifestations and incantations of both might be seen as serving their designated functions particularly well; after all, they have had a relatively tenacious existence in their respective worlds. Rather, the goal in reviewing these practices and their differences has been to illustrate that despite its historical precedence and continued presence, traditional measurement has not been synonymous with classroom assessment. Appeals to the measurement tradition as the sole means of improving teacher practice are therefore bound to fall short. It is suggested that continued research into classroom assessment and the variables that exert

an influence on these practices offers a possible avenue to greater understanding of what it is that teacher actually do. In turn, recommendations and future instruction may become better informed. However, before considering the recent calls for precisely this sort of research, it may be helpful to note briefly the state of understanding that exists regarding teacher assessment practices.

An incomplete understanding of classroom assessment. With the traditional focus of measurement research concentrating on matters of technical sophistication, large-scale testing and the other aforementioned issues, is not surprising that investigations of teacher practice had remained relatively uncommon until the past decade. For example, Stiggins et al. (1986) reported that due to the narrow focus of measurement research, little is known about assessment as it is developed and used by teachers in the classroom. Others have similarly pointed out that research on standardized testing has overshadowed investigations of teacher-developed assessments, educators' needs and the collection of direct evidence from teachers (Crooks, 1988; Gullickson, 1984; Lomax, 1996, Siggins & Bridgeford, 1985).

Essential information towards understanding teacher assessment practices, such as the link between assessment and learning; although thought to be crucial, are described as not having been widely addressed in the research (McCallum et al., 1993). Moreover, areas of practice that are known to be fundamental and widespread in classroom assessment, including the use of observation and teachers' reliance on their own judgements of student performance, have tended to receive little attention (Rogers, 1999; Stiggins & Bridgeford, 1985). Wilson and Maritinussen (1999) summarized the point when they stated, "Not much is understood about the processes teachers use to make judgements about students' abilities and potential performance" (p. 268). Thus, while some common assessment practices of classroom educators (for example, those described in this review above) have been described, a thorough knowledge base does not seem to have been established. This is reflected within a growing body of literature, whose authors have called for research directed towards an understanding of the processes and complexities that, for better or worse, constitute classroom assessment as it is currently practiced.

Calls for research on classroom assessment. The notion that assessment should address the classroom context and be rooted in a solid understanding of teacher practice has recently gained momentum. Measurement instruction for teachers, it is argued, ought to be informed by knowledge of what exactly it is that teachers do and why they do it (Airasian & Jones, 1993; Gullickson, 1984; Rogers, 1999; Stiggins & Bridgeford, 1985). Brookhart (1999), has underscored this by stating that classroom assessment skills should be developed as part of a larger repertoire of teacher activities, which are contingent on coming to understand precisely what content is important in preparing teachers for their work. Others have similarly expressed calls for strengthening the link between assessment and classroom practice through research that is aimed at understanding teacher application and classroom activities (Anderson, 1989, 1999; Crooks, 1988; Gullickson, 1985; McIntyre, 1990; Mclean, 1990; Shulha, 1999; Stiggins, 1990a, 1990b). Stiggins (1990b), has gone so far as to state that describing the status of classroom assessment methods is "...our most important research objective..." (p. 92). Recognition of the differences between traditional measurement and classroom assessment is reexamined through efforts towards understanding the unique nature of evaluation as practiced by teachers.

Taken together, these various initiatives and suggestions may be described as undercurrents in classroom assessment research. Gipps (1994) for example, described the tenets of assessment as moving from a psychometric approach to one that is based on a broader model of educational assessment; one that includes such factors as the influence of context. Although they disagree with Gipps over the role of psychometrics, Hattie and Jaeger (1998) also have noted a shift in measurement towards a greater interest in the interaction between assessment and classroom learning. While the integration of assessment, learning and teaching is a topic that is far from clear (Wilson, 2000a), such issues are being taken up within the research agenda (Rogers, 1999). The move is one in which assessment is considered as an undertaking to support learning rather than a summative evaluation of what has taken place (Broadfoot et al., 1991; McCallum et al., 1993). Recommendations to teachers are cited as shifting towards considering the utility and relevance of assessment for the educators who are most effected by them (Whittington, 1999). There is what might be described as an admonition that teacher

training and recommendations ought to be informed by an adequate understanding of classroom practice.

To date, it would be safe to state that a thorough and detailed understanding of the methods and processes of classroom assessment has not been attained. Although there has been a move among some researchers towards investigating these practices in the past 10-15 years, much remains unknown. Recent investigations have focused attention on the need to further examine such facets of assessment as the role of competing forces teacher face, the influence of contextual factors, teachers' implicit knowledge base, the development of grading practices and so forth (Anderson, 1999, 2000; Shulha, 1999). It is within this strand of research that an examination of educator gender was conducted for this thesis.

The Classroom Assessment Project - National

A joint team of researchers from Queen's University and the University of Victoria have carried out a long-term collaborative research program into classroom assessment, dubbed the Classroom Assessment Project - National (CAP-Nat). This involved a partnership, not only between researchers and graduate students from the two universities, but also with practicing classroom teachers (Shulha, 2000b). The overarching intent behind the project has been to gain further understanding into classroom assessment and improve teacher practice. A short description of the research carried out as part of the project is given, followed by a discussion of investigations into pre-service teacher practice that make up the direct background to this thesis.

A brief history of CAP-Nat. Shulha (2000b) has outlined the efforts at collaboration between the various components of CAP-Nat over the four years the project has been active (her timeline runs from September, 1996 through to May, 2000). She notes that the research endeavors have taken different and not necessarily synchronous forms across the various initiatives and sites, indicating the complexity of these relationships. They are described as reflecting a "multidimensional" joint inquiry that does not fit straightforward patterns of investigation (p. 4). However, for the sake of description, geography, and the particular nature of the research can be used to divide the various initiatives into paired components.

Research was divided between that conducted in Ontario and orchestrated by Queen's University and that carried out in British Columbia through the University of Victoria. At the same time, the investigations took two basic forms: direct collaboration with practicing teachers and examination of pre-service teacher practices. Of the direct teacher collaborations, a pair of educators worked with researchers at both universities, resulting in a total of four teachers involved in the project at any one time. For a number of reasons however, the ongoing collaboration with teachers at the University of Victoria was disbanded in June of 1998 (Bachor, Shulha, Anderson, Wilson, & Muir, 1998; Shulha, 2000b). This provided researchers at this university the opportunity to focus their efforts on the assessment practices of pre-service teachers.

At Queen's University, the teacher collaboration model continued for the duration of the CAP-Nat project, culminating in several recently presented papers (Locke, 2000; Lee, 2000; Notman, 2000, Petrick, 2000, Shulha, 2000a, Wilson, 2000b). These in-depth efforts focused on the rich description of practices as informed by individual educators and their respective experiences, strategies, aims and so forth. For example, topics addressed include, the effect of portfolio assessment on student led conferences (Notman, 2000), an elementary teacher's perspective of the relationship between learning and assessment (Petrick, 2000), and perceptions of knowledge and learning as they influence the classroom practices of two teachers (Locke, 2000).

Understanding classroom assessment through the active involvement of teachers as co-investigators provides a means to the direct link between research and practice that has been identified as desirable by educators (McIntyre, 1990). As part of CAP-Nat, it is one manifestation of research that was seen as promoting a new academics of assessment utilizing different approaches to research in order to come to a more comprehensive understanding of teacher's assessment practices (Shulha, 2000b). Another of these investigation strategies involved research with pre-service teachers.

Investigating pre-service teacher practice. As noted, the research components involving direct teacher collaboration, while undertaken at both universities, did not continue at the University of Victoria. Instead, the focus at the University of Victoria was concentrated on examining the practices of pre-service teachers. These investigations involved the study of individuals enrolled in the teacher training programs

offered at one of the two universities. Researchers at Queen's University carried out the initial investigation, with a modified version of the dataset being adapted by those at the University of Victoria. Together, these two research initiatives illustrate another means the CAP-Nat team have used to explore the topic of how teachers begin to make judgements about a student's performance and what influential variables may shape teacher practice. Due to their direct relation to this thesis, the projects are described in some length, beginning with the initial study at Queen's University.

Investigation of pre-service teacher assessment practices at Queen's University. A number of papers have outlined the methodology and background to the initial investigation of pre-service or novice teachers' assessment of student achievement (Anderson, 1999; Shulha, 1999; Wilson, 1999; Wilson & Martinussen, 1999). A summary of the investigation is given here to illustrate the nature of the research.

The participants consisted of 147 teacher candidates (100 females and 47 males) enrolled in the Faculty of Education at Queen's University. These pre-service teachers were all training to become educators of students in the grades 4 to 10 range. The investigation itself was introduced to the participants over five sections of a required course on teaching skills taught by three instructors at the university (Wilson & Martinussen, 1999). Each participant was given a portfolio containing various assignments or achievement products developed to represent the language arts work of a fictional student named 'Chris'. The materials were developed to simulate the real-life assignments of a grade 8 student, and included work adapted from actual students and teachers, as well as commercially prepared worksheets. In addition, the portfolios contained information on Chris's background, including, parents and siblings, school attendance, recent standardized achievement test scores and so forth. Over subsequent weeks, further information was added to the portfolios, fleshing out details about the fictional Chris with such information as his/her classroom and teacher, library readings, scores on group tasks and further assignments (see elaboration in Anderson, 1999; Wilson & Martinnusen, 1999). In the end, each participant had a portfolio containing eight language arts achievement products and several pieces of background information for the simulated student (Anderson, 1999; Anderson, 2000; Wilson, 1999).

Under the scenario that their judgements would be used by Chris's teacher as part of her/his progress assessment, the pre-service teachers were asked to read all materials contained in their portfolios, mark all the language arts tasks Chris had completed and issue a final grade for the term (Wilson & Martinussen, 1999). Thus, participants provided their views and assessments of Chris's progress in language arts (Wilson, 1999).

While all participants received the same instructions, the portfolio materials differed in a number of controlled ways. Using a randomized factorial design, the contents contained three levels of 'expectations' for the student, three levels of performance 'growth' and two levels of 'parental involvement' - gender was originally included as a variable but had to dropped due to an interpretive difficulty in the dataset, however, all other information was gender-neutral (Wilson & Marinussen, 1999). In brief, expectations for Chris were high, medium and low based on descriptions of her/his background, parent employment, and standardized test scores. Similarly, the growth variable took on the form of improving, steady or falling behind in performance, through achievement reports and variation in Chris's performance on the distributed assignments. For example, the 'improving' Chris exhibited work that began relatively poorly but progressively improved as more assessments were added to the portfolios. 'Parental involvement' was high or low, depending on information regarding the parents' participation in parent-teacher conferences. Thus, the portfolios patterned achievement results of an improving, steady or falling behind student, for whom there were high medium or low expectations and high or low parental involvement. Of the eight assignments, all but two remained consistent with the growth variable, showing improving, steady or falling behind performance. However, two of these eight (including the final exam) were identical across all Chrises. Wilson and Martinussen (1999) hypothesized that the variations in expectation, growth and parental involvement would each independently effect the reported grade Chris was given. In other words, it was thought that assessment would not be based on the singular factor of achievement as those in the measurement tradition have argued it ought to be.

Three different analyses of this investigation have been reported (Anderson, 1999; Shulha, 1999; Wilson & Matinussen, 1999) and their results are summarized here. Wilson and Marinussen found that objective evidence of Chris's performance did not in

itself determine the awarded grade. More specifically, both expectations and growth (but not parent involvement) were found to be significantly related to the final mark participants gave to Chris. Noteworthy is the finding that participants whose Chrises were falling behind in performance but high in expectation level were given scores exceeding those of Chrises with low expectations in the same falling behind category, as well as those Chrises with low expectations in the steady category. Therefore, regardless of the fact that actual performance was identical or even better, expectation skewed evaluative judgements significantly. The authors speculated the slowing performance of this subgroup was ignored or explained away by participants who were faced with disconfirming information in light of the relatively high expectations for the student. Wilson and Martinussen (1999) concluded that assessment by these pre-service teachers did not result in objective grading practices. As the authors stated, participants "...did seem to infer many of Chris's personal qualities from what they were given and used those inferred characteristics to shape their judgements about what Chris was doing and how adequate this was in terms of achievement and growth" (p. 276).

Using the data from this investigation, Anderson (1999) constructed a model of assessment that included the variables of achievement, growth and background. Anderson noted that despite the theoretical measurement perspective regarding the unidimensionality of achievement, factors other than achievement underlie teachers' scoring practices. This is reflected by the fact that the two achievement products which were identical for all Chrises showed substantial variation, to the effect that their scores tended to coincide with the products that did vary from one portfolio to another. The model therefore presented 'background' (which includes expectations and parent involvement) as a latent variable and student growth as a factor influencing marks awarded the various products. As Anderson wrotes, "The final model, then, had three components influencing the evaluation of student achievement products: background, growth and achievement" (p. 283).

Using structural equation modeling, Anderson (1999) concluded that overall, the model he had developed fit the data, and that the final grade awarded to Chris was accounted for well. Further, it was found that different products were influenced by different factors, with some seeming to be swayed more so by achievement and others to

a greater extent by growth. As Anderson put it "In fact, with these data, growth appears to have more influence over the marks awarded achievement products than does achievement. Yet the influence of an underlying factor does not have a consistent influence across all achievement products" (p. 286). The model represents student assessment as being influenced by various factors, which, as it turns out, exhibit inconsistencies that are not yet understood. The notion that a single factor such as achievement underlies the marks awarded by these pre-service teachers is not supported by this analysis. Anderson concluded that a major outcome of this investigation is the demonstration that portfolio products derived from stimulated students offer a useful approach to understanding novice teacher assessment practices, and that the use of deeper information structures (such as journals) may help address some of the unanswered questions. This recommendation is echoed in an additional analysis by Shulha (1999), as well as the subsequent research carried out at the University of Victoria.

Beyond the quantitative analyses carried out by Anderson (1999) and Wilson (1999), an examination of the qualitative data to come out of the project was undertaken by Shulha (1999). These data appeared in the form of comments and student feedback that pre-service teacher participants had written on the assignments they had marked, as well as information from an open-ended survey some individuals (N=82) completed at the end of the project. Shulha noted that although it was never asked of them, 136 of the 147 participants chose to comment on Chris's assignments, and that in fact, many considered it a professional responsibility. These comments served a variety of functions from encouragement and suggesting means of improvement, to relationship building (with the fictitious Chris) and communicating the marking criteria. Of interest is the finding that much of this data consisted of what might be described as a 'feeling factor', whereby these pre-service teachers went well beyond the given information and took into account a combination of factors in the grading process, including growth, effort, judgement about the appropriateness of an assessment instrument, a holistic impression of the student, and judgement about the utility of the school grading policy. Again, there is a sense in which assessment is arrived at through a combination of various factors, with achievement being only one of these variables.

Based on a cluster analysis of the survey data, the notion that a multiplicity of factors go into assessment was further supported (Shulha, 1999). For example, participants expressed a need to know the student personally and have a holistic understanding in order to assess them adequately. The participants indicated a strong desire to be 'fair' to Chris. As Shulha stated, the comments demonstrated that these participants found it difficult to assess Chris in isolation from the social context where learning takes place. They see assessment as shaped by the teacher, the student, the environment, the assessment strategy and so forth. Shulha reiterated that as conceptualized by these pre-service teachers, the quality of classroom assessment is multidimensional. Some of those dimensions are identified within the participant comments. However, she argued that efforts ought to persist into research that generates further insight into teachers' implicit knowledge and the conditions that situate assessment within a broader context of teaching, learning and schooling. To that end, an adaptation of this pre-service teacher investigation was carried out at the University of Victoria.

Investigation of pre-service teacher assessment practices at the University of Victoria. The analyses of the quantitative and qualitative data to come from the investigations of pre-service teacher assessment practices at the University of Victoria have been reported previously (Anderson, 2000; Anderson, Bachor & Baer, 2001; Bachor & Baer, 2000). Essentially the dataset described above is applicable to this adaptation of the research. However, some alterations are worthy of note. In particular, the portfolio assignments were modified to reflect the work of students in grade 5. In addition, there was not one student with three performance patterns, but rather three pupils (simply referred to as A, B and C), each representing one of the patterns. Thus, every participant received three portfolios to mark. Each portfolio established one of three expectation levels and contained assignments that remained consistent with this performance. The resulting portfolios presented one student who was struggling in language arts, another whose work was acceptable or average relative to the other two participants, and a third who consistently produced high quality work. In this case the students' backgrounds were consistent with their performance, creating high expectations for the high performing student, moderate expectations for the moderately performing student and so

on. Finally, there were a total of six, rather than eight, language arts assignments in each of the three portfolios. All 127 participants (108 female and 19 male) marked and established a grade for each of the three fictional students. An additional task had participants keep a diary during the course of this research project, in which they recorded thoughts and comments about the process of assessing these pupils. The analysis of these diaries is described below, following the results of the quantitative analysis.

Anderson (2000) has described the methodology, including details of the portfolio contents, in his analysis of the marks and grades awarded by each participant to the various students. From the summary statistics, it was found that there was considerable range and variation across pre-service teachers in the marks and grades they assigned. While the ranking of the three pupils was consistent with the expected pattern, there was overlap among the three students, with, for example at least five participants awarding the highest grade to the student classified as moderate achiever. Assessment, therefore, was not consistent across pre-service teachers, and again does not appear to be based purely on achievement product.

Looking at the correlation between the final mark and the letter grade awarded, Anderson (2000) noted that the relationship is not perfect. In fact, the final marks were found to account for only 38 to 75% of the variance of the assigned letter grades. Regression analysis of the six achievement products was therefore used to explore further this curiosity. It was discovered that the scores on the achievement products were better aligned with the final marks awarded to the students than with the letter grades assigned to the same pupils. Grading, therefore, was not a simple practice of mathematically translating achievement scores to final marks to letter grades, as has been suggested within the measurement tradition. Some other factors were confounding or intervening in this process. As Anderson has written, "...for the letter grade, the final, results are not well accounted for by the six achievement products" (p. 10). It appears that not only were marks not consistent for identical products across these pre-service teachers, factors others than these scores then entered into the final marks and letter grades. This is consistent with the findings reported from the pre-service teacher research at Queen's University (Anderson, 1999; Shulha, 1999; Wilson & Matinussen, 1999). In closing,

Anderson (2000) argued that an examination of the participant diaries should shed light on the unknown information these participants used in evaluating these simulated students.

Examining pre-service teacher's portfolio diaries. The analysis of the pre-service teachers' portfolio diaries was undertaken by Bachor and Baer (2000), with the intent of producing a different but complementary approach to the quantitative investigations. An examination of this data-source was seen as a possible means to accessing some of the more complex processes believed to be taking place in the assessments the participants were carrying out. The study is reported in some detail, as the second author has extended and reanalyzed portions of these data for this thesis.

As part of the University of Victoria study, all pre-service teachers were given a blank spiral bound diary or journal and asked to record their comments, thoughts and ideas regarding the assessment of the three portfolios each of them had received. It was requested that they make entries into these diaries throughout the duration of the investigation running from September to March. There were no further instructions given, and participants were free to write as much or as little as they wanted. The result, as might be expected, was a collection of 127 diaries varying in length from several lines and short paragraphs to many pages. These journals were subsequently transcribed and stored as a text file. This large bank of information was then downloaded as a single 'primary document' into *Atlas/ti* (Muhr, 1997), a qualitative data analysis program.

In order to conduct a content and pattern analysis of this relatively large collection of textual data (totaling upwards of 20,000 lines in *Atlas/ti*) Bachor and Baer (2000) developed preliminary codes on the basis of "...informed practice and the assessment literature" (p. 5). The authors then substantiated and further developed these initial codes through Glasser and Straus's 'constant comparison' method (Tesch, 1990), whereby data from the first three participants was subjected to repeated coding in order to ensure consistency of code application, comprehension in accounting for the data and elimination of redundant categories. The revision and refinement of code categories based on existing theory and practice was said to be necessary in order to produce codes that were accurately reflective of the diary content. The final codes were described as accurately and comprehensively representing the data in a manner that ensured

consistency and eliminated redundancy. Regarding consistency, the authors conducted reliability checks for the code categories. Using inter-judge agreement (Kazdin, 1982), both authors independently coded three randomly selected sections of text. Agreement rates of between 72 and 96% were reported for these reliability checks.

Three superordinate code categories were developed, along with five second-order categories and fourteen primary codes (see Bachor & Baer, 2000, Table 1, p. 6 - included here as Appendix A). These primary codes were then applied to the collected entries of all 127 participant diaries. Because the codes themselves reflected increased partition of larger clusters, several levels of analysis were possible, ranging from numerical tallies of specific categories to larger collections of related comments. Likewise, *Altas/ti* allowed questions to be asked across both codes and participants in order to construct a more complete picture of these pre-service teachers' assessment practices.

Bachor and Baer (2000) noted that the assessment of the three simulated students by these participants produced a number of noteworthy patterns in the journal data. While almost all participants (124/127) reported establishing some criteria for marking, relatively few (13/127) modified or revisited their criteria during the study. Reflecting the significance of context identified in the investigations by Shulha (1999) and Wilson and Martinussen (1999), Bachor and Baer found that 50 of these 127 pre-service teachers commented on contextual matters. These 50 participants expressed concern about the artificial nature of the assessment, a disconnection with the teacher, a lack of understanding concerning student knowledge and so forth. A variety of factors from knowing the student, teacher and classroom setting, to accounting for previous student learning, were described as being relevant to assessment by these participants. The notion that achievement ought to be the unitary basis for assessment, immune from contextual influences, appears to be questioned by many of these novice teachers. Perhaps more interesting, however, is the decision paths identified across all 127 participants regarding the interpretation of the simulated student assignments.

Based on their analysis, Bachor and Baer (2000) concluded that the pre-service teachers in their investigation tended to follow one of two main decision paths in the process of evaluating the portfolio assignments. The majority were termed "task

restricted participants" (TRP), in that they remained somewhat conservative in their assessment decisions, tending to make judgements about the student assignments and generally not extending beyond that task. Others, however, were labeled "student elaboration participants" (SEP) due to their tendency to make seemingly unwarranted judgements about the fictional students.

There were 100 participants described as TRP. These individuals tended to "...stick to the assignments, establishing criteria and evaluating task..." (p. 10). They did not make comments classifying students' abilities (such as inferring from the assignments that he/she might be special needs or learning disabled), nor did they infer that pupils' quality of life might be reflected in their work (such as considering assignments to indicate a poor home life). Of the TRP group, 22 are described as extremely conservative, simply setting criteria, marking assignments, and avoiding all comments and inferences about the students' general abilities, their affective state or other factors extraneous to the immediate marking of students' assignments. While this group comes closest to the type of practices advocated by measurement specialist, these participants reflect a minority of the individuals, for most TRP did express concerns about such matters as considering pupil feelings and emotional state -- only 22 of all 127 participants restricted their comments solely to the assessment of student product.

The individuals termed "student elaboration participants" (SEP) represent the opposite end of this group of pre-service teachers. While a majority of individuals were noted to express concerns about context, students' affective state and so on, these 27 participants made judgements that are described as "...exceeding the evidence provided" (Bachor & Baer, 2000, p. 11). This entailed inferences about the impact of such factors as a student's home and social life on their achievement products. Others suggested that one or more of the pupils might have a having special education needs based on the limited evidence provided in the portfolios. As Bachor and Baer concluded about this set of participants, "They seemed prepared to base their assessment decisions on some undefined assumptions. They appeared to have an intuitive basis for the judgements they made and speculated willingly about the three hypothetical learners and their families" (p. 15). This tendency was disconcerting, due to the potential impact of such far-reaching judgements. Educational decisions and evaluations based on scant evidence, limited

encounters with a student (the portfolios in fact offered no encounters and very limited information) or some intuitive sense of the child and his/her family history would not seem to offer any tangible foundation for such decision making.

The pre-service teacher diaries analyzed by Bachor and Baer (2000) contained comments consistent with the findings of other studies in this classroom assessment project (for example, Anderson, 1999; Shulha, 1999; Wilson & Martinussen, 1999). These participants identified such factors as classroom context, knowledge of the student and concerns about student feelings as important issues in their evaluation practices. Assessment does not appear to follow a straightforward translation of achievement to mark to grade (Anderson, 2000), and to some extent, these identified variables point to reasons why this may be the case. This again stands in contrast to the technical concerns and unidimensional nature of assessment advanced within the measurement tradition. The finding that some of these pre-service teachers made further judgements described as unsupported by the evidence, indicates the importance of understanding assessment and generating informed teacher practice. As these diaries offered a rich set of complex data, further analysis was thought to be worthwhile. It is with this goal in mind that a gender-based analysis of the data was conducted for this thesis.

Educator Gender as a Variable in Classroom Assessment

To this point in the review, it is hoped that the basic background to classroom assessment research has been elucidated. It has been argued that classroom assessment as conducted by teachers (particularly at the lower grade levels) is not synonymous with the practices of traditional measurement. Further, it has been proposed that research directed at increasing an understanding of classroom assessment as it takes place among educators represents a vital and necessary avenue of investigation. To that end, the calls for increased research in this direction were noted. The CAP-Nat investigations, described in some detail, are indicative of these undertakings. The pre-service teacher studies, which form the immediate background to this thesis, are the most relevant examples here, as they suggest novel directions for future research in this area. In particular, the large and complex amount of information contained in the diaries studied by Bachor and Baer (2000) appeared to invite further analysis. The second author noted that an examination of these data from the perspective of pre-service teacher gender would provide an

opportunity for a potentially unique analysis of a variable that is not well understood. However, before turning to this analysis (described in the following chapter) it is necessary to detour this review in order to describe the relevant literature on educator gender as a variable in classroom assessment. The details of the pre-service teacher investigations noted above will be returned to when describing the methodology of this thesis.

Questions regarding the influence of gender on assessment have typically focused on students as the source of variability and discrepancy in performance. For example, Okpala (1996) has cited numerous investigations of gender bias in classroom interactions, whereby pupil gender was found to be a factor in teacher feedback, positive reinforcement and so forth. Similarly, investigations of performance in mathematics, perhaps the most 'gender-examined' discipline, have concentrated on discrepancies of test scores between males and females of various ages (Fennema & Peterson, 1985; Wiles, 1992). While pupil characteristics no doubt contribute to the complexity of assessment, particularly in the natural setting of the classroom, this accounts for only a portion of the variance. Contextual factors as well as educator traits ought not to be overlooked as sources of variability acting on the classroom assessment process. In fact, teacher attributes such as gender have been specifically identified as influential variables in student/teacher interactions (Lindow, Marrett, & Wilkinson, 1985). Consideration of teacher gender asks whether male and female educators differ in significant ways when it comes to classroom assessment.

As noted above, the assessment literature contains a substantial amount of research on student gender as related to various outcomes and performance discrepancies. The same however cannot be said regarding investigations of gender differences among educators. Hopf and Hatzichristou (1999) stated, "Although there is a wealth of empirical studies examining the effect and the correlates of student gender in school, teacher gender has rarely been a research focus" (p. 1). Likewise, Goodwin and Stevens (1993) referred to the dearth of information about the differences between male and female teaching practices and methods in higher education. The apparent paucity of information concerning educator gender as a variable in assessment practices is taken as an information gap worthy of investigation.

As investigations of educator gender and assessment proved to be severely limited, research on a variety of topics thought to inform teacher practices and consequently assessment, were seen as possible sources of information. Seven categories were identified as being sufficiently related to classroom assessment that they might illuminate the role of teacher gender as a variable here. The categories include math/science education, classroom interaction, interaction with other variables, disruptive behaviour/special education referral, international investigations, higher education and performance evaluation. These seven categories and their relation to assessment are described in turn below.

Math/Science education. Gender differences associated with learning and achievement in mathematics have been the focus of a substantial number of investigations, and as Wiles (1992) described, this has typically involved differences in achievement outcomes across age groups. Common to these investigations is the inclusion of pupil gender as the fundamental variable acting upon the measured outcome. Rarely included is an examination of teacher gender. However, in an investigation of middle school mathematics classes, Wiles examined the possible interaction between student and educator gender on scores assigned to problem-solving tasks.

Proceeding from the hypothesis that teachers will reflect stereotypical expectations of male success in mathematics, Wiles (1992) had educators assign scores to problems solved by students identified as either male or female. Educator gender was included as a possible variable effecting ratings given the assignments. Wiles found that when asked to mark problems submitted by male or female students, gender of the teacher was not significant either as a main effect or in interaction with student gender. While the non-significant finding is notable, the involvement of a disproportionately small number of male teachers raises a call for further research, and as Wiles noted, a lack of gender bias was not necessarily demonstrated. Even in the realm of mathematics, where gender analyses has been most prominent, firm conclusions regarding assessment practices are not clear. This caveat regarding the cautious interpretation of results and the call for further research is a salient theme throughout the literature cited (in mathematics, see for example, Fennema & Peterson, 1985). The area of teacher-student questioning in science classes is no exception.

Following mathematics, science performance has received its share of gender based investigations, and here too the focus has been primarily on disparities between male and female outcomes on various tests and measures, as well as on interests in higher education courses and career areas (Morse & Handley, 1985). An adjunct to this work has examined the discourse and interaction that takes place in science classrooms, the argument being that this is at the heart of the learning experience and hence directly effects student performance (Barba & Cardinale, 1991; Morse & Handley, 1985). Again, however, educator gender has not been included in the majority of this research. In an exception, Barba and Cardinale observed teacher-student questioning interactions as they occurred across gender of both teacher and student. They found that although female students had fewer interactions and received less attention from science teachers overall, there were also significant differences between male and female educators. Males interacted more with students identified as 'targets' (pupils receiving four or more interactions per class) than fellow female teachers. In fact, male teachers interacted with 70% of the target students, while female teachers were observed to interact with only 30%. Although the results do not portray straightforward or easily interpreted interaction effects, the significant differences across educator gender indicate that this variable may play a role in classroom interactions and consequently the learning experiences of students.

Although math and science are the most heavily gender examined of the formal educational disciplines, conflicting and uncertain findings from the all too limited studies that have included teacher variables indicate the need for further investigation. Continuing with the stance that teacher-student interactions offer insight into possible gender based differences concerning educator attitudes and behaviours toward students, including those informing assessment practices, classroom interaction studies from other domains are thought to offer additional insight.

Classroom interaction. As noted above, classroom interactions are seen as cousins to assessment practices in that they may provide information regarding teacher attitudes and behaviours towards students. In addition, interactions are often a primary and immediate sources of feedback (Okpala, 1996), akin to informal assessment. The exchange of information of all sorts between students and teachers is described as an

essential element to the learning process (Barba & Cardinale, 1991), and is considered a fundamental facet of the larger assessment context. In terms of studies related to classroom interaction, student gender has been given primary attention, as it was in research concerning math and science (Brophy, 1985). Pupil gender has been found related to everything from seating position and question responding to participation level; with boys generally receiving the lion's share of both positive and negative attention (Brophy, 1985; Okpala, 1996; Stake & Katz, 1982). The inclusion of teacher gender in such investigations has been a rarity, although a few attempts have been made to cull out this factor.

In an examination of teacher attitudes and behaviours toward male and female pupils, Stake and Katz (1982) included gender as a variable in their observation of grade four, five and six classes. Citing inconsistencies in earlier investigations, the authors pointed out the lack of clarity regarding classroom atmosphere and interaction across teacher gender. From their own work they concluded that while boys may receive more reprimands than girls, female teachers were significantly more positive than males in their attitudes and behaviours towards pupils. Female teachers tended to give more sympathy responses and provided more encouragement than their male colleagues, who gave more soft reprimands and rated children higher on descriptions of poor achievement and noisy behaviour. Although by no means indicative of irrefutable gender differences in assessment, the findings hint that further examination of these factors seems appropriate.

Investigating interactions in seventh and eighth grade classrooms, Good, Sikes and Brophy (1973) also found that female teachers encouraged students more than their male counterparts. They stated, "Female teachers often praise students following correct answers, while male teachers were more likely to give process feedback following both correct and incorrect and answers" (p. 77). The type of feedback offered students would seem relevant to assessment at the unstructured level, and speculation might have it that this would influence more formal processes as well. Despite the findings that female educators appear to teach differently, Good et al. concluded that teachers of both genders essentially treat students similarly. In a synthesis of the work on teacher-student interaction, Brophy (1985) likewise concluded that male and female teachers tend to be

more similar than different in their interaction with students. Nonetheless, he stated that future research is required, with a specific view towards "thicker description with more attention to qualitative aspects of classroom events..." (p. 137). Investigations regarding the interaction of teacher gender with variables other than student gender may provide some background to further studies.

Interaction with other variables. Beyond the two by two analyses of teacher-student gender, various investigations have attempted to unfetter the role of such diverse characteristics as self-concept, locus of control and motivation as they might pertain to educator gender differences. Early investigations by Brandt and Hayden (1974) and Brandt et al. (1975) employed undergraduate students in the role of 'teachers' in order to examine ascribed performance judgements across gender and with other interacting variables.

Noting that previous investigations had seldom revealed consistent differences, Brandt et al. (1975) included gender in their study of attitudes and attributions of causation concerning student performance and ascription of motivation. They found that in general, students described as successful and highly motivated received more favourable ratings for skill, effort and personality traits. Across gender, male pseudo-teachers rated students' skills higher and gave "well motivated" pupils relatively higher ratings than their female cohorts. In addition, female subjects with a high internal locus of control assumed more responsibility for student performance than female subjects with an external locus of control. This difference was not found among male subjects.

Using similar methodologies, Brandt and Hayden (1974) had previously noted gender discrepancies with ratings of motivation, particularly concerning successful/unsuccessful students said to be over or underachievers. They concluded, "It appears likely that basic differences between male and female preferences or attitudes exist prior to the teacher-student interaction and that these differences influence teacher-student interactions"(p. 313). Despite these noted differences, and those described in their following investigation, the authors stated that they perceive teacher sex differences to be relatively unimportant overall (Brandt et al., 1975). While teacher attitudes towards performance levels undoubtedly inform the assessment process, the significance of gender differences does not appear to be clear-cut or well established.

Several investigations of further variables have produced significant, albeit mixed results when educator gender was included. For example in a study of sex roles and evaluation, Bernard (1979) found interactions between the gender of teachers and students when evaluating written performance. Overall, Bernard wrote that the pattern of interaction suggests a "...cross-sex bias when teachers evaluate the written performance of students" (p. 561). This played out at the level of sex-roles across curriculum, with female teachers evaluating male students' answers on a physics assignment more highly, while male teachers evaluated the identical assignments by female students higher. Other significant differences were noted, including the attribution of greater logic to an identified male student by female teachers. Male and female teachers are described as differing in their perceptions of student sex role behaviour, and these differences are revealed at the level of evaluation of student assignments.

Studying the interaction of grade level and teacher gender on beliefs concerning student decision making in physical education classes, DeVoe (1990) reported significant gender by grade differences. Grade level was found to be a factor only for male middle school student teachers, who reported that pupils should make fewer decisions than was reported by either female middle school student teachers or male high school student teachers. DeVoe argued that the differential effects are accounted for by an interaction between student teachers' gender and grade level of placement. Whether the beliefs teachers hold regarding the allocation of responsibilities towards students decision making is directly related to assessment attitudes and practices is not clear. The interaction of grade level differences adds further confusion.

Looking at self-concept among teachers of mathematics, Relich (1996) pointed out that educator variables have received relatively little attention when compared to investigations of student differences. Including gender in his study, Relich found that overall, both male and female teachers encouraged boys more than girls and that educator self-concept was a significant factor in the attitudes, teaching and self-perceptions of the participants. Gender differences did emerge regarding attitudes and pedagogical approaches to mathematics, however, Relich described these as relatively minimal in comparison to the impact the differences noted across self-concept profiles. The extent to

which these gender interactions can be dismissed is uncertain. This is particularly true when including such contentious factors as race/ethnicity.

Academic performance disparities, dropout rates and the under-representation of minorities in teaching/role model positions have made investigations and interventions addressing racial and ethnic concerns relatively high profile. Using longitudinal survey data, Ehrenberg and Goldhaber (1995) examined the interaction of teacher-student race, gender and ethnicity on evaluation. How teachers subjectively related to and evaluated students, as well as student learning measured by standardized tests was included. The match between teacher and student race, gender and ethnicity was not found to be related to pupil learning (standardized assessment outcomes), but was identified as a significant determinate of teachers' subjective evaluations. For example, white female teachers evaluated their white female students significantly higher than white male teachers. As noted by the authors, these results are open to conflicting interpretation, especially when considering how they might inform relationships and performance indicators outside of the school setting. The importance of subjective evaluations and potential gender differences acting upon them should not however be overlooked, particularly when considering the daily assessments teachers conduct in the confines of their classrooms. Teachers' perceptions, be they on the direct assessment of student outcomes, or on more complex social/behavioural interactions with students, would seem to be areas of understanding relevant to informing the classroom assessment process.

Examining teacher perceptions of the social behaviour of elementary students, Rong (1996) included the interaction of both race and gender on the assessment of pupil behaviours. Rong argued that teacher perceptions of students' social behaviour are important considerations because perceptions are so closely related to expectations. In turn, teacher perceptions and impressions are said to be predictive of not only future behaviour, but also achievement. As a main effect, female teachers were found to rate female students' social behaviour more positively, regardless of race. Concerning interactions, black female teachers were described as not being influenced by student race, while white female teachers rated black students considerably lower than white students. White male teachers tended to rate white male and female students relatively more equally than their black or white female colleagues did (black male teachers were

excluded from this analysis due to the small number). The main and interaction effects are not easily interpreted and the noted differences do not follow straightforward patterns, although the higher ratings female teachers gave female students echoes the results described by Ehrenberg and Goldhaber (1995). In the end, Rong stated that the "teachers' perceptions of student social behaviors are a result of complex interactions of students' and teachers' race and gender" (p. 261). Gender differences in the perception and interpretation of student behaviour are thought to have very real impacts, not only on future teacher/student interactions, but also on student achievement, and hence classroom assessment. This may be particularly notable in regards to the perception of 'problem behaviours' and special education referrals.

Problem behaviour/special education referral. As mentioned above, educator perceptions of student behaviour and the interactions between pupil and teacher are considered essential, albeit complex, elements of classroom assessment. By extension, the interpretation of problem behaviours and the decision of regular classroom teachers to make special education referrals are seen as related to this larger assessment context. Investigations of educator tolerance and perceptions of problem behaviour, as well as the variables involved in special education referral, have been spurred on, and undoubtedly made more relevant, by the move towards the inclusion of all students within the regular classroom. Of these investigations, a limited number have included educator gender as a possible interacting variable.

Using naturalistic observations in the active classroom setting, Ritter (1989) examined the behavioural ratings assigned by regular and special education teachers to pupils identified as severely emotionally disturbed. Not surprisingly, regular classroom teachers perceived greater problem behaviours and lower school competence than their special education colleagues. Regarding gender, a significant correlation was found whereby female teachers rated behaviour problems higher in general than did their male counterparts. More specifically, female teachers were more sensitive to 'externalizing' or overt physical types of behaviour, as well as to overall behaviours, in the regular education classroom. Ritter concluded with the caveat that statistical treatment beyond a correlation analysis was not possible, but considering the implications of increasing

'inclusion', further investigation seems justified. To some extent, this has been indirectly supported through two investigations of special education referral.

The decision to refer an individual for special education services typically follows a model by which student characteristics are the paramount, if not only, factors taken into consideration (McIntyre, 1990). However, student characteristics are noted to explain only a small portion of the variance in the identification of pupils with learning disabilities, while educator variables such as gender are thought to be influential factors requiring further investigation (McIntyre, 1988). To that end, McIntyre (1988, 1990) has included teacher gender as a fundamental variable in the study of referral rates.

In his initial investigation, McIntyre (1988) found that the level of problem student behaviour interacted with teacher gender, to produce significant differences in referrals. McIntyre wrote, "when students with high levels of problem behavior are considered for referral, male teachers are much more likely than females to decide *not* to refer [author's italics]. However, when students with low levels of problem behavior are considered for referral, the decisions of male and female teachers do not differ" (p. 382). In a reanalysis of the data to account for discrepancies in teacher standards and aggression of student behaviour, McIntyre (1990) extended the results to conclude that female teachers had higher referral rates for students with high levels of aggression. Again, it was only in regards to aggressive behaviours that gender differences were noted. Similar to the findings of Ritter (1989) described above, McIntyre (1988, 1990) concluded that differences in educator gender are significant for externalized problem/aggressive behaviours. How these differences may play out in regards to assessment practices is not at all clear, but in light of the increased inclusion of all pupils in the regular classroom, the question would appear to be a relevant one. Providing further fodder to the issue of educator gender and classroom assessment is the consideration of investigations from traditions outside of North America and the UK.

International investigations. The inclusion of international investigations must naturally be a cautious one considering the cultural differences and the dangers involved in any attempts at generalization. Described here are a limited number of studies that specifically attempted to address educator gender differences in the classroom context. They are included as further contributions to the limited information available on

educator gender and classroom assessment, and as additional sources indicating the need to investigate this topic.

Perhaps the most extensive recent research project addressing questions of gender related influences in the active classroom was conducted by Hopf and Hatzichristou (1999) on a sample of over 1800 public school teachers in Greece. Highlighting the findings of colleagues from other countries, the authors began by stating that "astonishingly little research has focused on the teacher gender variable" (p. 3). They concluded however, from their own review of the literature, that gender effects the perceptions and behaviours of teachers and students in "complicated and interrelated ways" (p. 3).

Dividing participant results by elementary and secondary status, Hopf and Hatzichristou (1999) identified several significant teacher gender differences in the assessment of student competence. At the elementary level, female teachers evaluated children's adjustment as less problematic on various aspects of academic and social functioning, while simultaneously considering boy's interpersonal behaviour more positively than male teachers did. In secondary schools, however, male educators assessed children's overall interpersonal behaviour as less problematic than did fellow female teachers. If student gender is included, secondary teachers evaluated the interpersonal behaviour of pupils of the opposite gender as less problematic, with male teachers evaluating female students as less problematic and female teachers assessing male students as less problematic. The interaction between educator gender and student grade level supports the hypothesis put forth by Hopf and Hatzichristou, that teacher gender may be more informative as part of an interaction effect with other independent variables. This complexity and interaction of variables is a prevalent theme, whether in Greece, North America or Pakistan.

An investigation of teacher gender and student achievement in Pakistan (Warwick & Jatoi, 1994) produced similar results concerning the interaction of various factors. Although significant main effects were found, with students of male teachers in mathematics having significantly higher achievement scores than students of female teachers, this was informed by the inclusion of other variables. For example, Warwick and Jatoi wrote, "teacher gender and its interaction with urban residence proved to be the

best predictors of mathematics achievement..." (p. 384). Thus, while teacher gender was found to be a significant variable, its interaction with other factors is described as being even more informative. The distinctive barriers identified for female teachers in rural Pakistan may not apply to educators from other countries, but this does serve as a reminder that simple notions to understanding student assessment often fall short when considering the complexity of pupil, teacher and contextual factors. Historically this has meant the examination of student variables, with relatively little regard for possible educator and contextual differences. Teacher gender is identified as one of the 'overlooked' differences in the research (Hopf & Hatzichristou, 1999; Relich, 1996), its absence noted not only in public schooling, but in investigations of higher education as well (Goodwin & Stevens, 1993).

Higher education. Moving up the ladder from elementary, middle and secondary school, higher education in colleges and universities offers further insight into pedagogical processes and assessment practices. Although instructors employed by higher learning institutions are often not educators per se, and direct comparison to public schooling is tenuous at best, information gleaned here offers another source of information into potential gender differences. However, consistent with the information described under other headings above, very little research has been conducted on educator gender issues (Goodwin & Stevens, 1993; Henderby & Diamond, 1998) and even the inclusion of student gender has produced conflicting results (Centra & Gaubatz, 2000). With limited investigations and cautious interpretation, some significant findings are worthy of mention.

How instructors teach and what they consider 'good' teaching presumably informs assessment practices, not only because assessment is often built into teaching style and process, but also because it affects student achievement and outcome. In an effort to examine teaching practices, Goodwin and Stevens (1993) sampled university faculty, asking them what they considered to be 'good' teaching and what were thought to be appropriate outcomes of that teaching. Included in the investigation were questions of grading practices that were used and favoured. The authors reported several significant gender differences, but stated that overall, they were surprised at the small number of variations found. Relevant to assessment practices, female respondents agreed

more strongly than their male colleagues that higher-order thinking skills, concern about student self-esteem and seeking a variety of learning levels via exams and discussions were of concern for 'good' teachers. Similarly, the mean of female instructors was significantly higher for the notion that competition for grades in a class should be avoided. There were, however, no significant differences for preferred teaching and grading practices, and similar views were reported regarding 'good' teaching and appropriate outcomes. The influence of the noted differences on assessment as related to possible outcome discrepancies and variation within methodologies was not addressed, and questions beyond the use of similar practices remain unanswered.

In a more recent investigation, Centra and Gaubatz (2000) included instructor gender and teaching practices in an examination of bias among student evaluations of teaching. They reported significant differences in classroom practices and methodologies for male and female professors. Female instructors taught differently, in that they lectured less and used discussion to a greater extent than their male counterparts, even when class size was held constant. In fact, male instructors were almost twice as likely to use lectures, although the majority of both groups used a combination of lecture and discussion. Unfortunately differences in assessment practices and outcomes were not included and further investigation is required to determine whether such differences in teaching practices are noted at the level of assessment.

Focusing on undergraduate and graduate managerial finance courses, Henerby and Diamond (1998) examined the impact of student and instructor gender on grade performance. Consistent with authors from related fields, they noted that the results of previous studies have been mixed. However, in their own investigation, significant differences were found, with students of female instructors demonstrating higher grades, lower withdrawal rates and greater percentages of pupils passing the courses. Henerby and Diamond considered their results an important initial step to exploring gender issues in finance education. Whether these findings are related to those describing female teachers as more positive regarding classroom interactions (Good et al., 1973; Stake & Katz, 1982) is unclear and remains purely speculative. In higher education there appear to be some identified gender differences in educational practices and outcomes, although the degree and significance of these differences is uncertain. Extending the examination

of university instruction, and business courses in particular, research on performance evaluation is seen as a further source of information that might inform queries concerning classroom assessment and educator gender.

Performance evaluation. Performance evaluation may be considered a 'real world' relative of classroom assessment in that it attempts to make judgements concerning performance on a task; often in regards to work related measures. The distinction being that information is collected from various outcomes (for example, sales records), or through direct observations of the task itself. For example, managerial evaluations frequently involve data regarding profit margins and sales as well as observation of interactions with clients. Nonetheless, potential evaluator gender differences in performance assessment are seen as a relevant source of information, possibly contributing the minimal and mixed results noted for classroom assessment investigations.

Staying within the university setting, Gundersen et al., (1996) included evaluator gender in their study of impression management influences upon performance appraisal ratings. As consistently identified in the research and reviews cited above, the authors here also stated that the literature to date contains conflicting results regarding gender and performance evaluation. In their own investigation, they reported significant differences across gender, with females giving higher ratings to high performers and lower ratings to low performers, when compared to male evaluators. These results were significant only when taken as interaction effects with employee performance level. Main effects for gender were not found. This would seem congruent with findings from other research areas already discussed, which describe gender as a complex interacting variable (for example, DeVoe, 1990; McIntyre, 1988; Warwick & Jatoi, 1994).

The findings reported by Gundersen et al (1996) are also consistent with those of an earlier investigation in which bank managers evaluated the performance of an assistant manager applying for promotion (Nevill, Stephenson & Philbrick, 1983). Although male and female bank managers were found to share common views of successful behaviour required for managerial positions, interaction effects were found concerning evaluator gender and the success of the applicant. Nevill et al. stated, "females clearly rated successful applicants more positively and unsuccessful applicants more negatively [than

male managers did]"(p. 169). Again, the influence of gender is found as an interaction effect and its role is more complex than simple assumptions might postulate. While not all interactions produce significant differences (for example Pulakos, Oppler, White & Borman (1989) reported minimal variance for the interaction of race and gender on performance evaluation), the difficulty is to identify relevant effects. In the complex and often varied world of classroom assessment, this is not an easy or straightforward task.

Summary In setting the background and historical framework for research into classroom assessment, the role of traditional measurement comes to the fore. Its prevalence and prominence in assessment research and the recommendations that have shaped practice are well known. However, it is argued that much of classroom assessment falls beyond the bounds of traditional measurement, and that indeed these are often unique undertakings with different aims, intentions and methodologies. It is proposed that research directed at furthering our understanding of classroom assessment as practiced by educators is both necessary and essential to providing a firm knowledge-base about what it is that teachers actually do. The calls for increased research in this direction were noted above.

Classroom assessment and the surrounding interactions between teachers and students are described as complex processes involving a range of interacting variables including: student factors, teacher attributes and environmental/contextual differences (Lindow et al., 1985; Mavrommatis, 1997). Educator gender is considered one of a multitude of variables that ought to be accounted for as part of the move towards understanding this complexity. To date, few investigations have considered the influence of educator gender on assessment and related activities (Goodwin & Stevens, 1993; Henebry & Diamond, 1998; Hopf & Hatzichristou, 1999), and several reviews have pointed out that the available research has historically generated conflicting and inconclusive results (Brandt et al., 1975; Centra & Gaubatz, 2000; Gundersen et al., 1996; Stake & Katz, 1982).

Across the seven categories of information considered in this section of the review, several common themes emerged. As noted above, educator gender has largely been overlooked, while the brunt of research has undoubtedly been directed at student variables (see for example the numerous citations in Brophy, 1985; Fennema & Peterson,

1985; McIntyre, 1988, 1990; Okpala, 1996; Stake & Katz, 1982). Although the inclusion of student attributes, including gender, is a necessary element to the understanding of classroom assessment, this considers only one among several sources of variability. Throughout, the need for further investigation, and in particular, Brophy's (1985) call for thick description and increased attention to the qualitative aspects of classroom events is noted. It is suggested that a greater understanding of the complexity of variables that effect teacher practices is a necessary aspect of assessment research. Educator gender is one of the variables that ought to be considered. It is with the goal of conducting exactly this sort of qualitative investigation that the data described in Bachor and Baer's (2000) study of pre-service teacher diaries are analyzed for this thesis.

Chapter 3

METHODS

Design

The research design for this study was based on a subset of data collected as part of an investigation of pre-service teachers' classroom assessment practices (Anderson, 2000; Bachor & Baer, 2000, Anderson, Bachor & Baer, 2001). A group of 127 undergraduate students enrolled in the teacher training program at the University of Victoria participated in the original investigation. Participants received portfolios containing background information and work samples of three simulated grade 5 students. Portfolio contents included the work of students whose expectation levels and performance were simulated to reflect high, moderate and low achievement levels. Each participant was asked to score all three students' assignments and submit a final grade for the fictional pupils. In addition, participants were asked to record a diary for the duration of the investigation. The diaries contained the thoughts, processes, criteria, concerns and so forth that these individuals had in regards to their assessment of the three simulated students. A selection of these diaries distributed across female and male participants (N=38) were used to conduct a gender-based analysis of assessment practices for this study.

The portfolio contents used to replicate the background and work of the three grade 5 students contained a variety of features and assignments. Background descriptors of the classroom and school were included, as were six language arts work samples. Both the work samples and background materials were consistent with a pattern in which the pupils represented high, moderate or low expectation and performance levels. The pupils themselves were simply referred to as students A, B and C, and all participants received a portfolio from each of the three students. Participants evaluated the work samples as well as a final exam and recorded marks and letter grades on a record sheet for each of the three students. Final marks and letter grades were also recorded. Throughout the duration of the investigation, participants recorded a diary of their assessment practices.

The investigation itself was introduced in one of three required courses taught by two different instructors at the University of Victoria. It ran from September to March, and participants were able to write in their diaries throughout this period. Instructions concerning diary contents were limited to encouraging the participants to record their comments, ideas, marking criteria or anything else they wished regarding the assessment task. The diaries or journals thus contain the thoughts, ideas and comments these pre-service teachers had about the assessment of the portfolio contents, the fictional students, grading practices, recommendations and so on. Data from a sub-sample of these diaries were analyzed here.

For this study, a gender-based examination of the pre-service teacher diaries was conducted. An 'interpretational analysis' (Gall, Borg & Gall, 1996) of the diary data was used to compare the journal entries of all male participants with those of a randomly selected sample of female cohorts. According to Huberman and Miles (1994) this is best typified as an investigation of 'descriptive understanding'. In particular, a series of stratified code categories, designed to allow for comprehensive and consistent content classification (Bachor & Baer, 2000) were established. Preliminary development of these categories was based on a combination of informed practice and extensive revision and modification through repeated coding and re-coding of the diary text in order to accurately represent the data. The gender-based study conducted here used various levels of analysis from general comparisons of diary length and content to an analysis of the stratified code categories, as well as further investigation of identified content clusters. Through the comparison of diary contents between male and female participants, questions regarding various assessment practices were examined.

Participants

The participants for this study were drawn from a larger investigation conducted as part of a Canadian classroom assessment project (see Anderson, 2000; Bachor & Baer, 2000; Anderson, Bachor & Baer, 2001). From a pool of 127 (108 females and 19 males) undergraduate students enrolled in the teacher training program at the University of Victoria, a sub-set of journals completed by 19 female and 19 male pre-service teachers were selected for inclusion in this investigation. While the journals of male participant reflect a 'convenient sample' of information from available individuals, the diaries of

female cohorts were selected using 'purposeful random sampling' (Gall et al., 1996). A table of uniform random numbers (Howell, 1995) was used to select the 19 female participants whose diaries would be compared to those of their male counterparts. This sampling methodology was chosen with the intent of avoiding selection bias among the disproportionate number of journals from female participants. Thus, diaries from a total of 38 pre-service teacher participants were included for analysis in this study.

As noted above, all participants were undergraduate teacher-training students enrolled in one of three courses at the University of Victoria. These individuals were in their 'professional' or fourth year of university. Although they were still taking courses, at this stage their program also involved school-based practicums. Therefore, these participants had some experience of classroom life and presumably some exposure to assessment and the other routines that teachers face.

Instrument

Data for this study consisted of the complete entries from sampled participants' portfolio diaries/journals. The diaries were written in spiral bound notepads, distributed one per participant as part of the original portfolio investigation. These notepads contained over 200 lined pages to allow for as many entries and as much writing as individuals wished -- no one used all the space provided. While the diaries were considered a required portion of the overall study, specific guidance and instruction were avoided in order to generate data as unfettered as possible by expectations and other content suggestions. Further, these diaries while part of the required task were not graded and participants did not receive marks for them.

As might be expected, the 38 diaries sampled for analysis in this study varied in regards to style, length and so forth. For example, they differed in length from those containing a few pages of detail, to others that ran on for many pages, describing extensive accounts and particulars. Actual length of text ranged from 309 to 54 lines when transcribed for analysis. Likewise, stylistic variations altered from point-form recommendations to personal narratives and lengthy discussions of various topics. The entries differed widely in content, as instructions guided participants to simply record their thoughts, processes, criteria and deliberations concerning the marking task. Scores allotted to work samples and final grades were recorded on a separate sheet, although

many participants also included them within the space of their diary entries (see Anderson, 2000 for an analysis of these scores). Although the diaries generally contained criteria for marking, the thought processes and concerns around the marking task, as well as descriptions about particular assignments, there was a diversity of detail and focus. This is consistent with the complexity expected from data collection methods and information sources that produce thick description and largely unstructured information (Gall et al., 1996; Miles & Huberman, 1984).

The handwritten journals were transcribed verbatim into a 'text file', delineating participant boundaries but storing the data as a single file for easier manipulation. Finally the diaries of all participants were imported into *Atlas/ti* (Muhr, 1997) -- discussed below -- for data management and analysis.

Limitations

Before outlining the methods of data analysis, it seems prudent to note some limitations to this study. In particular, the sample was limited to participants enrolled in the teacher training program at the University of Victoria, and therefore, generalizability -- or what Marshall & Rossman (1989) alternatively referred to as transferability -- beyond this population is suspect. This is underscored by Huberman and Miles (1994), who have stated that while identified patterns in qualitative data may be common across cases analyzed, they remain particularistic, and do not in and of themselves warrant generality.

A further caveat of this particular study stems from the fact that the data represent responses to an artificial task. That is, the assessment and grading participants were asked to reflect upon in their diaries was conducted for fictional students outside of the classroom context. The portfolio contents were based on actual assignments and adapted from student and teacher work, however, participants were aware that their evaluations would not actually be used in real classrooms or in genuine student reporting. Thus, while the portfolio data simulated authentic assignments and were based on actual practice, it cannot be known if the diaries reflect the thoughts and ideas that these individuals might have in an actual classroom assessing real students.

Data Analysis

Data analysis of relatively unstructured, thick and varied information, such as that contained in the pre-service teacher diaries, has been described as resistant to neat and tidy processing (Richards & Richards, 1994). As Marshall and Rossman (1989) have written, the underlying goal of analysis in this context is to bring order, structure and meaning to the data. They further described the procedure as one of moving from organization, to generating meaningful categories, to identifying prominent patterns/themes and testing hypotheses. Huberman and Miles (1994) have similarly described the process as a search for regularities and the comparative analysis of underlying similarities and systematic associations. Using analytic software and the development of code categories, organization and retrieval of information takes shape. This in turn allows for meaning to be constructed from complex information by pulling together material and permitting analysis (Miles & Huberman, 1984). The evolution of the diary contents into a manageable analyzable format -- by means of software -- through to the refinement of descriptive code categories -- the term favoured by Miles and Huberman -- provided the basis for the comparative analysis conducted in this study.

Analytic software. The use of computer technology has moved beyond data management for information storage and retrieval, although this remains a salient function of many programs. For this study, the qualitative data analysis program *Atlas/ti* (Muhr, 1997) was used, not only to manage the data, but also to assist in the identification of categories within the text and the comparison of patterns and themes across subjects and hence between genders. Richards and Richards (1994) have described *Atlas.ti* in their examination of computers in qualitative research. The program is specifically designed to support the code and retrieval process that the authors have called "...certainly the most widely recommended technique for management of rich and complex records" (p. 447). Further, the program aids in the pursuit of patterns and comparison of text segments that allow for theoretical links between various categories of information.

Essentially, *Atlas/ti* allows for large amounts of textual data to be imported and saved as 'primary documents' within the program. Coding, categorization and so forth take place 'on' the text, in a fashion analogous to a translucent template placed atop the

written records. The development of codes or categories refined and established to reflect the text are subsequently applied to the document, parsing the data into meaningful and comparable units. As Richards and Richards (1994) have noted, this is the basis of pattern identification and comparison, such as that which was at the center of this analysis. Of particular significance is the function *Atlas/ti* serves in generating a connected network of links among various entities (Huberman & Miles, 1994). For example, it is possible to examine relationships between text segments, participants, codes and so on through the use of Boolean (and other) operators that allow for queries such as; 'produce all A that are part of C', 'find instances of X that are not contained by Y' and 'how many participants stated P but not Q'. This extensive permutation of questioning renders pattern investigations and comparative analyses possible with complex data of the sort contained within the pre-service teacher diaries. Richards and Richards concluded that the main value of *Atlas/ti* is in investigations with a number of topics, some major identifiable characteristics and relationships between these. This was thought to be indicative of the data analyzed here.

Categorization and coding. Underpinning the investigation of patterns and comparative analyses are what Huberman and Miles (1994) have termed "...a set of conceptually specified analytic categories" (p. 431). The codes or categories of common themes identified serve both retrieval and organizational ends, and thereby set the stage for analysis (Miles & Huberman, 1984). Identification of these 'regularities' (Huberman & Miles) is a prerequisite to making sense of complex data.

Codes and code categories for the entire set of journal entries from the original investigation (n=127) were initially developed by Bachor and Baer (2000), and are listed in Appendix A. These 'descriptive codes' (Miles & Huberman, 1984), although based on the literature and informed practice, were solidly based on the data -- through repeated coding and re-coding of a sample of the diary text -- to accurately and comprehensively reflect textual content. Miles and Huberman have identified this combined approach as a good alternative to coding done exclusively from existing conceptual frameworks or from bottom-up development that does not consider past research.

In order to embed the codes firmly within the text, Bachor and Baer (2000) followed the 'constant comparison' method recommended by Glaser and Straus (Tesch,

1990), whereby a select portion of data was repeatedly examined and reworked in order to establish codes that accurately and comprehensively reflected textual content and meaning. This is consistent with the evolution of coding, refining and familiarity with the data discussed by others (see for example, Huberman & Miles, 1994; Miles & Huberman, 1984). In fact, Gall et al. (1996) have pointed out that revision of code categories leads to greater comprehension and accuracy. Further, 'inter-judge agreement' (Kazdin, 1982) was examined by Bachor and Baer in order to confirm the reliability of the categories. Independently coding randomly selected text segments, the authors demonstrated consistently high reliability rates for the code categories (ranging from 72-96% agreement). Having established overall codes exhibiting comprehension, accuracy and high reliability, the second author applied these categories to the entire dataset, representing all journal entries. These descriptive codes formed the basis of the pattern analysis and comparative evaluation conducted for this study.

Using the techniques of constant comparison and establishing categories within the text described above, further comparative analyses were possible through closer examination of higher order codes. That is, clusters were developed within code categories to address several key questions for this study. Fundamentally, this entailed the identification of categories within codes in order to allow for more detailed comparisons between the diaries of female and male pre-service teachers. Categorizing the content of comments within a particular code classification was necessary in order to focus particular queries across gender.

Gender-based analysis. As described above, a sample of 38 participant diaries were selected for comparative analysis across gender. Using *Atlas/ti*, individual codes were employed to separate and identify each of the specified diaries. Codes were further applied to distinguish the gender of all participants included for analysis. Thus, every diary entry was coded with both a participant number and a gender indicator, allowing for retrieval and organization of data by individual journal, gender or a combination of the two. As Miles and Huberman (1984) have noted, the ability to 'cluster' results in such a manner sets the stage for analysis.

The recommended 'familiarity' with the data (see for example Gall et al., 1996; Huberman and Miles, 1994) when it is of a thick qualitative nature such as that of the

diaries examined here, necessitates a 'feel' for the overall text itself. This helps to ensure that major themes, disparities and so on have not been ignored or overlooked. To that end, the initial step upon identification and coding of the journals was the careful and concise reading and rereading of the diaries. Although this researcher had previously achieved an understanding of the text and prevalent themes, this step was felt necessary in order to focus more specifically on the sample to be analyzed. Further, becoming familiar with the data from the perspective of gender-based comparisons had not been considered previously. The notion of becoming even more immersed with the textual information created a familiarity that allowed for more informed and focused comparative questioning and analysis.

At this stage, what might be called two prerequisites of data analysis had been met. General familiarity with the textual material was established and the diaries were identified and coded with overall categories to allow for organization of this relatively complex material. Using *Atlas/ti,* retrieval functions were similarly possible. That is, data could now be organized, retrieved and examined by gender, participant, code category, and even by particular quote or line segment. Using the operators in *Atlas/ti*, questions regarding 'relationships in the data' (Huberman & Miles, 1994) could be posed. In order to do this, further analysis of 'meaningful content categories' (Marshall & Rossman, 1989) was conducted whereby particular content clusters were identified within higher order categories. For example, comments classified as 'establishing criteria' for evaluation were individually examined then clustered together in order to explore what type, class or particular content these criteria expressed. Essentially, this process consisted of further categorizing code categories to obtain a deeper level of analysis. Gall et al. (1996) have described this as an 'interpretational analysis' in which salient constructs have been identified and significant patterns (in this case across gender) are explored and illustrated.

The inclusion of participant quotes highlights code and content categories with illustrative examples. For the majority of selected quotes, a basic pattern in which random sampling was used to extract examples was followed. Random sampling applied for categories with more than 10 cases, while purposeful selection was used in those instances with fewer than 10 examples. In a small number of cases with less than 10

examples, however, where no particular quote appeared more illustrative than another, random selection was also used. In one case, purposeful selection was used for categories with more than 10 examples where a particular quote was required to illustrate the breadth and content of a specified category. In addition, an attempt was made to balance quote selection across gender.

Research Questions

As noted above, this investigation is best described as one of attempting to achieve 'descriptive understanding' (Huberman & Miles, 1994). Although the description is focused on a specified comparative analysis, hypotheses testing and causal implications were not possible or considered desirable. The purpose was to conduct a gender-based examination of pre-service teacher assessment practices based on an analysis of the diary data. The information presented is thus descriptive in nature. A set of research questions were established to form the basis of the study. These questions were seen as moving from overarching inquiries to those of more focused interest. Thus, the following general queries were addressed by the analysis:

1) What are the similarities and differences in the overall contents of the participant diaries?

2) In what ways does gender influence the focus and concern these participants have regarding evaluation in general?

3) How do the means of evaluating (evaluation criteria) differ between gender?

4) How are the comments about the students and overall evaluation of these pupils different between female and male participants?

The more focused questions included the following five:

5) What are the differences between male and female participants regarding the significance of non-achievement variables including: the importance of knowing the student, concerns over the fictional pupils' affective state and desiring information on the subject or classroom in order to carry out assessment?

6) How do the interventions suggested as part of these pre-service teachers' evaluations differ between gender?

7) In what ways do the decision paths these pre-service teachers appear to follow in their assessment practices differ between gender?

8) What are the discrepancies between male and female participants in regards to the assessment of inferences that classify the fictional student(s) as having a possible special learning need?

9) What are the gender differences in the assessment of "extreme" inferences concerning the influence of student(s)' quality of life on their performance products?

Chapter 4

RESULTS

The results obtained from the analysis of the 38 pre-service teacher diaries sampled for this study are described in this chapter. To reiterate, the pre-service teacher diaries consisted of the journals maintained by a group of undergraduate teacher training students at the University of Victoria. The diaries reflect unstructured data, in that no particular directions were given and no questions were posed to the participants. Individuals were simply requested to record their thoughts, ideas, and processes as part of an investigation into student assessment practices. Thus, the textual information contained within the diaries, although thick and descriptive, was relatively unwieldy and cumbersome.

To provide some manageability to this multifarious set of data, the diaries were transcribed and imported into the qualitative analysis program *Atlas/ti*. Bachor and Baer (2000) had developed code categories in order to further the retrieval and organization of this complex collection of information, as well as to carry out an overall examination of the complete set of journals. The second author noted that a comparative analysis of the diaries across gender might inform this relatively unexamined variable. To that end, the diaries of all 19 males and a randomly selected sample of 19 female participants were identified and coded for analysis in this study. Within category content was examined and clustered in order to glean a more precise image of the diary data and to allow for comparative analyses.

While establishing 'familiarity' with the sample of diaries was a necessary step in the analysis, the original content of these 38 journals remained unedited and was analyzed verbatim. Therefore, a series of informed research questions was established in order to structure this comparative inquiry. Moreover, the questions were considered in such a manner so as to move from overarching queries to a series of more focused questions. In this way, it was possible to move from a general examination of similarities and differences across gender, to specified and narrowed comparisons of assessment. In particular, four general questions were posed, including those concerning overall diary content, the focus and concern about evaluation, criteria for evaluation and evaluative

comments made to pupils. This was followed by five focused questions centering on the significance of non-achievement variables, interventions, decision paths, assessments that classify students and inferences over pupils' quality of life. Taken together, this series of nine questions was used to guide this study.

Results are summarized here according to the research questions noted above. While the overall goal is to describe the similarities and differences between the assessment practices of female and male pre-service teachers, the analysis is structured by the questions posed. For this reason, the results are presented in order of these queries, although summaries and overlap are also discussed.

Question 1: Overall Diary Content

Through familiarity with the sample of diaries and by the analysis of various facets of the textual content, a general depiction of the data was gleaned. Several factors were considered in order to ascertain an understanding of these generalities, including: the length and style of information, as well as the content of the diaries by code and frequency.

It was thought necessary to conduct an initial examination of the diaries from the perspective of whether they were at all comparable. That is, to consider such basic features as the length of text, style of comments and so forth across the entries of male and female participants. Looking at sheer output in terms of the number of lines contained in the diaries, tremendous variability was recorded among both genders. This had been expected, due to the unstructured and complex nature of the data. The diaries written by female pre-service teachers ranged anywhere from a high of 309 lines to a low of 57 lines, with a mean of 151 transcribed lines of text. Likewise, the male participants recorded diary lengths ranging from 277 to 54 lines, with a mean of 140 lines. With a collection of data totaling 5523 lines (2869 among females and 2654 among males), the overall diary lengths of the two populations were taken to be relatively similar and comparable. In fact, the within population variability was more notable among both samples than that between the two samples. Further, this within sample variability was remarkably consistent across gender. In order to confirm this similarity, a *t*-test was run for length of diary contents across gender group. The *t*-test results proved non–significant at 0.05 alpha ($t(39) = 0.59 <$ critical value of 2.03). The diaries were thus taken

to be relatively similar in length and diversity across these two populations. This in itself, however, says nothing about actual content.

Regarding the general style of the diary content, overarching patterns were noted between the entries by establishing the portion of participants that fell into similar response patterns. In particular, some participants tended to write in a very 'objective' manner, avoiding personal descriptions, thoughts and anecdotes. These individuals offered straightforward assessment criteria and evaluated the tasks. There was minimal feedback, few evaluative comments to the students and no reflection about their own thoughts or musings concerning assessment. Point-form statements were the written means of choice. A randomly selected example by one participant stated:

Criteria:

1. Sentence mechanics & spelling 4
2. Clear sequence: beginning, middle, end 4
3. Original ideas & clarity of expression 4

Rating scale

4- superior work

3- good work

2- adequate

1- beginning

These diaries contrasted with those that contained individual reflection and personal ruminations about the activity of assessment. Such diaries tended to be as much about process as product. Contents described personal difficulties, concerns, speculations and so forth. Although criteria were established, large portions of text were devoted to descriptive narratives concerning all manner of individual thoughts about the assessment task, evaluation in general and so on. Detailed paragraphs imbued with personal pronouns are indicative of these diaries. For example, one participant (randomly selected) wrote:

> In marking this criteria, I've tried to be as simple and uncritical as possible. I think that this is a very creative writing assignment and it is very difficult to evaluate or place value on creativity. Who is to judge which word(s) is better than another? As well, if I had marked harder (i.e. more marks subtracted) I think it would discourage risk-taking to try new words that students don't use regularly.

In order to examine these two styles of diaries across gender, all entries were reviewed numerous times and identified as belonging to one or another of these classes, or fitting somewhere in between. Reliability checks using inter-judge agreement (Kazdin, 1982) were conducted to confirm these general writing styles, with point by point agreement rates of over 80%. Participants of both genders used a variety of writing styles to describe assessment and the task of marking the three portfolio assignments. Among the diaries of male pre-service teachers, 5 of the 19 or 26% were identified as being of an 'objective' style, while 12 or 63% contained a substantial amount of personal rumination and individual reflection. Two or 11% were not considered as clearly belonging to either category. Among female participants, 3 of the 19 or 16% of the journals were classified as objective, 12 or 63% were identified as personal or individually reflective and 4 or 26% did not belong to either category. As with length of the diary text, differences across gender regarding writing style were minimal. The most notable stylistic variation among the diaries of both male and female participants was indicated by personal writing and individual phrasing, with the other categories differing only slightly (by two journals in either group). Again, the differences within gender are greater than those across these sampled groups. In terms of both textual length and general content style, the diaries appear to be comparable, exhibiting little difference across gender. This leads to more particular questions concerning the comparison of categories identified within the individual diaries.

In order to further examine the overall contents of the pre-service teacher diaries sampled here, an analysis of the code categories was conducted (these categories are described in Appendix 1). This involved considering both the codes by participant (how many diaries within each group contained at least one example of the category) and

frequency counts for the number of times each code occurred per population. Results for the codes by participant across gender are presented in Table 1, while frequency counts for both female and male pre-service teachers are given in Table 2.

Table 1

Codes by Number of Participants Across Gender

CODE CATEGORIES	Females	Males
Assignment Based Context-classroom	6	5
Assignment Based Context-subject's background	1	2
Assignment Based Criteria-establishing	19	19
Assignment Based Criteria-reviewing/refining	1	0
Assignment Based Questions/Comments-concerns	17	16
Assignment Based Questions/Comments-positives	11	11
Intervention-comments	10	6
Intervention-student	10	8
Person Based Competency-performance on task	15	14
Person Based Competency-student	9	12
Person Based Competency-classification	4	1
Person Based Quality of Life	2	1
Person Based Student Comments-affective state	4	5
Person Based Student Comments-knowledge of	2	2

In general, the number of participant diaries found to contain each of the code categories one or more times was relatively similar across gender. This is particularly true in regards to the distribution of categories within each group. For example, among the diaries of both males and females, the greatest number of individuals discussed the category of 'establishing criteria', with all participants writing about this topic. Likewise, having questions or making comments of 'concern' and discussing a student's 'performance on a task' were the second and third categories noted by the majority of participants across both genders. As indicated in Table 1, the relative number of participants by code categories is distributed similarly in the counts for the diaries of male and female participants. While particular codes may exhibit differences (to be

discussed in more detail below), the overall number of participants writing about each of the code categories was similar for the diaries of the male and female pre-service teachers. This finding was confirmed by frequency counts of all categories.

While the number of participants per code category provides a useful indicator of the individuals discussing each of these topics, it does not account for the frequency or number of times a particular category occurs. In order to determine the occurrence of codes by participant, frequency counts were tabulated. Frequency counts for the diary contents generally support those of the participant by code category tallies. That is, the overall contents of the diaries of both female and male pre-service teacher groups tend to be relatively similar in regards to the topics discussed and categorized by the designated codes. Although there are some differences among particular categories, the frequency of occurrence for codes is similar across gender. For example, the three categories cited by the majority of participants ('establishing criteria', expressing 'concerns' and making 'performance on task' judgements) are also the three most frequently cited codes. While some categories (for example 'performance on task') tend to occur in greater numbers than is suggested from the participant counts, the relative frequency across gender remains similar. Overall, the diaries of both male and female participants exhibit similarity in regards to length, style, diaries per code category and frequency of codes.

Table 2

Frequency Counts by Code Category Across Gender

CODE CATEGORIES	Females	Males
Assignment Based Context-classroom	16	8
Assignment Based Context-subject's background	1	2
Assignment Based Criteria-establishing	108	104
Assignment Based Criteria-reviewing/refining	1	0
Assignment Based Questions/Comments-concerns	63	59
Assignment Based Questions/Comments-positives	33	23

(table continues)

CODE CATEGORIES	Females	Males
Intervention-comments	20	8
Intervention-student	28	16
Person Based Competency-performance on task	119	94
Person Based Competency-student	30	41
Person Based Competency-classification	6	1
Person Based Quality of Life	3	3
Person Based Student Comments-affective state	6	7
Person Based Student Comments-knowledge of	3	2

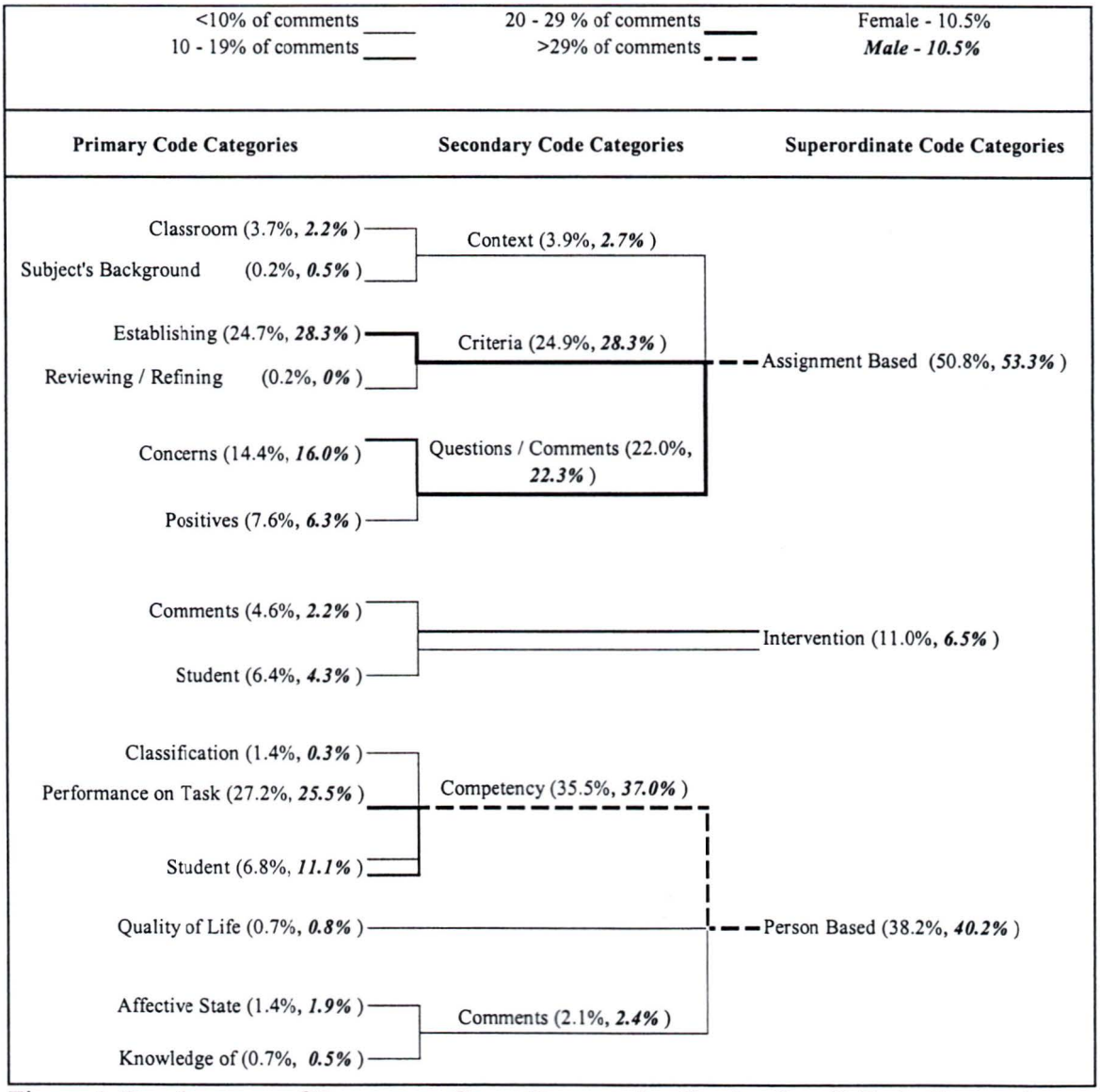

Figure 1 Dendrogram of Frequency Count Percents by Gender for all Code Categories.

Similarity across the diaries is further illustrated through a dendrogram of frequency count percents in Figure 1. As depicted, the overall clustering of comments is almost identical across gender, with notable variation occurring only for two categories; intervention comments and student competency statements. A comparative analysis of more focused questions may determine if these overall similarities hold when it comes to particulars.

Question 2: Focus and Concern Regarding Assessment

The distribution of code categories across gender implies that in general the focus of assessment as discussed by the participants was similar. All of these individuals (n=38) established criteria for marking, and the majority used those criteria to make judgements about performance on a task (n=29 including 15 females and 14 males). Further, most individuals expressed some concerns or had questions about assessment, the particulars of evaluation tasks and so on (n=33 including 17 females and 16 males). Other areas discussed by these participants and noted in the less frequently cited code categories are considered in more detail below.

Regarding the concerns or difficulties individuals had about assessment and marking the student portfolio work, it was thought that although the number of comments was similar across gender, a closer analysis of the types of comments was necessary. In total, there were 122 questions/comments of concern made in the diaries. Female participants accounted for 63 of these comments and males for 59. Examination of the comments revealed that they were of four overarching categories, including concerns over the subjectivity/validity/clarity of a task, questions regarding marking criteria/marking schemes/weighting, concerns over personal abilities, and general/unspecified concerns with assessment or an evaluation task. Table 3 contains the frequency counts for each of these categories of concern by gender.

Table 3

Frequency Counts of 'Questions/Comments of Concern' Across Gender

CATEGORIES OF CONCERN	Females	Males
Concerns Over Subjectivity/Validity/Clarity of a Task	18	18
Questions Regarding Criteria/Marking Scheme/Weighting	19	13
Concerns Over Personal Abilities with Assessment or Evaluating a Specified Task	2	2
General/Unspecified Concerns with Assessment or an Evaluation Task	24	26

The distribution of comment categories classified as matters of concern was very close across gender. The greatest number of comments among both sexes were those of an unspecified or general nature, such as these, randomly selected from a male and female participant:

> This was by far the hardest assignment to mark.

> After I finished marking I realized that although I thought it was clear in my head what I wanted, I don't think I could verbalize it to the student. Why some were 2 and others were 3? They just were. It was difficut [sic] to mark.

Questions of concern centering on the criteria, marking schemes or weighting of evaluations were also common in the diaries of both male and female participants, although the latter recorded slightly more of this category. Similarly, concerns over the subjectivity, validity or clarity of an assignment or evaluation were stated in equally high numbers among both genders. This category of comment is illustrated by the following example selected at random from a female participant regarding her evaluation of a particular task:

> The marking felt rather subjective though, and I ended up reading each question & answer together so as to rank the best answer to the less plausible... This poetic analysis assignment was very different to mark due to its subjective nature.

The least recorded category of concern was that of questioning personal ability, indicated by only four instances - two from either gender. Again, the distribution of comment category was the same for both female and male participants.

The general focus of the diary content as well as the questions or comments of concern expressed by the participants were similar across male and female pre-service teachers. Both the numbers of comments made by participants and the frequency of content categories analyzed, indicate that the distribution of questions or comments of concern expressed is similar. At this level of analysis gender differences in the participant diaries are not prevalent. Examination of other factors is therefore considered.

Question 3: Means of Evaluating

Two categories of comments were classified as 'criteria' for evaluating assignments: establishing criteria and reviewing/refining criteria. Due to the fact that there was only one identified case of reviewing/refining criteria noted among the sampled diaries, this category was not considered further. As noted in Table 1, a total of 212 instances of comments classified as 'establishing criteria' were recorded, with the distribution across gender being relatively even (108 by female and 104 by male participants). This distribution was divided among all participants, for the diaries of every member in both samples (n=38) contained at least one instance of this category.

Through examining the written content of cases in which participants established criteria, two clusters of comment types were noted. Individuals tended to discuss criteria for marking student assignments as 'straightforward keys' to evaluation, or as 'criteria including other considerations' such as justifications, personal reflections and student characteristics. This latter category was distinguished because the criteria were not presented as a recipe-style exercise, but as ones in which other concerns and considerations became embedded within the discussion of criteria. This contrasted sharply with the relatively formulaic statements of the straightforward key cluster. For

example, illustrating straightforward comments, the randomly selected writing from one male participant stated:

Criteria:	
Spelling/Language usage (punct, caps etc.)	5
1 mark off each mistake	
Understanding of Poem	10
Personal meaning	10
Total	25

In contrast, others established evaluation criteria in a manner that was much more reflective, descriptive and confounded with references to a variety of factors beyond the criteria themselves. Another randomly chosen male participant included the following comment in his diary:

> Multiple choice – I'm not a big fan of. Doesn't really show if students understand what they have just read. They could have guessed). I gave one mark for each response that fit the story. Some questions had more than one correct choice.
> ? Some the M.C. questions used here are open to personal interpretation. i.e. question 1 part 2.
> ? Since the story only hints at a supposed correct answer it leaves it open to personal interpretation. I would mark any response correct because these questions do not allow the children to write about what they might be thinking, so I am forced to guess at what they might be thinking.
> ? I chose to let some of the students have the opportunity to explain what they were thinking. If their thoughts could fit in with the story I would mark their questions correct.

Conducting frequency counts for these clusters across gender, male participants recorded 57 instances (55%) of 'straightforward key' comments and 47 examples (45%)

of 'criteria including other considerations'. Among the diaries of female participants, 'straightforward key' comments were recorded 50 times (46%), while 'criteria including other considerations' were noted in 58 instances (54%). The diaries of male participants contained a relatively greater number of instances of establishing criteria in a straightforward manner, while those of female pre-service teacher exhibited a higher number of criteria that included other considerations. However, across both genders there was an even split between the two cluster categories, with only 10 percent separating either group (ranging between a minimum and maximum of 45 and 55 percent among male participants).

Question 4: Evaluative Comments and Evaluations of Pupils

Three primary code categories were considered evaluative of either a student's work or of the student him/herself: 'performance on a task', 'student' performance and student 'classification'. These were all competency classifications, ranging from the simplest judgements of a particular piece of work to those commenting on a student's abilities, to the most 'extreme' form of evaluation, designating a pupil as having a special learning need on the basis of her/his performance. As depicted in Tables 1 and 2, slightly more female than male participants made a greater number of performance on task and student classification comments, while somewhat more male pre-service teachers made a higher amount of student performance evaluations. Due to the particular significance of the student classification category, this is analyzed in greater detail below. Results for the two remaining categories are considered here.

Of the competency evaluations, those focusing on a particular task performance were the most frequent, with 119 instances in 15 female diaries and 94 instances in 14 male diaries. The majority of comments across gender were short sentences about how a student performed on a particular task. Illustrative of these comments are the following two randomly selected statements made, by a female and male participant:

> Student C 17/18 - Very consistent and shows understanding but could do more in terms of detail in some parts of Jimmy's story.

Student A – Short, and kind of careless, doesn't show signs of editing or rereading, still A gives some personal anecdotes. 8.5/12

Examination of the performance on task comments indicated that they were relatively uniform. That is, there was little difference noted in content, style and so on between these identified codes of text. In fact, there was almost no diversity across these comments, either within or between gender. However, it was noted that in some instances, these evaluations were tied to interventions directed at the student or as general comments. In these cases, the evaluation of a task was followed, preceded or imbedded with an intervention suggestion. Essentially, evaluations and proposals for improvement went hand-in-hand in such instances.

An analysis of the 'co-occurrence' (defined as being embedded or within two lines preceding or following a comment) of performance on task evaluations and intervention comments, was conducted. A total of 18 cases of this co-occurrence were found among all performance on task comments. Fourteen of these occurred in the diaries of female pre-service teachers, while four were found among the writings of male participants. Female participant diaries contain a greater number of evaluations in which offering suggestions for improvement via interventions was related to the mark allocated for a task. This is reflected in the relatively higher number of intervention comments made by female participants overall – described below.

Regarding student evaluations, this class of comments was generally longer and more descriptive than evaluations of task. This cluster reflected expressions of judgement regarding a student's general ability in a particular area. For example, a student might be evaluated as having poor grammar skills, being good at spelling and so on. Likewise, the judgement about a student may be in the form of a general evaluation, such as the statement that this is an 'A' pupil, an average performer, bad at school or some other related comment. The following purposefully chosen quote from a male participant exemplifies this class of comments:

Student A

- Capitalization and punctuation is weak
- Apostrophes are not used
- Does not capitalize first person singular "I" at times
- Forgets to begin new sentences with capital letters
- Needs some instruction on the "comma"
- Some words are joined together
- Expresses ideas well!

Male participants recorded a slightly greater overall number of student evaluations than their female counterparts, accounting for 41 instances, relative to 30 cases among females. While there was somewhat greater variability in these comment types relative to performance on task evaluations, they were generally similar in form and content. However, it was again noted that in some cases, intervention suggestions were inextricably bound to the evaluations. An analysis of the co-occurrence of student evaluations with intervention suggestions was therefore conducted.

Although there were 11 more cases of student evaluation in the diaries of male participants, a higher number of co-occurring comments appeared in female pre-service teacher diaries. Specifically, 9 of the 30 cases of student evaluation occurred with intervention comments among females, while only 3 of 41 appeared with this proximity among males. Again, evaluations conducted by female participants tended to occur more frequently alongside interventions directed at addressing the assessment than is the case among the evaluations conducted by the male pre-service teachers sampled here.

Question 5: Non-Achievement Variables

A series of factors within the diaries were identified as non-achievement variables. These categories consisted of comments centering on issues that were not related directly to assessment or evaluation of the portfolio assignments. They include the sort of concerns about background and knowing the student that have typically been considered 'confounding' by measurement experts, but cited as influential among educators (see for example Airasian & Jones, 1993; Brookhart, 1991, 1993; Whittington,

1999). Four code categories were defined as non-achievement variables, including, comments about the classroom (including the teacher), the participant/subjects own background, having knowledge of the student and concerns regarding the student's affective state. Participant and frequency counts for the four categories are listed above in Tables 1 and 2. The categories are considered in turn here.

In terms of commenting on the classroom, requesting information about the pupils' classroom context and so on, female participant diaries contained 16 examples of this comment type, while the diaries of males contained 8 instances. Analysis of these comments revealed three general categories of focus: questions over what the classroom teacher had previously taught/told the students, the teachers intended outcome with the assignments and what work the students had previously done in the class. Table 4 contains the frequency counts for these three categories.

Table 4

Frequency Counts for Categories of Comments Concerning the Classroom

CATEGORIES OF CONCERN	Females	Males
Questions about what the classroom teacher had previously taught/told the students.	11	7
Concerns over the teachers intended outcomes or expectations of the assignments.	4	1
Knowing what work the student had previously done in the class.	1	0

As noted in Table 4, questions regarding what students had been taught in the classroom were most common, followed by concerns over the intended outcomes of an assignment. This pattern is illustrated across gender, although the fact that female participants recorded double the number of comments concerning the classroom is highlighted in these counts. The final category of previous student work was noted in only one instance.

Comments made within the subject diaries that reflect concerns or questions about their own backgrounds were cited by only three participants in three instances (2 among males and 1 among females). Examples within this category are instances of participants

feeling uncomfortable with their own background abilities as evaluators or with a particular aspect of assessment.

Comments concerning knowledge of the student as a variable in evaluation were found in two instances among male participants and in three cases within the diaries of female pre-service teachers. Both examples found in the male participants' diaries concern the fact that they do not know how representative a particular item is of a student's work. One of the female entries is similarly focused on the notion that judging work relative to a particular student's work history is vital. The final two comments from this category simply state that the participant either felt like they were getting to know the student or that they did not yet know her/him. No further differences were noted across gender of this small cluster of comments.

The final category of non-achievement variables concerns statements over evaluation as related to a student's affective state. These comments were relatively evenly divided across gender with 7 examples among males and 6 among females. Upon further analysis, two sub-categories became apparent within this cluster: inferences about a student's affective state and how it was influencing performance, and the impact of an evaluation on a student's affective state. The categories are illustrated in turn by the following set of comments purposefully selected from two female participants.

> If these students were really in my class, I would definitely meet with student B to determine whether he/she just had a bad day or was really missing the 'main' ideas.

> I know if this was a real student on these little assignments I would probably mark a little easier to give a boost to his self esteem.

Examination of these two sub-categories for male participants found 5 cases of comments in which affective state was considered as having an influence and 2 instances of concern that an evaluation might effect a student's affective state. Results for female participants were split evenly between these categories, with 3 examples of each. Although the diaries of male participants contained a relatively greater number of

comments regarding the affective state of a pupil influencing performance and a lesser number of instances expressing concern that an evaluation might effect a student's affective state, the differences were no more than two point across gender for either category.

Question 6: Assessment Interventions

As discussed above, the diaries of female pre-service teachers contained a relatively greater number of intervention comments co-occurring with performance on task and student evaluation statements. The intervention category itself consisted of two primary codes described as intervention 'comments' (those hinting at intervention suggestions directed at the task, the class, the teacher and so on) and 'student' interventions (consisting of suggestions for particular pupils). Across both of these categories, female participants recorded double the number of intervention comments (48 to 24). As noted in Tables 1 and 2, a relatively greater number of female participants made a larger number of comments across both categories of intervention statements.

Examination of these comments revealed that in and of themselves they were relatively uniform in nature. The intervention 'comment' sub-category or primary code contained general suggestions directed at the teacher, task or other agent of change. For example, two randomly chosen comments from male participants stated:

> Students could benefit from instruction on spell checker.

> I believe a better test would have been to give them the story first then test them using the multiple choice questions for understanding.

The intervention 'student' primary code contained comments of a similar ilk, the difference being that these were clearly directed at specified individuals rather than a task or general class of person. These comments focused the crux of the intervention at a pupil identified as the individual attached to an assignment or series of tasks. The following random examples from two female participants illustrate this class of statements:

> However, I believe he could benefit from continued vocabulary practice, as well as sentence meaning as it relates to word choice.
>
> Student B - I like to see this student exploring longer and more complicated sentence structures. Some teaching may be needed on commas however.

Although the general content of intervention statements was noted to be uniform across comments, differences were identified in the relationship between this category and other clusters. In particular, comments regarding the student in the form of evaluations or judgements were found to co-occur in the diaries of some individuals. Interventions were essentially paired with judgements about a pupil. In other words, the intervention was seen as addressing some need the participant had assessed from a student's submitted work.

Conducting a co-occurrence analysis of all intervention statements with the cluster of comments considered 'person based' (see Appendix 1), it was noted that a total of 33 instances of intervention occurred in tandem with person based statements. This represents 46% of all intervention comments (33 of 72). Across gender, the diaries of female participants contained 26 of these co-occurrences (representing 54% or 26 of 48), while those of male cohorts contained 7 (totaling 29% or 7 of 24). Breaking these results down further into the primary intervention codes, student interventions co-occurred in 24 of the 26 instances (92%) among female participants and in 6 of 16 cases (38%) among male pre-service teachers. The sub-category of intervention comment included only 2 of 20 examples (10%) of this co-occurrence in writings by females and 1 of 8 instances (13%) among males. The greater number of intervention comments by female participants is most noted among student intervention statements, particularly in the interaction or co-occurrence with person based categories.

Question 7: Decision Paths

Although both male and female pre-service teachers sampled for this study followed a general assessment pattern of establishing criteria and evaluating performance on a task, some participants made additional assessment comments. That is, while certain

participants took a very 'task restricted' path to evaluation (essentially staying at the level of establishing criteria and evaluating assignments), others were considered 'student elaboration' participants. Student elaboration participants went beyond straightforward criteria and evaluation and made inferences regarding a student's competency classification through such things as speculating that the pupil might be learning disabled. These qualities were inferred from the students' written work and background information and were in no way explicit in the portfolio contents available to the participants. As a starting point to considering the decision path individuals took, those termed 'task restricted participants' (TRP) were defined as participants making no classification or quality of life comments, while 'student elaboration participants' (SEP) were deemed those who made student competency classifications. These definitions represent overall means of examining variations in these strands of decision path.

Considering task restricted participants (TRP), 15 females and 17 males made neither student classification nor quality of life comments. Thus, 4 female and 2 male participants were excluded from the category defined as task restricted. Further analysis was conducted in order to remove additional categories and determine how many participants of each gender represented extremes in regards to restricting their diary contents to the task of evaluating assignments in the portfolios. In Table 5 this decision path is detailed, with an increasing number of variables parsed out of the initial definition moving down the table.

The majority of both female and male participants were classified as TRP, following an assessment pattern or path that did not venture into judgements about the classification or quality of life of the fictional students. Within the TRP cluster, it was possible to remove categories that went beyond the task of establishing criteria and commenting on performance. Individuals who avoided all comments concerning the students' affective state, intervention recommendations or person based competency judgements represented the extreme end of TRP. These individuals simply carried out a pattern of assessment that did not go beyond the items they were marking. There were 6 female and 2 male pre-service teachers who limited their diary contents to this level of comment.

Table 5

Task-Restricted Participants' Decision Path

Decision Path*	Females	Males
Task-Restricted Participants (TRP) – no 'classification' or 'quality of life' comments.	15	17
TRP who also made no 'affective state' comments.	14	13
-TRP who made no 'affective state' or 'intervention comments'.	9	8
-TRP who made no 'affective state' or 'intervention student' comments.	9	8
-TRP who made no 'affective state', 'intervention comment' or 'intervention student' comments.	6	5
TRP who made no 'affective state' comment and no 'person based competency-student' comments	10	6
TRP who made no 'affective state' comment nor either 'intervention' statement and no 'person based competency-student' comments.	6	2

* The decision path includes the progressive removal of extraneous categories.

Along the continuum of the TRP decision path, the number of participants of both genders declined, with for example less than half (6 female and 5 male) avoiding comments regarding affective state and either cluster of intervention category. The elimination of participants who made affective state, intervention or student competency statements resulted in the 6 female and 2 male participants at the far end of the TRP path. Those participants who recorded at least one or more comment for any of the parsed out categories, or those who discussed more than one of these categories represents the majority of the TRP cluster across gender. However, there were a slightly greater number of females who avoided the inclusion of any of the parsed out categories in their diaries.

Those individuals described as student elaboration participants (SEP) represent the other side of TRP. SEP went well beyond the task of establishing criteria and assessing student assignments. Based on the limited information provided in the students' submitted work, they made inferences about a variety of personal characteristics and attributes regarding the fictional students. The person based competency

classifications that defined this group's decision path included speculation that pupils were learning disabled, gifted, and in need of Reading Recovery. As noted in Table 6, there were 4 female and 1 male participant who made such classifications.

Although competency classifications served as the stipulative definition of this decision path, further extrapolation was possible by adding additional code categories. That is, by the inclusion of categories that made inferences about a student's quality of life, affective state or through the extension into various interventions, the number of participants who carried their elaborations to the greatest extremes were identified. Table 6 proceeds from the original number of SEP and adds queries for further comments to establish the additional comment types these participants made.

Table 6

Student Elaboration Participants' Decision Path

Decision Path	Females	Males
Student Elaboration Participants (SEP) ['Person Based Competency-classification']	4	1
SEP who made 'quality of life' comments	2	0
SEP who made 'affective state' comments	3	0
SEP who made 'quality of life' and 'affective state' comments	2	0
SEP who made 'Intervention-comments'	4	0
SEP who made 'Intervention-student' comments	4	0
SEP who made 'Intervention-comments' and 'Intervention-student' comments	4	0
SEP who made 'quality of life', 'affective state', 'Intervention-student' and 'Intervention comments'	2	0

Among the male participants, the single individual identified as SEP did not make further comments regarding such things as quality of life, affective state or intervention. Among the four female participants identified along this decision path, two of them ventured to make comments across all categories considered in this analysis. These participants speculated not only about the students' classification, but commented on

their assumed quality of life and affective state. In addition, they offered suggestions for interventions both as general comments and as directed at students. For example, one comment -- intentionally selected for its illustrative point-- stated:

> I am wondering if this student has a learning disability or not one of the greatest home lives. The last sentence concerns me "I could probably tell you more but it would take to long". This student needs a great deal of encouragement and assisstance (sic). I hope that s/he gets it. I can see that this student does not enjoy school all that much or writing from the sentence I mentioned above.

Later in her diary, the same participant expressed her concerns about the student's affective state when she wrote:

> I guess I am scared of hurting a child's self-esteem, because I can remember how devastated I would feel at times if I did not get the mark I felt I had achieved.

The two female participants identified as commenting across all categories considered in Table 6 represent the extreme cases of SEP. The single male participant made no further comments of any of the classes analyzed, while all of the female participants offered both types of intervention comments and half made comments across all categories.

Question 8: Student Classifications

Examination of participant decision paths above highlight the distinction between individuals who tended to restrict their diary comments and those who elaborated about various characteristics of the fictional students. Comments termed student classification statements were considered to represent the most far-reaching form of elaboration due to the speculations made and the potential impact of such comments.

In all, there were a total of 7 student classification comments made by 5 of the 38 participants (13%). Six of these comments were contained within the diaries of 4 female

participants, while one was written by an individual male. The contents of the comments themselves were divided into four specific sub-types. Three of the samples from female participants' diaries and the single male example described a fictional student as having - or possibly being - learning disabled or in need of learning assistance. Among female participants, two further comments proclaimed that a student might be gifted or imbued with exceptionalities. A single female participant considered a student as a candidate for a Reading Recovery program. Examples of these student classification comments are illustrated in the following randomly selected citations from the single male participant and two of the female participants.

> He/she may be learning disabled.

> Need to really work with this student in the area of language arts or get the child the learning assistance s/he may need.

> Ongoing notes- student may be gifted > provide enrichment opportunities.

As noted in the results across SEP above, among the 4 female participants who made student classification comments, all of them offered intervention suggestions, 3 expressed concerns about a student's affective state and 2 included notes on the pupil's quality of life as well as the other categories discussed. The male participant who made a student classification comment did not include any instances of these other categories in his diary.

Student classification statements were found in 13% of the participant diaries across a total of 7 instances, which represents 0.9% of all comments coded (every line of diary text was coded). Although the inferences made in this comment category were far reaching, the number of cases was in the minority both in regards to individuals and in relation to the total number of comments coded. Females recorded somewhat more of these comments than males, with 22% of sampled female participant diaries containing one or more student classification comment and 5% of the male diaries exhibiting this code category.

Question 9: Influence of Pupil Quality of Life

Another of the categories of comments that appeared in relatively small numbers but represents claims that extend beyond the task of assessing student work, is quality of life. Quality of life statements took the form of making inferences about a student's home and social life and how that might be reflected in or influence a particular pupil's school work. There were a total of 6 quality of life comments made across the diaries, 3 by participants of each gender. A single individual recorded all three of the cases among male participants, while one female pre-service teacher made two of these comments and another recorded one. As a percentage, 11% of females and 5 % of males made quality of life comments. In regards to frequencies, this totals 0.7% of all comments coded in the sampled diaries.

The three quality of life statements recorded by the male participant occurred in direct succession; each within the assessment of the three identified fictional students. The following quotes illustrate this

> Student A
> I get the impression that this person has a frustrating home-life. "Dad needs a job", for a student to write this means to me there is probably a lot of tension at home. This will transfer to the school environment. It's hard to be positive and motivated when its [sic] not very positive at home. I sense "anger and frustration" with this student.

> Student B
> I get the impression that this person is very happy with his/her life. A very positive home-life with lots of parental support is what I see. With the comment made about "pressure" I fell this person is a perfectionist, one must be careful and watch for this student putting too much pressure on him/herself.

> Student C

> I get the impression that this student is your typical grade 5 student except he/she has much older bother and sisters. This person likes to be close to his/her friends. Needs to find a place to fit in – be part of a group.

The quality of life statements within the diaries of the two female participants reflect similar concerns regarding home and/or social life. The two examples recorded by one female participant did however differ in that they were written in a somewhat abbreviated fashion relative to the other noted instances of this comment type. This participant wrote

> Student B:
> Impressions – very expressive. – social person. – likes to do well. – great family support. – active busy life. – high self-expectations.
> Student C:
> Impressions – not very close family. – self conscious> wants and needs to be cool. – wants to be independent> resists authority. – does look toward future (good sign). – good literacy level.

The small number of participants and the low occurrence of quality of life comments did not allow for considerations of potential gender differences for this category of comment type. However, the inclusion of such inferences concerning an individual's quality of life in the assessment of pupil work raises questions and concerns about the possible implications of assessment practices. Recommendations for further research along with a discussion of the results from this study are considered in the following chapter.

Chapter 5

DISCUSSION

The 19 female and 19 male pre-service teacher diaries analyzed in this investigation contained a broad range of thoughts and processes relevant to understanding the assessment practices of these individuals. The purpose of this investigation was to carry out a gender-based examination of these descriptive data in order to explore possible differences between the assessment practices of female and male participants as noted within the diaries. The results of the analysis are discussed in this chapter. The chapter is divided into three overall sections. First, the results are interpreted and summarized across the research questions. Second, an effort is made to situate the findings within previous research carried out in the field. Finally, future directions and recommendations are offered.

Interpretation of Results

Discussion of the results is guided by the nine research questions that formed the basis of analysis for this investigation. These questions represent a shift from more general or overarching queries about the diary contents, to those of a more focused and narrowed interest regarding specific comment types made by participants. In order to provide structure to the discussion, the questions are clustered into the two main categories of overarching questions and focused questions.

Overarching questions.

A series of four broad questions were posed in order to structure the overall analytic comparison of the diary contents. The questions directed attention towards a general comparison of the written content, questions or comments of concern, criteria for assessment and evaluative comments. These are considered in turn.

Initial examination of the diary contents revealed that the length of entries varied greatly across individual participants. This was expected, due to the unstructured nature of the task. Diaries ranged in length from over 300 lines, to less than 60 lines of recorded text. This variability was consistent across gender, with the mean number of lines written by the male and female comparison groups differing by only 12 lines. Within gender variability was found to be greater than between group differences. Likewise, the writing styles -- noted as objective versus those imbued with individual reflection and rumination

-- were found to occur in relatively equal numbers across gender. The most frequently noted class of writing style among both female and male participants (63% for both genders) discussed assessment practices in a manner that included 'personal' content and concern about the process rather than purely objective evaluations of student product. This is consistent with the findings of other investigations, noting that educators tend not to approach student assessment as an objective enterprise free from intuitive and informal components (for example, Airasian & Jones, 1993; Bachor & Anderson, 1994; McCallum et al., 1993; Mavrommatis, 1997; Wyatt-Smith, 1999).

Further comparison of the general diary contents was conducted through participant and frequency counts for all defined code categories (listed in Appendix 1). Although some differences for individual code categories were noted (these are discussed under the focused questions below), the distribution of comments across categories was comparable for the male and female participant diaries analyzed. The greatest frequency of comments for both participant groups centered on establishing assessment criteria and marking student tasks. In terms of the number of participants recording each comment category at least once, the greatest occurrences for both females and males were noted for the categories of establishing criteria and questions/comments of concern. The general distribution of the diary contents throughout the categories is relatively similar among the participant sample groups examined.

The majority of participants (33 of 38) expressed some concerns or had questions regarding assessment. Participant counts as well as frequency counts were distributed relatively evenly among female and male participants in regards to this category of comment. In fact, there was a discrepancy of only four comments across gender groups out of 122 identified instances of assessment concern or question. Further analysis of these comments revealed four general classifications of questions and concerns that participants had regarding assessment of the assigned student portfolios – identified as concerns or questions over subjectivity/validity/clarity, criteria/marking scheme/weighing, personal abilities and general or unspecified concern. Distribution across the four clusters produced similar results for both groups of participants. The diaries of male and female pre-service teacher groups contained an almost even distribution of comments when queried by this deeper analysis of categories.

Two classes of comments were considered means to evaluating student work: establishing criteria and reviewing/refining criteria. As there was only one identified case of the latter, it was not considered further. All participant diaries contained at least one instance of criteria establishment, and the frequency counts for this comment category were relatively evenly distributed across gender groups (108 for females and 104 for males). The large number of comments noted for this code was a result of the portfolio task, in which participants had been instructed to evaluate the submitted assignments and record their processes, thoughts and so forth. Bachor and Baer (2000) had also noted the high frequency of evaluation criteria statements in their earlier analysis. Across gender group, this large cluster of comments was remarkably similar, not only in regards to frequency counts, but also upon further content analysis. Male participants did include a relatively greater number of criteria that were stated in a straightforward manner, while females had a higher ratio of criteria that included other considerations such as justifications and personal reflections. However, the discrepancy between these two categories across gender was under 10 percent. In terms of both frequency and criteria related comments, the two participant samples were relatively similar.

A final collection of code statements considered in the overarching analysis of the diary contents focused on three categories of evaluative comments. These evaluative comments were directed either at a student's work, the student her/himself, or at a classification of the student. Across the three categories of evaluation type, slightly more females commented on task performance and student classification, while males recorded a greater number of student evaluation statements. However, the differences across gender groups altered by no more than three participants for any single code category and the distribution of frequency counts across the codes followed an identical pattern by gender group, with counts decreasing from task to student to classification. Likewise, there were no identified differences in content, style or other factors noted between the sampled participant groups. Discrepancies however were identified in regards to the interaction between evaluations and interventions.

By conducting a co-occurrence analysis, whereby intervention statements could be identified as immediately preceding, embedded or following an evaluation statement,

instances of this interaction could be pulled from the diaries. For both task-based and student-centered evaluations, female participant diaries contained a greater co-occurrence of intervention comments. The sample of female participants accounted for over 75 percent of all instances of these evaluation-interaction co-occurrences. This was the case even within student evaluation statements, where there was a greater frequency of comments made by males. However, it must be added that for male and female groups these co-occurrences represent the minority of evaluation comments overall, as they are noted for 15% of female and 5% of male cases. Although other investigations have suggested that gender differences may be noted as complex interaction effects rather than straightforward discrepancies (for example, DeVoe, 1990; Ehrenberg & Goldhaber, 1995; Rong, 1996), interpretation of these results must be considered within the relatively small number of interactions identified. Differences in intervention comments discussed below are also directly relevant.

Focused questions. A series of five focused questions were used to structure a more detailed gender-based examination of the diary contents. These queries included analysis of non-achievement variables, interventions, decision paths, student classifications and quality of life statements. Discussion of these analyses follows in turn.

Non-achievement variables consisted of comments that were not directly related to the assessment of portfolio assignments, and included such factors as statements regarding the students' classroom, the participant's own background, having knowledge about a student and concerns about a student's affective state. These are the types of factors that advocates of the measurement tradition have typically attempted to control, but are known to be influential in the classroom (Airasian & Jones, 1993; Brookhart, 1991, 1993; Whittington, 1999).

Overall, the distribution of non-achievement variables across gender group was relatively similar, however, in some instances it was reflected by only a few examples. Expression of concern regarding participant's own background and knowledge of a student were noted in only 2 and 5 cases respectively. These comments were distributed in relatively even frequencies across gender group, as were those pertaining to the affective state of a student. In all of these cases, the frequency counts differed by only

one comment per gender group. The exception to this was found within comments concerning the fictional students' classroom. Female participants recorded double the number of classroom statements, accounting for 16 of the 24 cases, or 67% of this cluster. Further analysis, breaking down the category into more refined content classes, did not identify additional differences. While there appeared to be a greater frequency of classroom comments from the female sample group, the number of instances was relatively small, differing by 8 cases. In addition, participant counts revealed that 6 females and 5 males were responsible for all of these comments. Females who expressed classroom concerns did so a greater number of times per diary, not with a substantially larger number of participants per sample group.

The greater co-occurrence of evaluation and intervention statements described for female participants above, is related to the higher number of intervention comments made by female participants. Female participants recorded a greater number of intervention comments and intervention statements directed at students – accounting for 71% and 64% of each category. By participant counts, this represents a total of 20 females and 14 males, indicating that the number of comments per identified diary are higher among female participants, while the number of individuals discussing this topic one or more times did not differ as substantially across gender group.

Additional co-occurrence analysis for all intervention comments with the cluster of codes termed 'person-based' (those including student evaluations and judgements) was conducted. Female participants accounted for 79% of all co-occurrences, with student directed interventions being the focus of 92% of these interactions. In contrast, only 21% of intervention comments made by males co-occurred with person-based statements. The coupling of student intervention and evaluative statements suggests that female participants may offer a greater number of suggestions or potential solutions to issues they identify from assessment and evaluation. This however, must be considered within a context where intervention comments make up less than 10% of the total diary categories and are contained within the writings of approximately half of all participants.

Both female and male participants tended to follow a general assessment pattern of establishing criteria and evaluating performance on a task. However, some individuals restricted their discussions to this level, while others elaborated into areas that made

inferences about student characteristics and variables beyond the portfolio assignment. The majority of individuals from both gender groups were deemed task restricted, and did not make judgements classifying or commenting on the quality of life of the fictional students. This is consistent with the earlier findings of Bachor and Baer (2000).

Although a greater number of female participants were categorized as extremely task restricted (6 females to 2 males), individuals of both groups declined in number as categories were parsed out in order to identify the extent that individuals remained restricted to the task of marking assignments. The pattern of participant counts in examining this decision path remained similar across gender groups until student directed competency comments were removed. The slightly greater number of males recording such statements was largely responsible for the four person discrepancy at the extreme end of task restricted participants. Overall however, these differences were not noted in regards to the overall decision path.

Concerning participants who followed a decision path of elaborating on student characteristics in their diaries, the number was relatively small. Four female and one male participant fit this path, and at the extreme end of this category, two female participants were identified. Although this decision path is represented by only 13% of the sampled participants, it was considered important to analyze this further due to the ramifications of classroom assessment practices based on personal inferences and resultant possible judgements about student classifications and quality of life. Results, however, must be considered within bounds of the small number of participants (21% of females and 5% of males).

A greater number of female participants followed a decision path that was classified as one of elaborating beyond the information available to them in the portfolio assignment. In particular, two female participants (10% of the sample) went to the extreme of including several categories of inference about the students whose work they were marking. No male participants were noted at this end of the decision path. Although the number of participants is small and gender differences must be considered cautiously, the possible impact of such decisions on classroom assessment practices suggests that they not be overlooked. Unlike the decision path of task restricted participants, the pattern of student elaboration was found to contain a greater number of

females from the initial analysis. The lone male participant did not include additional elaboration comments in his diary, while all of the female participants discussed some further cluster of comment. The limited number of individuals identified in this decision path appeared to differ across the participant groups. This is noted in the definitional category of elaboration: student classification.

As described above, there were a total of 5 participants who made 7 student classification statements -- 4 females (22%) and 1 male (5%). The 4 female participants made 6 of these comments, while a single instance was recorded by the male participant. Although this category of comment is represented by a limited number of examples, gender differences are identified both in the category itself and in its connection to further elaborative comments. The limited number of classification statements naturally hinders the ability to make strong claims regarding this category. As Bachor and Baer (2000) note, however, the suggestions made in these comments highlights the possible impacts of such inferences on classroom assessment practices and students who are considered in this manner. Further investigations of these sorts of suggested inferences in assessment should not ignore the possible differences found in this investigation.

A final category of comment considered in this investigation centered on making inferences regarding a student's quality of life. This category of comment was also considered elaborative in regards to evaluative statements made on the basis of assignment collections available to participants. Across gender group, there was little difference in either participant or frequency counts, with a total of only 6 recorded instances divided evenly between the two sample groups.

Summary. The overall analysis of the diaries recorded by female and male participants revealed that the general content, style, comment types, concerns, assessment criteria and evaluations were comparable across groups. Differences between the gender groups were typically minor. In general, comment categories exhibited like distributions and within group differences were overall more apparent than between group comparisons.

An exception to the similarities across gender was found for the co-occurrence of evaluations and interventions. Female participants recorded a generally greater number of intervention statements, and these occurred together with both evaluation comments

and person based variables in greater numbers than was noted among male participants. While the relatively small number of co-occurring instances in this study is a cautionary note, approximately 50% of the female participants' diaries sampled contained at least one occurrence of this. This suggests that the propensity to offer intervention suggestions and solutions to evaluative outcomes may be greater among female than male participants.

Female pre-service teacher diaries analyzed in this investigation contained a relatively greater number of comments regarding classroom statements and concerns about the classroom context in relation to the assessment of student assignments. This however was only the case for frequency counts, as participant numbers differed by only one individual across gender group. Although female participants expressed more concern about understanding this contextual variable, differences are not necessarily suggested in light of the lack of distinction noted in the participant counts.

Although the number of participants who made classification statements about a fictional student was limited to 13% of the total population, the potential implications of these suggestions drew attention to this category of comment. More females than males made student classification statements both in terms of frequency and participant counts. In turn, a small number of female participants extended beyond this category to infer qualities that were considered reflective of student elaboration decision paths. While slightly over 20% of females and 5 % of males made classification statements, the relationship between inferences of this nature and possible gender differences would require additional investigation to draw more firmly rooted conclusions. Recommendations for future research are suggested following a brief discussion of the relationship of these findings to previous research.

Relation to Previous Research

Consistent with efforts to further understanding of classroom assessment practices, this investigation underscores and builds upon a variety of findings noted in previous research. There are a number of outcomes that related to identified differences between classroom assessment as practiced by teachers (and pre-service teachers) and recommendations made within the measurement tradition. Likewise, the gender-based analysis -- which forms the basis of this investigation -- builds upon findings described in

previous studies. Through the intersect of these strands of research, it is hoped that further understanding into classroom assessment and recommendations for future research emerge.

Assessment practices in general. Although the intent of this investigation was to conduct a gender-based analysis of the diary contents for the sample of pre-service teachers considered, some assessment practices identified are consistent with and lend support to those noted in othe0r investigations of teacher practice. Essentially, the practices articulated within the diaries of these teachers are in line with those described in several previous investigations of educator assessment. In particular, the pre-service teacher diaries analyzed imply that assessment is by no means a straightforward enterprise concentrated on matters that have traditionally informed the measurement perspective. That is, the deliberations over technical sophistication and standardization that have been so ubiquitous within the measurement tradition were not found to be a paramount focus in the writings of these individuals (Airasian & Jones, 1993; Cross & Frary, 1999; Stiggins 1986; Wilson, 1989, 1990a). However, similar to the identified practices noted in other investigations, these pre-service teachers appeared to follow practices that were complex, diverse and descriptive of an undertaking fraught with implicit, intuitive and personally evaluative inclinations (Airasian & Jones, 1993; Bachor & Anderson, 1994; Chase, 1986; McCallum et al., 1993; Mavrommatis, 1997; Wyatt-Smith, 1999). Of further interest was the finding that student characteristics (as inferred from the submitted tasks) were stated as being relevant factors of assessment by several participants. Although this had been cited in previous investigations (Brookhart, 1991, 1993; Whittington, 1999), its occurrence here is particularly telling in light of the fact that the students were fictional and all inferences were based purely on textual information. It might be expected that these sorts of concerns would be more prominent in the classroom, where educators interact with students and engage in exchange that is much more likely to foster a personal and emotional context.

Regarding the investigations of classroom assessment that are the direct predecessors to this study, the findings here tend to corroborate those of previous work. As noted by Bachor and Baer (2000), the majority of participants concentrated the greatest number of comments on establishing criteria and evaluating the submitted tasks.

While this may have been an artifact of the study itself, participants of both gender groups also tended to concentrate at this level of assessment. However, as both Bachor and Baer and Shulha (1999) also describe, much of the data included descriptions of student feedback, feeling factors, the desire to know the participant and so forth. Social and interpersonal notions of assessment were included by several participants despite the relatively sterile and fictional nature of the information available to them.

While most participants across gender group did concentrate their described efforts on marking the submitted assignments, as Wilson and Martinussen (1999) found, this tends not to be done in a straightforward manner. In fact, various inferred characteristics were found to shape judgements or otherwise influence evaluation in their investigation as well as in this study. Bachor and Baer (2000) further found that a minority of individuals recorded elaborative inferences regarding the fictional students. This was noted in this analysis too. Assessment across the participant groups analyzed in this investigation was a complex enterprise consisting of a combination of criteria and evaluations coupled with inferences of various sorts that centered on several factors. This generalization is not only consistent with earlier investigations, it also appears to hold across the two gender based groups examined.

Gender-based analyses. The limited number of investigations including teacher gender as a variable of analysis have so far offered conflicting and inconclusive results (Brandt et al, 1975; Centra & Gaubatz, 2000; Gundersen et al., 1996; Stake & Katz, 1982). Assessment itself has been portrayed as a process of multiple interacting variables, of which educator gender is considered to be one of a possible number of influential factors (Lindow et al., 1985). In this analysis, gender was similarly found to be a complex factor that was not noted to influence assessment in a straightforward manner. This is consistent with the findings of previous studies, which have generally not produced main effects, and describe gender differences as minimal relative to overall variation within groups (for example, Brandt et al., 1975; Gunderson et al., 1996; Relich, 1996). The pre-service teacher diaries across gender group were comparable in regards to general contents, code clusters and other overarching considerations. Within group differences were more notable than those between groups, and possible gender differences did not emerge until a more detailed level of analysis was conducted.

The greater number of intervention statements recorded by female participants and the more frequent co-occurrence of these comments with evaluation and person-based variables may at least in part be congruous with investigations of classroom interaction. If classroom interactions are considered part of the information -- albeit subjective -- that informs assessment and often acts as pseudo-assessment in itself, there is at least tenuous ground for allowing this comparison. Studies of teacher/student interactions have reported that female teachers are more positive towards pupils, offer more sympathy responses, give more encouragement and provide a greater amount of praise (Good et al., 1973; Stake & Katz, 1982). In this investigation, female participants were more likely to provide interventions or solutions to evaluations of pupil work. Interventions can be seen as analogous to encouragement and praise in that they offer a means to improved or continued outcomes and a possible avenue to remedy undesirable assessments. There is arguably an emotive aspect to intervention statements, in the sense that they not only present solutions to the student, they may also provide educators with a means to build on and/or ameliorate the ramifications of an evaluation. While this comparison is speculative at best, the notion that female educators appear to differ in the types of feedback they offer students is noted.

Much of the evaluation and assessment that goes on in the classroom is subjective, informal and intuitive (see for example Airasian & Jones, 1993; Chase, 1986; Mavrommatis, 1997). In regards to gender based investigations, Ehrenberg and Goldhaber (1995) found gender to be a significant factor of teachers' subjective evaluations. Similarly, in the participant diaries analyzed here, portions of the content centered on discussions beyond objective assessment practices. While gender differences for comments pertaining to the classroom context were noted, differences for other areas of non-achievement statements were not found. However, it is suggested that the more subjective aspects related to classroom assessment (such as knowing the student, understanding the classroom, personal background and concerns over students' affective state) may deserve greater attention when attempting to unravel the role of gender.

A final area noted within the analysis concerns the elaborative inferences and student classification statements made by a select number of participants. As described, these inferences and the scant nature of the evidence offered for them underscore their

possible impact. Investigations of special education referral are related to this group of comments by the fact that student classifications pertain to statements focused on special needs categories. Overall, Ritter (1989) found that in the classroom setting, female teachers rated behaviour problems higher in general than did their male counterparts. Likewise, McIntyre (1988, 1990) concluded that female teachers exhibited higher referral rates for externalized problems/aggressive behaviours. In the diaries sampled here, more female participants made student classification statements. Although the connection between referral rates and classification statements is suggestive only, the greater number of female pre-service teachers suggesting possible special education classifications or interventions would appear to draw attention to the role of such implications in classroom assessment.

Future Directions

The role of gender as a variable in classroom assessment is not straightforward and does not appear to exhibit main effects on classroom assessment practices. This investigation supports the conclusion of previous research in articulating that gender tends to have minimal influence relative to overall practice, but that that it may be exhibited as an interaction or more particular effect (for example DeVoe, 1990; Gunderson et al., 1996; Neville et al., 1983; McIntyre, 1988). The analysis of the participant diaries highlights possible strands that may guide further research and investigation.

In particular, further research directed at the nature of interventions and content related to suggestions and recommendations may offer insight into this aspect of assessment. The relatively greater number of intervention statements offered by female participants, as well as previous investigations into classroom interaction (see Good et al., 1973; Stake & Katz, 1982), suggest possible differences here. However, the small sample size in this investigation and the limited number of previous studies conducted to date does not allow for well-substantiated conclusions. Rather, educator intervention and assessment based suggestions are seen as areas of student evaluative feedback that future examinations of classroom assessment and gender may consider. In addition, such investigations may include potentially mitigating factors that were beyond the scope of this investigation, including grade level. A research design exploring the multiple

interaction of teacher and student gender as well as teacher and student grade level could serve to elucidate possible effects of these variables.

Subjective or non-achievement aspects of evaluation, such as concern over the classroom context or a student's emotional state, are another avenue that further research may follow. The differences found in this study regarding classroom centered comments, along with Ehrenberg and Goldhaber's (1995) findings concerning gender, race and subjective evaluation, support the call for a greater focus on these aspects of assessment. The inclusion of ethnicity and student gender are additional variables that were not included in this study (student gender was not identified in the study). Such factors may be explored in further investigations examining interaction effects.

The inclusion of classification statements suggests another area that future research may examine. The greater number of classification comments made by female participants, coupled with findings from other investigations in which some clusters of special education referral rates differed across gender (Ritter, 1989; McIntyre, 1988, 1990), highlights this component of assessment. Further examination of such inferences is recommended because of the sample size in this investigation and the tentative connection suggested between this and special education referral rates. A greater understanding of the minority of educators who make classification and other such inferences is proposed, not only to examine possible gender differences, but also because of the potential impact such suggestions may have regarding assessment in general.

In addition to the suggestions for further research identified from the outcomes of this analysis, several limitations within this study implicate additional directives. The limited number of male participants across the entire sample of diaries collected (19 of 127) resulted in a relatively small comparative group of participants across gender. Increase in sample size, particularly the inclusion of males, is suggested. As noted above, the addition of other variables such as student gender, ethnicity and so on are further factors that may be included in efforts to understand such interaction effects on assessment practices.

The descriptive nature of this analysis does not allow for the positing of hypotheses concerning underlying causes, and further research would certainly be required to begin discussions that extend beyond correlations. Moreover, the simulated

nature of the study, although a novel means towards understanding assessment, would require 'real world' exploration to connect these practices to the classroom. While some of these connections have been attempted through the literature review, other areas such as the role of student and classroom concern may alter in actual teaching contexts. Future investigations employing simulations of this sort would benefit from a more intimate connection with actual practice, for example in the form of observation, teacher records and work samples.

While the analysis conducted here appears to both corroborate some of the findings from previous studies, as well as offer possible directions for further research, its descriptive nature naturally fosters more questions and suggestions than answers and solutions.

REFERENCES

Airasian, P.W. & Jones, A.M. (1993). The teacher as applied measurer: Realities of classroom measurement and assessment. *Applied Measurement in Education, 6,* 241-254.

Anderson, J.O. (1989). Evaluation of student achievement: Teacher practices and educational measurement. *Alberta Journal of Educational Research, 35*, 123-133.

Anderson, J.O. (1990). Editorial: Assessing classroom achievement. *Alberta Journal of Educational Research, 36*, 1-3.

Anderson, J.O. (1999). Modeling the development of student assessment. *Alberta Journal of Educational Research, 45*, 278-287.

Anderson, J.O. (2000, May). *The evaluation of student achievement: Preliminary analyses in modeling teacher decisions.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Anderson, J.O., Bachor, D.G., & Baer, M.R. (2001, April*). Using portfolio assessment to study classroom assessment*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Bachor, D. (1990). Towards improving assessment of students with special needs: Expanding the data base to include classroom performance. *The Alberta Journal of Educational Research*, *36*, 65-77.

Bachor, D. & Anderson, J. O. (1990). *Assessment practices in the primary program: Description of observed practices, functional factors, and recommendations as to some general principles: Final Report*. Victoria, BC: Province of British Columbia, Ministry of Education.

Bachor, D.G., & Anderson, J.O. (1994). Elementary teachers' assessment practices as observed in the province of British Columbia. *Assessment in Education, 1*, 63-93.

Bachor, D.G., Anderson, J.O., & Walsh, J. (1994). Classroom assessment and the relationship to representativeness, accuracy, and consistency. *Alberta Journal of Educational Research, 40*, 247-262.

Bachor, D.G. & Baer, M.B. (2000, May). *An examination of pre-service teachers' portfolio diaries.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Bachor, D.G. & Crealock, C. (1986). *Instructional strategies for students with special needs*. Scarborough, ON: Prentice-Hall, Canada.

Bachor, D.G., Shulha, L.N., Anderson, J.O., Wilson, R.J., & Muir, W. (1998, October). Developing and Maintaining Collaborative Research Partnerships in Classroom Assessment. Paper presented at the Measurement and Evaluation: Current and Future Research Directions for the New Millennium Conference, Banff, AB.

Bacon, D. (2000). The money in testing: Testing one, two, three. *Z Magazine, 12(9),* 40-44.

Barba, R. & Cardinale, L. (1991). Are females invisible students? An investigation of teacher student questioning interactions. *School Science and Mathematics, 91*, 306-310.

Bateson, D.J. (1990). Measurement and evaluation practices of British Columbia science teachers. *The Alberta Journal of Educational Research, 36*, 45-51.

Bernard, M. E. (1979). Does sex role behavior influence the way teachers evaluate students? *Journal of Educational Psychology, 71*, 553-562.

Boxall, W. & Gilbert, J. (1999). Developing a holistic assessment stance in student teachers. *Assessment in Education: Principles, Policy and Practice, 6*, 247- 261.

Brandt, L.J. & Hayden, M.E. (1974). Male and female teacher attitudes as a function of students' ascribed motivation and performance levels. *Journal of Educational Psychology, 66*, 309-314.

Brandt, L.J., Hayden, M.E. & Brophy, J.E. (1975). Teachers' attitudes and ascription of causation. *Journal of Educational Psychology, 67*, 677-682.

Broadfoot, P., Abbott, D., Croll, M.O., Pollard, A. & Towler, L. (1991). Implementing national assessment: Issues for primary teachers. *Cambridge Journal of Education, 21*, 153-168.

Brookhart, S.M. (1993). Teachers' grading practices: meaning and values. *Journal of Educational Measurement, 30*, 123-142.

Brookhart, S.M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice, 18*, 5-13.

Brophy, J. (1985). Interactions of male and female students with male and female teachers. In L.C. Wilkinson & C.B. Marrett (Eds.), *Gender Influences in Classroom Interaction* (pp. 115-142). Orlando, FL: Academic Press Inc.

Butterfield, S., Williams, A., & Marr, A. (1999). Talking about assessment: Mentor-student dialogues about pupil assessment in initial teacher training. *Assessment in Education, 6*, 225-246.

Centra, J.A. & Gaubatz, N.B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 70*, 17-33.

Chase, C.I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement, 23*, 33-41.

Crooks, T.J. (1988). The impact of classroom evaluation practices on student. *Review of Educational Research, 58*, 438-481

Cross, L.H. & Frary, R.B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*, 53-72.

DeVoe, D.E. (1990). The influence of teacher gender and student grade level on student teachers' beliefs concerning student decision making. *Journal of Instructional Psychology, 17*, 197-205.

Ehrenberg, R.G. & Goldhaber, D.D. (1995). Do teacher race, gender and ethnicity matter? Evidence from the national educational longitudial study of 1988. *Industrial and Labor Relations Review, 48*, 547-561.

Fennema, E. & Peterson, P. (1985). Autonomous learning behavior: A possible explanation of gender-related differences in mathematics. In L.C. Wilkinson & C.B. Marrett (Eds.), *Gender Influences in Classroom Interaction* (pp. 17-36). Orlando, FL: Academic Press Inc.

Gall, M.D., Borg, W.R & Gall, J.P. (1996). *Educational Research*. White Plains, NY: Longman Publishers.

Gipps, C. (1994). Development of educational assessment: What makes a good test? *Assessment in Education, 1,* 283-291

Good, T.L., Sikes, J.N. & Brophy, J.E. (1973). Effects of teacher sex and student sex on classroom interaction. *Journal of Educational Psychology, 65*, 74-87.

Goodwin, L.D. & Stevens, E.A. (1993). The influence of gender on university faculty members' perceptions of "good" teaching. *Journal of Higher Education, 64*, 166-185.

Gullickson, A.R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research, 77*, 244-248.

Gullickson, A.R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research, 79*, 96-100.

Gullickson, A.R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23*, 347-354.

Gundersen, D.E., Tinsley, D.B. & Terpstra, D.E. (1996). Empirical assessment of impression management biases: The potential for performance appraisal error. *Journal of Social Behavior and Personality, 11*, 57-76.

Hattie, J. & Jaeger, R. (1998). Assessment and classroom learning: A deductive approach. *Assessment in Education, 5*, 111-122

Henebry, K.L. & Diamond, J.M. (1998). The impact of student and professor gender on grade performance in managerial finance courses. *Financial Practice and Education, 8*, 94-101.

Hopf, D. & Hatzichristou, C. (1999). Teacher gender-related influences in Greek schools. *British Journal of Educational Psychology, 69*, 1-18.

Howell, D.C. (1995). *Fundamental Statistics for the Behavioral Sciences.* Belmont, CA: Duxbury Press.

Huberman, M.A & Miles, M.B. (1994). Data management and analysis methods. In N.K. Denizin and Y.S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 428-443). Thousand Oaks, CA: Sage Publications.

Kazdin, A.E. (1982*). Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York: Oxford University Press.

Lee, M. (2000, May). *The emergence and anatomy of collaboration in a teacher/researcher assessment project: A cognitive and cultural view.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Lindow, J., Marrett, C.B. & Wilkinson, L.C. (1985). Overview. In L.C. Wilkinson & C.B. Marrett (Eds.), *Gender Influences in Classroom Interaction* (pp. 1-15). Orlando, FL: Academic Press Inc.

Locke, C. (2000, May). *Student control and responsibility: The influence of two teachers' classroom assessment practices.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Lomax, R.G. (1996). On becoming assessment literate: An initial look at pre-service teachers' beliefs and practices. *Teacher-Educator, 31*, 292-303.

McCallum, B., McAlister, S., Brown, M. & Gipps, C. (1993). Teacher assessment at key stage one. *Research Papers in Education, 8*, 305-328.

McIntyre, I. (1990). Classroom assessment : What research do practitioners need? In D.J. Bateson (Ed.), *Classroom Testing in Canada* (pp. 82-90). Vancouver, BC: Center for Applied Studies in Evaluation, University of British Columbia.

McIntyre, L.L. (1988). Teacher gender: A predictor of special education referral. *Journal of Learning Disabilities, 21*, 382-383.

McIntyre, L.L. (1990). Teacher standards and gender: Factors in special education referral. *Journal of Educational Research, 83*, 166-172.

McLean, L.D. (1990). Time to replace the classroom test with authentic measurement. *The Alberta Journal of Educational Research, 36*, 78-84.

Maguire, T.O. (1990). Grounded authentic assessment and teacher education. In D.J. Bateson (Ed.), *Classroom Testing in Canada* (pp. 82-90). Vancouver, BC: Center for Applied Studies in Evaluation, University of British Columbia.

Marshall, C. & Rossman, G.B. (1989). *Designing Qualitative Research.* Newbury Park, CA: Sage Publications.

Mavrommatis, Y. (1997). Understanding assessment in the classroom: Phases of the assessment process – the assessment episode. *Assessment in Education, 4*, 381-399.

Mayo, S.T. (1964). What experts think teachers ought to know about educational assessment. *Journal of Educational Measurement, 1*, 79-88.

Miles, M.B & Huberman, M.A. (1984). *Qualitative Data Analysis: A Sourcebook of New Methods*. Newbury Park, CA: Sage Publications

Morse, L.W. & Handley, H.M. (1985). Listening to adolescents: Gender differences in science classroom interactions. In L.C. Wilkinson & C.B. Marrett (Eds.), *Gender Influences in Classroom Interaction* (pp. 37-56). Orlando, FL: Academic Press Inc.

Muhr, T. (1997). *Atlas/ti: The Knowledge Workbench.* Version 4.1, Berlin: Scientific Software Development.

National Association for the Education of Young Children. (1988). NAEYC position statement in the standardized testing of young children 3 through 8 years of age. *Young Children, 43*, 42-47

Nagel, N. & Driscoll, A. (1992, April). *Dilemmas caused by discrepancies between what they learn and what they see: Thinking and decision making of preservice teachers*. Paper presented at the annual meeting of the American Research Association, San Francisco, CA.

Neville, D.D., Stephenson, B.B. & Philbrick, J.H. (1983). Gender effects on performance evaluation. *The Journal of Psychology, 115*, 165-169.

Notman, D. (2000, May). *The effects of portfolio assessment and student-led conferences on ownership and control.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Nuttall, D. (1987). The validity of assessment. *European Journal of Psychology of Education, 2*, 109-118.

Okpala, C.O. (1996). Gender-related differences in classroom interaction. *Journal of Instructional Psychology, 23*, 275-285.

Petrick, B. (2000, May). *The relationship between learning and assessment in the elementary classroom.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Pulakos, E.D., Oppler, S.H., White, L.A. & Borman, W.C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology, 74*, 770-780.

Relich, J. (1996). Gender, self-concept and teachers of mathematics: Effects on attitudes to teaching and learning. *Educational Studies in Mathematics, 30*, 179-195.

Richards, T.J. & Richards, L. (1994). Data management and analysis methods. In N.K. Denizin and Y.S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 445-461). Thousand Oaks, CA: Sage Publications.

Ritter, D.R. (1989). Teachers' perceptions of problem behavior in general and special education. *Exceptional Children, 55*, 559-564.

Rogers, T.W. (1990). Current educational climate in relation to testing. *Alberta Journal of Educational Research, 36*, 52-64.

Rogers, T.W. (1999). Measurement and evaluation: current and future research directions for the new millennium. *Alberta Journal of Educational Research, 45*, 329-332.

Rong, X.L. (1996). Effects of race and gender on teachers' perception of social behavior of elementary students. *Urban Education, 31*, 261-290.

Schafer, W.D. (1989, March). *Assessment essentials in professional education of teachers.* Paper presented to the annual meeting of the American Educational Research Association, San Francisco, CA.

Shulha, L. M. (1999). Understanding novice teachers' thinking about assessment. *Alberta Journal of Educational Research, 45*, 288-303.

Shulha, L.M. (2000a, May). *Implementing newer principles of assessment into planning and delivery instruction.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Shulha, L.M. (2000b, May). *Understanding collaboration: Lessons from a large scale research project in teachers' classroom assessment practices.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Edmonton, AB.

Shulha, L. M., Wilson, R.J., & Anderson, J.O. (1999). Investigating teachers assessment practices: Exploratory, non-foundationalist, mixed method research. *Alberta Journal of Educational Research, 45*, 304-313.

Stake, J.E. & Katz, J.F. (1982). Teacher-pupil relationships in elementary school classrooms: Teacher-gender and pupil-gender differences. *American Education Research Journal, 19*, 465-471.

Stiggins, R.J. (1990a). Making assessment training relevant for teachers. In D.J. Bateson (Ed.), *Classroom Testing in Canada* (pp. 82-90). Vancouver, BC: Center for Applied Studies in Evaluation, University of British Columbia.

Stiggins, R.J. (1990b). Towards a relevant classroom assessment research agenda. *Alberta Journal of Educational Research, 36*, 92-97.

Stiggins, R.J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*, 23-27.

Stiggins, R.J., & Bridgeford, N.J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*, 271-286.

Stiggins, R.J., Conklin, N.F., & Bridgeford, N.J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5*, 5-17.

Tesch, R. (1990). Qualitative Research: Analysis Types and Software Tools. Bristol, PA: The Falmer Press.

Warwick, W.P. & Jatoi, H. (1994). Teacher gender and student achievement in Pakistan. *Comparative Education Review, 38*, 377-399.

Whittington, D. (1999). Making room for values and fairness: teaching reliability and validity in the classroom context. *Educational Measurement Issues and Practice, 18*, 14-22.

Wiles, C.A. (1992). Investigating gender bias in the evaluations of middle school teachers of mathematics. *School Science and Mathematics, 92*, 295-298.

Wilson, R.J. (1989). Evaluating student achievement in an Ontario high school. *Alberta Journal of Educational Research, 35*, 134-144.

Wilson, R.J. (1990a). Classroom procedures in evaluating student achievement. *Alberta Journal of Educational Research, 36*, 4-17.

Wilson, R.J. (1990b). The context of classroom procedures in evaluating students. In D.J. Bateson (Ed.), *Classroom Testing in Canada* (pp. 82-90). Vancouver, BC: Center for Applied Studies in Evaluation, University of British Columbia.

Wilson, R.J. (1999a). Aspects of validity in large-scale programs of student assessment. *Alberta Journal of Educational Research, 45*, 333-343.

Wilson, R.J. (1999b). Classroom assessment investigations: introduction. *Alberta Journal of Educational Research, 45*, 263-266.

APPENDIX A

Codes Assigned to Participants' Diary Data

Code	Definition
	Superordinate Category
	Assignment-Based
Context	
Classroom	Points raised about the task, teacher, classroom, et cetera
Subject's Background	Comments made about the pre-service teacher's own background
Criteria	
Establishing	Process of establishing assessment criteria
Reviewing/ Refining	Subsequent reviewing and refining of initial criteria
Questions/Comments	
Concerns	Queries raised about the assignment/task
Positives	Comments made about the assignment/task
	Intervention
Comments	Hints of an intervention, such as suggestions directed at task, class, teacher, et cetera
Student	Specific suggestions for an intervention, directed at either student A, B, or C
	Person Based
Competency	
Performance on Task	Statements about performance on task, directed to Student A, B, or C indicating how well he/she did on an assignment
Student	Statements directed at the student going beyond task comments, designating the student, eg. Student A is poor speller
Classification	Statements directed at the student going beyond assignment comments. Designating one of the students as having a special educational need, eg. Learning Disabled, gifted, et cetera
Quality of Life	Statements directed at the student's family life or social life, such as commenting about their socio-economic status
Comments	
Knowledge of	Comments indicating that knowing the student was important to participant's understanding of his/her progress as a learner
Affective State	Statements made about the emotional state of either Student A, B, or C

VITA

Surname: Baer Given Names: Markus Rene

Place of Birth: Vancouver, BC, Canada Date of Birth: October 28, 1968

Educational Institutions Attended:

University of Victoria	1998-2001
Simon Fraser University	1987-1995

Degrees/Certificates Awarded:

B.G.S. Simon Fraser University	1995
P.D.P. Simon Fraser University	1995

Honours and Awards:

University of Victoria Graduate Fellowship	1998-2000
President's Honour Role	1992-1994
Simon Fraser Open Scholarship	1990-1994

Publications:

Bachor, D.G. & Baer, M.R. (1999). Illustrations of special needs education service provisions across canada. In C. Brock & R. Griffin (Eds.), International Perspectives on Special Needs Education. London: John Catt Educational Ltd.

Bachor, D.G. & Baer, M.R. (May, 2000). An examination of pre-service teachers' portfolio diaries. Paper presented at the Canadian Society for Studies in Education Conference, Edmonton, Alberta.

Baer, M.R. & Roberts, J.J. (in press). Complex HIV treatment regimens and patient quality of life. Canadian Psychology.

Anderson, J.O, Bachor, D.G & Baer, M.R. (April 2001) Using Portfolio Assessment to Study Classroom Assessment Practice. Paper to be presented at the American Educational Research Association Annual meeting, Seattle, Washington.

Bachor, D.G. & Baer, M.R. (in press). Pre-service Teachers' Portfolio Diaries: An examination of simulated assessment practices. Alberta Journal of Educational Research.

PARTIAL COPYRIGHT LICENSE

UNIVERSITY OF VICTORIA PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain by the University of Victoria shall not be allowed without my written permission.

Title of Thesis:

Pre-service Teacher Assessment Practices: a Gender-based Analysis.

Auth[redacted]

Markus R. Baer

August 20, 2001