ISSUES OF RELIABILITY AND VALIDITY
IN USING PORTFOLIO ASSESSMENT TO MEASURE
FOREIGN LANGUAGE TEACHER PERFORMANCE

DISSE RTATION

Presented in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Anna-Marie Gelines, B.A., M.A.

*****

The Ohio State University
1998

Dissertation Committee:

Professor Gilbert Jarvis, Adviser

[Signature]
Co-Adviser

Professor Ayres D’Costa, Adviser

[Signature]
Co-Adviser

Professor Cora Kaylani

College of Education
ABSTRACT

Portfolio assessment is an increasingly common trend in teacher education programs because the multifaceted nature of portfolios allows for the demonstration of a variety of strengths and competencies essential to effective teaching. Portfolios provide preservice teachers a means with which they can demonstrate their conceptual knowledge and their ability to apply that knowledge in the classroom (pedagogical content knowledge). The reliability and validity of the portfolio scores in making professional decisions concerning the competence of preservice teachers, however, is a difficult issue because of the variability found among individual portfolios.

The current study investigated issues of reliability and validity in using portfolios to evaluate the pedagogical abilities of preservice teachers in an elementary foreign language methods course. A correlational analysis among holistic portfolio scores and other measures used to evaluate students in the course demonstrated a positive relationship between portfolio scores and two of the independent variables: performance-based tasks (r=.47) and years of teaching experience (r=.35). The regression analysis indicated that the performance tasks were the best predictor of portfolio scores (22% of the variance explained) followed by years of experience (16%). These results established the predictive and criterion-related validity of the portfolios as an accurate method of measuring pedagogical content knowledge in the course. A principal components factor analysis was conducted on the holistic data to examine construct validity. The results yielded two factors with eigenvalues of 2.37 (conceptual knowledge) and 1.14 (pedagogical content knowledge), and explained 70% of the variance in the instruments. The final exam, GPA, and years of experience loaded on factor one; portfolios and tasks loaded on factor two.
Inter-rater reliability and score reliability were investigated through a Rasch model FACETS analysis and a generalizability study to determine the relative impact of factors such as raters and items on the portfolio scores. The level of rater agreement (r=.81) and the score reliability (r=.90) established by the FACETS analysis showed that the portfolios were rated consistently and provided an accurate measure of teaching competency. The results of the generalizability study indicated that while the rater and item facets alone did not account for a large portion of the variance (1% of total variance for each facet), the item by person (12%) and the rater by person interactions (36%) did account for a larger portion. It was concluded that these interactions might have been due to the individuality and format of the portfolios when combined with the personalities and preferences of the raters. The subjectivity inherent in rating portfolios therefore continues to be a limitation in the implementation of portfolio assessment as well as an area for further research.

Instructing preservice teachers to implement portfolio assessment through personal experience with portfolios may foster their use of portfolios in the future. Subjects were surveyed about their understanding of portfolio assessment and rating scales previous to the course and after developing a portfolio. The results showed that the students had an increased understanding of the nature of portfolio assessment after their experience with portfolios in the course. The students also demonstrated a positive attitude toward portfolios and a willingness to use them with future students.
DEDICATION

TO MY LOVING HUSBAND ALBERTO

I appreciate your love and support in getting me this far!

and

FOR MARIA ANNELIESE, MY BEAUTIFUL BABY GIRL

who literally accompanied me through it all.
ACKNOWLEDGMENTS

I wish to thank my advisers, Dr. Gilbert Jarvis and Dr. Ayres D’Costa, whose insights and intellectual support challenged me to do my best. I appreciate their patience while I learned.

I wish also to thank Dr. Cora Kaylani for stimulating discussions and multiple draft readings. Much appreciation to Dr. Debbie Wilburn Robinson who read the initial draft and allowed me to implement portfolios in her methods class.

I am grateful to my five raters and friends: Teresa Benedetti, Michele Bowman, Macel Braxton, Major Rein Foerg, Ph.D., and Dr. Annela Teeman. Without your help, the portfolios would never have been rated!

I am also grateful to the students in the Spring 1996, 617 methods course that suffered through the work with a smile. A special thank you to Hanaa El-Youssef, Laura Mickelson-Barrett and Kimberly Schlichting for the interviews.

Finally, I want to thank my family, especially my parents and my husband, Alberto, and all my friends for always supporting my endeavors, for providing me with every opportunity to learn, and for being with me through it all.
VITA

March 6, 1968 .................. Born - Springfield, Massachusetts

1990 .......................... B.A. Hispanic Studies, Connecticut College
CT Certification 7-12 Spanish

1991 .......................... M.A. Hispanic Studies, New York University
Madrid, Spain

1991-1993 ...................... Instructor, English as a Foreign Language
Universidad de Madrid, Spain

1993-1994 ...................... Spanish Teacher, Westfield H.S.
Westfield, MA

1994-1997 ...................... Graduate Teaching Associate, The Ohio State
University

1997-1998 ...................... Spanish Teacher, Williams M.S.
Longmeadow, MA

PUBLICATIONS

Research Publication

a Foreign Language Teacher Education Course. In DiDonato, R. (Ed.) Building

FIELDS OF STUDY

Major Field: Education
(Foreign and Second Language Education)

Minor Fields: Spanish Culture and Civilization
Testing and Measurement
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Statement of the problem</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Research questions</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Theoretical bases</td>
<td>8</td>
</tr>
<tr>
<td>1.4 Assumptions</td>
<td>15</td>
</tr>
<tr>
<td>1.5 Definition of key terms</td>
<td>15</td>
</tr>
<tr>
<td>1.6 Limitations of the study</td>
<td>18</td>
</tr>
<tr>
<td>1.7 Conclusion</td>
<td>19</td>
</tr>
<tr>
<td>2. Review of the literature</td>
<td>20</td>
</tr>
<tr>
<td>2.1 Historical perspective of assessment</td>
<td>21</td>
</tr>
<tr>
<td>2.2 Alternative assessment</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Overview of assessment principles</td>
<td>26</td>
</tr>
<tr>
<td>2.4 The theory of multiple intelligences</td>
<td>32</td>
</tr>
<tr>
<td>2.5 Problems with current preservice teacher evaluation</td>
<td>36</td>
</tr>
<tr>
<td>2.6 Portfolio assessment in teacher education</td>
<td>38</td>
</tr>
<tr>
<td>2.7 Teacher portfolios in practice</td>
<td>43</td>
</tr>
<tr>
<td>2.8 Limitations of teacher portfolios</td>
<td>46</td>
</tr>
<tr>
<td>2.9 Issues of reliability in portfolio assessment</td>
<td>48</td>
</tr>
<tr>
<td>2.10 Issues of validity in portfolio assessment</td>
<td>52</td>
</tr>
<tr>
<td>2.11 Conclusion</td>
<td>55</td>
</tr>
<tr>
<td>3. Methodology</td>
<td>57</td>
</tr>
<tr>
<td>3.1 Context of the study</td>
<td>57</td>
</tr>
<tr>
<td>3.2 Subjects</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.3 Instrumentation</td>
<td>59</td>
</tr>
<tr>
<td>3.4 Procedures</td>
<td>62</td>
</tr>
<tr>
<td>3.5 Variables</td>
<td>63</td>
</tr>
<tr>
<td>3.5.1 Dependent variable</td>
<td>63</td>
</tr>
<tr>
<td>3.5.2 Independent variables</td>
<td>65</td>
</tr>
<tr>
<td>3.4 Hypotheses</td>
<td>67</td>
</tr>
<tr>
<td>3.5 Data analysis</td>
<td>68</td>
</tr>
<tr>
<td>4. Results and Discussion</td>
<td>73</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>73</td>
</tr>
<tr>
<td>4.2 Results</td>
<td>75</td>
</tr>
<tr>
<td>4.2.1 Validity</td>
<td>75</td>
</tr>
<tr>
<td>4.2.2 Reliability</td>
<td>91</td>
</tr>
<tr>
<td>4.2.3 Preservice Teachers' Attitudes</td>
<td>108</td>
</tr>
<tr>
<td>4.3 Discussion</td>
<td>112</td>
</tr>
<tr>
<td>4.3.1 Validity</td>
<td>112</td>
</tr>
<tr>
<td>4.3.2 Reliability</td>
<td>114</td>
</tr>
<tr>
<td>4.3.3 Preservice Teachers' Attitudes</td>
<td>119</td>
</tr>
<tr>
<td>4.4 Conclusion</td>
<td>120</td>
</tr>
<tr>
<td>5. Summary, Recommendations, and Implications</td>
<td>122</td>
</tr>
<tr>
<td>5.1 Overview of the Study</td>
<td>123</td>
</tr>
<tr>
<td>5.2 Summary of Findings</td>
<td>125</td>
</tr>
<tr>
<td>5.2.1 Question One: Validity</td>
<td>125</td>
</tr>
<tr>
<td>5.2.2 Question Two: Reliability</td>
<td>131</td>
</tr>
<tr>
<td>5.2.3 Secondary Questions: Attitudes</td>
<td>138</td>
</tr>
<tr>
<td>5.3 Pedagogical Implications</td>
<td>138</td>
</tr>
<tr>
<td>5.4 Limitations to the Study</td>
<td>141</td>
</tr>
<tr>
<td>5.5 Recommendations for Further Research</td>
<td>143</td>
</tr>
<tr>
<td>List of References</td>
<td>147</td>
</tr>
<tr>
<td>Appendices</td>
<td>157</td>
</tr>
<tr>
<td>Appendix A: Student information sheet</td>
<td>157</td>
</tr>
<tr>
<td>Appendix B: Assessment survey</td>
<td>158</td>
</tr>
<tr>
<td>Appendix C: Portfolio guidelines</td>
<td>161</td>
</tr>
<tr>
<td>Appendix D: Analytic rating scale</td>
<td>164</td>
</tr>
<tr>
<td>Appendix E: Holistic rating scale</td>
<td>165</td>
</tr>
<tr>
<td>Appendix F: Performance tasks</td>
<td>167</td>
</tr>
<tr>
<td>Appendix G: Final exam</td>
<td>169</td>
</tr>
<tr>
<td>Appendix H: Portfolio Interview Questions</td>
<td>170</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Descriptive Statistics of Key Variables ........................................ 78</td>
</tr>
<tr>
<td>4.2</td>
<td>Pearson Correlation Matrix of Key Variables ..................................... 79</td>
</tr>
<tr>
<td>4.3</td>
<td>Regression Analysis of Portfolio Scores on Performance Tasks ............... 82</td>
</tr>
<tr>
<td>4.4</td>
<td>Regression Analysis of Portfolio Scores on Tasks and Years .................. 83</td>
</tr>
<tr>
<td>4.5</td>
<td>Regression Analysis of Portfolio Scores on Tasks, Years, and Exam .......... 83</td>
</tr>
<tr>
<td>4.6</td>
<td>Regression Analysis of Portfolio Scores on Tasks, Years, Exam, and GPA ..... 84</td>
</tr>
<tr>
<td>4.7</td>
<td>Multiple Regression Analysis Showing Change in $R^2$ ........................ 85</td>
</tr>
<tr>
<td>4.8</td>
<td>Factor Loadings and Percent Variance Explained for Principal Variables ...... 87</td>
</tr>
<tr>
<td>4.9</td>
<td>Descriptive Statistics of Holistic Ratings of Portfolios ...................... 93</td>
</tr>
<tr>
<td>4.10</td>
<td>Pearson Correlation Matrix of Holistic Ratings .................................. 93</td>
</tr>
<tr>
<td>4.11</td>
<td>Rater Frequency and Mean Scores .................................................... 97</td>
</tr>
<tr>
<td>4.12</td>
<td>Rasch Model Rater Measures and Fit Statistics .................................... 99</td>
</tr>
<tr>
<td>4.13</td>
<td>Rasch Model Student Measures and Fit Statistics .................................. 101</td>
</tr>
<tr>
<td>4.14</td>
<td>Rasch Model Submission Measures and Fit Statistics ............................ 103</td>
</tr>
<tr>
<td>4.15</td>
<td>Generalizability Study: Variance Component Estimates ......................... 104</td>
</tr>
<tr>
<td>4.16</td>
<td>Mean Ratings of Items on Assessment Survey ...................................... 109</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Scatterplot of Performance Tasks and Portfolio Scores</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>Scatterplot of Teaching Experience and Portfolio Scores</td>
<td>80</td>
</tr>
<tr>
<td>4.3</td>
<td>Scatterplot of Final Exam Grades and Portfolio Scores</td>
<td>81</td>
</tr>
<tr>
<td>4.4</td>
<td>Scatterplot of Overall GPA and Portfolio Scores</td>
<td>81</td>
</tr>
<tr>
<td>4.5</td>
<td>Scatterplot of Final Exam Grades and Overall GPA</td>
<td>90</td>
</tr>
<tr>
<td>4.6</td>
<td>Calibrations of Rater, Student, and Submission Facets</td>
<td>95</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Educators have been reconceptualizing the goals of teacher education in terms of recent theories of what constitutes effective teaching (Barton & Collins, 1993). In line with this reevaluation of the objectives of teacher education is an important trend that may shape evaluation of teacher performance in the future: alternative assessment. Alternative forms of assessment have become widespread in educational contexts in the past decade (Barton & Collins, 1993; Seldin, 1991), and attempts to measure the teaching competencies of preservice teachers through performance-based evaluations have created a variety of alternatives and supplements to the conventional tests that are often used in final program evaluations of future teachers. Portfolio assessment is one of the performance-based assessment tools that is often chosen for implementation in teacher preparation programs because the multifaceted nature of portfolios allows for the demonstration of a variety of strengths and competencies essential to effective teaching. An impetus behind the shift toward using teacher portfolios in the assessment of preservice teachers is the need to evaluate this variety of teaching abilities in the educational context. Portfolios provide preservice teachers with an arena in which they can demonstrate their understanding of conceptual knowledge and their ability to apply that knowledge in the classroom.

As proposed by Gardner (1983) in his theory of multiple intelligences, a complex assessment tool is required to measure an individual's cognitive profile. Portfolios allow
teacher educators to assess the teaching competence of preservice teachers through activities that require the use of a variety of abilities and intelligences. As the field of teacher education attempts to professionalize the preparation of future teachers, there is a need to consider all aspects of an individual’s ability in both training and evaluation. Teacher portfolios reflect the complexity of teaching and recognize the difficulty of measuring all competencies through one standardized score, such as is done with the current state certification exams employed throughout the country. The use of portfolio assessment with preservice teachers is especially relevant to the field of teacher education because it allows for in-depth monitoring of the development of effective teachers (Millis, n.d.). The portfolio also functions as a link between theory and practice (Barton & Collins, 1993), allowing the future teacher to observe and reflect on connections between course work and field experiences (Hartle & DeHart, 1993). Performance-based assessment, including teacher portfolios, are being incorporated into programs of teacher preparation across the country at an increasing rate. At least two states (Connecticut and Kentucky, to date) have implemented teacher portfolios as part of the process of assessing teacher competency for state certification; more states will soon follow, including Ohio (Marilyn Troyer, personal communication, 1997). State departments across the nation are realizing the potential of teacher portfolios in the overall assessment of teachers. In the specific context of this study, the School of Teaching and Learning at The Ohio State University plans to implement a system of teacher portfolios to assess preservice teachers in the Foreign and Second Language Education program.

The use of performance-based assessment in teacher preparation programs, however, can be problematic. As the United States moves toward a national certification system for preservice teachers, the need for standardized criteria with which all teachers can be evaluated becomes more important (Carnegie Forum, 1986). The use of alternative forms of assessment, such as portfolios, that are not standardized to specific norms and
that have varying criteria for effective pedagogy make standardization of national
compétencies in education a difficult task. The reliability and validity of the portfolio scores
in making professional decisions concerning the competence of preservice teachers
becomes an issue because of the variability found among the criteria for portfolio creation,
the individuality inherent in portfolios, and the dependency on subjective ratings versus
objective scores. In essence, it is the very nature of portfolios that makes it difficult to
standardize the criteria for their creation and evaluation.

Although reliability and validity are crucial to the acceptance of portfolio assessment
as an accurate measure of ability and knowledge, limited research has been conducted in
these areas. The majority of the articles published about portfolios include procedural
information, but few address the more important issues of reliability and validity. Of the
46 publications listed in the 1993 Center for Research on Evaluation, Standards, and
Student Testing (CRESST) Alternative Assessments in Practice database, only 13 reported
data on rater agreement (CRESST, 1993); and the numbers are even smaller for
publications concerning the validity of portfolios. The lack of validated alternative
assessment instruments, and the absence of an agreement about what constitutes effective
teaching, reflect the absence of studies of the validity of portfolio assessment; without a
consistent model of teaching competency against which to measure future teachers, and
without validated instruments with which to compare portfolio scores, it is difficult to
establish the validity of portfolio assessment in teacher education. Messick (1994) argues,
however, that “performance assessments must be measured by the same validity criteria as
are other assessments” (p. 13), and Terwilliger (1997) claims that the authenticity of
performance-based assessment, including portfolios, is not a sufficient measure of validity.
He claims that validity of an instrument is “fundamentally linked to how well the device
reflects an underlying construct” (p. 26). Reckase and Terwilliger are among many
researchers that are calling for studies that provide evidence that portfolios are a valid assessment tool before they are incorporated into educational evaluation.

Inter-rater reliability, as shown by the research (CRESST, 1993), is the most commonly studied aspect of portfolio assessment when reliability is investigated, but the reliability of the scores themselves is not always scrutinized. According to Reckase (1997), inter-rater reliability is a measure of the extent to which the scores could be replicated when the portfolios are assessed by different raters using the same rating scales. This is not the same as score reliability which “refers to whether students can perform on a similar set of assignments designed to meet the educational objectives” (Reckase, 1997:30), and he calls for the use of generalizability theory to investigate the reliability of the scores as indicators of a student’s ability to meet educational objectives.

In order for portfolio assessment to become one of the assessment tools used in the certification of future educators standards of reliability and validity in the collection and evaluation of portfolio samples must be achieved. It is therefore imperative that data concerning the reliability of the scores from portfolios and the validity of their use in evaluating preservice teachers be collected and scrutinized. The results from this study contribute to the growing body of literature in support of the use of portfolio assessment in foreign language teacher education, and may open the doors for its use in the foreign and second language teacher education program at Ohio State. It provides a basis for the guidelines, criteria, and scoring rubrics necessary to evaluate the portfolios in an accurate and timely fashion, as well as providing data regarding the reliability of the portfolio scores and the validity of their use in evaluating future teachers.

Statement of the problem

Portfolio assessment has become a popular alternative or supplement for evaluating students in teacher education programs (e.g., Tierney, Carter, & Desai, 1991; Wiggins,
1989; Zollman & Jones, 1994). Current research has investigated the value of using portfolio assessment as an alternative form of evaluation in elementary and secondary classrooms (e.g., Valencia, 1990; Wolf, 1989), and it can be argued that student development cannot be accurately measured solely through the use of standardized tests and single letter grades (Moeller, 1994). The multidimensionality and self-reflection that characterize portfolio assessment appeal to teacher educators, and may be particularly applicable in the evaluation of preservice teachers (Schrier & Hammadou, 1994). As demonstrated above, many of the issues of reliability and validity of this type of assessment, however, are often neglected in the publications regarding portfolios (Calfee & Perfumo, 1993; Reckase, 1995).

The purpose of this study, therefore, is to investigate the issues concerning the reliability and the validity of using portfolio assessment as an evaluation component in a foreign-language educational methods course, with particular emphasis on establishing reliable and valid procedures. This narrower purpose extends to the broader plans for the incorporation of teacher portfolios at the program and state levels in the near future. As a secondary aspect in the study, descriptive data concerning preservice teachers’ attitudes toward portfolio assessment were collected and analyzed. Finally, pedagogical implications for applying the findings of this study are presented.

Research Questions

Current research suggests that the use of portfolios to assess preservice teachers is an attractive alternative or supplement to traditional forms of assessment (Ryan & Kuhs, 1993; Zubizaretta, 1994). Many of the reports about portfolio assessment, however, leave the issues of reliability and validity unaddressed. This study investigates the following broad questions in the use of portfolio assessment: To what extent is portfolio assessment a reliable and valid measurement tool for preservice elementary-school foreign language
teachers? What are the beliefs that future teachers enrolled in a course that implements portfolio assessment have toward assessment and testing? What implications does the investigation of reliable and valid procedures in using portfolios to assess teaching competence have for teacher education programs? With these overall issues in mind, the following questions are the principal focus of the research:

1. As a measure of the validity of portfolio assessment in assessing and predicting foreign language teacher competence, to what extent do the portfolio scores correlate with, and predict scores on, other established measures of:
   a. application of pedagogical content knowledge as measured by performance-based tasks completed in the course?
   b. pedagogical content knowledge as measured by reported years of teaching experience?
   c. achievement of the cognitive objectives of the teacher education course as measured by the course instructor’s final exam?
   d. academic achievement in teacher education program as measured by overall GPA?

2. To what extent is portfolio assessment a reliable measure of teacher competence as determined by:
   a. the consistency with which raters agree in their ratings of each portfolio?
   b. the relative impact of sampling due to persons, raters, and tasks in evaluating preservice teachers’ competence?
   c. the consistency with which ratings assigned to the overall portfolio using the holistic rating scale correlate with the ratings assigned to individual pieces within the portfolios using the analytical rating scale?
The instruction of preservice teachers in the use of portfolio assessment may have an effect on those teachers' notions of assessment and also may foster their use of this assessment tool in future classes (Adams, 1995; Ford & Olhausen, 1991; Krause, 1996). An important question in the current study, therefore, concerns the pedagogical implications for the use of portfolio assessment by the students in the methods course with their future foreign language classes. The following questions are secondary aspects of the study:

1. As reported on a survey instrument developed by the researcher:
   a. What beliefs about assessment and familiarity with assessment tools do the preservice teachers report prior to the methods course?
   b. What beliefs about assessment and familiarity with the tools do the students report following the experience with portfolio assessment in the methods course?
   c. To what extent is there a difference between the reported beliefs and familiarity with assessment tools on the pre-course survey and the post-course survey?
   d. Do the students indicate an intention to use portfolio assessment in future classes?

2. Through subsequent focus interviews with students enrolled in the class, what is the attitude of preservice teachers toward the use of portfolios in evaluating both teachers and students?

3. What implications can be drawn from this study for educating and assessing future foreign and second language teachers through the implementation of portfolio assessment in teacher education programs?
Theoretical Bases

Traditional assessment tools that measure the achievement of a student in methods courses are the primary means of determining the competence, and future performance of preservice foreign language teachers. Many of these conventional exams and tasks assigned in methods classes require students to rely heavily on logical reasoning skills to determine the solutions to problems and then to employ their linguistic skills to express these solutions clearly to the professor. Gardner (1983) claims that each person’s blend of competencies produces a unique cognitive profile that may not be adequately measured by traditional testing, but that could be more thoroughly evaluated through a combination of assessment tools. He proposes a theory of multiple intelligences that covers seven areas of competence including: linguistic, musical, logical-mathematical, spatial, bodily-kinesthetic, intrapersonal, and interpersonal. Each intelligence is useful in acquiring and demonstrating a variety of knowledge and competencies. A complex theory of intelligences such as that proposed by Gardner requires a multidimensional assessment tool to capture the true profile of the preservice teacher. Portfolio assessment is one such tool because it allows the preservice teachers to rely on their stronger intelligences to demonstrate competence in the areas of expertise necessary for effective teaching while identifying the weaker areas to the teacher educator to be considered for further education and development. Many advocates of portfolio assessment are increasingly promoting portfolios as a way of measuring multiple intelligences and of accommodating diverse learning styles and preferences (Hancock, 1996).

There is a demonstrated shift toward multidimensional measures of teaching competence in teacher education and development. Issues of reliability and validity, however, cannot be ignored as often occurs in descriptive publications of the implementation and benefits of portfolio assessment. The investigation of reliability and validity can be seen as complementary objectives in the development of assessment tools
because both issues seek to measure the true ability of the examinee, not extraneous factors such as testing context, items, or raters. Investigating the reliability and validity of portfolio-based assessment has proven problematic, however, because of the multidimensionality of the criteria in setting guidelines for portfolio creation, the lack of agreement on standards and models of excellence in the teacher preparation field, and the subjectivity inherent in the scoring of portfolios and the subsequent interpretation of the scores.

A principal concern in the development of assessment tools is to identify potential sources of error and to minimize the effects of these factors on the test scores; in other words, to establish the reliability of the scores assigned. According to Bachman (1990), the investigation of reliability is determining the extent to which test performance is affected by measurement error or factors other than those intended for assessment. The majority of current research on the reliability of portfolios focuses on a discussion of inter-rater reliability (CRESST, 1993). Reckase (1997), however, calls for the investigation of both score and scorer (inter-rater) reliability. He agrees that past research has demonstrated scorer reliability, proof that the rating scales can be employed consistently by different, trained raters, but argues that studies have neglected to report evidence that the portfolio scores are reliable. In a recent discussion of the reliability of performance assessment, he argued that the reliability statistics reported in the research "only supported inferences about the accuracy of the scoring process" (p. 30) and did not indicate that the portfolio scores themselves were reliable. Reckase claims that score reliability refers to the consistency of student performance on similar measures of student ability. Inter-rater reliability, therefore, provides evidence that the scoring process is reliable, whereas further evidence is needed to demonstrate that the scores are reliable. The use of generalizability theory may provide a means for investigating the issue of score versus scorer reliability. Whereas correlation analysis of the scores given by each judge can indicate a consistency among the raters, a
generalizability study can provide data regarding the impact of the raters, tasks, and the examinees on the portfolio scores. The results would provide sufficient evidence for the claim that the scores and the scorers are reliable. According to Brennan (1997), generalizability theory may be particularly suited to establishing the reliability of performance assessment instruments because of the need to understand the results from two or more facets simultaneously, especially tasks and raters. Further investigation into issues of scorer and score reliability can also be conducted through means of item response theory. The many-facet Rasch model, a program developed by Linacre (1990), separates error that is attributable to judge severity from the actual scores of the portfolios. The program generates a score as well as scorer reliability coefficient. In response to criticism of research that investigates scorer reliability only, this study provides evidence of scorer reliability through inter-rater reliability coefficients, and evidence of score reliability by means of a generalizability study as well as the use of item response theory.

Another crucial area of concern in developing assessment instruments is the investigation of the validity of the measure. Bachman explains validity as the extent to which an individual’s performance as measured with an assessment tool is due to the abilities intended to be assessed. The lack of reference to validity issues when discussing portfolio assessment in the literature is an indication of the problems that researchers have had in establishing validity. Many researchers claim that the authentic nature of portfolio assessment is sufficient evidence to establish validity. Terwilliger (1997) argues, however, that the authenticity of portfolios focuses on the face validity of the instrument and to do so “is to concentrate only on the surface features of the assessment (and) misses the point that the validity of an assessment device is fundamentally linked to how well the device reflects an underlying construct” (p. 26). Messick (1994) adds that “performance assessments must be evaluated by the same validity criteria as are other assessments” (p. 26). He states that validity indicates the extent to which the performance represents the ability being measured.
(1988b). According to Hatch and Lazaraton (1991), there are three types of validity that are central to research: (a) content validity; (b) criterion validity; and (c) construct validity. Content validity refers to how well the instrument measures the ability based on the representativeness of the sampling of items or tasks. Content validity is not measured statistically, rather information from the literature and from experts is used to determine the validity of the types of tasks or items included in the test. Criterion validity refers to the accuracy of the instrument in predicting an individual’s ability in a current or future situation as based on the results from the performance on the test. It is often measured through use of multiple regression analyses to determine the factor that best predicts performance on the assessment instrument. Construct validity is the most difficult to measure because there is no direct measure of many constructs. Instruments can measure an individual’s propensity to demonstrate certain traits that have been indicated through theory and research as valid characteristics of a certain construct. An example of this would include the use of teacher tests to measure knowledge of content, curriculum and pedagogy in an attempt to assess ability to perform in the classroom. Factor analysis can be used to establish the factorial or construct validity of assessment tools by demonstrating that the items or tasks load on the expected scales, thus indicating that the tasks are measuring the construct that they claim to be measuring.

Based on the definition provided by Hatch and Lazaraton (1991) of content validity as the representativeness of the sampling of tasks in an instrument, the documentation portfolio implemented in the teacher preparation course for the purposes of this study reflected the content knowledge expected of preservice teachers. The tasks suggested for inclusion in the portfolio and provided as a guideline for the preservice teachers were chosen from a larger domain of tasks characterized as necessary abilities of preservice teachers. The professional standards created by the American Council of Teachers of Foreign Languages (ACTFL, 1988), the American Association of Teachers of French
(Goepper, 1989), the American Association of Teachers of German (AATG, 1992), and the American Association of Teachers of Spanish and Portuguese (AATSP, 1990) as summarized in Lipton (1996), provided the framework for the list of suggested activities. These activities were also presented in a manner that is consistent with Gardner's (1983) theory of multiple intelligences by allowing for the use of different intelligences in each type of task (See Appendix C for a description of tasks). The tasks were divided into three categories from which the student needed to choose. The first category consisted of tasks that required library-based research (linguistic and logical intelligences). The second category included tasks that were instruction-based such as the development of lessons based on movement (application of bodily-kinesthetic intelligences), activities centered around songs (musical intelligence), and development of materials for the classroom (integration of spatial intelligence). The final category consisted of tasks that were field-based, including interviews with teachers, site-based observations, and professional development (inter and intrapersonal intelligences).

Recent research in the teacher education field has identified principles for effective teaching that are used to create the preparation programs and assessments of future teachers. Shulman (1987) organizes the knowledge required for effective teaching into the following categories: (a) content knowledge; (b) general pedagogical knowledge; (c) curriculum knowledge; (d) pedagogical content knowledge; (e) knowledge of learners; (f) knowledge of educational contexts; and (g) knowledge of educational purposes and values. He argues that the category of principal concern in teacher preparation is pedagogical content knowledge because "it represents the blending of content and pedagogy"; it is the aspect of teacher preparation that distinguishes content specialists from pedagogues. In other words, pedagogical content knowledge is what defines the teaching profession more than other categories, and what needs to be promoted in teacher education and evaluation. A problem that researchers have encountered, however, is that experienced teachers know a
great deal more than they are able to articulate, thus making pedagogical content knowledge difficult to operationalize. Studies that investigate “wisdom-of-practice” (Shulman, 1987) are being conducted for the purpose of identifying and codifying principles of effective teaching to be able to establish standards of practice in education. Most of these studies are based on experience in the teaching field (years of practice) and performance in the classroom. The standards of practice that emerge from these studies shape the knowledge base for teacher preparation, and are based principally on the understanding of both content and pedagogy, performance in the educational context, and years of teaching experience.

The “wisdom-of-practice” research described by Shulman (1987) provides a model for the current study of the validity of portfolio assessment. The categories used to identify and codify principles of teaching provide guidelines for the principal variables with which the correlation and regression analyses can be conducted to determine the criterion and predictive validity of the portfolio scores in assessing preservice teachers’ competence in the classroom. An understanding of content (conceptual knowledge) can be measured through grade point average (GPA) and final course exam scores, whereas pedagogical content knowledge can be measured through years of previous teaching experience and performance in the classroom. Linear relationships between the portfolio scores and the other variables (performance tasks, years of teaching experience, final exam grade, and GPA) on scatterplots, as well as significant correlation coefficients in the correlation matrix, indicate the existence of a positive relationship between portfolio scores and the other principal variables. These relationships can help to establish the accuracy of portfolio scores in evaluating the competence of teachers.

Regression analysis provides a manner for predicting performance on the dependent variable (portfolio) via one or more independent variables. Linear regression can predict performance from one variable or test score whereas multiple regression can predict
performance from a variety of independent variables. This is done through weighting of the independent variables to determine which of the variables is more effective in predicting performance on the dependent variable (Hatch & Lazaraton, 1991). In this study, the regression analysis shows which independent variable(s) or combination of variables (performance tasks, years of teaching experience, final exam grade, or GPA) can best predict performance on the dependent variable (portfolio). The regression analysis also indicates those variables that do not add information to the prediction of scores. Current evaluation of preservice teachers determines the competency of future teachers by predicting performance in the classroom from academic grades (GPA) and scores on standardized tests. The results from the regression analysis lead to establishing the validity of portfolio scores as an accurate measure of the pedagogical competence of preservice teachers.

Factorial validity, construct validity through factor loadings, can be used to establish the existence of the desired construct(s) in the assessment tool. Factor analysis determines the number of factors that are present in the measures thus better defining the constructs and maximizing the explained variance. By choosing items or tasks that belong maximally to their scale or construct, internal consistency is improved thereby increasing the reliability of the scores on the instrument. In this study, factor analysis can establish the existence of the two constructs in the various measures: (a) conceptual knowledge as measured by the final exam and overall GPA; and (b) pedagogical content knowledge as measured by the portfolios, performance tasks, and years of teaching experience.

Educational research and theory reflect a shift toward the implementation of alternative assessment instruments, but psychometricians demand evidence that these assessment tools are reliable and valid. In response to this demand, the current study investigated reliable and valid procedures for implementing portfolios as an assessment tool in a foreign-language teacher education course. The use of portfolio assessment in one
methods course in a program of teacher education can be seen as an example of the impetus
toward the implementation of portfolio assessment on a larger scale in the program itself. It
contributes to the establishment of portfolio assessment as a viable supplement to traditional
forms of assessment in the evaluation of preservice foreign language teachers.

Basic Assumptions

1. Assessment practices are integral to instruction.
2. Providing a variety of assessment methods enhances the quality of evaluation.
3. Teacher awareness of assessment practices has an impact on classroom instruction.
4. Preservice programs are a primary source in preparing teachers in all facets of education,
   including student assessment.
5. Teacher competencies can be measured and evaluated.
6. Portfolio assessment is currently used in teacher education courses and is therefore an
   important field of study.
7. Students made a reasonable and consistent effort to create a portfolio.
8. The students provided honest answers on the survey instrument.

Definition of Key Terms

The following operational definitions of terms are presented as they are to be understood in
this study:

Alternative Assessment: measures of achievement or competence that include methods other
than traditional tests, such as projects, portfolios, and performance-based tasks.

Conceptual Knowledge: the knowledge of theories and concepts that pertain to a specific
field of study; in this study, the knowledge of theories, methodologies, and program
models relevant to the teaching of foreign languages in the elementary school as measured
by the final exam.
Content Validity: the representativeness (Hatch & Lazaraton, 1991) of the portfolio tasks in measuring teacher preparation; in this study, the extent to which the portfolio activities represent the domain of pedagogical content knowledge in which preservice foreign language teachers need to be competent.

Documentation Portfolio: a compilation of student work (Goodman, Goodman, & Hood, 1989) that focuses on teacher competence and professional development through the completion of activities designed to engender the acquisition of pedagogical knowledge. Selected activities were chosen as being representative of the standards established by ACTFL (1988) for the foreign language teaching profession. For the purposes of this study, the portfolio must include a minimum of six selections, two from each of the following three categories: library-based research activities, instruction-based activities, and field-based research activities. It must also contain an introductory letter describing both the preservice teacher and the portfolio, and anecdotal reports regarding the importance of each piece to their demonstration of teaching competence and professional development.

Factorial Validity: a measure of construct validity as determined through factor loadings.

Generalizability Study: a statistical analysis of portfolio scores based on a theory that emphasizes the estimation of random effects variance components (Brennan, 1997) which enables the researcher to isolate several sources of measurement variability (e.g., persons, tasks or raters) and to estimate their magnitude through analysis of variance procedures (Shavelson & Webb, 1991).

Inter-rater Reliability: the degree to which judges agree in their holistic and analytical scoring of the portfolios (Richards, Platt, & Platt, 1992); also referred to as “scorer reliability” (Reckase, 1997).

Many-facet Rasch Model Analysis: a statistical model that allows for the objective estimate of the ability of each examinee to be determined free from the level of severity of the
judges, from the difficulties of the items, and from the arbitrary manner in which the categories of the rating scale have been defined (Linacre, 1993).

**Pedagogical Content Knowledge**: the ability to organize and present subject matter in a learnable format (Shulman, 1986); in this study, pedagogical content knowledge is measured by the self-reported years of teaching experience accumulated before entering the certification program (Naizer, 1993) and by the completion of performance-based tasks for the course.

**Performance-based Tasks**: tasks in which the examinee applies mastered skills and knowledge; in this study, these tasks include daily and unit lesson plan writing, presentation of lesson plans, peer-teaching activities, creation of assessment activities, and a personal journal of written reactions to course readings. These tasks reflect the domains of pedagogical knowledge as summarized by Ryan and Kuhs (1992) in which a teacher should be competent before certification.

**Portfolio Assessment**: the use of a portfolio in the evaluation of teaching competence and future performance as an elementary school foreign language teacher through a compilation of work that includes samples from a variety of domains, and that allows for student involvement in learning and assessment processes. Although portfolios are often referred to as measures of product and process, in this study the portfolio is intended to measure the products of learning and not the processes of becoming a teacher.

**Preservice Teacher**: a student enrolled in a teacher education and certification program that is designed to develop a person’s pedagogical content knowledge before he/she begins teaching in primary and/or secondary schools.

**Score Reliability**: the consistency with which students can perform on a similar assessment designed to meet the same educational objectives (Reckase, 1997).

**Validity**: an evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test
scores (Messick, 1988b); in this study, a measure of the degree to which portfolio scores correlate with other established criterion measures of teaching competence and/or achievement in the assessment of preservice foreign language teachers’ performance.

Limitations of the study

1. The study concentrates on the use of portfolio assessment with preservice foreign language teachers and may not be generalizable to all areas of instruction that use portfolio assessment.

2. The study investigates only one instance of portfolio use, namely that experienced by participants in a single course within the sequence of required courses in the preservice foreign language teacher certification program. Students may have had experience with portfolio assessment in other courses.

3. The study is limited to the experience of the preservice foreign language methods course and does not involve a follow-up study once the students are given their first jobs as foreign language teachers.

4. The size and the characteristics of the sample used in this study may make it difficult to generalize to a larger population of preservice teachers.

5. The study focuses on one type of portfolio, specifically the compilation of samples to demonstrate competency in teaching languages at the elementary school level, and may not be generalizable to other types of portfolio use, such as those prepared to demonstrate competence in teaching of other content (e.g., science, mathematics) or at different levels (e.g., secondary, university).

6. The portfolio used in the study is a product-based portfolio and is not intended to measure the process-oriented aspects of portfolio assessment.

7. Differences among raters on the judgments of the portfolios may be due to learning style, cultural, and/or personality conflicts between the raters and the students. Issues of
learning style, cultural background, and personality type were beyond the scope of this study and are not included as possible causes of variance among ratings of the portfolios.  

8. The lack of validated performance-based measures against which to compare teacher portfolio scores required the use of quantitative, traditional measures of teaching competency to be used in an attempt to validate the use of portfolio assessment for evaluating preservice foreign language teachers.

Conclusion

Because of the nontraditional nature of portfolio assessment, validity and reliability issues are difficult to address, and are therefore often glossed over or ignored (see CRESST, 1993). These issues, however, need to be examined if portfolio assessment is to become a viable supplement or alternative to traditional methods of assessment. Calfee and Perfumo (1993) assert that portfolio assessment is likely to fail if valid data and reliable information are not provided through this type of evaluation. The present study, therefore, adds to the knowledge base on the use of portfolio assessment in teacher preparation and evaluation by determining the reliability of the scores and rating scales in judging teacher competency, as well as the validity of these decisions as based on portfolio scores. This study also responds to the demand for additional research and development in pedagogical assessment (Martinez & Lipson, 1989). The background, aims, and significance of the research to document the need to investigate reliability and validity issues, as well as the research questions were presented in this section. In the next section, related literature and studies on the issues of traditional and alternative assessment measures, the use of portfolio assessment in preservice teacher education, and the importance of establishing reliability and validity in alternative assessment are reviewed as background information for this study.
CHAPTER 2

REVIEW OF THE LITERATURE

Introduction

The importance of assessment is documented by the extensive writing that has been
done on the topic. It is one of the areas targeted for reform by a number of educational
organizations as well as leaders in the field of education (Naizer, 1993). An initial review
of the literature demonstrates a widespread shift toward alternative assessment measures in
the educational field, but also indicates the lack of research on the reliability and validity of
these measures. An overview of studies and reports presents a brief history of traditional
assessment tools, the need to develop multidimensional tools for the evaluation of teacher
competencies and future performance, and the emergence of an array of alternative
assessment measures, including portfolios. Gardner’s (1983) theory of multiple
intelligences serves as the springboard for the discussion of the need for multidimensional
assessment tools that can measure an individual’s strengths. An introduction to portfolio
assessment and its use in the field of preservice teacher education is then provided. Finally
this literature review addresses the importance of reliability and validity issues in
assessment, and the issues of establishing reliable and valid procedures in using portfolios
to evaluate preservice teachers.
Historical Perspective of Assessment

Ward and Murray-Ward (1993) provide a historical account of the roots of current testing practices in which they claim that a body of knowledge and theory has developed in the history of educational measurement that spans more than 5,000 years. During this time, testing methods have undergone numerous changes and continue to be evaluated and adjusted to better guide current practice. The origins of student assessment began as early as three thousand years ago with the civil service testing program in China. There are also records of ancient performance testing in both the Bible and early American history, and of the university exams given during the Middle Ages in Europe (Ward & Murray-Ward, 1993).

Prior to this century, academic achievement was measured in the form of subjective judgments about student responses on oral exams or essays. Galton’s research on individual differences and the normal curve, the emergence of experimental psychology, and the creation of the Stanford-Binet test of intelligence all led to the development of norm-referenced achievement and aptitude measures during the twentieth century. There are four principal areas in the development of testing methods in this century (Ward & Murray-Ward, 1993): (a) the period from 1900-1915 in which early tests of intelligence and educational achievement began to appear; (b) a period of rapid development from 1915-1930 during which the growth of behavioral psychology impacted the structure of test development; (c) a rapid expansion in testing and the beginning of criticisms of traditional tests occurred in the period from 1930-1960; this period was also marked by the creation of specific aptitude tests, a return to more global assessment, and the development of psychometric theory from which the concept of testing standards was derived; (d) the final period, 1960-present, has seen a demand for more global and varied assessment, and the
development of both norm-referenced and criterion-referenced testing as well as the emergence of alternative assessment. It is this last period of educational measurement that needs to be described in further detail for the current study.

Norm-referenced tests were developed based on trait theory which claims that individuals differ on identified characteristics that are normally distributed in the population. These tests are appropriate for comparing student achievement, but are problematic in that they often do not fit the curriculum that is being assessed and limited forms of learning are measured (Gredler, 1996). Criticism of norm-referenced assessment measures and the development of new curriculums that emphasized the mastery of skills led to the development of criterion-referenced testing in the 1960s. Criterion-referenced tests were designed to provide information about the mastery of certain skills. These tests, which include the Educational Testing Service’s National Teacher Examinations for determining teaching competency, are appropriate for describing skills of the students and assigning individuals to varying levels of mastery, but criterion-referenced testing often led to discrete point assessment of skills with no measure of the ability to synthesize material (Ward & Murray-Ward, 1993). Haertel (1991) summarizes the concerns expressed by the profession in his claim that the teacher tests currently in use have been criticized for the “absence of criterion-related validity evidence supporting their use...for treating pedagogy as generic...and for failing to measure many critical teaching skills” (pp. 3-4). Additional criticism includes debates about the reliability of test scores across cultures, among testers and test situations, and lack sensitivity as indicators of competence (Cole, Messner, Swonigan, & Tillman, 1992; Grady, 1992; Touzel, 1993). A report from the National Center for Fair and Open Testing (1992a) argues that standardized tests lack cultural sensitivity and fail to measure critical thinking skills such as those needed in problem solving.
The criticisms of norm-referenced and criterion-referenced assessment measures led to the development of alternative assessments in the 1980's. Cognitive psychologists argued that content and processes needed to be grouped together in order to provide for meaningful learning. Contemporary learning theories contend that meaningful learning is reflective, constructive, and self-regulated (Hancock, 1996). Individuals acquire new information using diverse learning styles and incorporate this information into their already existing schemata. Traditional tests fragmented learning and compared achievement to standardized criteria or other students, and thus did not accurately assess what students know (Ward & Murray-Ward, 1993). Formal assessments report minimal information about the contextual factors that may impact performance as many tests can measure only instances of competence (Cole et al., 1992). In line with these contemporary learning theories is Gardner's theory of multiple intelligences which serves as the theoretical basis for this study. Intelligence is no longer seen as a linear and univariate concept; rather it has become a complex and multivariate concept (Ducette, Sewell, & Shapiro, 1996) that requires a multidimensional assessment tool, such as portfolios, to accurately measure all competencies in context.

Currently, educational researchers and practitioners argue against the sole use of traditional forms of assessment in teacher education (Cole et al., 1992; Shulman, 1988; Valencia, 1990). Shulman (1988) suggests using a variety of assessment techniques including portfolios and tests that reflect the complexity of teaching and the contexts in which teaching occurs. The criticisms of traditional assessment measures are summarized in Zollman and Jones (1994), who argue that objectives have become defined in terms of competitive success, norms, or standards, rather than authentic learning, and that traditional assessments are often measures of accumulated knowledge, test-taking skills, and socioeconomic status rather than accurate measures of learning. Tierney et al. (1991) criticize the use of standardized testing and provide a list of the negative outcomes of the
use of standardized test in assessing student achievement. Their claims include the following: (a) tests assess students across a limited range of abilities; (b) tests are often mechanically scored with little input from the teacher; (c) tests address achievement only, not progress and growth; (d) tests separate learning from assessment; (e) tests do not allow for collaborative assessment of students in which the student plays a role in her own learning and assessment.

Alternative forms of assessment, including the use of portfolios and performance-based measures, have developed in response to the increasing criticisms of traditional forms of assessment, especially the standardized test (e.g. Baron, 1991; Shavelson, Carey, & Webb, 1990; Shepard, 1989; and Tierney et al., 1991). Costa (1989) argues: “If we are to change education to meet the needs of the information age, we must overcome our habit of using product-oriented assessment techniques to measure process-oriented education. We need to redesign assessment to fit the goal of the restructured school” (p. 2). In response to claims like this, traditional testing methods are becoming less popular and more widely criticized. Although educational and psychological measurement of the effects of education is a relatively new concept (Gredler, 1996), the use of alternative assessment has become a focal point in debates about American education and one of the main topics in current educational reform movements.

Alternative Assessment

Alternative forms of assessment have emerged in the past two decades in response to a need for holistic evaluation of student achievement (Moeller, 1994). Early exploration of the use of alternative assessment measures began with the establishment of the National Assessment of Educational Progress (NAEP) in 1969 to monitor achievement in a variety of subject areas (Fisher, 1993). Dissatisfaction with traditional forms of assessment such as standardized achievement tests and teacher-designed multiple-choice tests reflects the
need for alternative assessment measures (Zollman & Jones, 1994). The expected results of alternative forms of assessment are summarized in Arter and Spandel (1992). They claim that educational measurement needs to assess more than knowledge of facts to include learning strategies such as the ability to think critically and reflect on learning. Assessment and instruction must be integrated, and active student involvement in assessment should be encouraged. The National Board for Professional Teaching Standards has called for a system that examines a multidimensional knowledge base and is flexible, comprehensive, and integrated in its approach to assessment (Baratz-Snowden, 1993; Ryan & Kuhs, 1993). Shepard (1989) believes that future teachers need practice in solving real problems and complex tasks not provided in traditional testing. Haertel (1991) addresses the use of alternative assessment in teacher evaluation and suggests that the assessment of teaching needs to reflect the complexity of the task and recognize the impossibility of specifying a single correct answer or acceptable procedure. He contends that there is little or no evidence provided by traditional assessment to predict classroom performance. He claims that “the teacher tests now in common use have been strenuously and justifiably criticized for their content, their format, and their impacts, as well as the virtual absence of criterion-related validity evidence supporting their use” (Haertel, 1991:3-4). Studies of the relationship between the National Teacher’s Examination and preservice teachers’ performance are mixed because this test does not have proven validity in predicting teaching performance (Howey & Zimpher, 1996). Evertson, Hawley, & Zlotnik (1984) also concluded that various standardized tests used for evaluating competency in teacher education programs were not consistently related to measures of effective teacher performance.

Alternative assessment has as its main goal to gather evidence about how preservice teachers are “approaching, processing, and completing real-life tasks” (Huerta-Macias, 1995:9). This is of particular importance in the field of teacher education where context and learners can impact the performance of teaching tasks. Gredler (1996) also presents three
critical goals of alternative forms of assessment: (a) they should measure complex intellectual capabilities; (b) they should focus on important learning processes; and (c) they should inform teaching practice about students’ strengths and weaknesses in relation to cognitive skills.

A popular form of alternative assessment is the use of portfolios. Portfolios have their origins in professions where people showcase their abilities: commercial art, photography, modeling. In the 1970s, research in first-language (L1) literacy saw a need for authentic assessment to measure the development of a student’s reading abilities that could not be measured through discreet skill testing (Kingore, 1993). This need, combined with the whole language movement of the late seventies, fostered the creation of authentic and holistic approaches to teaching and assessment in the elementary and secondary grades. At this time, portfolio assessment is most frequently implemented in composition courses to measure student development in the writing process (Reckase, 1995). Its use in evaluating preservice teachers, however, is becoming widespread as practitioners realize that traditional measures of assessment may ignore an individual’s prior pedagogical knowledge and test critical teaching skills in isolation (Zollman & Jones, 1994).

Overview of Assessment Principles

Assessment is referred to in the literature as “the objective and accurate measure of performance in relationship to the construct of the interest area” (Andrews & Barnes, 1990:572). There are, however, limitations to objective measurement. Performance on tests can be affected by one or all of the following factors: the tasks, the testing context or time allowed, and learning style or personality of the test-taker. In order to minimize the effects of these limitations and maximize reliability and validity of test scores and their interpretation, Bachman (1990) suggests three fundamental steps in the development of an assessment tool: (a) provide clear theoretical definitions of the abilities to be measured; (b)
specify the conditions to be followed in observing performance; and (c) quantify our observations so that measurement scales have the necessary properties.

Reliability is defined by Bachman (1990) as the quality of the tests scores. A reliable score would be one that is relatively free from measurement error. The extent to which scores remain consistent across instances of test use and across individuals taking the test is also a measure of reliability. Clearly identifying the factors that are potential sources of error measurement are principal in the investigation of reliability. There are several methods for establishing the reliability of test scores: classical true score (CTS) measurement theory, generalizability theory, and item response theory.

The CTS model of estimating reliability looks at the observed score as having two components: the true score and the error score. The true score is the individual’s ability and the error score is due to factors other than ability. One of the main limitations with the CTS model, however, is that it considers all error to be homogeneous and random. It is impossible to know what factors are causing the variance in the scores to be able to control for those factors in a subsequent administration of the test. CTS is not very useful when establishing the reliability of measures such as portfolios because the variance due to raters, for example, cannot be separated from the variance due to tasks. This could be problematic because rater variance could be easily controlled by rating instrument calibration and rater training.

Generalizability theory separates the magnitude of sampling variability due to raters, tasks, methods, occasions, etc. and provides estimates of the measurement error in the form of variance components (Shavelson & Webb, 1991). It then presents a summary coefficient reflecting the reliability of generalizing the sample scores to a larger domain of hypothetical measures. Variability due to raters, for example, deals with the traditional concern in portfolio and performance assessments of inter-rater reliability. Sampling variability due to tasks refers to internal consistency of the measure, whereas variability due
to occasions corresponds to the classical notion of retest reliability of an assessment tool. The fact that the variance in a score is divided into components allows for easier management of the factors that cause the error. If a large portion of the variance is due to tasks, for example, it would be feasible to reconstruct the test to reflect more consistency among the items. According to Brennan (1997), the use of generalizability theory in judging the reliability of performance assessments may be especially helpful because of the need to look at several facets simultaneously, namely raters and tasks.

Finally, item response theory includes another dimension in its estimation of the reliability of assessment scores: the individual taking the test. Psychometricians have developed a number of models that relate an individual’s performance on an assessment measure to both the individual’s level of ability and the difficulty of the task or test. The primary advantage of item response theory over the CTS model and generalizability theory is its precision of measurement. Item response theory does not treat error variance as homogeneous, the estimation of reliability is independent from the sample scores, and the individual’s estimate of ability is independent of the set of items taken or tasks performed (Bachman, 1990). The many-facet Rasch model developed by Linacre (1990) provides a statistical means for estimating an individual’s ability level independently from the sample scores and from the raters who assigned the scores. The model separates error due to the severity or leniency of a particular judge by assigning a severity score to each judge. This score accounts for the measurement error that may be due to subjective use of the rating scales by a particular judge. The same is done for error that may be due to given items on the test, in this case, the samples in the portfolio. This is particularly important in establishing the internal consistency of portfolio samples as portfolios tend to be individual compilations of knowledge and ability. The lack of standardization of the items included in a portfolio is at the same time the appeal and the complexity in using portfolio assessment to evaluate teaching competency.
Issues of reliability are crucial in the evaluation of knowledge and achievement. Important decisions are often made on the basis of test and performance scores, and the scores must be reliable across contexts and individuals. Alternative assessment tools are considered problematic because of the difficulty in establishing the reliability of scores that are derived from individual and subjective materials and performance. Reliability, however, is only a part of the technical accuracy of assessment measures. Validity is also an important issue in the development of alternative assessment tools.

Validity is defined by Messick (1988b) as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). Whereas the investigation of reliability is concerned with the extent to which an individual's test performance is due to factors other than the ability that is being measured, validity looks at the extent to which the performance reflects the ability being measured. The technical accuracy (i.e. reliability and validity) of an assessment tool can be seen as having two concerns: minimizing the effects of measurement error and maximizing the effects of the ability being measured.

There are several types of validity that are looked at in test development. Content validity is based on the degree to which a test accurately measures the skills or knowledge it intended to measure (Hatch & Lazaraton, 1991; Richards et al., 1992). In the use of teacher portfolios, representative tasks include any tasks that a teacher would be expected to perform in the instructional context. Examples may include lesson plans, unit creation, materials development, mini-research projects, etc. This can be established through comparison of the universe domain, in this case nationally accepted foreign language pedagogical competencies (e.g., ACTFL, 1988), and the sample domain, in this study, the portfolio submissions, to determine the extent to which the selected tasks represent the universe domain.
Criterion validity investigates the accuracy with which a tool measures the competency that it intends to measure as compared to other established tests that claim to measure the same ability (Richards et al., 1992). It can be measured through a correlational analysis to demonstrate the relationship between two or more tests. Comparing the results from a performance-based test and scores from a portfolio would be an example of a method of investigating the validity of the portfolio scores. The problem that researchers have encountered when establishing the validity of alternative assessment tools is the lack of validated instruments against which they can compare the portfolios. Although not a conventional manner of investigation, criterion validity, therefore, has been studied in the converse sense: comparing two instruments that are thought to measure different abilities to observe the differences among them (Naizer, 1993; Nweke, 1991). An example of this is using a teaching portfolio to measure pedagogical content knowledge and then comparing the portfolio scores to the scores received on the National Teacher’s Examination (NTE) from the Educational Testing Service. These two tools are very different in that a portfolio is said to be performance-based and multidimensional measure of teaching competence whereas the NTE is a logical and linguistic-based traditional exam that measures teaching ability through multiple choice questions about pedagogical theories. Comparing these two measures will likely result in little or no relationship, which may indicate that they are each measuring distinct abilities. The danger in establishing criterion validity through comparison with other measures is that one must assume that the measure against which the portfolios are being compared, for example, the NTE, is a valid measure of the construct in itself. This is not always possible, though; for example, Nweke (1991) compared teaching portfolio scores against results from the American College Test, a measure of high-school achievement. The correlation coefficient was low (r=.11) which he claimed to be an indication that the two instruments were measuring different abilities.
Establishing the validity of performance-based assessment instruments, namely portfolios, poses a problem for researchers. The lack of valid instruments against which to measure portfolio scores has forced researchers such as Nweke (1991) to resort to comparing alternative assessment tools against traditional instruments to determine if there is a difference rather than a relationship between the instruments. The CRESST (1993) report indicates a neglect in validity studies which is most likely due to the challenge of finding comparable instruments against which portfolios may be measured. Several researchers have attempted to work around this lack of validated performance-based instruments by comparing portfolios to such constructs as pedagogical content knowledge as measured by the number of years of teaching experience (Naizer, 1993) and college achievement as measured by grade point average (Nweke, 1991).

A form of validity that is similar to criterion validity, and often classified under the same title, is predictive validity. This type of validity indicates the accuracy with which a test can be used to predict performance in future contexts (Hatch & Lazaraton, 1991). The Graduate Record Examination (GRE), for example, is an indication of how a student will perform in graduate school based on linguistic, logical, and mathematical abilities as measured through a multiple-choice test. By the same token, portfolios may predict performance in the classroom based on the ability to create appropriate materials, devise lesson plans, compile resources for language teaching, etc. that students demonstrate in their portfolio submissions. Predictive validity is commonly measured through the use of multiple regression procedures. As mentioned above, the assumption that needs to be made in investigating criterion or predictive validity is that the instruments against which one is comparing portfolios are valid measures of the construct. This leads to the discussion of construct validity, the degree to which the assessment measures the construct or ability it is designed to measure. Construct validity is difficult to measure, however, because there are no direct measures of many of the constructs used to define abilities in education. Often,
traits that are representative of the construct can be measured as evidence of the construct. Pedagogical content knowledge, for example, is a construct that cannot be measured directly. It is defined in theory as the ability to transform knowledge into a teachable format (Shulman, 1986), and may be measured indirectly through the number of years of teaching experience (Naizer, 1993) or the ability to apply knowledge in the completion of tasks required in the daily context of teaching (Ryan & Kuhs, 1992). Factor analysis, however, is a useful tool for determining the existence of desired constructs in the instrument. Performance on a variety of measurements that are shown to include the same underlying construct will load on the same factor thus confirming the existence of the construct in the tools.

It is critical in the implementation of assessment tools to establish their reliability and validity in accurately measuring the abilities that we claim they can measure. The decisions that are made based on the scores received on assessment tools require that the ability being measured is maximized and that the variance due to confounding factors such as raters and tasks that are impacting the ability is minimized. It is important that preservice teachers are judged as competent to teach based on their ability to perform in an educational context without interference that may be caused by confounding variables such as linguistic ability, rater subjectivity, and test construction.

The Theory of Multiple Intelligences

This construct was chosen for inclusion in the literature review because it represents a notable shift in educational thinking and assessment in recent years: a change from a unitary concept of ability to a multivariate concept. This shift in the concept of intelligences as multiple rather than singular has been reflected in assessment through the shift toward measures that observe performance and multiple instances of ability. Alternative assessment instruments such as performance-based tasks and portfolios represent the many dimensions
of a person’s ability, intelligence, and learning style. The integration of these multidimensional assessment tools into educational measurement represent a clear movement away from a singular conceptualization of ability toward a model with multiple intelligences.

The original test of intelligence was first created by Alfred Binet to determine a child’s intellectual competence as a basis for effective teaching. The concept of intelligence was expanded by the contributions of a general intelligence theory of Spearman (1927) and the primary mental abilities theory of Thurstone (1938). The contrasting views of Thurstone and Spearman led to the debate between a unitary concept of intelligence and the concept of several intelligences, such as that proposed by Gardner (1983). The Multiple Intelligence (MI) theory is based on the belief in the existence of different intellectual competencies. Gardner and his associates (Gardner, 1983; Gardner and Hatch, 1989) argue that these intellectual competencies can be identified through activities in a given context. Gardner proposed the seven areas of intelligence through investigation of criteria in several studies including those of normal development, of exceptional children and brain-damaged patients, and empirical findings in neuropsychology and psychometric studies.

The seven intellectual competencies are outlined in Frames of Mind (Gardner, 1983). They are as follows:

Linguistic intelligence: Verbal ability such as the language skills needed to deal with verbal symbols and functions of language.

Logical-mathematical intelligence: Ability to manipulate mathematical and scientific symbols, and to use logical and numerical skills in reasoning.

Spatial intelligence: Visual-spatial ability to manipulate or mentally rotate an object to transform it from its original structure.

Interpersonal intelligence: Ability to understand the behavior of humans in complex
social environments.

**Bodily-kinesthetic intelligence:** Ability with which one is able to control bodily movements.

**Intrapersonal intelligence:** Ability to use one’s self-knowledge to guide one’s behavior.

**Musical intelligence:** Musical expressiveness.

Gardner argues that these seven intelligences are fundamentally independent of each other and that a unitary measure of intelligence is questionable. A teacher may use one or all of these intelligences in the instructional context. It is, therefore, necessary for teacher educators to determine an individual’s teaching competency based on the whole person not only on one or two areas of intelligence. The inter- and intra-personal intelligences, for example, are critical to the interactional nature of instruction and to the context of teaching. The ability to demonstrate competence in this area, however, cannot necessarily be achieved through conventional written exams. Although simulations of teaching experiences may be included on written exams in an attempt to measure a person’s ability to react to given situations, a more complex tool is necessary to measure these types of intelligence. Portfolios allow for inclusion of tasks such as video-taped teaching sessions in which an individual’s interaction with students can be evaluated.

A central point in Gardner’s argument is the need to change from conventional tests, in which the linguistic and mathematical intelligences are the core components of a single measure of competence, to multidimensional measures of a number of competencies. Gardner and his colleagues have shifted the paradigm in assessment and are calling for the use of authentic practices, such as portfolios, to measure competence (Ducette, Sewell, and Shapiro, 1996). The MI theory has challenged the conventional testing paradigm by
suggesting a contextual approach to assessment that focuses on competencies applicable to practical experiences.

The use of portfolio assessment allows preservice teachers to draw on all of the intellectual competencies that they possess in the demonstration of teaching competence. Tasks such as a video-taped teaching session can demonstrate the ability to apply pedagogical knowledge in context as well as the manner in which the teacher interacts with students. This format moves away from the need to rely on linguistic ability alone to describe how one would apply this knowledge such as in the example of simulated teaching experiences in written exams. Conventional tests often ask preservice teachers to answer questions about hypothetical teaching situations. These types of questions emphasize the ability to use logical and linguistic skills in the response and cannot accurately assess all of the intellectual competencies proposed by Gardner. By allowing preservice teachers the opportunity to demonstrate their pedagogical knowledge and abilities through the collection of various tasks and projects, the teacher portfolio provides a more complete picture of the preservice teacher’s abilities.

In the teacher education field, the need for complex and multidimensional assessment tools is becoming increasingly important as the framework outlining the professional competencies of future teachers expands to include a multitude of abilities. Portfolio assessment can help teacher educators evaluate preservice teachers’ ability to use the concepts and knowledge they have learned in creating products or performing tasks. It is no longer effective to evaluate preservice teachers on their logical and linguistic ability alone; it is crucial that assessment measures change as the concepts of intelligence and pedagogical competence are expanded.
Problems with Current Preservice Teacher Evaluation

Attention has recently been focused on identifying and improving methods for evaluating teachers and on improving teacher education programs (Boileau, 1993; Nagel & Engel, 1992; Shulman, 1988; Tyler, 1985). The teaching profession has been reconceptualizing models of effective teaching to reflect information gathered from qualitative studies done in the field. Shulman (1987) provides a model of the knowledge base for teachers that identifies essential knowledge in teacher preparation. Teachers are no longer being trained in the craft or art of teaching, but are being educated in the profession of teaching. Many current methods used to evaluate preservice teachers, however, do not reflect the recent changes that have been made in teacher education. Developing a system for teacher evaluation may be difficult because of the lack of agreement on what constitutes effective instructional performance (Grady, 1992). Shulman (1987) and his colleagues have gathered information about what constitutes effective teaching and have organized it into categories of knowledge, but there is still a lack of consensus in the field of teacher education concerning the knowledge base and how to evaluate future teachers. Current certification exams focus on general pedagogical knowledge, basic skills, and knowledge of content. Some state and national certification exams add a performance component that is based on hypothetical teaching contexts (e.g., the PRAXIS examinations developed by the Educational Testing Service). Teaching portfolios are another assessment tool that is used to evaluate teachers. Portfolio assessment provides a multidimensional instrument for measuring teacher ability, but there are limitations to using it as a sole indicator of ability because of the lack of models regarding content and scoring methods (Wolf, 1991b). Methods that rely solely on written examinations and grade point averages, however, may fail to completely evaluate a preservice teacher’s competency. A combination of both conventional testing methods and newer performance-based instruments appears to be a more holistic method of evaluating teachers.
According to Schrier and Hammadou (1994), traditional assessments are seldom valid predictors of future teaching performance because they do not measure adequate and representative samples of potential for performance as teachers. The assessments employed to evaluate preservice teachers include written examinations to test basic skills and content knowledge, direct observation of performance to evaluate pedagogical skills, and standardized tests (Elbow & Balanoff, 1986; Paulson & Paulson, 1990). Scores on these tests are assumed to measure knowledge of preservice teachers, but they do not indicate if skills needed to perform in the classroom are present. Rafferty (1994) claims that standardized tests cannot measure the range of abilities that are critical to the assessment of preservice teachers. Cole et al. (1992) argue that the complexities associated with teaching are too diverse to be reflected by a single test score.

Conventional written assessments may provide multiple views of instructional preparation, but most do not reflect authentic assessment of classroom performance. Clarification of the purposes of evaluation can improve evaluation of preservice teachers. Employing tools that do not effectively measure preservice teacher’s performance compounds the problem of how to predict instructional competency (National Center for Fair and Open Testing, 1992b). Shulman (1988) suggests using a combination of methods to evaluate preservice teachers, including portfolios. Reviewing the contents of a teacher portfolio can document a preservice teacher’s competency and knowledge. Through a variety of submission types, a preservice teacher can demonstrate content knowledge, pedagogical knowledge, and instructional competence. Evaluating a collection of work from multiple sources (e.g., research tasks, instructional activities, observation and reflection in the field) provides authentic assessment of a preservice teacher’s capabilities. At the same time, the results from more conventional measures of achievement can provide insight into preservice teachers’ understanding of educational concepts and theories.
Portfolio Assessment in Teacher Education

A new trend has emerged in teacher education that demands a move away from traditional testing and an emphasis on authentic assessment that requires multiple sources of data for evaluation (Anderson & Burns, 1989; Grady, 1992; Parker, 1993; Wolf, 1991a). This trend involves the use of portfolio assessment in evaluating teacher competency and in monitoring the development of pedagogical skills during teacher training courses. Portfolios emerged in education as a latent response to interests generated by the Competency-Based Teacher Education reform movement that began in the 1970s which emphasized evaluating teachers to improve instruction (Roth & Pipho, 1990). Currently, portfolios are used to document instructional competence for inservice and preservice teachers (Cole et al., 1992; Edgerton, Hutchings, & Quinlan, 1991; Elbow & Balanoff, 1986; Weinberger & Didham, 1987) and to document growth and development of teacher competency over time (Biddle, 1992; DeFabio, 1993). Gardner (1983) contends that to assess abilities, the actions and judgments of the learners must be observed. Portfolio assessment allows for insight into the development of the thought processes that generate teaching strategies (Schrier & Hammadou, 1994). Portfolios also provide a way for students to become self-reflective and involved in their own learning (Zollman & Jones, 1994).

Portfolio assessment is a student-centered approach to assessment that integrates current theories of testing and instruction (Tierney, et al., 1991). The portfolio provides an assessment of preservice teacher competencies through the observation of a variety of work samples that demonstrate a full range of a person's intelligence. A basic tenet of portfolio assessment is the ownership and responsibility that is placed on each individual. Unlike traditional forms of assessment, students in courses that use portfolios as a means of assessment have control over their learning and assessment because they choose what is to be included in the portfolio. According to Paulson, Paulson, and Meyer (1991), portfolio
development is the “intersection of instruction and assessment” (p. 61), thus allowing individuals to observe the purpose of assessment as integrated with learning rather than isolated from instruction. Student participation in the guidelines for content selection and the criteria for evaluating portfolios are principal requirements in the successful implementation of portfolio assessment. The authors maintain that portfolios can be a powerful tool for preservice teachers to take charge of their learning because it requires them to collect and reflect on examples of their own work, rather than fulfilling requirements extended by the instructor.

Self-assessment is another basic tenet of this approach as the portfolio provides a context for reflection on content selection, teacher strengths and weaknesses, and quality of work. Rybozynski (1991) reports that reflective statements in the portfolio become part of the process for developing reflective skills in the preservice teacher, which in turn fosters the development of teacher competency. Zubizaretta (1994) claims that portfolio development serves as a catalyst for improvement in one’s teaching philosophy, methods, and goals. The portfolio also allows preservice teachers to see the link between theory and practice and reflect on connections between course-work and field experiences (Hartle & DeHart, 1993). By including observations of other teachers, sample lesson plans, created materials, and similar documentation in the portfolio, preservice teachers can find the connection between the ideas provided in textbooks and the teaching skills that are necessary for effective transmission of knowledge. Paulson, Paulson, and Meyer (1991) claim that portfolios offer the opportunity to observe preservice teachers as they learn to make judgments about their own competence and performance. “Teachers must be continually involved in professional development, reflection on teaching, and self-improvement” (Jeffries, 1996:8). Tedick and Walker (1994) find that teacher education programs suffer from fragmentation and focus primarily on methodology not application and reflection. Formal assessments used in these programs fail to account for the impact of
contextual factors such as subject matter or age and level of learners (Shulman, 1988). Zollman and Jones (1994) argue that an important theme in current assessment research is that “preservice teachers must be involved in evaluating their growth as learners” (p. 9). One of the underlying principles in portfolio assessment is the ability to document the progress of learners toward a goal; another is the provision for self-assessment in portfolio evaluation.

Portfolio assessment can also lead to diversity and authenticity in the assessment of preservice teachers because the portfolio is a representation of each individual’s development. Students in methods courses are allowed to choose the samples and presentation format of their portfolio, thus providing for various learning styles and personalities. When used with preservice teachers over the course of their studies, the portfolio becomes a collection of work that best represents each individual’s style of teaching. Cognitive theorists argue that achievement should be studied as an individual’s behavior in and of itself (Drury, 1994; Gifford & O’Connor, 1992). Portfolio assessment allows for the observation of higher-order thinking skills and the development of each individual’s schemata as the student builds on previous knowledge in the incorporation of new concepts and ideas.

Stahle and Mitchell (1993) provide a final beneficial aspect to the use of portfolio assessment in teacher education courses. Through personal experience in portfolio assessment, the preservice teacher is educated in the implementation of this form of assessment, and therefore is able to use portfolio assessment with future foreign language students. Barton and Collins (1992) asserts that an increased awareness of issues in student assessment by future teachers will positively affect assessment practices in public schools. Stroble (1992) supports that argument through her claim that if preservice teachers are required to assemble a portfolio for means of their own assessment, it will allow them to understand the process, and thus increase the possibilities that they will use this assessment
tool in the future. Krause (1996) contends that if teacher educators expect preservice teachers to be purposeful about how they evaluate students, then teacher educators must also demonstrate this intentionality in their assessment of future teachers. Finally, Adams (1995) suggests that preservice “education programs provide the opportunities for teacher educators to participate in assessment reform efforts by exemplifying models of assessment” (p. 568).

Although there is no universal definition of teacher portfolios (Bird, 1988), Arter and Spandel (1992) provide a description of potential portfolios: “a purposeful collection of student work that tells the story of the student’s efforts, progress, or achievement in a given area. This collection must include student participation in selection of portfolio content; the guidelines for selection; the criteria for judging merit; and evidence of student self-reflection” (p. 36). Other educational researchers promoting the use of portfolio assessment agree that the portfolio contents are determined primarily by the purpose of the portfolio. Nweke (1991) defines a teacher portfolio as a collection of documents over time to provide evidence of competence, knowledge, skills, and abilities. Adams (1995) provides a number of suggestions to include in a teacher portfolio, including evidence of subject-matter knowledge, knowledge of instructional methodology, awareness of school cultures and student populations, evidence of teacher’s ability to organize and present information, and examples of academic and professional tasks completed. Ryan and Kuhs (1993) provide a framework of six interrelated components as the basis for focus in teacher assessment. These include a knowledge of subject matter, intellectual abilities and problem-solving skills, pedagogical skills, curriculum knowledge, insight, and skill, knowledge about learners and learning, and attitudes and dispositions of the teacher. The complexity of the frameworks for assessment indicate that a multidimensional tool is necessary to evaluate the relevant domains of a teacher’s knowledge. Portfolio assessment, according to numerous practitioners and researchers, holds particular promise in the assessment of

The advantages to portfolio assessment appear to be numerous but there do exist weaknesses in this approach. Just as advances in the pedagogical field have challenged traditional views of assessment, advances in psychometrics are challenging the more modern forms of assessment (Bachman, 1990). Advocates of alternative assessment claim that portfolios are a valid form of assessment because they are an authentic, real-life, process-oriented approach alternative to testing. There are, however, solid arguments against alternative assessment that include questions of validity and reliability. The relative novelty of portfolio assessment in teacher education has fostered a lack of agreement among practitioners and researchers regarding their reliability, validity, and utility (Cole et al., 1992; Edgerton, Hutchings & Quinlan, 1991; Johns & VanLeirsburg, 1991; Smolen & Newman, 1992; Shulman, 1988; Tierney, 1993; Wolf, 1991a). Shulman (1988) reports that using portfolio assessment in conjunction with other evaluative methods improves the reliability and validity of information collected through traditional means. A lack of a universal definition for teacher portfolios and clearly delineated means with which to evaluate them, however, foster a mistrust in the reliability, validity, and utility of portfolios as evaluative tools (Tierney, 1993; Wolf, 1991a). The current trend toward using portfolio assessment as a supplement to or alternative for traditional assessment may be useless if there is no agreement regarding the constructs used to define teacher competence nor how a portfolio can document this. Clarification in the evaluation of portfolios is also needed is this mode of alternative assessment is to flourish in teacher education.
Teacher Portfolios in Practice

Portfolios are now being accepted as assessment tools for preservice teachers in teacher education programs (Bird, 1988; Elbow & Balanoff, 1986; Olhausen & Ford, 1990; Shulman, 1990). Tierney (1992) reports an increasing number of teacher education programs that are implementing portfolios. A review of the field studies that impact the use of teacher portfolios is summarized here.

The Stanford Teacher Assessment Project (TAP) is a project funded by the Carnegie Foundation and headed by Lee Shulman that began as a national effort to develop and evaluate teacher assessment tools. Preservice teachers begin developing a teacher portfolio at the initiation of their teacher education program to document growth throughout the training. The portfolios include the following areas: introduction, professional responsibility, content mastery and content pedagogy, classroom management and organization, student specific pedagogy, and a conclusion. The project developed a scoring system to rate teacher portfolios holistically to avoid the dissection of teacher performance. The studies conducted by TAP are the most programmatic to date.

The Ohio Consortium for Portfolio Development (OCPD) is a collaborative research project that was initiated by Lee Shulman, the director of TAP, in 1987. Three universities in Ohio (Wright State University, University of Dayton, and Central State University) studied alternative models of teacher assessment using portfolios to document preservice teachers’ content mastery and competence. The Wright model was a process-oriented portfolio begun by the preservice teachers at the onset of the teacher education classes. It was used to document professional growth and to demonstrate problem-solving skills. The Central State model was used to promote teacher reflectivity. The members of the consortium proposed that portfolios complement traditional teacher assessment (Cole & Uphoff, 1992) and encourage the development of reflective practices, which leads to an achievement of higher standard of practice (Bird, 1990).
Nweke (1991) conducted a study of the portfolios created by preservice teachers enrolled in the Alabama Consortium for Minority Teacher Education. The project investigated the relationship between teacher portfolios and traditional measures of achievement such as the American College Test and college GPA. The results of the study showed a minimal correlation between performance measured by portfolios and performance measured by traditional tests. Naizer (1993) also conducted a study of portfolios in a preservice methods course for prospective mathematics teachers. He investigated issues of reliability and validity through a generalizability study and discriminant analysis. The results of the generalizability study indicated good agreement among raters, accounting for a maximum of 2.5% of the variance. The discriminant analysis revealed that students with more strategic knowledge and pedagogical experience scored higher on the portfolio thus helping to establish the portfolio as a valid measure of the processes of teaching.

Another relevant example is the experiment with portfolios in the evaluation of preservice teachers at the University of South Carolina (Ryan & Kuhs, 1993). Traditional exams have been replaced by the development, presentation, and defense of a portfolio in the early childhood and elementary certification programs. The guidelines and criteria are discussed with students at the beginning of the program. The procedures for the portfolio reflect what the faculty believe to be crucial to the development of effective teachers. The portfolio is used to reflect the process of becoming a teacher as well as for formal, final evaluation of teaching competency.

Work is also being done at Wichita State University where prospective teachers have been developing portfolios since 1991. A content analysis conducted by Potthoff, Carroll, Anderson, Attivo, and Kear (1996) showed that there are commonalities across self-selected portfolio submissions. The program requires that students include certain items in the portfolio such as a philosophy of education and a list of coursework, but the
final portfolios also include four to six self-selected entries. The authors report that the self-selected entries are often unit plans, lesson plans, and miscellaneous classroom activities. There were few statistically significant differences across demographic variables such as gender and race in determining the types of submissions included in the portfolio.

A final example of portfolio use is the National Teacher Assessment, the first assessment by the National Board for Teaching Standards. This program attempts to reward teachers with the title of outstanding teacher which can mean enhanced prestige and monetary gains. Of the 300 teachers that participated, only 81 were recognized as outstanding on the basis of a professional portfolio. Although the program reached a high level of inter-rater reliability (Hammadou, 1996), it was fraught with problems, especially management of time and funds. Bradley (1995) reports that the largest expense was the scoring of each candidate’s work; the cost of evaluating the first set of candidates was $4,000. Hammadou (1996) reports that the National Board for Professional Teaching standards are not the only ones using portfolios to license teachers; several states have mandated some form of a summative portfolio for licensure. They include: Alabama, Connecticut, New York, and Vermont. Proposals are also pending for performance-based assessments in Kentucky, Oregon, and Pennsylvania.

As demonstrated by the examples above, portfolio assessment is quickly becoming a new trend in the assessment of prospective teachers. It can be beneficial in the evaluation of preservice teachers because of the multidimensional and individual nature of the samples included. There are, however, several limitations to the use of portfolios for the evaluation of teaching competency. These limitations include the lack of a model for content and scoring, the amount of time and funds needed to develop and score the portfolios, the subjectivity inherent in the scoring of portfolios, and most importantly, the issues of reliability of the ratings assigned and the validity of the decisions that are made based on the ratings.
Limitations of Teacher Portfolios

Developing a system for teacher evaluation using either standardized tests or portfolio assessment may be difficult because of the lack of agreement on what constitutes effective instructional performance (Grady, 1992). Researchers in the field of teacher preparation are consistently revising what it means to be an effective teacher. This in turn becomes problematic at the evaluation stage also because it is not clear what aspects of a preservice teacher’s knowledge need to be measured. The important questions become: What do future teachers need to know how to do? followed by, How do we measure that knowledge? In order to make reliable and valid judgments about teacher preparedness from portfolio scores, it is necessary to know what constitutes teacher competency and what tasks measure that competency.

Researchers investigating portfolio utility in the Stanford Teacher Assessment Project (TAP) identified three limitations to portfolios (Shulman, 1988). The first was the difficulty that raters had in rating portfolios without a bias created by the portfolio packaging; portfolios that were aesthetically pleasing may have been rated higher. The second problem reported was the determination of what was sufficient evidence to include in the portfolio. The raters could not agree on the amount of content necessary to document teacher competency. The final limitation concerned the provision of structured guidelines in portfolio creation. Too few guidelines led to a wide variety among portfolios but highly structured guidelines removed the individuality of the teacher portfolio.

Common problems that lead to the unreliability and invalidity of portfolio scores are outlined by Stiggins and Bridgeford (1985). They include: (a) failure to inform students of the scoring criteria; (b) failure to prepare scoring procedures in advance; and (c) failure to check performance on the portfolio against other measures. Wolf (1991b) also claims that teacher portfolios are problematic because of the lack of models regarding content and
scoring methods. Controlling for these problems may lead to better reliability and validity in using teacher portfolios to assess teaching competency. Instructors or programs that intend to implement portfolio assessment must provide students with clear guidelines as to the type and amount of appropriate submissions as well as the methods of scoring to be used. Portfolio scores should also be monitored for scoring reliability and judge severity.

Schrier and Hammadou (1994) discuss the problem of the amount of time needed to develop and score portfolios as a challenge to portfolio assessment. They argue, however, that comprehensive assessment of a complex field like teaching naturally requires extensive time in evaluation. Ford and Olhausen (1991) agree that portfolios are time-consuming to evaluate but claim that clear criteria and training opportunities for raters can aid the implementation of teacher portfolios. Johns and VanLeirsburg (1991) echo the concern that portfolios are problematic in terms of the amount of time needed to manage and review the contents.

Hammadou (1996) discusses the incorporation of portfolio assessment in the teacher education program at the University of Rhode Island and the problems that were encountered there. She discusses the extensive amount of time that was required in the evaluation of the portfolios. More importantly, she describes the frustration of the faculty upon discovering that only one of the students linked theory to practice in the portfolio. This aspect is one of the principal reasons behind using portfolios in the evaluation of prospective teachers. Hammadou concluded that a well-articulated curriculum and clear guidelines of what is expected in terms of entries are essential to successful portfolio implementation.

Reliability of ratings can be a critical limitation to portfolio use in evaluating preservice teachers. Tierney (1994) cautions against using portfolios to evaluate prospective teachers because of the legal issues of the reliability of the scores, especially relative to cutoff scores for certification. Recent studies involving using portfolios to assess
future teachers report difficulties in achieving even moderate inter-rater reliability (Koretz, et al., 1992). This may be an indication that there is a need to simplify the rating process and provide training for raters. Limited information is available, however, on the reliability of ratings for teacher portfolios, and few studies report methods of training raters (Gibson, 1995).

Researchers support the use of teacher portfolios as part of a larger evaluation system in determining teacher competency (Cole et al., 1991; Nweke, 1991). Nweke (1991) claims that portfolios provide more reliable and valid data because of the multidimensional and comprehensive nature. Others support this view of using multiple sources of information such as those included in a portfolio for making decisions regarding teacher competency. Minimal work, however, has been conducted on establishing reliable and valid procedures for the implementation of portfolios in the assessment process. Some of the limitations listed can be overcome through investigation of issues of reliability and validity in using teacher portfolios to measure future teaching competence.

Issues of Reliability in Portfolio Assessment

Bachman (1990) addresses the importance of establishing the validity and reliability of assessment instruments and suggests an array of techniques for achieving reliability and validity. Reliability, as defined by Richards, Platt, and Platt (1992) is the degree to which a test gives consistent results when used on different occasions and by different people. The most common forms of reliability established in current research on portfolio assessment are inter-rater and intra-rater reliability (CRESST, 1993). It is important to demonstrate that scores assigned on a given measure are consistent across raters and within raters across assignments. Stiggins (1987) claims that reliability is maximized by using clear scoring criteria, training raters, gathering enough samples of performance, and providing for standard assessment conditions. Reckase (1997) discusses the importance of distinguishing
between score reliability and scorer (inter-rater) reliability. Scorer reliability refers only to the consistency with which judges agree on their ratings. He claims the only inference that can be made from this is that the scoring would be the same if other trained raters were used. While scorer reliability is helpful in establishing reliable rating scales and training procedures, Reckase argues that inferences about the score reliability, and thus the examinees' performance, cannot be made based on the inter-rater reliability coefficient. He claims that score reliability refers to whether the ratings are a "reasonable representation of student capabilities" (p. 30). Generalizability studies are useful in this aspect (Reckase, 1997; Brennan, 1997) because they can separate the variance in a model into components thus providing a measure of the impact due to the ability that is being measured and the impact due to other factors, such as raters, tasks, and occasions. Item Response Theory can also distinguish the error component in a model and separate the variance due to confounding factors so that the person's score becomes the true measure of ability.

Reckase (1995) estimated a theoretical reliability of scores on portfolios to demonstrate the possibility of achieving internal-consistency and inter-rater reliability coefficients. There are, however, few research studies which address this important issue. Of 46 publications concerning portfolio assessment in 1993, only 13 reported information about reliability in the evaluation of portfolios (CRESST, 1993). Several of those studies are discussed here. Herman and Winters (1994) investigated the technical quality of portfolio assessment and found that although evidence exists to support the validity and reliability of portfolio assessment, more research needs to be conducted before claims of the accuracy and usefulness of portfolios can be supported. In their investigation of reports that discuss the reliability of portfolios through the establishment of inter-rater agreement, three principal studies were discussed.

Vermont's statewide portfolio assessment project demonstrated disappointing results. Students in grades four and eight developed mathematics and writing portfolios that
were rated by classroom teachers. Koretz, Stecher, and Deibert (1993) reported inter-rater reliabilities ranging from 0.28 to 0.60; a level of agreement considered not sufficient enough to permit reporting of the scores to the state. Klein, McCaffrey, Stecher, and Koretz (1995) reported the reliability of mathematics portfolio scores based on the first two full years (1992 and 1993) of implementation of the program. Unfortunately, the scores from both years were not considered reliable enough to report to students. Reliability of a student’s total score ranged from 0.45 to 0.49 in 1992; in 1993, the values increased to a range of 0.63 to 0.71. The level of reliability sought, however, was 0.85. Inter-reader agreement rates were reported at 55%; only 10% above the agreement rate due to chance alone. Kappas ranged from 0.02 to 0.51 with a median of 0.28. The authors conclude that the low reliability of student scores and the low reader agreement rate were caused by insufficient reader calibration, scoring guides that were too open to reader discretion and interpretation, and the freedom with which students were allowed to choose the pieces included in the portfolio.

Research in local settings, however, suggests that portfolios can be scored reliably (Herman, Gearhart, & Baker, 1993; LeMahieu, Gitomer, & Eresh, 1995). LeMahieu et al. (1995) achieved an average 0.84 inter-rater reliability coefficient in the study of the Pittsburgh Public School District’s use of portfolio assessment in evaluating writing performance. In a previous study of the same program, LeMahieu, Gitomer, and Eresh (1993) as cited in Herman and Winters (1994) achieved inter-rater agreement correlations that ranged from 0.60 to 0.70. Herman et al. (1993), in a study of an elementary school-based study of writing portfolios, reported average correlations between scores given by pairs of raters was 0.82 and the percentage of agreement for all pairs of raters averaged 0.98.

Another related study by Brookhart (1995) on the reliability and validity of student annotations in portfolio assessment demonstrated an acceptable internal consistency
reliability. The study investigated the use of a student annotation form used to collect student reflections on work samples in portfolios. A relationship between student perceptions of the task and their ability to handle it would be demonstrated through each student’s willingness to complete the task as reflected in the annotation. The students were asked to report their perceptions of the difficulty of the task, their effort in responding, and their willingness to do more work at the same level of difficulty. The author reported alpha values of 0.62 for difficulty ratings, 0.71 for do-more ratings, and 0.69 for academic self-efficacy, thus demonstrating consistency among the student annotations and the performance of the task.

Other studies include several dissertations that investigated the reliability of scores and reader agreement rates. Naizer (1993), in a study of the use of performance-portfolios with preservice science and math teachers at Texas A &M university, reported a high degree of inter-rater agreement (48% to 86%). Gibson (1995), in a study of teacher portfolios in special education, also conducted at Texas A &M university, reported a 78.26% agreement rate among the judges of the portfolios.

Despite the promising results reported in several of the studies that have investigated the reliability of portfolio scores more work is still needed. There is evidence that inter-rater reliability is achievable when raters are extensively trained in the use of the scoring rubrics. This task, however, often discourages classroom teachers from using portfolio assessment because of the overwhelming and time-consuming task of rating the portfolios and then calibrating with other teachers. Linacre (1993) has developed a statistical model that allows for objective estimates of a student’s ability to be made free from judging error, item difficulty, and the randomness of rating scale categories. The many facet Rasch model analysis may provide the means that advocates of alternative assessment tools need to satisfy the psychometricians demands for valid and reliable procedures. The model can establish reliability in the use of what are essentially
individualistic and qualitative tools without requiring extensive rater training and calibration. Establishing the reliability of portfolio scores through Rasch-model analysis may increase the utility of portfolio assessment in teacher education.

Issues of Validity in Portfolio Assessment

Although the reliability of portfolio scores is a necessary condition, it is not a sufficient enough prerequisite to the core issue in technical quality: validity (Herman and Winters, 1994). Stiggins (1987) suggests that validity can be ensured by providing for clarity of purpose, clearly delineated characteristics for evaluation, the use of tasks that represent a range of performance contexts, and comparison of performance ratings with other data. The National Board for Professional Teaching Standards establishes validity as one of the four criteria to be used in selecting assessment measures for certification of teachers (Baratz-Snowden, 1993). Richards et al. (1992) define validity as the degree to which an instrument measures what it is intended to measure. Bachman (1990) gives further details as to the various kinds of validity, for example content validity and criterion validity, that can be established and the procedures available to demonstrate each type.

Content validity, according to Bachman (1990) can be established through ensuring the representativeness of the sample. Researchers argue that content validity of portfolios may be achieved by including a variety of documents in the portfolio that capture a true measure of future performance (Gredler, 1996; Hammadou & Schrier, 1994; Ryan & Kuhs, 1993). Martinez and Lipson (1989) contend that real-world tasks, such as those found in portfolio assessment, which relate subject matter to the complex world of teaching offer greater validity than traditional measures. Wiggins (1990) suggests that content validity should depend in part on whether the assessment tool simulates real-world tests of ability. Ryan and Kuhs (1992) summarize the domains of pedagogical knowledge in which a preservice teacher should be competent before achieving certification. These include a
knowledge of subject matter, problem-solving skills, pedagogical skills, knowledge about learners and their attitudes, and insights into curriculum and planning. To the extent that portfolio assessment can document the achievement of those objectives, it is a valid measure of pedagogical content knowledge and a reliable indication of a preservice teacher’s ability to perform in the classroom.

The criterion and predictive validity of portfolios is not so easily attained, however. Herman and Winters (1994) advocate looking at relationships between the results of portfolio assessment and other indicators of student performance or achievement. The problem with using other measures of achievement to establish the criterion validity of portfolios is precisely the argument that advocates of alternative assessment employ to convince practitioners and researchers of the need to supplement traditional assessments: portfolios measure abilities and performance that cannot be measured through traditional assessment tools. The problem in establishing validity, therefore, lies primarily in determining that traditional measures of achievement have little or no relationship with scores provided by alternative measures. Fewer studies have been conducted that investigate the validity of portfolio scores due to this problem; Herman and Winters (1994), however, discuss two such reports. Performance tasks, such as those used in the methods course that served as the context for this study, can be considered similar in their measurement of teaching competence because they require that the teacher apply content and pedagogical knowledge within the educational context. A positive relationship, therefore, might be expected because both the performance tasks and the portfolio scores evaluate the application of knowledge, not only the knowledge itself. Naizer (1993), for example, found that pedagogical content knowledge as measured by education courses and years of teaching experience best predicted portfolio scores in his study of the use of portfolio assessment with preservice math and science teachers.
Other researchers have also investigated issues of validity in terms of the relationships between portfolio scores and other measures of ability. Koretz et al. (1993) investigated the relationships between portfolio scores and measure of both similar and different capabilities in the Vermont program. They found moderate correlations ranging from 0.47 to 0.58 between writing portfolio scores and direct writing assessments. Similarly, Gearhart, Herman, Baker, and Whittaker (1993) found virtually no relationship when comparing results from standard writing assessments and those from writing portfolios.

Nweke (1991) conducted a study in which scores from the portfolios of 30 college students enrolled in teacher education programs were correlated with other established measures of academic achievement. Results indicated that performance measured by portfolios is different from performance measured by traditional forms of assessment such as the American College Test and overall GPA. The correlations reported for the portfolio scores and college GPA were $r=0.11$, and the portfolio scores and the American College Test $r= -0.12$.

Naizer (1993) investigated the validity of portfolio scores in a math and science preservice teacher education course. In a measure of association analysis of the final exam and portfolio scores, a low correlation ($r=0.22$) was achieved, thus suggesting that the two instruments were measuring different knowledge and performance dimensions. Concurrent validity was established, however, with pedagogical content knowledge which was considered to be a similar measure to the portfolios. Through a discriminant analysis, Naizer demonstrated that students with more strategic knowledge and more pedagogical experience as measured by the number of education courses and years of teaching experience scored higher on the performance portfolios.

Establishing the validity of portfolio scores is a necessary aspect in ensuring the technical quality of portfolio assessment, and in advocating its use as a supplement to a
traditional assessment framework. The problem that many researchers encounter, however, is the lack of established measures of performance with which to compare portfolio scores. It is therefore often necessary to compare the relationships between traditional measures of assessment and alternative assessment tools with the expectation that these relationships will prove to be minimal.

Conclusion

One of the most problematic areas in portfolio assessment is the lack of reliability of the scores and the validity of the measure in making judgments concerning the evaluation of students. Barton and Collins (1993) argue, however, that “the lack of clearly established procedures and criteria for assessing the evidence that students include in their portfolios is indicative of the relative novelty of the portfolio concept” (p. 205). Stiggins and Bridgeford (1985) contend that many educators ignore important practices when using alternative assessment measures thus contributing to the invalidity and unreliability of alternative assessment.

Attempts to improve the reliability and validity of the use of portfolio assessment in making educational decisions have been made (Gibson, 1995; Herman and Winters, 1994; Koretz et al., 1995; Naizer, 1993; Nweke, 1991; Reckase, 1995), but more studies of this nature need to be undertaken if portfolio assessment is to become a viable supplement to testing in education. Calfee and Perfumo (1993) assert that portfolio assessment is likely to fail if valid data and reliable information are not provided through this type of evaluation. This study, therefore, is proposed to investigate the issues of reliability and validity of the use of portfolio assessment in preservice foreign language teacher education.

This section has provided relevant background information for the present study. A brief history of traditional and alternative assessment measures describes the point of conflict between the two in terms of issues of reliability and validity. The implementation of
portfolio assessment can be useful to assessment in teacher education to document teacher performance, allow for self-assessment, and to integrate the complex nature of teaching with a multidimensional form of assessment such as portfolios. The research methods used in this investigation of reliability and validity of the use of portfolios in the assessment of preservice foreign language teachers are described in the next section.
CHAPTER 3

METHODOLOGY

Introduction

In the previous sections the purpose of this investigation and the background information relevant to the study were presented to provide a rationale for the research and to describe the problem that is addressed in this study. This section gives an outline of the research process and methods that were used in the collection and analysis of the data. The underlying purpose of this study is to provide baseline data about the issues of reliability and validity in using portfolio assessment in a foreign language teacher education course. The portfolios developed by the students in the foreign language teacher education course used as the context for the study served as the dependent variable for the statistical analyses to be conducted in this research. The following areas are addressed here: subjects and context of the study, instrumentation and data collection, the research methods employed, and the analysis and presentation of the data.

Context of the study

The current study was conducted in a foreign language methods course, Teaching Foreign Languages in the Elementary School, at The Ohio State University. The course is designed to introduce preservice teachers to methodology that is appropriate for foreign and second language instruction in an elementary school setting. The expected learning
outcomes of the course include an awareness of FLES programs and instructional approaches, an understanding of child psychology and child cognitive development, and the ability to develop instructional strategies and techniques, teaching and learning materials, and assessment instruments that are appropriate for use with elementary school students. Following the notion of Shulman’s “pedagogical content knowledge” (Shulman, 1986; 1987), the students are encouraged to develop instructional strategies that reflect the contextualized and communicative aspects of learning a language. Traditional assessment, in the form of a final, written exam, is used in conjunction with the performance-based assessment to provide a broader basis for evaluation of the students. Whereas a traditional exam measures the students’ knowledge of FLES programs and instructional approaches (conceptual knowledge), the performance tasks and the portfolio include activities that allow for the application of content knowledge and pedagogical knowledge in the creation of activities and materials that are appropriate in the teaching of languages to young children.

Subjects

The subjects for the study consisted of the 37 students enrolled in two sections of the methods course previously described. Demographic information was solicited from the students on the pre-course survey and is used here to describe the sample. Of the 37 students, 14 were international students; 18 were graduate students, 14 were undergraduates, and 5 were nontraditional (post-BA certification or continuing education students). The average GPA for students in the course was 3.5; the students reported having taken an average of 9.2 education courses prior to this one. The average number of years of teaching experience (including international settings, university TAs, and tutoring) was 3.4 years; 10 of the students reported having no teaching experience (0 years). Major
languages included Spanish, French, German, Japanese, Arabic, and English as a Second Language (ESL).

A stratified, purposeful sample of ten students was chosen for further investigation. Based on the demographics of the class, the ten subjects were selected according to academic status, resident classification, and primary language taught. The sample consisted of five undergraduate, three graduate, and two nontraditional students. There were two Spanish majors, two German, two French, two ESL, one Japanese, and one Arabic. Three of the ten subjects were international students. The stratification variables of academic status, resident classification, and primary language taught were chosen in order to adequately represent the larger population of preservice teachers to which this study generalizes. The representative population of foreign and second language teachers is one that includes individuals from a variety of backgrounds with different educational experiences. In order to account for possible differences in the portfolios attributable to language, country of origin, and educational history, it was determined that the above stratification variables should be employed to select a representative sample from the class.

Instrumentation

Data was collected over the duration of one quarter (ten weeks) through the use of the instruments described below.

Student information sheet: The students were asked to self-report data including: country of origin, gender, age, number of years of teaching experience, number of education courses taken, student status (undergraduate, graduate, or professional), and overall GPA. Students were also asked to grant permission for use of their portfolios in this research (See Appendix A).
Assessment survey: An instrument developed by the researcher that asks students to report their beliefs about the objectives of assessment, and to rate their understanding of a number of assessment tools, including portfolios and rating scales. The instrument was piloted in two educational courses prior to administration and changes were made for clarification of the items (See Appendix B).

Portfolio: The portfolio model used in the course is best described as a documentation portfolio (Goodman, et al., 1989), which is a compilation of student work that focuses on providing a multidimensional view of teaching competence and professional development. Students developed a documentation portfolio throughout the course in which they included a variety of independently completed assignments that demonstrated competence in foreign language pedagogy appropriate for children. The nature of the portfolio allows for the use of multiple intelligences such as those proposed by Gardner (1983). Tasks were identified that represented the abilities outlined in Gardner's theory. For example, one of the instruction-based tasks asked students to create a lesson based on movement (bodily-kinesthetic) or song (musical). Another instruction-based task asked students to develop a unit that was appropriate for a particular age group (interpersonal). The library-based tasks required students to conduct research (linguistic and logical) whereas the field-based tasks necessitated on-site observations and interviews with teachers (inter and intrapersonal). The portfolio accounted for 25 percent of the total course grade. A review of the literature of the major competencies required by teachers of foreign language at the elementary school level was completed to identify the activities to be included in the portfolio. The professional standards created by the American Council of Teachers of Foreign Languages (ACTFL, 1988), the American Association of Teachers of French (Goepper, 1989), the American Association of Teachers of German (AATG, 1992), and the American Association of Teachers of Spanish and Portuguese (AATSP, 1990) as summarized in Lipton (1996),
provided the framework for the list of suggested activities (See Appendix C for the guidelines and activities).

**Analytic rating scale:** A scoring rubric developed by the researcher that was used by the five raters in the anonymous rating of the ten portfolios in the random sample (See Appendix D). Data from this rating scale was used in the Rasch-model analysis and the generalizability study to determine the reliability of the raters and the portfolio scores.

**Holistic rating scale:** The criteria used by the researcher to assign grades to the portfolios. This scale was also used by the blind raters in an overall assessment of the ten sample portfolios (See Appendix E). Data from this scale was used in the correlation, regression, and factor analyses to determine the validity of the portfolio scores.

**Performance tasks:** The students were required to complete several assignments that applied the skills and knowledge acquired through course readings and lectures. These tasks included daily and unit lesson planning, presentation of lessons, peer-teaching activities, creation of assessment activities, and a personal journal of written reactions to course material. Ryan and Kuhs (1992) summarized the domains of pedagogical knowledge in which teachers should be proficient before certification (e.g., curriculum and planning, problem-solving, pedagogical skills). The performance-based tasks in the course provided students with an arena to demonstrate ability in these domains. The tasks accounted for 50 percent of the total grade for the course (See Appendix F for the guidelines to these tasks).

**Final exam:** The students in the course were also required to complete a traditional, written exam that asked them to synthesize and present the knowledge that they acquired in the course. The exam was essay format and was rated on a 100 point basis; it accounted for 20 percent of the total course grade (See Appendix G).
Procedure

The student information sheet and the pre-course administration of the assessment survey were completed during the first class meeting of the ten week course. Students were given a course syllabus and the guidelines for the portfolio submissions, as well as an outline of the criteria used in the evaluation of the portfolios. Students began working on the portfolio and added submissions to the portfolio over the duration of the course. The researcher provided feedback at midterm (five weeks into the course) to help the students in developing the portfolio.

The post-course assessment survey and take-home final exam for the course were distributed during the final course meeting. The following question was added to the end of the survey: ‘Based on your experience with portfolios this quarter, would you be willing to implement portfolio assessment in future classes that you teach? Please explain why or why not.’ The portfolios from both sections of the course were collected and rated by the researcher using the holistic rating rubric for purposes of assigning a course grade. The course instructor for each section graded the final exams of her own students. The exam grades and portfolio grades, as well as the grades for the performance tasks completed throughout the course, were collected by the researcher for the correlational and multiple regression analyses.

A stratified, purposeful sample of ten portfolios was chosen for further investigation. The portfolios chosen for the sample were based on the following variables: (1) academic status of the student (undergraduate, graduate, professional/non-degree), (2) residence status (international versus domestic students), and (3) primary language taught (commonly taught versus less-commonly taught). These stratification variables were chosen in order to best represent the field of foreign language teachers to which the study generalizes. Five raters, familiar with the field of foreign language teacher education, rated each of the ten portfolios anonymously using both the analytic and the holistic scoring
rubric. Although the raters did not have extensive experience in using portfolio assessment to evaluate the teaching competence of preservice teachers, all were familiar with portfolio assessment and scoring rubrics. The raters were given explicit, written instructions for rating the portfolios, as well as the guidelines and criteria for evaluation that had been provided to the students. The course instructors did not participate in this evaluation of the portfolios to ensure objectivity in the scoring of the portfolios and to attempt to exclude rater bias. Each rater evaluated the portfolios analytically by individual submissions, and then assigned a holistic score to each of the portfolios. These scores were not used in determining the overall course grade of the students, but were used in establishing reliable procedures for integrating portfolios in foreign language teacher education. Finally, of the ten subjects chosen for further investigation, three were interviewed in a focus group regarding their overall reaction to the portfolio experience. Appendix H contains the questions used to guide the interview.

Variables

**Dependent variable**

According to Shulman’s (1987) model for the knowledge base in effective teaching, teacher preparation includes educating teachers to reason about their teaching as well as to perform skillfully. He argues that teachers must learn to use their knowledge base to make choices and determine their actions in the classroom. Shulman (1987) claims that the standardized assessments used to measure subject matter knowledge and basic skills trivialize teaching and ignore the complexities of the educational context. In particular, he identifies pedagogical content knowledge as of special interest in the education and assessment of teachers because it represents the combination of content and pedagogy that is necessary for effective presentation of material. Portfolios have been described as an attractive method of evaluating preservice teachers (Tierney, Carter, &
Desai, 1991; Wiggins, 1989; Zollman & Jones, 1994) because they allow for a demonstration of a variety of abilities that are necessary for effective teaching. Portfolio assessment, however, can be subjective in its evaluation of teaching competence because of factors such as individual submissions and raters, and reliable and valid procedures for assessing students are needed.

Gardner's (1983) theory of multiple intelligences proposes that each individual has a unique profile of competencies that cannot be measured solely through the use of tests. The reasoning behind this assumption is that the majority of traditional tests ask students to rely heavily on linguistic and logical-mathematical skills. There are, however, five other areas of intelligence that Gardner believes can better be reflected by a multidimensional assessment tool such as portfolios. The teaching portfolio developed in the course, therefore, served as the dependent variable for this study in order to determine if the scores assigned are reliable and valid in the assessment of teachers. In this study, performance-based tasks such as creating materials and activities, observing other teachers and children, and visiting resource centers allowed students to draw on abilities other than linguistic and logical reasoning. The portfolio was designed with the intention of allowing preservice teachers the opportunity to demonstrate pedagogical knowledge and explore areas of professional development. An important component of the portfolio was the self-reflection piece that accompanied each work sample in the portfolio. For each of the six submissions, students were required to discuss the choice of submission, the procedures used to complete the task, and their overall satisfaction with the submission. There was also a self-reflecting component in the initial letter to the reader found at the beginning of each portfolio. The portfolio scores were a composite of scores assigned to each portfolio by the five raters using both holistic and analytic rating scales.
Independent variables

Performance tasks: These activities allowed the student to use a number of intelligences as well as apply content knowledge and pedagogical reasoning to create an activity, teach a lesson, and design a unit with materials. The tasks were similar to the creation of a portfolio in that they asked the preservice teachers to demonstrate teaching competence through application of pedagogical content knowledge. The principal difference between the performance tasks and the portfolio is the reflection required in completion of the portfolio; the students were asked to reflect on their choice of submissions to the portfolio and their perceived performance on the activities. These self-reflections were also rated and combined with the ratings given on each submission. Because both the performance tasks and the completion of the portfolio required students to apply knowledge of pedagogy, content, and educational theory, the scores received on the performance tasks should therefore correlate with the portfolio scores.

Years of teaching experience: “Teaching is, essentially, a learned profession” (Shulman, 1987). Teaching experience generally fosters the acquisition of pedagogical skills that in turn leads to teaching competence. According to Shulman (1987), teaching begins with the understanding of content and how to teach it. After a series of activities designed to transmit the knowledge from teacher to student, teaching ends with new comprehension by the student and the teacher. Studies of inexperienced and experienced teachers have provided a concept of effective teaching. This “wisdom of practice” is a principal source of the knowledge base for teaching, but is only available through experience in the classroom. Years of teaching experience, therefore, is a manner of codifying of pedagogical content knowledge. The students reported the years of teaching experience on the information sheet at the beginning of the course. Although many of the students in the course are classified as preservice teachers in the American public school system, they may in fact have had teaching experience in other countries or settings. Based
on Shulman’s model (1987), it was expected that students that reported a higher number of years of experience would receive a high score on the portfolio. This relationship was found to be positive in the study conducted by Naizer (1993).

Course final exam: Content knowledge, and curriculum and pedagogical theory, are important aspects in the education of preservice teachers (Shulman, 1987). A written exam may be reliable in measuring the curriculum and pedagogical knowledge acquired in a course (conceptual knowledge) but not necessarily the ability to perform in the classroom (pedagogical content knowledge). The exam in the course used in this study asked the students to synthesize the readings completed in the course and present a logical argument in support of elementary school programs. The students had to rely heavily on logical and linguistic reasoning in responding to this request. The portfolio, however, asked the preservice teachers to combine general curriculum and pedagogical knowledge with content knowledge to demonstrate the ability to transfer the material into learnable format through activities such as the creation of sample lessons and materials, and observation of and discussion with inservice teachers. Because the portfolio and exam were assessing different abilities in the course, little or no relationship between the exam scores and the portfolio scores was expected. This would support findings in the literature that portfolio performance is not reflected in traditional exams such as the American College Test (Nweke, 1991).

Grade point average: A student’s GPA is a reflection of achievement in the courses completed during a number of years of study. The high number of courses that require a final, written exam or paper typically result in a student’s GPA being a reflection of the ability to use linguistic and logical intelligences, rather than intelligences such as musical, spatial or bodily-kinesthetic. By the same token, GPA often indicates an ability for scholarship, but may not indicate an ability to transfer knowledge into learnable format for students, also known as pedagogical content knowledge. Shulman (1987) claims that
content knowledge, general pedagogical knowledge, and curriculum knowledge are important components of the knowledge base for teachers but they are not a sufficient base for effective instruction. Pedagogical content knowledge is what he argues distinguishes content specialists from teachers. The portfolio format used in this study required demonstration of pedagogical content knowledge as well as knowledge of content and curriculum. Because of the multidimensional nature of the portfolio in contrast with a general score of achievement, it was expected that the GPA of the preservice teachers would have little or no relationship with the portfolio (dependent variable). This expectation is reflected by studies in the field that indicated no relationship between teacher performance on portfolios and GPA (Howey & Zimpher, 1996; Nweke, 1991).

Hypotheses

The following are the null hypotheses that were investigated in the present study.

H01 As a measure of validity, there is no relationship, neither correlational nor predictive, between portfolio scores and the following measures:

a. application of pedagogical content knowledge as measured by performance tasks completed in the course.

b. pedagogical content knowledge as measured by reported years of teaching experience.

c. achievement of the course objectives as measured by subjects’ grades on the course final exam.

d. academic achievement in a teacher education program as measured by the subjects’ overall GPA.
As a measure of reliability:

a. There is no agreement among raters in the ratings of each portfolio.

b. There is no impact on portfolio scores due to raters, persons, or tasks in evaluating preservice teachers' competence.

c. There is no relationship among the grades assigned to individual pieces within the portfolio using the analytical rating scale and the overall grade assigned to the portfolio using the holistic rating scale.

Data analysis

In order to overcome the limitations of portfolio assessment outlined previously in the literature review (lack of agreement in content and scoring models, amount of time and funds needed, subjectivity inherent in scoring, reliability of scores, and validity of decisions based on the scores), steps were taken to establish a means of investigating the validity of the portfolio scores through correlation analysis, multiple regression, and factor analysis, and the reliability of both scores and scorers through correlation analysis, item response theory, and a generalizability study. Because of the nature of the data collected, both quantitative and descriptive measures were used. The first research question investigated the criterion and predictive validity of the portfolios. To establish baseline data concerning the validity of portfolio assessment as a measure of the preservice teachers' pedagogical competence, intercorrelations among portfolio scores (dependent variable) and the independent variables (performance tasks, years of teaching experience, course final exam, and overall GPA) were calculated using Pearson's Product Moment Correlation. A multiple regression analysis was conducted to determine which of the independent variables best predicted performance on the dependent variable (portfolio). A principal components factor analysis was also conducted to confirm the validity of the two constructs that were measured by the different variables: (a) conceptual knowledge as measured by the final
exam and GPA; and (b) pedagogical content knowledge as measured by the portfolios, performance tasks, and years of teaching experience.

The first limitation cited in the literature as problematic for portfolio assessment is the lack of a model for the content of a portfolio. Researchers claim that highly structured guidelines affect the individuality of the portfolios, whereas unclear, or loosely structured guidelines lead to a wide variety in portfolio content. In this study, this problem was accounted for in several ways. Firstly, through examination of accepted guidelines of teaching competency in foreign language teacher education, a series of tasks for possible inclusion in the portfolio were drafted that would enable the preservice teachers to demonstrate their teaching competency. Secondly, the students were given a list of suggested activities that could be completed to fulfill the required submissions. The preservice teachers had a choice of activities but were provided with clear guidelines in order to complete each of the tasks. Thirdly, the portfolios were collected halfway through course to check the students’ understanding of the guidelines and criteria for completion. Informal, written feedback was provided to each student in order to assure that the materials included in the portfolio were sufficient in quantity and quality for rating purposes. In terms of the validity of the decisions based on the portfolio scores, this limitation was overcome by not basing the final grade entirely on the portfolio scores. The course grade was a combination of grades received on both alternative and traditional assessment instruments, thus assuring the validity of the grades.

A second limitation in portfolio assessment is the lack of clear models for scoring. Stiggens and Bridgeford (1985) claim that many problems of reliability stem from the failure to prepare scoring rubrics in advance and the failure to share the scoring criteria with the students. In this study, those limitations were overcome by creating the scoring rubric prior to portfolio creation, and by sharing the rubric with the students before they began assembling a portfolio. The students in the course were given a copy of the holistic scoring
rubric at the same time as they were given the guidelines for portfolio content. The guidelines for portfolio content and the scoring rubric were discussed in class the first week of the course to assure a mutual understanding. The model for the holistic scoring rubric was adapted from models used in other courses in the foreign language education program area at The Ohio State University and was tested with the raters and with professors in the teacher preparation program to assure the clarity of the holistic rubric. As a measure of assurance against subjectivity from the use of only the holistic rating scale, an analytic rating scale was developed and used to rate the portfolios a second time. The two scores were then analyzed by means of a t-test to determine if there were any significant differences among the scores produced by using the holistic rubric versus using the analytic rating scale. It was expected that the individual submissions within each portfolio were consistent in quality so as not to cause a significant change in scores when using either of the rating scales.

The subjectivity inherent in scoring portfolios is another principal limitation in the use of portfolio assessment. The second research question in this study investigated the reliability of the procedures used in evaluating the portfolios. Through a variety of analyses, both score and scorer (inter-rater) reliability were examined. Initially, an inter-rater reliability coefficient was calculated using the holistic rating scale data from the Pearson Product Moment correlation matrix and Fisher’s transformation table. These data were the holistic scores assigned by each of the five raters to the ten portfolios selected for the sample. Item response theory, which presents an estimate of an individual’s ability that is free from measurement error due to factors such as raters and tasks, was also employed using the data from the analytical rating scale to ensure the reliability of the portfolio scores. A many-facet Rasch model analysis of the data from the stratified, random sample of ten portfolios was conducted using the computer program FACETS. Data collected from the five judges’ analytical scale ratings of the ten portfolios were entered in a three-facet model
to determine the consistency with which the raters used the rating scales in assessing the portfolios. Based on the judge severity score of each rater, it was possible to separate the variance due to raters and thus determine the ability of each student to apply pedagogical content knowledge in the completion of the portfolio activities. The reliability of the portfolio scores was also investigated using generalizability theory. Traditional reliability (KR-20) measures only internal consistency and is not recommended for performance-based assessment, including portfolios (Trevisan, 1990). Brennan (1997) and Reckase (1997) recommend using generalizability theory to analyze score reliability. Generalizability theory (Shavelson & Webb, 1991) analyzes the various sources of measurement error and provides a reliability coefficient for the portfolio scores. Data were collected from the raters using the analytical rating scale; a Likert scale of 1-6 with 6 representing superior work. A generalizability study was used to determine the relative impact of measurement error that was due to the raters, the items, and the individuals creating the portfolios (desired variance), as well as the interactions among these variables.

The secondary research questions concerning the students’ beliefs about portfolio assessment were investigated through qualitative methods: a pre- and post-course survey and a post-course focus-group interview. The results from the initial and final surveys of the students' notions of assessment were calculated and compared to determine if there was any significant difference among the pre and post course student ratings of understanding of assessment purposes and familiarity with assessment tools. The interview was conducted with three of the subjects from the stratified sample of ten so as to better represent the students in the course. The discussion centered around the students’ reactions to using portfolios and their beliefs about what the portfolio demonstrated relative to teaching competence. Significant statements from the interview were transcribed for further analysis and are included in the results section.
Conclusion

The current study investigated the reliability and validity of portfolio scores in evaluating the teaching competence of preservice foreign language teachers. This section described the overall design of the study and the procedures for data collection and analysis. The following chapter provides the results of the study and a discussion of these results using the research questions and hypotheses as an organizational format.
CHAPTER 4

RESULTS AND DISCUSSION

Introduction

In recent years, teacher education programs have begun to integrate portfolio assessment into their evaluation of preservice teachers. The rationale behind this is strong: portfolios are a multidimensional tool that can measure the complex nature of teaching more holistically than traditional tests (Zollman and Jones, 1994). Few studies, however, have been reported in the literature that investigate reliability and validity of portfolios in assessing students. Calfee and Perfumo (1993) claim that research of the reliability and validity of portfolio assessment is necessary to ensure the success of this tool in evaluation. The results of the studies that have been done indicate that reliable and valid procedures can be established for portfolio assessment (e.g., Gibson, 1995; Koretz, et al., 1995; Naizer, 1993; Nweke, 1991). This study also contributes to the growing body of research that is indicating that portfolio assessment can be a reliable and valid method of evaluation.

Issues of reliability and validity in using portfolio assessment to evaluate teacher education students were the principal focus of the current study. As a secondary aspect, information regarding preservice teachers' attitudes toward portfolio assessment were also investigated. Both quantitative and qualitative data were collected to address the research questions. Validity issues were investigated through intercorrelational and multiple
regression analyses among the principal variables (portfolio scores, performance tasks, GPA, final exam grades, and years of teaching experience). Factor analysis was used to establish the validity of the constructs being measured by the variables: conceptual knowledge and pedagogical content knowledge.

Reliability was investigated using a correlational analysis, item response theory, and generalizability theory. Rater consistency or agreement was examined using the holistic rating scale data through a Pearson correlation matrix, and then transformed to an inter-rater reliability coefficient using a Fisher conversion table. The analytical rating scale data were subjected to a Rasch model FACETS analysis to examine score reliability and to determine rater severity and consistency in using the rating scales to assess each submission in the portfolios. Score reliability was also investigated using a generalizability study. The nature of the study and the data that were collected warranted the use of two rating scales to evaluate the portfolios (holistic and analytical). A t-test was conducted to determine if the two rating scales were applied according to the same criteria; in other words, if there was a significant difference between the scores given using either of the two rating scales. Finally, the attitudes of preservice teachers toward portfolio assessment, and their understanding of the evaluation process were investigated through surveys involving all of the students in the course, as well as a focus-group interview with three of the students.

This chapter contains the results of a number of statistical analyses performed using both the holistic and analytical data from the portfolio ratings. The results from the pre- and post-course assessment surveys and transcribed statements from the interview with students are also presented. Results will be organized into three sections according to the research questions and hypotheses. Part One addresses the issues of validity that were investigated in this study (Null Hypothesis 1). Part Two presents the results in the establishment of reliability in using portfolios including both the reliability of the portfolio
scores and of the raters (Null Hypothesis 2). Finally, Part Three addresses the understanding and beliefs that students in the course demonstrated both pre- and post-course through responses on the assessment survey. Information from the interview concerning students’ attitudes toward portfolio assessment is also included in Part Three. Discussion of the results follows the presentation of all analyses at the end of the chapter.

Results

Part One: Issues of Validity

The first research question of this study addressed the establishment of validity of portfolio assessment for the evaluation of preservice teachers. Several types of validity were addressed in this study including content validity, criterion and predictive validity, and construct validity. Data for the analyses were collected from the 37 students in the class through means of self-reported questionnaires and from the grades given by both instructors. The researcher was the only rater for the portfolios at this level of the study and she used the holistic rating scale to grade the portfolios; the analytical scale was not used for this portion of the study. The content validity of the portfolios was addressed by the researcher prior to the beginning of the course and the implementation of portfolio assessment. Criterion and predictive validity were investigated through correlational and regression analyses using the data from the holistic ratings of the portfolios. Construct validity was addressed through a principal components factor analysis also using the holistic data.

The researcher established the content validity of the portfolios through a survey of preservice teacher education syllabi and literature to determine the types of activities that allowed future teachers to demonstrate the ability to apply content knowledge. The national standards established by foreign language teaching associations such as ACTFL (American Council of Teachers of Foreign Languages) served as the framework for creating many of
the suggested tasks. Observation of elementary school foreign language classrooms and teachers throughout the public school system in Columbus, Ohio provided further insight into the pedagogical knowledge that a preservice teacher needs to perform in the elementary-school foreign-language classroom. An ensuing list of suggested activities to include in the portfolio was created by the researcher, and approved by the course instructor as being a representative sample of the types of activities an elementary-school foreign-language teacher is expected to be able to perform. These activities included lesson planning, creation of classroom materials, observations of elementary school children and teachers, visits to resource centers to collect materials, and investigation of current topics using on-line networks and current educational journals (see Appendix C for a complete list of activities). One of the limitations of portfolio assessment that is cited in the literature is the difficult decision to provide strict guidelines for portfolio submissions thus inhibiting creativity, or loosely structured guidelines which often lead to a wide variety in the submissions. According to Gibson (1995), many researchers establish guidelines for items that are required in the portfolio, but do not restrict items not on the list from being included in the portfolio (Cole et al, 1992; Nweke, 1991). The guidelines provided to the subjects in this study were explicit enough to ensure mutual understanding of the necessary requirements, but not so restrictive as to limit individuality.

A content analysis of the portfolio submissions for the 37 students indicated a preference for certain types of submissions on the part of the students. Under the library-based category, the majority of the submissions that students chose to include were an ERIC search of a current topic relative to foreign language in the elementary schools; a reaction to the English-only issue currently being debated in many state legislatures; an annotated bibliography of children’s books; and a mini-research report on a topic in early foreign language learning. For the instruction-based section, students generally chose to include a cultural unit including lesson plans; a translation of a book to their L2 with
accompanying lessons plans; and a learning activity or game that incorporated a current approach to communication (e.g., TPR). The field-based submissions were primarily focused on interviews with teachers in the field; visits to local stores to compile resource lists for teaching foreign languages to children; and a rationale for early language learning in schools.

The validity of the portfolios used in the study was investigated using the holistic rating scale data from the researcher’s evaluation of the portfolios. Correlation, multiple regression, and factor analyses were used to establish the criterion/predictive and construct validity of portfolio assessment. The results are discussed here.

*Research Question One*: As a measure of the validity of portfolio assessment in measuring future teacher performance, to what extent do the independent variables (*performance tasks, years of teaching experience, final exam, and GPA*) correlate with and predict performance on the portfolio?

To answer Question One, data were collected from the various course requirements. Students developed a portfolio over the duration of the quarter. The portfolios contained six submissions that were intended to be a representation of each individual’s pedagogical content knowledge. Final exam scores and performance task grades were obtained from the course professors. Data for GPAs were collected from the students’ information sheet. Descriptive statistics of these variables are presented in Table 4.1
<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Course Grade</td>
<td>37</td>
<td>90.57</td>
<td>6.0</td>
</tr>
<tr>
<td>Portfolio Score</td>
<td>37</td>
<td>22.81</td>
<td>1.8</td>
</tr>
<tr>
<td>Grade for Tasks</td>
<td>37</td>
<td>49.94</td>
<td>3.9</td>
</tr>
<tr>
<td>Years of Teaching Experience</td>
<td>33</td>
<td>3.36</td>
<td>4.9</td>
</tr>
<tr>
<td>Final Exam Grade</td>
<td>37</td>
<td>90.70</td>
<td>10.7</td>
</tr>
<tr>
<td>GPA</td>
<td>26</td>
<td>3.53</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 4.1: Descriptive Statistics of Principal Variables. Possible total scores for final course grade=100; portfolio=25; tasks=55; exam=100; GPA=4.0.

Criterion validity can be investigated through correlational analyses to determine if there is a criterion-relatedness among the variables that would indicate they were measuring the same ability. A correlation is a “functional relationship between two measures” (Bachman, 1990:259) that indicates how separate measures tend to vary in the same way with respect to each other. This type of analysis does not indicate a causal relationship among variables, but it can indicate if there is a positive relationship. A Pearson-Product Moment correlation analysis was conducted to determine the relationship among the variables. The results from the analysis among the principal variables (portfolio scores, performance tasks, years of teaching experience, final exam scores, and overall GPA) are presented in Table 4.2.
<table>
<thead>
<tr>
<th></th>
<th>Exam</th>
<th>GPA</th>
<th>Years</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portfolio</td>
<td>.292</td>
<td>.296</td>
<td>.353*</td>
<td>.468*</td>
</tr>
<tr>
<td>Exam</td>
<td></td>
<td>.541*</td>
<td>.286</td>
<td>.164</td>
</tr>
<tr>
<td>GPA</td>
<td></td>
<td></td>
<td>.459*</td>
<td>.234</td>
</tr>
<tr>
<td>Years</td>
<td></td>
<td></td>
<td></td>
<td>.120</td>
</tr>
</tbody>
</table>

*p < .05

Table 4.2: Pearson Correlation Matrix of Principal Variables

Scatterplots can also be used to determine the strength of the relationship, including the linearity, between the variables. The scatterplots from the analysis of the independent variables and the dependent variables in this study reflect a similar relationship among the variables as the relationship demonstrated in the correlation matrix. The independent variable, portfolio scores, was plotted with each of the dependent variables, and the results indicate that there is some linearity among the variables, although the magnitude of the relationship is not high in all cases. The scatterplots between portfolio scores and each of the independent variables are shown in Figures 4.1 through 4.4.
Figure 4.1: Scatterplot of Performance Tasks and Portfolio Scores

Figure 4.2: Scatterplot of Teaching Experience and Portfolio Scores
Figure 4.3: Scatterplot of Final Exam Grades and Portfolio Scores

Figure 4.4: Scatterplot of Overall GPA and Portfolio Scores
In continuation, a multiple regression analysis, based on the correlation matrix, can indicate if there is a predictive value to the independent variables for the dependent variable. In this study, a significant F statistic in the regression analysis and significant Beta values would indicate that one or more of the independent variables was able to predict performance on the dependent variable, the portfolio. Stepwise regression was utilized to determine the degree to which each of the variables was able to predict performance on the portfolios. The results from the multiple regression analysis of the principal variables are presented in Tables 4.3 through 4.6.

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>B</th>
<th>Beta</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Tasks</td>
<td>0.22</td>
<td>0.07</td>
<td>3.13*</td>
</tr>
</tbody>
</table>

$R^2 = .22, \ F (1, 35) = 9.82, \ p < .0035$

* $p < .05$

Table 4.3: Regression Analysis of Portfolio Scores on Performance Tasks
<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>B</th>
<th>Beta</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Tasks</td>
<td>0.23</td>
<td>0.07</td>
<td>3.47*</td>
</tr>
<tr>
<td>Years of Teaching Experience</td>
<td>0.11</td>
<td>0.05</td>
<td>2.01</td>
</tr>
</tbody>
</table>

$R^2 = .38$, $F (2, 30) = 9.02$, $p < .0009$

*p < .05

Table 4.4: Regression Analysis of Portfolio Scores on Tasks and Years

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>B</th>
<th>Beta</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Tasks</td>
<td>0.21</td>
<td>0.06</td>
<td>3.24*</td>
</tr>
<tr>
<td>Years of Teaching Experience</td>
<td>0.08</td>
<td>0.05</td>
<td>1.51</td>
</tr>
<tr>
<td>Final Exam</td>
<td>0.05</td>
<td>0.02</td>
<td>1.91</td>
</tr>
</tbody>
</table>

$R^2 = .45$, $F (3, 29) = 7.75$, $p < .0006$

*p < .05

Table 4.5: Regression Analysis of Portfolio Scores on Tasks, Years, and Exam
<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>B</th>
<th>Beta</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Tasks</td>
<td>0.27</td>
<td>0.09</td>
<td>3.10*</td>
</tr>
<tr>
<td>Years of Teaching Experience</td>
<td>0.28</td>
<td>0.12</td>
<td>2.35*</td>
</tr>
<tr>
<td>Final Exam</td>
<td>0.05</td>
<td>0.03</td>
<td>1.73</td>
</tr>
<tr>
<td>GPA</td>
<td>-0.76</td>
<td>0.83</td>
<td>-0.91</td>
</tr>
</tbody>
</table>

\[ R^2 = .50, \ F (4, 21) = 5.34, p < .004 \]

*p < .05

Table 4.6: Multiple Regression Analysis of Portfolio Scores on All Principal Variables

The results of the full regression model demonstrate that 50% of the variance in portfolio scores is accounted for by the independent variables \( R^2 = .50 \). Entering the four independent variables (performance tasks, years of teaching experience, final exam grades, and GPA) resulted in a significant model: \( F (4, 21) = 5.34, p < .004 \). The best predictors of portfolio scores were the scores on performance tasks \( t = 3.10, p < .005 \) and years of teaching experience \( t = 2.35, p < .03 \). The other variables in the model were not significant in the prediction of portfolio scores but did contribute significantly to the model as a whole. Table 4.7 illustrates the change in the \( R^2 \) in the multiple regression analysis of the full model.
<table>
<thead>
<tr>
<th>Steps</th>
<th>$R^2$</th>
<th>Change in $R^2$</th>
<th>$F$-Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Performance Tasks</td>
<td>.22</td>
<td>.22</td>
<td>9.82*</td>
</tr>
<tr>
<td>2 Tasks, Years of Teaching</td>
<td>.38</td>
<td>.16</td>
<td>9.02*</td>
</tr>
<tr>
<td>3 Tasks, Years, Exam</td>
<td>.45</td>
<td>.07</td>
<td>7.75*</td>
</tr>
<tr>
<td>4 Tasks, Years, Exam, GPA</td>
<td>.50</td>
<td>.05</td>
<td>5.34*</td>
</tr>
</tbody>
</table>

* $p < .05$

Table 4.7: Multiple Regression Analysis of Full Model Showing Change in $R^2$

On the basis of the results presented in Table 4.7, it can be observed that the performance tasks account for the largest portion of the model (22% of the variance explained by the model). Including the other principal variables adds substantially to the model. The second largest portion of the model is accounted for by the years of teaching experience (16%). The other two variables, final exam grades and GPA, accounted for 7% and 5% of the variance respectively.

Construct validity was also a significant area for investigation in the establishment of portfolios as a valid measure of foreign language teacher performance. In this study, construct validity was established through a principal components factor analysis on the scores from the five principal variables. The theoretical basis for the portfolios used in the course was the literature regarding principles for effective teaching. Shulman's (1987) model for effective teaching identifies pedagogical content knowledge and knowledge of
content and pedagogy (conceptual knowledge) as important categories in teacher preparation programs. A two-factor solution and orthogonal rotation were specified because of theory, and because it was observed in the correlation analysis that the two factors (conceptual knowledge as measured by the final exam and GPA, and pedagogical content knowledge as measured by the portfolio, performance tasks, and years of teaching experience) were not correlated. The results yielded two factors with eigenvalues of 2.37 and 1.14. The two-factor solution explained 70% of the variance among the measures. The two factors reflected the same underlying constructs that had been originally specified; factor one, conceptual knowledge, included the final exam and overall GPA; whereas factor two, pedagogical content knowledge, included the portfolios and performance tasks. Years of teaching experience, however, loaded on factor one rather than factor two as was originally proposed. Table 4.8 presents the results of the factor analysis.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor One (Conceptual Knowledge)</th>
<th>Factor Two (Pedagogical Content Knowledge)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Years of Experience</td>
<td>82*</td>
<td>-8</td>
</tr>
<tr>
<td>2 Overall GPA</td>
<td>80*</td>
<td>18</td>
</tr>
<tr>
<td>3 Final Exam</td>
<td>68*</td>
<td>35</td>
</tr>
<tr>
<td>4 Performance Tasks</td>
<td>-5</td>
<td>93*</td>
</tr>
<tr>
<td>5 Portfolios</td>
<td>43</td>
<td>74*</td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>2.37</td>
<td>1.14</td>
</tr>
<tr>
<td>Percent of Variance</td>
<td>47%</td>
<td>23%</td>
</tr>
</tbody>
</table>

Table 4.8: Factor Loadings and Percent Variance Explained for Principal Variables

In light of the results presented in response to Research Question One, the corresponding null hypotheses relative to the investigation of validity are discussed here.

**H01a:** There is no relationship between portfolio scores and the other performance measures used to evaluate students in the course.

Portfolios accounted for 25% of the final course grade; other similar performance measures of the ability to apply pedagogical knowledge accounted for another 55% of the final grade. These performance measures included unit and daily lesson plans, peer teaching, and a student journal of personal reactions to course readings and material.
The results from the correlation analysis indicate that portfolio ratings correlate moderately well with the scores for the performance tasks ($r=.47$, $p<0.004$). The scatterplot shows a similar trend in the data: a moderate, positive correlation between portfolio scores and performance tasks for the low to average portfolio scores and performance task scores. In the regression analysis, performance tasks were the best predictor of scores on the portfolio accounting for 22% of the variance in the model. The factor analysis also confirmed that the two variables were measuring the same construct; both portfolios and performance tasks loaded on the same factor. The null hypothesis is therefore rejected; there was a positive relationship between portfolio scores and the other performance measures used in the course.

**H01b: There is no relationship between reported years of teaching experience and portfolio scores.**

Based on Naizer’s (1993) study, a moderate correlation was expected between years of teaching experience as a measure of practical application of content knowledge and portfolio scores as a measure of the ability to apply content knowledge. The correlation coefficient did prove to be significant, albeit moderate ($r=.35$, $p<.04$). The scatterplot also shows a moderate linearity between the two variables. The results from the regression analysis indicated a positive relationship between years of teaching experience and portfolio scores. Years of teaching experience accounted for 16% of the variance in the full model indicating that the number of years of experience is a good prediction of scores on the portfolio. The factor analysis did not confirm the results of the correlation and regression analysis; years of teaching experience loaded on the same factor as the final exam and GPA (conceptual knowledge). Although the factor loadings do confound the definition of the two constructs, they do not detract from the underlying, positive relationship between portfolio scores and teaching experience. The null hypothesis is therefore rejected; there
was a positive, but moderate, relationship between years of teaching experience and the portfolio scores.

**HO1c:** There is no relationship between portfolio scores and subjects' grades on the course final exam.

The final exam accounted for the remaining 20% of the course grade and required students to write a ten-page synthesis of the course readings. A significant correlation coefficient was not achieved between portfolio scores and final exam grades ($r = .29$, $p < 0.08$). The scatterplot reflected a similar trend; there was no linearity between the variables thus suggesting a lack of correlation between them. The regression analysis reflected the same conclusion; the final exam, while significant to the full model, only accounted for 7% of the variance and is not a good predictor of portfolio scores. The factor analysis also confirmed these results; the final exam loaded on a different factor from the portfolios thus indicating that they are each measuring a separate construct. This hypothesis is therefore retained; there is no relationship between portfolio scores and grades on the final exam.

**HO1d:** There is no relationship between portfolio scores and subjects' overall GPA.

This hypothesis is retained based on the results of the correlation, regression, and factor analyses. There is not a significant correlation ($r = .29$, $p < .14$) between portfolio scores and overall GPA. The scatterplot and the results of the regression analysis reflect the same relationship between the scores on the portfolio and GPA. Although the inclusion of GPA in the full regression model enhances the $F$ statistic, the change in the $R^2$ only accounts for 5% of the variance. Both the $B$ and the $t$-test for this variable are negative indicating an inverse correlation between the two factors (portfolio scores and GPA) which
may be due to bias in the variable. The factor analysis results also indicate that GPA is measuring a different construct from portfolios because they load on separate factors.

Although not planned for investigation in the initial research proposal, the relationship between GPA and the final course exam proved to be positive; the final exam grade correlates well with the overall GPA ($r=.54$, $p<0.004$). This relationship may indicate that the final exam is more a measure of achievement, as is the GPA, than a measure of student progress and development. A scatterplot (See Figure 4.5 below) between GPA and final exam grades also shows a linear relationship between the two variables. The factor analysis confirmed that the underlying construct (conceptual knowledge) measured by the exam and GPA was different from the construct (pedagogical content knowledge) measured by the portfolios and performance tasks.

![Figure 4.5: Scatterplot of Final Exam Grade and Overall GPA](image-url)

90
In conclusion, validity issues were investigated through means of correlation, multiple regression, and principal components factor analyses. The results demonstrated that there were significant, positive relationships between portfolio scores and performance tasks, and between portfolio scores and years of teaching experience in the correlation and regression analyses. Positive relationships were not found between portfolio scores and final exam grades, nor between portfolio scores and GPA. The factor analysis confirmed the existence of two underlying constructs among the variables: conceptual knowledge and pedagogical content knowledge. An interesting result was the positive relationship between final exam grades and GPA as indicated by all three analyses.

Part Two: Issues of Reliability

Issues of reliability have been more commonly investigated in the recent transition to alternative assessment tools. The need for consistent and accurate results from assessment measures and from raters requires the establishment of the reliability of both the raters and the scores. The data for this part of the study were collected from the stratified, purposeful sample of ten students whose portfolios were assessed by five raters. Reliability was investigated in several forms in this study. Inter-rater agreement was established through a Pearson correlation matrix and Fisher transformation table using the holistic scores. A Rasch Model FACETS analysis using the analytical scores was conducted to investigate rater consistency, judge severity, and score reliability. A generalizability study was also conducted to determine the impact of factors other than teaching competence (i.e., persons, raters, and tasks). Two types of rating scales, a holistic and an analytical, were used in this study. A t-test analysis was conducted to determine if there were significant differences in the ratings among portfolios using either of the two scales. The following section answers Research Question Two in the investigation of issues of reliability.
Research Question Two: To what extent is portfolio assessment a reliable measure of teacher competence as determined by: (a) the consistency with which raters agree in their ratings of each portfolio; (b) the relative impact of sampling due to persons, raters, and tasks in evaluating preservice teachers’ competence; and (c) the consistency with which ratings assigned to the overall portfolio using the holistic rating scale correlate with the ratings assigned to individual pieces within the portfolios using the analytical rating scale.

To answer the first part of this question concerning the rater consistency, data were collected from the five raters’ judgments of the portfolios. Each rater was asked to rate ten portfolios using both the analytical scale and the holistic scale. Examples of these scales are found in Appendices D and E. An inter-rater reliability coefficient was calculated using a Pearson-Product Moment correlation matrix of rater agreement on the holistic scoring, and the Fisher transformation table as recommended in Hatch and Lazaraton (1991), and was found to be $r=0.65$. This inter-rater or scorer reliability coefficient reflects a similar level of inter-rater correlation to that achieved by LeMahieu et al. (1993) in the original study of Pittsburgh Public Schools’ use of portfolios. Descriptive statistics for the holistic ratings given by each of the five raters are presented in Table 4.9. The data from the holistic ratings of the portfolios show that Rater Five is the most severe judge and Rater Three is the most lenient judge. The correlational matrix of rater agreement using the holistic rating scale is presented in Table 4.10.
<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>10</td>
<td>20.65</td>
<td>3.92</td>
</tr>
<tr>
<td>Rater 2</td>
<td>10</td>
<td>21.20</td>
<td>1.93</td>
</tr>
<tr>
<td>Rater 3</td>
<td>10</td>
<td>22.25</td>
<td>1.65</td>
</tr>
<tr>
<td>Rater 4</td>
<td>10</td>
<td>21.83</td>
<td>3.03</td>
</tr>
<tr>
<td>Rater 5</td>
<td>10</td>
<td>20.05</td>
<td>2.73</td>
</tr>
</tbody>
</table>

Table 4.9: Descriptive Statistics of Holistic Ratings of Portfolios. Possible total score was 25.

<table>
<thead>
<tr>
<th></th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>.34</td>
<td>.57</td>
<td>.54</td>
<td>.65*</td>
</tr>
<tr>
<td>Rater 2</td>
<td>.50</td>
<td>.31</td>
<td>.64*</td>
<td>.63*</td>
</tr>
<tr>
<td>Rater 3</td>
<td></td>
<td></td>
<td>.82*</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05

Table 4.10: Pearson Correlation Matrix of Holistic Ratings of Portfolios

Using the data from the analytical ratings of the portfolios, a Rasch model FACETS analysis was conducted to determine the consistency with which the judges used the rating scales to evaluate the portfolios. The Rasch model allowed the portfolio scores to be
reported free from judge severity in the ratings, from the difficulty of the portfolio submission, and from the random nature of the rating scale categories. The output from a FACETS analysis provides a severity rating for each judge that illustrates the rater's tendency to score lower or higher that is warranted by student performance (Saal, Downey, & Lahey, 1980). It also provides a reliability coefficient that indicates the consistency with which the raters used the scales to judge the portfolios, as well as a score reliability based on the aggregate ratings of the portfolios by all of the raters. Another useful statistic that is provided by the FACETS analysis is the reliability of separation index which gives information on how well the elements in each facet define that facet. This index is similar to traditional indices of reliability in that it reflects the ratio of the true score variance to the observed score variance (Englehard, Jr., 1994).

According to Wright and Linacre (1994), “the ideal for measurement construction is that the data fit the Rasch model” though they admit that all data will vary to some extent. The chi-square fit statistics provided by the FACETS analysis indicate the extent to which the data fit the model, and give an indication of the validity of the mean square fit statistics for each of the facets in the analysis. The global fit statistics from the three-facet Rasch model in this study indicate that the data fit the model well (log-likelihood chi-square=610.1, d.f.=278, p < .00). The residual chi-square was not significant (299.4, d.f.=278, p < .18) also showing a good fit of the data to the model. Overall calibrations of raters, submissions, and students using the data from the analytical ratings of the portfolios are presented graphically in Figure 4.6.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Raters</th>
<th>Students</th>
<th>Submissions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Severe</td>
<td>More Able</td>
</tr>
<tr>
<td>2.0</td>
<td></td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0.0</td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>-1.0</td>
<td>Rater 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rater 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rater 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rater 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2.0</td>
<td>Rater 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.6: Calibrations of Rater, Student, and Submission Facets on Logistic Scale
As can be observed from Figure 4.6, the portfolio scores given using the analytical rating scale were slightly different from those scores assigned when the judges applied the holistic rating scale. Rater Four, for example, was the most lenient judge when using the analytical rating scale, but Rater Three was the most lenient judge when using the holistic rating scale. Rater Five continued to be the judge that rated the most severely regardless of the type of rating scale being applied. The two rating scales were employed in this study because of the type of analyses that were performed. The holistic rating scale is the optimal method with which to judge portfolios; analytical scores, however, were necessary for the FACETS analysis and the generalizability study.

Table 4.11 presents the rating scale frequencies of each judge from the analytical scoring of the portfolios. From the data, it can be observed that the judges tended to be lenient in their ratings of the portfolio submissions. The analytical rating scale required the raters to assign a score from one to six, with six being the highest, to each submission in the portfolios. All five raters judged each of the six submissions in the ten portfolios; in other words, each rater gave 60 scores. The majority of the scores were in the intervals of four, five, and six; few submissions were rated below a four. Rater Four had the highest mean score and was considered the most lenient judge whereas Rater Five had the lowest mean score and was therefore considered the most severe judge. The other three raters were consistent in their scoring of the portfolios.
### Table 4.11: Rater Frequency of Analytical Rating Scale Use and Mean Scores Assigned to Portfolios

<table>
<thead>
<tr>
<th>Rater</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
<th>Mean Score</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>13</td>
<td>16</td>
<td>26</td>
<td>60</td>
<td>5.1</td>
<td>0.08</td>
</tr>
<tr>
<td>Two</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>34</td>
<td>13</td>
<td>60</td>
<td>5.0</td>
<td>0.02</td>
</tr>
<tr>
<td>Three</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>29</td>
<td>18</td>
<td>60</td>
<td>5.1</td>
<td>0.08</td>
</tr>
<tr>
<td>Four</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>15</td>
<td>34</td>
<td>60</td>
<td>5.3</td>
<td>0.28</td>
</tr>
<tr>
<td>Five</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>16</td>
<td>37</td>
<td>2</td>
<td>60</td>
<td>4.6</td>
<td>0.42</td>
</tr>
</tbody>
</table>

FACETS provides separate analyses of fit statistics for each of the facets entered in the model. The raters measurement report provides data on the judge severity measure (in logits) that indicates the leniency or severity with which the raters applied the analytical rating scale. The student measurement report presents the ability of the students after accounting for judge severity and provides a score reliability coefficient. The submissions measurement report indicates the difficulty of the tasks and the reliability of the ratings of each submission.

The first report presented (Table 4.12) is the rater measurement report. This report gives a judge severity measure in logits that provides information about how each judge applied the rating scales. Positive measures indicate judge severity, whereas negative measures mean judge leniency. The measures in this study were all negative because of the overall leniency of these raters in their assessment of the portfolio submissions. The average rating should have been about 3.0 if the raters had used the full range of the scale.
These judges, however, generally used the higher end of the rating scale in their assessment of the portfolios (See Table 4.11), and the severity measure reflects these higher ratings. The data indicated that Rater Four was the most lenient in her rating of the portfolios; Rater Five was the most severe.

The mean-square infit and outfit statistics, based on the ratio of observed error variance to modeled error variance, are reported below. Infit statistics are an information-weighted version of outfit statistics. Thus, while high infit values reflect noise or unexpected performance near a person’s ability level, high outfit values reflect noise at the extreme levels. Wright and Linacre (1994) claim that reasonable mean square ranges for infit and outfit statistics depend upon the type of test being administered. They recommend a range of 0.4 to 1.2 when there is an instrument that is being judged by raters and agreement is encouraged. Conventional statisticians prefer a range of acceptable values between 0.8 and 1.2. The expected mean-square value for a fit statistic is 1.0, and variance from that indicates noise or ambiguity in the data. The acceptable values for rater agreement in this study were set in the range of 0.8 to 1.2 to provide a stricter limit on what was considered noisy (mean-square > 1.2) or muted (mean-square < 0.8) data.

Descriptive statistics for the rater facet are presented in Table 4.12. The rater severities range from -2.18 logits (SE=.18) for Rater Four who is lenient to -0.87 (SE=.17) for Rater Five who is the most severe. The overall differences between the raters are significant chi-square (4)=28.3, p < .00 with a high reliability of separation index (R=.81). Infit and outfit mean squares are presented with the standardized values to indicate the significance levels of these statistics. The modeled standard errors also presented in the table indicate the precision or reliability of the severity measures (Linacre, 1993). The size of these errors is primarily determined by the number of ratings a judge has made; the more ratings a judge makes, the smaller the standard error.
<table>
<thead>
<tr>
<th>Rater</th>
<th>Examinees Rated</th>
<th>Severity Measure</th>
<th>Standard Error</th>
<th>Infit Statistics MnSq</th>
<th>ZStd</th>
<th>Outfit Statistics MnSq</th>
<th>ZStd</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>10</td>
<td>-1.70</td>
<td>.18</td>
<td>1.2 1</td>
<td></td>
<td>1.2 1</td>
<td></td>
</tr>
<tr>
<td>Two</td>
<td>10</td>
<td>-1.57</td>
<td>.18</td>
<td>0.8 -1</td>
<td></td>
<td>0.8 -1</td>
<td></td>
</tr>
<tr>
<td>Three</td>
<td>10</td>
<td>-1.77</td>
<td>.19</td>
<td>1.1 0</td>
<td></td>
<td>1.1 0</td>
<td></td>
</tr>
<tr>
<td>Four</td>
<td>10</td>
<td>-2.18</td>
<td>.20</td>
<td>1.8 3*</td>
<td></td>
<td>1.5 2*</td>
<td></td>
</tr>
<tr>
<td>Five</td>
<td>10</td>
<td>-0.87</td>
<td>.17</td>
<td>0.4 -4*</td>
<td></td>
<td>0.4 -4*</td>
<td></td>
</tr>
</tbody>
</table>

Note: Asterisks indicate misfitting raters with mean square errors less than 0.8 and greater than 1.2.

Table 4.12: Rasch Model Rater Measures and Fit Statistics for Portfolio Scores

In Table 4.12, Raters One, Two, and Three have mean-square infit/outfit statistics that are close to the expected value of 1.0. Raters Four and Five, however, demonstrate considerable misfit. The mean-square outfit of Rater Four (1.5) indicates this judge is giving more extreme scores than the other judges. This rater gave more ratings of six than expected considering the scores given by the other judges. The same rater was also more variable at the expected score level of each than the other judges thus accounting for the considerable misfit in the infit mean square (1.8). Rater Five shows a different trend; the mean-square infit and outfit statistics of 0.4 indicate that this judge is not using the rating scale to its fullest and is relying on average scores. This rater shows a preference for the two high central categories (4 and 5) and in essence has reduced the scale to a dichotomy. The ratings by this rater are problematic because they are unusually muted. Rater Five, therefore, is showing an unusual consistency in her use of the rating scale, but is not
supplying as much independent information about the portfolios as those raters that use
more of the scale. Predictable ratings such as those given by Rater Five do not provide as
much information regarding differences among students as do the ratings given by judges
who use more of the scale.

The students measurement report discloses information about the examinees;
descriptive statistics for the student facet are presented in Table 4.13. The student ability or
performance measure had a range of 1.17 logits (SE=.32) for Student Five who is the most
able to -1.78 logits (SE=.22) for Student Ten who is the least able. The overall differences
between the students are significant, chi-square(9)=104.5, p < .00 with a high reliability of
separation index (R=.90). The average scores presented in Table 4.13 represent the actual
composite ratings given by the judges (observed) and the scores that the students would
have received had the judges rated the portfolios in a less lenient manner (fair). The fair
average, in other words, is a true representation of the students' ability or performance after
accounting for the judge severity measure. The mean observed score in this study was
M=5.0 (SD=0.4) and the mean fair score was M=4.1 (SD=0.5). The ability measure
indicates the teaching competence of the students as based on portfolio scores. The Rasch
estimates of teaching competence ranged from -1.78 to 1.17 (M=.00, SD=.83).
<table>
<thead>
<tr>
<th>Student</th>
<th>Avg. Score</th>
<th>Ability Measure</th>
<th>Standard Error</th>
<th>Infit MnSq.</th>
<th>ZStd.</th>
<th>Outfit MnSq.</th>
<th>ZStd.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.1</td>
<td>4.1</td>
<td>.05</td>
<td>.26</td>
<td>1.1</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>5.3</td>
<td>4.4</td>
<td>.48</td>
<td>.27</td>
<td>0.7</td>
<td>-1*</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>4.9</td>
<td>3.9</td>
<td>-.33</td>
<td>.25</td>
<td>0.6</td>
<td>-2*</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td>4.9</td>
<td>3.9</td>
<td>-.27</td>
<td>.25</td>
<td>1.1</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>5.5</td>
<td>4.8</td>
<td>-1.17</td>
<td>.32</td>
<td>0.5</td>
<td>-2*</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>5.5</td>
<td>4.7</td>
<td>1.07</td>
<td>.31</td>
<td>1.1</td>
<td>0</td>
<td>1.2</td>
</tr>
<tr>
<td>7</td>
<td>4.7</td>
<td>3.6</td>
<td>-.68</td>
<td>.24</td>
<td>0.9</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td>8</td>
<td>4.9</td>
<td>3.9</td>
<td>-.27</td>
<td>.25</td>
<td>1.1</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>9</td>
<td>5.3</td>
<td>4.4</td>
<td>.55</td>
<td>.28</td>
<td>1.0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>4.0</td>
<td>2.9</td>
<td>-1.78</td>
<td>.22</td>
<td>1.7</td>
<td>2*</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Note: Asterisks indicate students with mean square errors of less than .8 and greater than 1.2.

Table 4.13: Rasch Model Student Measures and Fit Statistics for Portfolio Scores

The results from the student measurement report showed that there were several misfitting students. Students Two, Three, and Five had a muted rating indicating an unusually consistent rating on their portfolios. Student Ten had a noisy infit and outfit ratings which indicate significant variations at all levels of her performance. On recommendation from Engelhard, Jr. (1994), misfitting scores were examined more in detail to determine if rescoreing was necessary. It was concluded that the ratings given by the judges using the analytical scale were similar to those that were given by the instructor.
researcher using the holistic scale. Student Five, for example, received a score of 24 of a possible 25 points using the holistic rating scale; Student Ten received a score of 21.

The final report presented in the FACETS analysis is the submissions measurement report, which provides information about the difficulty level of the submissions. In this study, the submissions were randomly organized in the students’ portfolios, therefore no conclusive findings about one activity or task being harder than any other can be made. The descriptive statistics do indicate that the overall differences for the submissions facet are significant, chi-square (5) = 12.5, p < .03 with a moderate reliability of separation index (R=.51). The difficulties of the submissions range from .42 (SE=.19) for Submission One, which is the most difficult to -.33 (SE=.21) for Submission Four, which appears to be the easiest. All of the fit statistics are between 0.8 and 1.2 which indicates that the data fit the model well. An interesting finding, however, is that the first submission was rated as the most difficult. This may indicate that the raters began judging the portfolios more severely at the beginning, and then became more lenient toward the middle of each portfolio, which would also explain why Submission Four appears to be the easiest item. Table 4.14 presents the data from the submissions report.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.42</td>
<td>.19</td>
<td>1.1 0</td>
<td>1.1 0</td>
</tr>
<tr>
<td>2</td>
<td>.12</td>
<td>.20</td>
<td>1.0 0</td>
<td>1.0 0</td>
</tr>
<tr>
<td>3</td>
<td>.28</td>
<td>.19</td>
<td>.9 0</td>
<td>.8 0</td>
</tr>
<tr>
<td>1</td>
<td>-.33</td>
<td>.21</td>
<td>1.2 0</td>
<td>1.2 0</td>
</tr>
<tr>
<td>1</td>
<td>-.25</td>
<td>.21</td>
<td>1.0 0</td>
<td>1.0 0</td>
</tr>
<tr>
<td>1</td>
<td>-.25</td>
<td>.21</td>
<td>.9 0</td>
<td>.8 0</td>
</tr>
</tbody>
</table>

Table 4.14: Rasch Model Submission Measures and Fit Statistics for Portfolio Scores

The data presented from the Rasch model FACETS analysis provided information concerning the rater reliability, score reliability, and submission difficulty. These results will be discussed relative to the null hypotheses after the presentation of the data from the generalizability study.

For the final analysis of reliability, a generalizability study using the data from the analytical scale ratings was conducted to determine the relative impact of raters, persons, tasks, and other factors on the portfolio scores. Generalizability theory provides researchers with information concerning the impact of various factors on the ability being measured. It does this by separating the reported variance into components for each of the facets entered in the design. In this study, the facets were portfolio scores, items (tasks), and raters. The desired variance is among the portfolio scores which are a measure of the
teaching competence of the subjects in the study. Variance due to item difficulty or rater severity is not desired but can be separated out by a generalizability study, as it was in the FACETS analysis. To answer this question, therefore, the data collected from the raters' analytical scoring of the portfolios was used in a P x I x R random design, where P represents the portfolios or persons (N=10) and is crossed with I representing items or submissions in the portfolio (N=6 per portfolio for a total of 60 items) and R representing the raters (N=5). This design was chosen based on the rating of the portfolios; each of the five raters scored all six submissions in each of ten portfolios. Variance due to persons, items, raters, and their interactions were calculated. The results from the G-study are presented in Table 4.15.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>df</th>
<th>Mean Square</th>
<th>Estimated Variance Component</th>
<th>Percentage of Total Variance$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons (p)</td>
<td>9</td>
<td>6.030</td>
<td>0.1397</td>
<td>17</td>
</tr>
<tr>
<td>Raters (r)</td>
<td>4</td>
<td>3.213</td>
<td>0.0229</td>
<td>1</td>
</tr>
<tr>
<td>Items (i)</td>
<td>5</td>
<td>1.819</td>
<td>0.0191</td>
<td>1</td>
</tr>
<tr>
<td>pr</td>
<td>36</td>
<td>1.837</td>
<td>0.3062</td>
<td>36</td>
</tr>
<tr>
<td>pi</td>
<td>45</td>
<td>0.752</td>
<td>0.1005</td>
<td>12</td>
</tr>
<tr>
<td>ri</td>
<td>20</td>
<td>0.359</td>
<td>0.0109</td>
<td>1</td>
</tr>
<tr>
<td>pri,c (residual)</td>
<td>180</td>
<td>0.250</td>
<td>0.2500</td>
<td>30</td>
</tr>
</tbody>
</table>

$^1$The percentages do not sum up exactly to 100 due to rounding.

Table 4.15: Estimated Variance Components for Portfolio Scores
The G-study results indicate that the rater component accounted for very little of the variance (1% of the total variance) thus providing evidence for rater agreement and consistency. The variance for persons (17% of total) indicates that the subjects differed in their portfolio scores and consequently in their demonstration of level of teaching competence. The variance component for the interaction between persons and raters, however, is quite large (36% of the total), which reduces to some extent the reliability of the rater variance component. The raters were consistent overall in their ratings of the portfolios but disagreed in the relative standing of persons on the scale. This difference may have been indicative of the subjectivity in rating portfolios. The individuality of a portfolio and the personality of its owner are inherently linked to the performance on the portfolio. When combined with a rater’s personality and preference for style or appearance, the ratings become more subjective. The greater variance caused by the interaction between raters and persons may have been due to this subjectivity. The combination of the rater component and the rater by person component, therefore, suggest that some of the variability in the portfolio scores is in fact due to the raters.

The variance due to items is also very small (1% of the total variance) but the fact that this component is greater than zero indicates that the tasks varied somewhat in the level of difficulty (Brennan & Johnson, 1995). The person by item variance component is higher (12%) relative to the other components and reveals that the task component contributes to score variability to some extent (Brennan & Johnson, 1995). The person by item interaction also reflects that the subjects differed in their relative standing from one item to another. This may be explained by subjectivity in the ratings of certain tasks in comparison to other tasks; judges may have rated tasks differently based on the type of submission versus the student’s performance on the task. The rater by item component (1% of total) is also small indicating that raters used the same part of the scale when evaluating items, and
did so consistently across items. Finally, the variance component for the residual (30% of the total) reflects a large portion of the variance that was due to the “three-way interaction between persons, raters, and items and/or other unsystematic or systematic sources of variation that were not measured in this study” (Shavelson & Webb, 1991:35).

In light of the data presented from the correlational analysis, the FACETS analysis, and the generalizability study, Research Hypothesis Two is discussed here.

**H02a** *There is no agreement among raters in the portfolio scores assigned.*

Based on the inter-rater reliability coefficient that was derived through use of the Pearson correlation matrix and the Fisher transformation table (r=.65), the rater consistency results from the FACETS output (r=.81), and the variance due to the rater component in the G-study (raters=1% of the total variance) this hypothesis is rejected; the raters were both consistent and reasonable in their use of the rating scales. The results from the FACETS analysis indicated that three of the raters were consistent in their ratings of the portfolios. Raters One, Two, and Three demonstrated use of more than two categories on the scale of one to six, although the tendency was to score the portfolios in a lenient fashion assigning primarily scores of four, five, and six to the submissions. The mean-square outfit statistics demonstrated that these raters were consistent in their application of the rating scale. Rater Four, however, showed considerable misfit in her scoring which reduced the scorer reliability coefficient. Rater Five was consistent in her ratings but reduced the scale to a dichotomy by assigning scores of nearly all four and five.

**H02b** *There is no impact on portfolio scores due to raters, persons, or tasks in evaluating preservice teachers’ competency.*

The null hypothesis is rejected based on the results from both the G-study and the Rasch model analysis. As individual components, rater variance and item variance in the
generalizability study amounted to a very small percentage of the total variance (less than 1% each), and the variance due to persons was moderate (17%). The interactions between the components accounted for more of the variance, however, and thus affected the reliability of the scores overall. The rater by person component accounted for 36% of the variance; when added to the rater component, this amount (37%) indicates that the raters did have an impact on the portfolio scores. The persons by items interaction also showed that the tasks had an impact on the portfolio scores. Although the scores assigned to the portfolios by the judges were demonstrated to be consistent among the raters, the impact of the raters, items, and persons on the portfolio scores caused variability among the relative standing of the subjects. This hypothesis, therefore, is rejected; whereas the individual components had little effect on the portfolio scores, the interactions among the components caused score variability. The results of the FACETS analysis, however, indicate that the scorer (inter-rater) reliability is 0.81 and the score reliability is 0.91 which demonstrates that the portfolios are a reliable measure of preservice teachers' competency.

**HO2c**  
*There is no relationship among the grades assigned to individual pieces within the portfolio using the analytical rating scale and the overall grade assigned to the portfolio using the holistic rating scale.*

This hypothesis was examined through a t-test analysis of the total scores for each rater using the holistic scale and the total scores using the analytical scale. Two rating scales were necessary in this study because of the nature of the analyses that were completed. On first observation, it appeared that the scores assigned using the holistic scale were generally higher (Mean score=84%) than those assigned using the analytical scale (Mean score=82%). Rater Three, for example, was the most lenient judge when using the holistic rating scale (Mean=22.25 for portfolio score), but Rater Four was the most lenient judge when applying the analytical scale (Mean=5.3 for submission score). A t-test procedure run
on SAS System to determine if there was a significant difference between holistic portfolio scores and analytical portfolio scores, however, was not significant (F=1.84, p < .57). This hypothesis is retained; there is no significant difference between the analytical ratings and the holistic ratings of the portfolios.

Part Three: Preservice Teachers' Attitudes

In the investigation of the students' attitude toward, and understanding of portfolio assessment, the following questions were examined: (a) what beliefs about assessment and familiarity with assessment tools do the preservice teachers report prior to the methods course?; (b) what beliefs about assessment and familiarity with the tools do the students report following the experience with portfolio assessment in the methods course?; (c) to what extent is there a difference between the reported beliefs and familiarity with assessment tools on the pre-course survey and the post-course survey?; (d) do the students indicate an intention to use portfolio assessment in future classes?

Student notions and understanding of assessment tools were explored using a survey instrument. Data were collected from the students regarding their beliefs about and understanding of assessment tools at the beginning of the course and at the end of the course to determine if there was a change in the students' understanding of assessment, especially portfolios and rating scales. The results of the pre- and post-course survey are reported in Table 4.16.
<table>
<thead>
<tr>
<th></th>
<th>Pre-course</th>
<th>Post-course</th>
<th></th>
<th>Pre-course</th>
<th>Post-course</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Item 1</td>
<td>3.89</td>
<td>1.26</td>
<td>3.97</td>
<td>1.06</td>
<td>Item 1</td>
</tr>
<tr>
<td>Item 2</td>
<td>4.70</td>
<td>0.62</td>
<td>4.82</td>
<td>0.39</td>
<td>Item 2</td>
</tr>
<tr>
<td>Item 3</td>
<td>2.69</td>
<td>0.98</td>
<td>3.03</td>
<td>1.21</td>
<td>Item 3</td>
</tr>
<tr>
<td>Item 4</td>
<td>4.43</td>
<td>1.04</td>
<td>4.67</td>
<td>0.60</td>
<td>Item 4</td>
</tr>
<tr>
<td>Item 5</td>
<td>3.38</td>
<td>1.30</td>
<td>3.61</td>
<td>1.12</td>
<td>Item 5</td>
</tr>
<tr>
<td>Item 6</td>
<td>4.76</td>
<td>0.43</td>
<td>4.80</td>
<td>0.42</td>
<td>Item 6</td>
</tr>
<tr>
<td>Item 7</td>
<td>2.16</td>
<td>1.12</td>
<td>2.75</td>
<td>1.37</td>
<td>Item 7</td>
</tr>
<tr>
<td>Item 8</td>
<td>4.30</td>
<td>0.84</td>
<td>4.60</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Item 9</td>
<td>4.03</td>
<td>0.91</td>
<td>4.12</td>
<td>1.02</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.16: Mean Ratings of Items on the Assessment Survey. Pre-course N=37  Post-course N=33

The mean scores of the survey results were determined by averaging the ratings given by all of the students on each item. There is a slight increase in the means between the pre- and post-course survey for the items in Part I of the survey, which asked students to report their beliefs about the purposes of assessment. There is more of a difference in the means between the pre-course and post-course group means for Part II, which asked students to report their understanding of assessment tools. Especially interesting are the
mean differences for items five, six, and seven of Part II which asked students to rate their understanding of student self-assessment reports, portfolio assessment, and rating scales. The group mean for each of these questions rose at least one interval on the rating scale in terms of student understanding. These results support the belief that students will better understand alternative methods of assessment if they are evaluated using the tool itself, rather than simply reading about it. The use of portfolios in the course may also have impacted students' willingness to use portfolios with future students.

The question at the end of the survey: 'Would you be willing to use portfolios with students in the future?' received positive replies from 30 of the 33 students that responded to the survey. Two students did not answer the question; one student said she would not use portfolios due to the overwhelming amount of time involved. Reasons reported by the students for using portfolios included the following: (1) portfolios allow for individuality and creativity; (2) learning style differences can be expressed; (3) portfolios show student progress and development; (4) portfolios give an overall perspective of student ability; (5) portfolios allow students to participate in assessment; and (6) it is fun to create a portfolio. The only concern that many of the students voiced was that portfolio assessment is time-consuming, especially in terms of grading and evaluation.

Further investigation of the beliefs that students hold regarding portfolio assessment in the evaluation of teaching competency was conducted through a focus-group interview. Three subjects from the stratified sample of ten were chosen for the interview. The students were asked to reflect on the overall experience of using portfolios in the course. All of them commented that it was a positive experience, but that it would have been more worthwhile and useful to have completed a professional portfolio as an exit instrument from the teacher education program rather than for the purposes of one course. The students all agreed that suggested activities, guidelines for submissions, and criteria for evaluation were a necessary component in the implementation of the portfolio in the course. They expressed a
concern for students that would not know what was expected of them without the
guidelines. All three agreed that the portfolio was useful in demonstrating the ability to
perform in the classroom. They commented that the final, written exam was helpful in
synthesizing the information transmitted in the course but that it was the performance tasks
and the portfolio that allowed them to demonstrate the ability to put theory into practice.
The students also commented that the portfolio afforded them the opportunity to integrate
content knowledge of a language into pedagogical knowledge, thus demonstrating their
pedagogical content knowledge. The exam, on the other hand, they thought was generic to
language content.

Overall, the preservice teachers that were interviewed held the opinion that portfolio
assessment was a positive manner of evaluating students, but they expressed the same
concerns about portfolios as are cited in the literature: time and cost involved and
subjectivity of scoring procedures. They indicated an interest in implementing portfolios in
their future classrooms, but indicated that it would require dedication on their part to
maintain portfolios for all of the students in their classes. In terms of using portfolios for
evaluating future teachers, all three preservice teachers indicated that portfolios allowed a
more holistic view of a person’s competence in language and pedagogy as well as ability to
perform in the classroom context.
Discussion

Validity

One of the purposes of this study was to examine issues of validity in using portfolios to assess preservice teachers' pedagogical competence. The results from a Pearson correlational analysis, a multiple regression analysis, and a principal components analysis suggest that portfolios are measuring teaching competence in a manner that is qualitatively different from traditional tests and measures of achievement. The results from the correlational analysis indicate a moderate correlation between portfolio scores and performance tasks completed by the students in the course ($r= .47$). The factor analysis results also indicated that the portfolios and the performance tasks measured the same underlying construct: pedagogical content knowledge. There was a moderate to low correlation between years of teaching experience and portfolio scores ($r= .35$). It was expected that a stronger correlation would have been found between years of teaching experience and portfolio scores because of the performance nature of the portfolio submissions. As suggested by Shulman (1987) in conducting “wisdom-of-practice” studies, the time spent in the classroom is what fosters pedagogical content knowledge. Many experienced teachers cannot describe what they do in terms of theory, but are able to perform effectively in the classroom and transmit knowledge to the students in a learnable format. The factor analysis, however, indicated that the years of teaching experience was not measuring the same construct as the portfolios and performance tasks. It is possible that because of the number of subjects in the study that were true preservice teachers (i.e., zero years of teaching experience) that this variable was biased toward those subjects with some years of teaching experience. The correlation between portfolio scores and years of experience did not suggest that those subjects with many years of experience performed significantly better on the portfolio that those students with little or no experience.
The correlations between the portfolio scores and measures of achievement in the course (final exam grades) and overall measures of achievement (GPA) were relatively low ($r = .30$ and $r = .29$ respectively). The factor analysis yielded similar results; the final exam and GPA loaded on a different factor from the portfolios and performance tasks, thus suggesting that the variables were measuring different constructs. The correlation between final exam grades and GPA was higher ($r = .54$) indicating a relationship between those variables. This may be due to the fact that the final exam and the GPA are essentially measuring the same thing: achievement in course content. The factor analysis confirmed this; both the final exam and GPA loaded on the same factor: conceptual knowledge. The low correlations between portfolio scores and the measures of achievement (course exam and GPA) were expected. As suggested in the literature, teacher portfolios measure competencies and performance that traditional instruments cannot always evaluate (Zollman & Jones, 1994).

The scatterplots demonstrated similar trends to the correlation analysis; the plots between the performance tasks and portfolio scores and the years of teaching experience and portfolio scores showed a positive, albeit moderate, linearity between the variables which indicates that there is a similarity in what the variables are measuring. A slight negative trend that was also presented in the upper segment of the scatterplot of performance tasks and portfolio scores indicates that those students that scored very highly on the portfolio may have scored less highly on the performance tasks. This discrepancy could be due to the difference in raters of the tasks and portfolios. The performance tasks were graded by the course professor and the researcher whereas the portfolios were graded by only the researcher. It is possible that the course instructor and the researcher had different standards for the higher levels of performance and therefore created variability among the scores for the tasks and portfolios. The dominant trend, however, is a positive, linear relationship between the portfolio scores and performance tasks.
The results from the multiple regression analysis demonstrate that the performance tasks were the best predictor of portfolio performance, accounting for 22% of the total 50% variance for the model. The years of teaching experience also added substantially to the model (16%). The final exam grades (7%) and GPA (5%) added to the model in terms of a significant F statistic, but they did not account for as much of the variance as the first two variables. An explanation for the portion of the variance that is accounted for by overall GPA and the final exam grades is that many students tend to receive similar grades in all of their education courses. It is possible that the course instructor and the researcher, knowing the owners of the portfolios and the answers they gave to the exam question, were influenced in their grade assignment. If the portfolios had been rated blindly for the purposes of assigning a grade, the results may have been different for this part of the analysis. In this study, however, the blind ratings were used only to establish the reliability of the scoring process and not to establish the validity of the portfolio scores in determining achievement in the course and teaching competence.

Reliability

Issues of reliability were another focus to the current study. Inter-rater reliability and consistency in rating scale use, as well as score reliability and the impact of other factors such as raters or tasks on portfolio scores were the main areas investigated. Inter-rater reliability was found to be quite high considering that the raters were not trained in a group prior to rating the portfolios, although all of the raters were provided with explicit, written instructions on how to score the portfolios using the rating scales. The decision to not train and calibrate the raters was grounded in the knowledge that teacher educators are often in a position where they do not have the resources to train raters to assess portfolios. Time and cost effectiveness is often cited as a limitation of portfolio assessment; rater training and calibration are often what requires extra time. Using a Rasch model FACETS
analysis can account for errors due to raters; most rater error can be found by using the calibration reports presented in a FACETS analysis (Engelhard, Jr., 1994). Once these errors have been detected, adjustments to student score reports can be made based on judge severity measures, thus providing reliable and accurate scores. In this study, using written instructions and ample time to rate the portfolios individually, the raters reached a level of agreement among their holistic ratings of the portfolios that is acceptable in the field, $r = .65$ as indicated by the correlational analysis and the Fisher transformation table. The results from the FACETS analysis using the analytical rating scale data also indicated that the raters were generally consistent with each other in the scores that they gave each portfolio; the inter-rater reliability coefficient reported for consistency among raters was $r = .81$. Similarly, the results of the generalizability study indicated that only about 1% of the variance was due to the raters.

The FACETS analysis investigated the consistency with which the raters employed the analytical rating scale across portfolios. It was observed that three of the five raters were consistent in their use of the rating scales. Two raters, Rater Four and Rater Five, were reported as having a mean-square infit and outfit coefficients with considerable misfit. Rater Four gave much higher ratings on many of the portfolios in relation to the scores given by the other judges. At the same time, however, this judge gave more scores in the lower range of the scale in comparison to the other judges. Rater Five essentially reduced the rating scale to a dichotomy by judging most of the portfolios in the range of four and five, which thus reduced the variance in her ratings. In general, the raters were very lenient in their assessment of the portfolios, utilizing the upper half of the rating scale almost exclusively. The severity measures for all of the judges were reported as negative, indicating that they assigned grades that were higher than expected considering the range of the rating scale. It is possible that this is a true reflection of the prospective teachers' ability to perform teaching tasks. The other possibility is that the raters assigned inflated grades.
Due to the fact that all of the raters assigned consistently high grades is a more probable indication of the superior quality of the portfolio submissions than of inflated grading. The modeled error which indicates the precision or reliability of the severity measures for each judge were similar for all raters, as were the outfit statistics, indicating a consistency and accuracy in the ratings.

The FACETS analysis also presented data regarding the student score reliability and the difficulty of the submissions. The score reliability of the analytical ratings of the portfolios as determined through a FACETS analysis was \( r = .90 \), indicating that the students received consistent and reliable grades on their portfolios. The student measurement report (Table 4.13) and the calibration of the three facets presented in Figure 4.6 indicate that the students are receiving high ratings on the portfolio scores from all judges. This may be due to the fact that the raters are lenient in their judging of the portfolios, or to the fact that the students are highly competent. When the average scores are studied from the remainder of the assessments in the course (Table 4.1), it can be observed that the students on average performed well in the class. The mean score on the exam was \( M = 90.7 \), the average grade in the class was \( M = 90.6 \), and the average GPA was \( M = 3.5 \). These numbers are indicative of the fact that the students in the program are generally high-performing and competent students. The one student that was rated as less able, Student Ten, might have received lower scores due to the nature of the organization of her portfolio. This student chose to organize the portfolio in a nontraditional format in that she combined submissions and then reflected on the portfolio as a whole rather than on the individual submissions. The raters may have had more trouble rating this portfolio because of the organization and therefore were more severe in their scoring. This would be an indication of the need for research that investigates student and rater personalities or learning styles relative to ratings of portfolios. The individualistic nature of portfolios can make it difficult to measure student ability objectively, but portfolios are useful in assessing
the holistic profile of prospective teachers in terms of the ability to perform in the classroom.

The results of the generalizability study demonstrated that the raters and items contributed very little to the total variance of the portfolio scores. Each of these two variable accounted for less than 1% of the total variance. This reflects a similarity in the results of the Rasch model analysis; the score reliability was \(r=.90\) indicating that there is little variance affecting the ratings of the portfolios. The variance component for persons was larger (17%) meaning that the subjects differed in their portfolio scores, and subsequently, demonstrate different competence in teaching as shown by the portfolio. The variance due to raters was expected to be low in order to establish the reliability of the portfolio scores in determining teaching competence. It was also expected that the items that students chose to submit for evaluation in the portfolio would not contribute significantly to the variance as they are all representative of tasks that are expected of prospective foreign language teachers, and therefore similar in their difficulty level. The relatively higher variance due to persons provides initial evidence of the ability to differentiate among teachers using portfolio scores. The fact that the scores were also clustered in the high end of the scale shows that all of the students were performing at or above expected levels, so fewer differences among them may have been evident.

The large variance component for the interaction between persons and raters (36%) shows that although the raters were consistent in their ratings, they disagreed on the relative ability or standing of some of the subjects. This may reflect an inherent subjectivity in the portfolio rating process; a portfolio is a demonstration of individuality and may conflict or coincide with the raters personality. The raters may also interpret the tasks and rating scales differently. This type of discrepancy could be problematic if it identifies a possible problem with unclear scoring criteria, or different ratings due to the style of the portfolios rather than the contents. Given that 30% of the sample were international students, possible language
barriers, different cultural expectations and beliefs of what is quality work, and personality or attitudinal differences among raters and subjects may have caused more variation in the person by rater interaction. This issue continues to be a limitation of portfolio assessment; standardization of portfolio formats to avoid subjectivity in scoring, however, would reduce the qualitative nature of the assessment. Standardized portfolios are no more than collections of independent work samples that are evaluated according to traditional standards of right or wrong answers.

The person by item variance component (12%) indicates that the relative standing of persons differed from one item to another. This may have been due to the fact that students were allowed to choose from among a variety of submissions the ones they would include in their portfolio. Some students, therefore, may have included submissions that were rated more severely. Based on the FACETS analysis results and the results of the G-study, it appears that the difference in item ratings within each portfolio may be due more to rater fatigue or initial unfamiliarity with the rating scale. The results of the FACETS analysis indicated that Item One of the portfolios was the most difficult; however, the students all have different submissions for Item One. It is possible that the raters evaluated the first item of each portfolio more severely than the latter ones because they were unfamiliar with the type of work they would see or with the rating scale. Item Four was consistently rated higher than the others; again, Item Four is not the same for each person. The rater by item interaction does not support this hypothesis; the variance component for the rxi interaction was only 1% indicating that there was little variance due to raters on different items.

The variance component for the residual (30% of the total variance) shows that the remaining portion of the variance was due to the interaction among the three factors: raters, items, and persons. It is possible also that this variance may be due to other systematic or nonsystematic variation sources that were not examined in this study. An example of these sources is preservice teachers' learning styles and personalities versus rater personalities.
and learning styles. Although considered an important aspect in the establishment of portfolio assessment as a reliable and valid method for evaluating teachers, time limitations did not permit the investigation of this issue in the current study.

A final aspect of the study included the investigation of types of rating scales. Two rating scales were used to assess the portfolios: an analytical scale of one to six, and a holistic scale of 25 total possible points. Based on the observed means of the scores, it was expected that the grades assigned using the holistic rubric would be significantly higher than those assigned using the analytical scale. The results of the t-test indicate, however, that the difference between the two means (holistic mean score=84, analytical mean score=82) was not significant (F=1.84, p<.057). This information may be useful to teacher educators that are interested in assessing portfolios based on both types of scales or one or the other. The analytical scale is often easier to apply because it is more objective, but it tends to reduce each individual’s portfolio to a series of discrete point scales. The holistic scale, on the other hand, although generally more nebulous in rating because of the descriptors and subjectivity, provides an integrated, qualitative profile of the prospective teacher (Schrier & Hammadou, 1994).

**Preservice Teachers’ Attitudes**

The final aspect of this study was the attitude that the preservice teachers demonstrated toward the use of portfolio assessment to judge their ability to teach. It also served as a manner of measuring students’ understanding of alternative assessment tools before and after their experience with portfolios. As reflected by the change in mean scores on the pre-and post-course assessment survey, prospective teachers reported a better understanding of portfolio assessment, rating scales, and anecdotal reports after having experienced portfolio assessment firsthand. Most of them also reported a willingness to implement portfolio assessment into their future classrooms in some aspect. Based on
information in the literature (Adams, 1995; Ford & Olhausen, 1991; Krause, 1996), it was expected that students in the course would better understand the portfolio as an assessment instrument if they were provided with the opportunity to develop their own portfolios.

The interview with three of the students from the class shed light on the preservice teachers' attitudes toward creating a portfolio as a way to demonstrate their ability to teach a foreign language to young children. The students that were interviewed reported that they saw the portfolio as an opportunity to integrate content and pedagogy; one student commented that it was the first time she had used her native language in preparing assignments for an education course. Another student commented that she remembered all of the submissions in the portfolio in detail, but that she could not remember the question on the final exam. This was an interesting reflection on the perceived importance of the material tested by the traditional assessment. Finally, the students reported that the portfolio provided the opportunity to reflect on their work and to consider aspects of their development as teachers. They reported that although they considered the creation of the portfolio and the reflection on each submission time consuming, they understood the value of the portfolio as a part of their professional development.

Conclusion

This chapter provided the results of the statistical analyses that were conducted to investigate the issues of validity and reliability of portfolios in the evaluation of the pedagogical competence of preservice teachers. Answers to the research questions and discussion of the research hypotheses were also presented. The results from the surveys of student attitudes toward alternative assessment were also discussed and selected information from the focus-group interview was provided.

The results from the correlation analysis and the multiple regression analysis indicated that the portfolios used in this study were a valid assessment of the competence of
preservice teachers. The factor analysis confirmed the existence of two underlying constructs among the variables: (a) conceptual knowledge as measured by the final exam and overall GPA; and (b) pedagogical content knowledge as measured by the portfolios and performance tasks. It was also demonstrated that the portfolios were a reliable source of information for teacher educators in terms of the pedagogical content knowledge of the preservice teachers. These conclusions were based on the inter-rater agreement coefficient and the student score reliability reported in the FACETS analysis, as well as the variance percentages provided by the generalizability study.

The next chapter provides a summary of the findings in this study relative to the results of other research previously reported in the literature. It also discusses the pedagogical implications of the results from this study, and makes recommendations for further research in this area.
CHAPTER 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Introduction

Teaching is a multifaceted activity that requires a multidimensional assessment tool. Preservice teachers are often called upon to utilize a variety of competencies in performing tasks, and these competencies may be difficult to measure through the use of conventional tests. It is often argued that portfolios are multidimensional and comprehensive collections that can demonstrate the teaching competence of preservice teachers more holistically than do traditional tests (Zollman & Jones, 1994). Teacher educators nationwide are realizing the need for alternative assessment tools and the potential of performance-based tasks and teacher portfolios in this aspect. The Educational Testing Service has developed a series of examinations (Praxis) that are intended to measure teaching competencies through tests of basic skills, content knowledge, and a newer addition, a performance-based test that requires students to answer questions based on hypothetical teaching contexts. The State of Ohio has implemented the Praxis examination system of certifying its teachers on the state level. Other states are using performance-based measures and portfolio assessment to evaluate future teachers as well. Kentucky and Connecticut are two examples of states that have implemented portfolios as a means of assessing the competency of preservice teachers.
Issues of reliability and validity, however, have often been ignored in the descriptive publications about the benefits of portfolio and performance-based assessments. Messick (1994) and Terwilliger (1997) call for the investigation of validity issues with respect to alternative assessments claiming that these newer tools must be measured by the same criteria for validity as the traditional ones. In terms of reliability, CRESST (1993) Alternative Assessments in Practice Database reported that only 13 of 46 publications discussed data on reliability, and all of those were relative to rater agreement. According to Reckase (1997), inter-rater reliability is not the same as score reliability which is a true measure of the reliability of an instrument. This study examined the issues of both validity and reliability in implementing portfolio assessment as a supplement to traditional assessment tools in a teacher education program. It addressed several of the limitations to portfolios cited in the literature, namely the subjectivity inherent in the scoring procedures and the need to address the validity of the decisions made about teaching competency using portfolios.

Overview of the Study

This study was conducted during the Spring Quarter of 1996 to investigate issues of reliability and validity in assessing the teaching competencies of prospective foreign language teachers. Portfolios developed during a foreign language methods course were used to investigate several areas of concern: inter-rater reliability, score reliability, and the reliability of the rating scales, as well as the validity of portfolios in making professional decisions concerning teaching ability. A further aspect of the study included an investigation of students’ attitudes toward, and understanding of portfolio assessment.

Two research questions were the focal point of the study. The first question dealt with issues of validity in using portfolio assessment to determine teaching competency. Content validity, criterion/predictive validity, and construct validity were investigated.
Content validity was examined through a content analysis of requirements for pedagogical knowledge as established by several professional organizations, including ACTFL. Statistical analyses to investigate criterion and predictive validity included correlation of principal variables (portfolio scores with performance tasks, years of teaching experience, final exam grades, and GPA) and multiple regression analyses to determine which of the variables best predicted scores on the portfolio. A principal components factor analysis was conducted to confirm the existence of two underlying constructs in the variables (conceptual knowledge and pedagogical content knowledge).

The second question looked at the reliability of portfolio scores in terms of inter-rater agreement and score reliability of portfolio submissions. A correlational analysis was conducted to determine an initial inter-rater reliability coefficient. A Rasch model FACETS analysis was also conducted to examine inter-rater reliability, judge severity, and score reliability. Finally, a generalizability study was conducted to look at issues regarding the reliability of portfolios in terms of variance components (raters, items, and persons).

Several secondary questions investigated the beliefs and understanding of the students in the course concerning the use of portfolio assessment in evaluating teachers. These questions were answered by means of pre- and post-course surveys and through a focus interview with three members of the class. The pedagogical implications concerning the implementation of portfolio assessment in a teacher education program were also examined.
Summary of Findings and Conclusions

Research Question One

As a measure of the validity of portfolio assessment in measuring future teacher performance, to what extent do the independent variables (performance tasks, years of teaching experience, final exam, and overall GPA) correlate with, and predict performance on the portfolio?

One purpose of this study was to investigate issues of validity in using portfolio assessment to evaluate teaching competency. Minimal research has been reported in the literature that looks at the validity of portfolio scores in assessing students. Even fewer studies provide statistical evidence that portfolio assessment is a valid indicator of the competency of prospective teachers. The research that has examined issues of validity is reviewed relative to the results of this study.

Nweke (1991) conducted a study in which scores from the portfolios of 30 college students enrolled in teacher education programs were correlated with other established measures of academic achievement. Results indicated that performance measured by portfolios is different from performance measured by traditional forms of assessment such as the American College Test and overall GPA. The correlations reported for the portfolio scores and college GPA were $r=0.11$, and the portfolio scores and the American College Test $r=-0.12$. He concluded that portfolios were measuring development and progress rather than achievement.

Naizer (1993) investigated the validity of portfolio scores in a math and science preservice teacher education course. In a measure of association analysis of the final exam and portfolio scores, a low correlation ($r=0.22$) was achieved, thus suggesting that the two instruments were measuring different knowledge and performance dimensions. Concurrent validity was established, however, with pedagogical content knowledge which was considered to be a similar measure to the portfolios. Through a discriminant analysis,
Naizer demonstrated that students with more strategic knowledge and more pedagogical experience as measured by the number of education courses and years of teaching experience scored higher on the performance portfolios.

Other researchers have also investigated issues of validity in terms of the relationships between portfolio scores and other measures of ability. Koretz et al. (1993) investigated the relationships between portfolio scores and measure of both similar and different capabilities in the Vermont program. They found moderate correlations ranging from 0.47 to 0.58 between writing portfolio scores and direct writing assessments. Similarly, Gearhart, Herman, Baker, and Whittaker (1993) found virtually no relationship when comparing results from standard writing assessments and those from writing portfolios.

The current study also sought to establish the validity of using portfolio assessment to measure foreign language teacher performance in preservice teacher education. As concluded in other studies of the validity of portfolio assessment to measure teacher competence, moderate to high correlations and predictive values were expected between the portfolio scores and both the performance tasks and years of teaching experience. Experience in the classroom and the ability to perform tasks such as lesson planning and materials development were expected to provide similar information about the teaching competence of a preservice teacher as were the submissions in the portfolios. On the other hand, relatively low correlations were expected among the achievement measures (overall GPA and the course final exam) and the grades assigned to the portfolios. Similar results in previous studies indicated that traditional or standardized instruments do not correlate well with portfolio scores due to the difference in the ability being measured (Koretz, et al., 1993; Nweke, 1991). Traditional exams tend to measure the achievement of an individual in a given subject while the portfolio aims to present a holistic portrait of an individual’s ability.
In the investigation of the validity of the portfolios used in this study, a significant, positive correlation was expected among the performance tasks used in the course and the portfolio scores, as well as between years of teaching experience and portfolio scores. It was also expected that the performance tasks and the years of teaching experience would account for a higher percentage of the variance in the multiple regression analysis than the traditional measures. Performance measures such as peer teaching, lesson planning, and unit creation required prospective teachers to utilize theoretical knowledge in the presentation of content, and provided a forum for the opportunity to apply pedagogical content knowledge. The portfolio in this study also asked the students to integrate pedagogy and content in the presentation of submissions that were a demonstration of their ability to teach foreign language to young children. Shulman’s (1987) model for the knowledge base for effective teaching requires that teachers be able to perform a certain number of tasks in the classroom in order to transmit knowledge effectively. Preservice teachers must be understand their content and be able to present it in a way that is learnable (pedagogical content knowledge). Teachers must also be able to manage a classroom, have an understanding of the learning process, be familiar with curriculum content, and be adept at planning daily activities. All of these tasks are performance-based activities that students completed in the course used in this study; either through the performance tasks or through the portfolio submissions.

Pedagogical content knowledge (Shulman, 1986; 1987) is listed above as one of the necessary components in the knowledge base for effective teaching. Wisdom-of-practice studies conducted by Shulman (1987) and colleagues investigated the behaviors displayed by effective teachers. Many of the teachers reported that they know more about teaching than they are able to articulate which makes it difficult to operationalize, and therefore transmit pedagogical content knowledge. The two-factor orthogonal solution specified in the factor analysis yielded results that confirmed the existence of two underlying constructs.
in the principal variables. Conceptual knowledge as measured by the final exam and overall GPA presented as factor one and pedagogical content knowledge as measured by the portfolios and performance tasks presented as factor two.

The correlation analysis reflected results similar to those of the factor analysis. There was a significant, positive relationship between grades received on the performance tasks and portfolio scores given (r=.47, p<.01). This correlation is due in part to the fact that the majority of the portfolio submissions were similar tasks as those required in other areas of the course (i.e., lesson plans, materials creation, etc.). The results from the multiple regression analysis indicated that the performance tasks were also the best predictor of the portfolio scores in the course (22% of the total variance). It is possible that the correlation and predictive value of the performance tasks are enhanced because the performance tasks and portfolio submissions are developed from the same philosophical viewpoint albeit by different people (the course professor and the researcher respectively).

There were several submissions by students, however, that required them to go outside the course resources to extend their professional network. Examples of this included visiting stores to compile a list of available materials, researching current issues in foreign language education for young children, interviewing/observing teachers in the field, etc. Submissions such as these afforded students the opportunity to demonstrate their willingness to participate in the field of foreign language education and/or to demonstrate their future performance as members of the teaching profession.

Pedagogical content knowledge may be fostered through the years of teaching experience; it is possible that there is a relationship between the number of years of experience and the level of teaching competence. It was expected in this study that there would be a positive relationship between the two measures: portfolio scores and years of teaching experience. Years of teaching experience predicted portfolio performance relatively well (16% of the total variance). These results reflect the findings of other studies such as
that of Naizer (1993), who found that years of teaching experience best predicted performance on the portfolio. The moderate correlation coefficient ($r = .35$) between years of teaching experience and portfolio scores in this study may be due to the fact that many of the students (77%) in the course were undergraduate students with little or no teaching experience in classroom settings and international students with teaching experience in a foreign setting. The average number of years of experience among students in the class was 3.4 years; 30% of the students reported having no prior teaching experience. Also, the differences among the teaching abilities required in tutoring, for example, or lecture-based classes such as in the foreign settings, may have affected the relationship between teaching experience and portfolio scores. The factor analysis showed years of experience as loading on the same factor as the the final exam and overall GPA. Although this finding does confound the definition of the two constructs, the results of the factor analysis do not detract from the significant, positive underlying relationship between portfolio scores and years of teaching experience.

The traditional measures used in the study to evaluate the students, the final exam and overall GPA, were not expected to have a positive correlation to portfolio scores. It was also expected that the exam grades and GPA would have little predictive power in determining preservice teachers' portfolio scores. The results of the analyses indicated that a significant correlation coefficient was not achieved between portfolio scores and final exam grades, nor between portfolio scores and GPA. The low correlations ($r = .29$ in both cases) between portfolio scores and these two measures of achievement may indicate that the portfolio is measuring a construct other than achievement in the course. The factor analysis results confirmed that the traditional measures shared the same underlying construct of conceptual knowledge whereas the performance tasks and portfolios loaded on the factor measuring pedagogical content knowledge. There was also a significant correlation achieved between the final exam and overall GPA ($r = .54$, $p < .004$). This
relationship indicated that the final exam is a measure of achievement like the GPA, and not a measure of ability to perform in the classroom. The exam asks the students to summarize conceptual knowledge gained in the course in the form of a rationale for early language learning; the portfolio, on the other hand, allows students a forum to display a variety of abilities necessary for effective teaching such as lesson planning, materials development, and interaction in the teaching context.

Although the final exam grades and GPA contributed to the regression model in terms of the significance of the F statistic, they showed minimal predictive power in the model. The final exam accounted for 7% of the variance and GPA accounted for 5% of the variance. These results may be explained by the fact that the data for this part of the analysis was provided by the instructors of the course who rated the portfolios and exams knowing who the owners were. It is possible that, to some extent, the grades on the portfolio and exam reflect the general pattern of grades that the student receives in education classes, which would also explain the relationship of the final exam grades to overall GPA. In other words, the data used for these analyses may have been biased to some extent in that the grading was not done blindly; the instructors may have graded differently because they knew the identity of the student.

Based on the results of the analyses used to establish validity (correlation, regression, and factor), it can be concluded that the portfolio is measuring an ability that is different from that being measured by the traditional exam and overall GPA. It is the researcher’s opinion that the portfolio scores reflect prospective teachers’ pedagogical content knowledge and the ability to perform authentic teaching tasks whereas the final exam grades and overall GPA indicate achievement in the course/program and the ability to synthesize material but not necessarily to apply it (conceptual knowledge). The factor analysis results confirmed this conclusion; the traditional measures loaded on one factor whereas the performance-based measures loaded on a different factor. There were two
underlying constructs measured by the principal variables: conceptual knowledge and pedagogical content knowledge.

Research Question Two

To what extent is portfolio assessment a reliable measure of teacher competence as determined by: (a) the consistency with which raters agree in their ratings of each portfolio; (b) the relative impact of sampling due to persons, raters, and tasks in evaluating the competence of preservice teachers; and (c) the consistency with which ratings assigned to the overall portfolio using the holistic rating scale correlate with the ratings assigned to individual pieces within the portfolios using the analytical rating scale.

The second question in this study investigated issues of reliability in using portfolios to determine teaching competence. Inter-rater reliability was first calculated using a Pearson correlation matrix and Fisher transformation table. A Rasch model FACETS analysis and a generalizability study were also conducted to examine inter-rater reliability, score reliability, and the impact of other factors on the portfolio scores. A t-test was conducted to assess the consistency with which the judges employed the two rating scales. Previous studies conducted on issues of reliability and portfolios are discussed relative to the findings of this study. There were no studies reported in the literature that investigated the reliability of portfolio scores using Item Response Theory; and only one dissertation discussed the results of a generalizability study to examine rater agreement in portfolio scores. This study employed a more traditional inter-rater reliability coefficient investigation as discussed above, as well as both a Rasch model FACETS analysis to examine rater agreement and a generalizability study to determine variance in portfolio scores due to raters, items, and persons.

Koretz, Stecher, and Deibert (1993) reported on Vermont’s statewide portfolio assessment project in math and writing. The inter-rater reliabilities ranged from 0.28 to
0.60, a level of agreement considered not sufficient enough to permit reporting of the scores to the state. Klein, McCaffrey, Stecher, and Koretz (1995) reported the reliability of mathematics portfolio scores based on the first two full years (1992 and 1993) of implementation of the program. Unfortunately, the scores from both of these years were not considered reliable enough to report to students. Reliability of a student’s total score ranged from 0.45 to 0.49 in 1992; in 1993, the values increased to a range of 0.63 to 0.71. The level of reliability sought, however, was 0.85. Inter-reader agreement rates were reported at 55%; only 10% above the agreement rate due to chance alone. Kappas ranged from 0.02 to 0.51 with a median of 0.28. The authors concluded that the low reliability of student scores and the low reader agreement rate were caused by insufficient reader calibration, scoring guides that were too open to reader discretion and interpretation, and the freedom with which students were allowed to choose the pieces included in the portfolio.

Research in local settings, however, suggests that portfolios can be scored reliably (Herman, Gearhart, & Baker, 1993; LeMahieu, Gitomer, & Eresh, 1995). LeMahieu et al. (1995) achieved an average 0.84 inter-rater reliability coefficient in the study of the Pittsburgh Public School District’s use of portfolio assessment in evaluating writing performance. In a previous study of the same program, LeMahieu, Gitomer, and Eresh (1993) as cited in Herman and Winters (1994) achieved inter-rater agreement correlations that ranged from 0.60 to 0.70. Herman et al. (1993), in a study of an elementary school-based study of writing portfolios, reported average correlations between scores given by pairs of raters was 0.82 and the percentage of agreement for all pairs of raters averaged 0.98.

Studies that investigated the reliability of portfolio assessment for evaluating prospective teachers include several dissertations that investigated the reliability of scores and reader agreement rates. Naizer (1993), in a study of the use of performance-portfolios
with preservice science and math teachers at Texas A & M university, reported a high
degree of inter-rater agreement (48% to 86%). Gibson (1995), in a study of teacher
portfolios in special education, also conducted at Texas A & M university, reported a
78.26% agreement rate among the judges of the portfolios. Both of these studies, however,
discussed portfolio scores in terms of the data on inter-rater agreement and did not discuss
the reliability of the scores themselves.

Based on the results of the research studies discussed above (e.g., Gibson, 1995;
Herman et al., 1993; Klein et al., 1995; Koretz, Stecher, & Deibert, 1993; LeMahieu et al.,
1995), it was expected that the inter-rater coefficients in this study would be about r=0.75.
The inter-rater reliability was first calculated using a Pearson-Product Moment correlation
matrix of rater agreement on the holistic scoring and was found to be r=0.65. This
coefficient reflects a similar level of correlation to that achieved by LeMahieu et al. (1993)
in the original study of Pittsburgh Public Schools’ use of portfolios. The differences
among the raters in the present study could be due to a lack of familiarity with using
portfolio assessment and scoring rubrics, as well as the absence of a formal training and
calibration session. All of the raters reported a knowledge of portfolio assessment but
indicated that they personally had not implemented it in their teaching. The coefficient that
was achieved, however, is considered acceptable in the field of education; further training
in the use of the rating scales and the criteria for evaluation would increase the agreement
among the raters and might be warranted depending on the decision that is to be made from
the scores.

In order to strengthen the results from the traditional rater agreement analysis and to
investigate the student score reliability, a many-facet Rasch model analysis was conducted
using the FACETS program (Linacre, 1990). Due to the nature and assumptions of the
many-facet Rasch model, the results of the FACETS analysis can provide an objective
estimate of the students’ ability to apply skills and knowledge acquired in the course.
Establishing the reliability of the portfolio ratings is crucial in promoting portfolios as an accurate measure of the multiple competencies of preservice teachers. It was observed through the FACETS analysis that the judges were consistent in their ratings of the portfolios ($r=0.81$). Two of the raters showed considerable misfit in their ratings in comparison to the other raters; one rater was more lenient than the others, and the other rater used the rating scale as a dichotomy rather than a six-interval scale which reduced variance in her ratings. All of the judges tended to be lenient in their ratings; the judge severity measures which indicate the leniency or severity of the raters in assessing the portfolios were negative indicating a high frequency of ratings on the upper end of the rating scale (intervals 4, 5, and 6). The fact that all of the raters were using this end of the rating scale did not effect the consistency of the ratings, but it did reduce the rating scale by half. This could be problematic if the portfolio scores were intended to illustrate major differences among students in their pedagogical content ability.

The FACETS analysis also provided information regarding the student and item (submission) domains. The student measurement report in the FACETS analysis indicated that the portfolio scores are a reliable measure of student ability ($r=0.90$). After accounting for the judge severity measure, fair or average scores were reported for the students that indicated that the students were performing well on the portfolios (observed mean=5.0; fair mean=4.1). The student ability index ranged from Student Five who was considered the most able to Student Ten who was considered the least able. Student Five had an average rating on the portfolio of 5.5 on a scale of 6, whereas Student Ten had an average rating of 4.0 on a scale of 6. These scores correlate to grades on other measures used in the class; Student Ten, for example, received an 86% on the final exam and a grade of 80% in the class. The average grade in the class was 90.5%. Student Five, considered the most able according to the raters, received a 95% on the exam and a 93% in the course. These results demonstrated that the portfolio scores were consistent with the other grades received in the class.
class. The submissions report indicated that the first item rated in each portfolio was judged as the most difficult whereas the fourth item in each portfolio was rated as the least difficult. Because the submissions in each portfolio were not in the same order, these results were interpreted as the first item was judged the most severely by the raters and the fourth the least severely. This might have been due to the initial reaction of the judges to the first item in each portfolio; by the fourth item, the raters had become familiar with the student's work and rated that one more leniently.

Generalizability theory has as its strength the ability to separate variance into several components to better analyze the origin of the error. Only one study (Naizer, 1993) has investigated issues of reliability of teacher portfolios using generalizability theory. Reckase (1993) calls for the use of generalizability theory in the investigation of score versus scorer reliability in his rebuttal of the results of the research done by LeMahieu, Gittomer, and Eresh (1995). Reckase claimed that the investigation of inter-rater reliability alone was not enough to claim that portfolio assessment was a reliable measure of competency; score reliability must also be investigated. It was, therefore, considered an important task to support the findings of Naizer's study, and answer to Reckase's criticism of previous reliability research in the attempt to demonstrate that portfolio assessment can provide reliable results. The design used in this study was PxlXR (persons crossed with items crossed with raters) because all persons in the sample (N=10) completed all of the items (N=6) which were rated by all of the raters (N=5).

Naizer (1993) used generalizability theory in his investigation of the reliability of performance-portfolio scores in measuring teacher competence in a math and science teacher education course. He reported that a very small percentage of the total variance was due to different raters (2.5% initially; 0.8% at fourth rating session). The nature of the portfolio used in his study required the items to be considered fixed variables, and the variance due to individuals decreased after the initial scoring of the portfolios.
The generalizability study results from this study supported the FACETS analysis. It was observed that the variance component due to raters was minimal (1%) relative to the total variance, as was the variance due to items (1%). This provided evidence that the impact of the raters and the items on the portfolio scores was relatively small indicating a consistency among the raters in the judging of the items in the portfolio. These results reflected the findings by Naizer (1993). The desired variance, that due to persons, was only moderate (17%) indicating that the scores on the portfolios did differentiate among the students but not significantly. This could be problematic if the scores were clustered at the lower end of the scale; however, because the majority of the scores were at the high end of the scale, the fact that portfolios did distinguish between persons as well may be due only to the fact that all of the students in the program are quite competent. The average GPA (3.53) and mean grade in the class (91%) indicate that the students in the course are generally high-performing. It may be difficult to differentiate among them using any type of instrument.

The variance component for the interaction between persons and raters was higher than expected (36% of the total). This indicates that although the raters used the rating scales consistently, they did not always agree on the standing of the person. Naizer (1993) found similar results in his investigation; approximately 15% of the total variance was accounted for by the interaction between raters and persons. This finding may indicate a problem with imprecise scoring criteria or a lack of explicit rater training and calibration. There may also be a variation in ratings due to the differences in style and format of the portfolios. The individuality and appearance of a portfolio is explicitly linked to the person creating the portfolio. The rater brings to the judging process his own personality and preference for style which may conflict with the examinee’s style. This can cause subjectivity in the ratings. An example of this subjectivity may have been found in this study: the ratings assigned to Student Ten, who was judged as the least able. It is possible
that the lower ratings of her portfolio were due to the unusual format and organization of her portfolio rather than a lack of ability.

The persons by items interaction variance component (12%) shows that there was a difference among students based on the item that was being rated. This percent of the variance may be explained by the subjectivity of the ratings of certain tasks relative to others. For example, the field-based task that required students to make a list of foreign language resources available at local stores may have been rated differently than a submission that required students to express their opinion of the English-only movement.

Both of the variance components for the interactions (rater x person and person x item) may be due to the variation in the items submitted by students, to the variety of cultural and linguistic backgrounds of the students and raters, and to the variety in style and format of the portfolios. The residual variance was 30%, and is accounted for by the interaction between persons, raters, and items, or by some other factor not accounted for in this study. It is possible that personality and/or learning styles of the raters and the students are affecting the scores that their submissions receive. This study would have been strengthened by the inclusion of information on the personality of the students and raters perhaps measured by a personality index. Investigation of personality and learning styles is thus warranted in further research of the reliability of portfolio scores.

Finally, a t-test was conducted to determine if the scores assigned by the raters when using the holistic scale were the same level of scores that were assigned when using the analytical scale. Two rating scales were necessary in this study because of the nature of the analyses that were conducted. The holistic rating scale was the scale used to grade the portfolios in the course and the one that was used to conduct the correlational and regression analyses. The FACETS analysis and the generalizability study, however, required that the data be provided in terms of items or submissions, not a holistic measure of the portfolio. In order to determine if the results from the FACETS analysis and the
generalizability study, both of which used the data from the analytical scoring of the portfolios, might be applicable to the scores assigned when using the holistic scale, a t-test analysis was conducted to determine if there was any significant difference between the scores from each of the two rating scales. A non-significant difference was expected between the overall grades assigned to the portfolios using the holistic rating scale and the grades assigned to individual pieces within each portfolio using the analytical rating scale. A t-test analysis of the mean scores for each rater and each rating scale indicated no statistically significant differences (F=1.84, p < .57) among scores assigned with the holistic scale and those given using the analytical scale. This is beneficial information for teacher educators that may prefer to implement one of the two types of rating scales. The results essentially demonstrate the ability to use either type of scale and still attain reliable ratings.

Secondary Questions

To what extent is there a difference between the reported beliefs and familiarity with assessment tools on the pre-course survey and the post-course survey?

There was a positive difference in the means between the ratings students gave on their understandings of assessment tools prior to the course and those ratings they gave after developing their own portfolios. The students reported a higher level of understanding of portfolio assessment and rating scales after their experience with both tools in the methods course. This difference was determined by a comparison of the overall means for each of the survey questions presented in Table 4.16.

Did the students indicate an intention to use portfolio assessment in future classes?

Students in the course reacted positively to the use of portfolio assessment and indicated an intention to use this form of assessment with their future students. These
results confirmed findings by Ford and Olhausen (1991) who reported a 19% increase in use of portfolio assessment by teachers who participated in a graduate literacy course in which portfolios were used for assessment.

Pedagogical Implications

Portfolios are currently used in several teacher education programs across the nation, and they are quickly becoming an accepted form of evaluation for prospective foreign language teachers. The multidimensional nature of the portfolio allows teacher educators to assess preservice teachers’ competency in teaching through an array of activities. The issues of reliability and validity are critical in the implementation of assessment measures, but they are often ignored in the descriptive publications about portfolio use with future teachers. In the local context, the program in foreign and second language education at The Ohio State University expects to move to the use of teacher portfolios in evaluating its M.Ed. students in the upcoming academic year. This study will provide the program coordinator with information concerning reliable and valid procedures for implementing portfolio assessment as an evaluative component in the assessment of the preservice teachers. The results from this study add to the current knowledge base concerning the use of portfolio assessment by determining the feasibility of its use in teacher education in terms of the reliability and validity of the scores, and by responding to the demand for additional research and development in pedagogical assessment (Martinez & Lipson, 1989).

Messick (1994) and Terwilliger (1997) claim that validity issues surrounding performance assessments such as portfolios cannot rely only on the authenticity of the assessment; these assessments must be measured against the same validity criteria as other instruments. Issues of validity were more difficult to establish in this study because of the individualistic and qualitative nature of portfolio assessment, the lack of validated
alternative assessment tools, and the absence of what constitutes effective teaching. Steps toward the creation of a model for the knowledge base for effective teaching are being taken by researchers such as Shulman (1987). That research will lead to more unified models of what constitutes effective teaching behaviors and thus lead to a codification of pedagogical content knowledge, the essential component of Shulman’s model for education. Teacher educators can establish portfolio assessment as a valid means of evaluating preservice teachers through an investigation of the representativeness of the portfolio contents in relation to course or program objectives, and through comparison to individual performance of students on other types of assessment. Low correlation coefficients between traditional measures of achievement and portfolio scores, as found in this study and others that investigated the issue (Naizer, 1993; Nweke, 1991), demonstrate that portfolios may be measuring abilities other than achievement in the course or in academics overall. Factor analysis can also indicate the existence of separate, underlying constructs that demonstrate the difference between traditional and performance-based instruments. Until further research on the validity of portfolios in evaluating teachers can be conducted, portfolio assessment can be used as a supplement to traditional measures of evaluating preservice teachers to provide a holistic profile of the prospective teacher.

Reliability issues surrounding portfolios have been more thoroughly investigated, especially in terms of the rater agreement levels that have been reached. Previous studies (e.g., Herman et al., 1993; LeMahieu et al., 1995) have indicated that a high level of inter-rater reliability can be achieved in the evaluation of portfolios, in particular if there is explicit rater training prior to the judging. The moderate inter-rater reliability coefficient achieved (r=0.65) in this study indicates that explicit rater training and orientation may have been necessary for more reliable use of the rating scales. It is possible, however, as shown by the results of this study, to obtain a reasonable level of rater agreement without extensive training, thus making portfolios a more feasible option for course instructors that
do not have time to corroborate with other instructors in the use of rating scales. The results of the FACETS analysis and the G-study indicate that the raters were able to rate the portfolios consistently without extensive training. The coefficient for rater consistency from the FACETS analysis was r=.81; and the G-study results demonstrated that only 1% of the variance was due to the raters. Engelhard, Jr. (1994) claims that one of the features of performance assessments like portfolios is that they depend on the quality of expert judgments. He posits that rater errors can be identified and the potential effects minimized through the use of rating scale models based on item response theory, such as the FACETS model used in this study. He states that most rater errors can be detected and adjustments can be made to the estimates of student competence thus providing reliable and true measures of ability. The FACETS model used in this study provides a promising approach for investigating rater errors that can lead to unfair assessments of student competence.

A final and important question in the current study concerned the pedagogical implications for the use of portfolio assessment by the students in the preservice teacher education class with their future foreign language classes. The instruction of preservice and in-service teachers in the use of portfolio assessment may have an effect on those teachers’ notions of assessment, and also may foster their use of this assessment tool in future classes (Adams, 1995). As shown by the increase in mean ratings of student understanding of portfolio assessment and rating scales as reported by students on the assessment survey, it can be concluded that through the use of portfolio assessment students do gain a better understanding of the process, and indicate a willingness to implement the tool in the future.

Limitations of the Study

The current study investigated issues of reliability and validity in using portfolio assessment to measure preservice foreign language teacher performance. It was expected that the portfolio scores would correlate well with, and be able to predict scores on, the
other performance-based variables used to evaluate students in the course. It was also expected that a factor analysis would show that portfolios, performance tasks, and years of teaching experience measure the same underlying construct: pedagogical content knowledge. Conversely, it was expected that there would be little or no relationship, either correlational or predictive, between the traditional measures used in the class and the portfolio scores. The factor analysis was expected to show a separate factor for the traditional measures thus indicating the existence of a separate construct.

The results of the validity study reflected the original expectations of the study, except in one area: the relationship between portfolio scores and years of teaching experience. The correlation analysis resulted in a moderately significant correlation coefficient ($r = .35$). In the regression analysis, the portfolio scores accounted for 16% of the variance in years of experience. The factor analysis, however, demonstrated that the years of teaching experience loaded on the same factor as the traditional measures rather than loading on the factor that measured teaching performance. These results confound the definition and separation of the two constructs specified in the study: conceptual knowledge and pedagogical content knowledge. It was expected that the number of years of teaching experience would measure the same construct as the performance-based evaluations. A possible limitation to this study may have been in measuring pedagogical content knowledge by means of self-reported years of teaching experience. A lack of clarification or standardization of what constituted teaching experience may have created variability in the measure; students reported private tutoring in the same manner as classroom teaching. There were students who had international, university-level experience and others with secondary U.S. public school experience. Controlling for these factors may have standardized the variable in order to give a more accurate measure of teaching experience in relation to the construct of pedagogical content knowledge.
A second limitation of this study was the variance that resulted in the generalizability study between persons and raters as well as in the overall variance that is due to factors not measured in the study. In establishing the reliability of scores, it is expected that the main portion of the variance will be explained by the factors or constructs that are being measured. In this study, however, the person x rater component and the residual component accounted for a combined total of 66%, more than half of the explained variance. These results indicate a limitation in the rating of portfolios that perhaps could be adjusted through further training and calibration of raters. The fact that the person x rater component is so large indicates that the raters, although consistent in their ratings of all the subjects, did not agree among each other on the relative standing of the subjects on the scale. This discrepancy reflects variability in the ratings that may be due to rater personality and/or expectations, the subjects’ linguistic and cultural background, and possible misinterpretation of the guidelines for the portfolio. In a future study, this limitation could be overcome with more extensive rater training and calibration. Another possible solution might be to investigate the influence of rater personality on ratings as well as subject personality on the raters’ evaluation of the portfolio.

Recommendations for Further Research

An impetus behind the shift toward using teacher portfolios in the assessment of preservice teachers is the need to evaluate a variety of teaching abilities that cannot be assessed through conventional tests. Measures of teaching competency that allow preservice teachers to demonstrate understanding of conceptual knowledge and apply that knowledge in the teaching context are needed. Teacher portfolios reflect the complexity of teaching and recognize the impossibility of measuring all competencies through one score. As the field of teacher education attempts to professionalize the training of future teachers, there is a need to consider all aspects of an individual’s abilities. Performance-based
assessment and teacher portfolios are increasingly implemented in programs of teacher education across the nation. Several states, including Kentucky and Connecticut, have implemented teacher portfolios as a means of assessing the competency of their teachers for state certification. The Educational Testing Service has developed a series of examinations (Praxis) intended to measure teaching competencies through tests of basic skills, content knowledge, and a newer addition, a performance-based test that requires students to answer questions based on hypothetical teaching contexts. Ohio has implemented the Praxis examination system of certifying its teachers on the state level. Teacher educators nationwide are realizing the need for alternative assessment tools and the potential of portfolios in this aspect. The present study corroborates results from other research that investigated the reliability, validity, and utility of portfolios in assessing students (Gearhart et al., 1993; Herman & Winters, 1994; Naizer, 1993; Nweke, 1991) but further investigation of the issues of reliability and validity of the procedures and criteria used in the creation and evaluation of portfolios is warranted.

Issues of criterion and predictive validity need further examination. Comparing scores from traditional measures and scores from portfolios is like comparing apples with oranges. Proponents of alternative assessment claim that instruments like portfolios measure constructs that are not identified by conventional exams. In the case of preservice teachers, supporters of portfolio assessment posit that teacher portfolios allow prospective teachers to demonstrate their teaching competence and give a strong indication of how these teachers will perform in authentic classroom contexts. Criterion and predictive validity of portfolios have been investigated through comparison of performance measures to traditional assessments (Naizer, 1993) but more work needs to be done. It should not be expected that portfolio scores correlate highly with many established testing methods, thus illuminating the problematic issue of establishing the validity of portfolio assessment. With what standardized criterion measures can portfolio scores be validated if not with traditional
tests? This is a question that warrants further research in the establishment of alternative assessment methods as viable supplements or alternatives to traditional tests. It may be possible to compare portfolio performance with assessments of later teaching success to establish criterion and predictive validity, but a reliable and valid measure of that performance will also need to be created. Further research on the construct validity of portfolios is also essential to their integration as a practical tool in teacher education programs.

Although issues of reliability have been addressed more often in the literature, most of the studies have focused on inter-rater agreement. As Reckase (1997) argues, inter-rater, or scorer, reliability only measures the ability of different raters to assign the same scores over a number of different judgments. It does not reflect the reliability of the scores relative to the person’s ability. This study investigated issues of scorer and score reliability, but further investigation of the score reliability of performance assessments is needed. Other sources of portfolio score reliability, such as score stability over time, stability across different rater pairs, and the effect of task or context on the portfolio scores have not yet been investigated (Herman & Winters, 1994). The effects of student and rater personality and learning styles would be an interesting addition to a replication of this study. It is possible that the person x rater and person x item variance components (36% and 12% respectively) or the residual variance component (30% of the total) in the generalizability study were due to affective factors among the raters and the students.

According to Engelhard, Jr. (1994), “when performance assessments are used in education, rater errors reflect potential factors that can lead to unfair and biased assessments of student competence” (p.110). Investigation of the reliability and validity of the assessment process must be investigated for portfolio assessment to become an alternative to traditional assessments. Steps have been taken to establish portfolios as a reliable and
valid instrument in the evaluation of preservice teaching competence, but further research is needed.

Conclusion

The current study presented information relative to the investigation of the issues of reliability and validity of performance assessments namely portfolio assessment. The study examined portfolios used in a teacher education course at The Ohio State University to determine if they were reliable and valid measures of the students’ competence in teaching foreign languages to young children. It was determined that the portfolios were a valid measure of the students’ pedagogical content knowledge (ability to present content in a learnable format) based on results from correlational and regression analyses. The factor analysis confirmed the existence of two separate underlying constructs: (a) conceptual knowledge as measured by the traditional tools, the final exam and overall GPA; and (b) pedagogical content knowledge as measured by the performance tasks and the portfolios. It was also indicated that the portfolio scores were a reliable and accurate measure of the students’ competence. The inter-rater agreement level was high (r=.81) as was the score reliability index (r=.90) in the Rasch-model FACETS analysis. Item response theory and generalizability theory were used as the principal analyses in the investigation of reliability; however, further research into the issues of reliability of other performance measures using these analyses are recommended.
LIST OF REFERENCES


National Center for Fair and Open Testing (Fairtest), (1992a). What's wrong with standardized tests? Cambridge, MA.


Dear 617 students:

As a means of collecting the maximum data for a study on the use of portfolios in a methods class, I need to ask your help in completing the questions on this survey. Please answer the questions honestly and to the best of your ability. They will only be used as a means of studying the effectiveness of portfolios in evaluating students such as yourself in the future and will not be made available to anyone but the researcher. Thank you for your time.

Name: __________________________________________
Student Status: _____Undergraduate _____Graduate _____Professional
Country of Origin: ________________________________
Language you will teach: __________________________
Current GPA in that language (if taking courses): _________
Current overall GPA: __________
Number of education courses you have taken: __________
Number of years of teaching experience (if any): __________

In order to be able to study the effectiveness of portfolio use, but still be able to return the portfolios to you in a timely manner, it will be necessary to photocopy all or part of many of the portfolios. Please sign below if you give permission to Anna-Marie Gelines of the Department of Educational Studies at The Ohio State University to photocopy the contents of your 617/FLES methods portfolio for research purposes only.

________________________________________________________________________

Spring Quarter, 1996
APPENDIX B

ASSESSMENT SURVEY

Dear Participant:

A critical issue that will determine the manner in which we evaluate students' performance is alternative assessment. In order to direct alternative assessment techniques and practices it is necessary to collect data on teacher awareness of alternative assessment. The purpose of this survey is to investigate preservice teachers' notions of educational assessment of foreign language students. I would appreciate a few minutes of your time to complete the attached questionnaire. Your answers are anonymous, and will be used in my dissertation as a requirement for an advanced degree in Foreign and Second Language Education at The Ohio State University. Thank you for your participation.

General Information:

1 Sex: female _____ male _____

2 Age Group: 20-25 _____ 26-30 _____ 31-35 _____ 36-40 _____

        41-45 _____ 46-50 _____ 50+ _____

3 Country of Origin: ________________________________

4 Number of years of teaching experience: _____

   a. Where: __________________________ b. What levels: ______________

     c. Age group of the students: ____________________

5 Student Status: undergraduate _____ graduate _____ professional _____

158
6 How much college course work in evaluation, assessment, measurement, or testing have you completed? (circle one)

A none at all  
B one course  
C part of one or more courses  
D two or more courses

7 What grade levels do you plan to teach when you finish your studies?

K-6 _____ 7-12 _____ University _____

8 How much control do you think you will have over assessment decisions?

100% ____  80% ____  50% ____  30% ____  10% ____  0% ____

PART I
Rate the following reasons for assessment in foreign language classes:
(not important) 1 2 3 4 5 (very important)

1 Find out what the students do not know 1 2 3 4 5

2 Provide students with feedback 1 2 3 4 5

3 Satisfy an administrative directive 1 2 3 4 5

4 Check students’ ability to apply knowledge 1 2 3 4 5

5 To keep students on task 1 2 3 4 5

6 Monitor student progress over time 1 2 3 4 5

7 Compare student performance to others 1 159

2 3 4 5
8 Help students organize their learning  | 1  2  3  4  5
9 Identify students’ misconceptions  | 1  2  3  4  5

**PART II**

*Indicate your understanding of the following assessment terms:*

<table>
<thead>
<tr>
<th>(not familiar)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 (very familiar)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Recorded informal observations</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2 Student writing journals</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3 Student attitude inventories</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4 Performance-based assessment tasks</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5 Student self-assessment reports</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6 Portfolio assessment</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>7 Scoring/rating scales</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

**PART III**

*Based on your experience with portfolios this quarter, would you be willing to implement portfolio assessment in future classes that you teach? Please explain why or why not.*
APPENDIX C

PORTFOLIO GUIDELINES

The creation of a portfolio is intended to allow the student to demonstrate professional development in an individual manner. The portfolio must include varied documentation that shows interaction with, exploration of, and reflection about the teaching of foreign languages at the elementary school level. Each student will decide how he/she wants to demonstrate this reflection, interaction, and exploration through the submissions included in the portfolio. The portfolio will be collected for evaluation at midterm and again at the end of the quarter (see attached criteria for evaluation). The midterm portfolio does not need to include any finished products but it should demonstrate progress in many of the activities.

The first page of the portfolio is a letter to the reader that explains who you are, what he/she will find in your portfolio, and why the portfolio is a reflection of your development as a future foreign language teacher. Each submission must include a reflection by the student that states why the student chose to include that activity in the portfolio and a self-evaluation of the quality of the work. Each student must include six submissions (two from each category) to the portfolio. The following are suggestions for items that could be included in the portfolio:

Library-based Items:
1. Choose a topic of interest to you in the field of foreign language instruction at the elementary school level and create an annotated bibliography of 3-4 articles that deal with that theme. The annotations should be approximately 50 words in length.

2. Do an ERIC search of current research (1993-1996) in the field of FL instruction at the elementary school level. Print out the abstracts of 10-12 articles under one theme (eg, FLES, Immersion) and write a one page reflection on the type of research that is being done.
3. Do a brief research report on a topic of interest not covered in the class (e.g., L1 reading instruction in immersion schools, the use of authentic materials in elementary schools). This submission should be 2 pages long with at least two references.

4. React to the English Only movement. Include at least three citations from other authors to support your opinion.

5. Review a book that deals with FL instruction at the elementary school level. The submission should be 2 pages and should conform to Modern Language Journal guidelines for book reviews.

6. Collect several readers that are used to teach literacy skills in the elementary school, and review the material included. Limit: 2 pages.

7. Create a brief annotated bibliography of 5-7 children's books that are appropriate in content and level for a particular grade (choose K-6). Explain why and how you would use these books in an elementary classroom.

**Instruction-based Items:**
1. Plan a fantasy experience. Create the materials and include a lesson plan of activities.

2. Create the materials for a unit plan—include a brief outline of the lessons/use of the materials in the classroom.

3. Create an assessment activity, either a traditional test, or a game that evaluates learning. Include background information on lessons leading up to evaluation, and rationale behind the approach that you chose.

4. Create a cultural unit (2-3 lessons) that could be presented to an L1 class to familiarize them with another culture. Include the lesson plans and a description of the materials that could be used.

5. Create a learning activity/game that incorporates at least one of the approaches presented in the text (e.g., TPR, Gouin series). Include the lesson plan and materials.
6. Create a lesson based entirely on multimedia (computer program, video, etc.) Explain how you would incorporate the multimedia materials into the lesson and why you chose the materials that you did.

7. Create the materials to accompany the unit plan that you created with your learning group. Include a brief explanation of the use of each of the materials, and their incorporation into the unit plan.

8. Translate a children’s book from English to your L2. Write a lesson plan that incorporates the use of this book. Describe or create accompanying materials. Explain why you chose this book.

Field-based Experience:
1. Interview a teacher in the field. Focus on specifics, like management/discipline strategies, lesson planning, etc. Report findings in an informal paper of 2 pages.

2. Observe young children at play and write up your notes on their social skills, speech, play habits. Reflect on the cognitive and social aspects of their development.

3. Visit various (2-3) sites for materials and resources (e.g., Lau Center downtown, Teacher’s Helper stores, and others) and write up what materials you were able to find and how you might use them in teaching.

4. Do a brief case study on one child. Observe instances of speech and writing/reading. Reflect on your observations of the cognitive development of a young child. Compare L1 acquisition to learning a foreign language. 2-3 pages

5. Write up a rationale for FL instruction in the elementary school (2 pages). Include a letter to the parents of the children and an outline of the arguments you would make before the school board.

6. Interview an expert in the field on an issue that interests you. Consult Curtain & Pesola to identify an important issue, determine current lines of thinking and summarize findings. Write up summary of interview and react to findings. 2 pages
APPENDIX D

ANALYTICAL RATING SCALE

Each submission and anecdotal report is to be rated on a Likert scale of 1-6 with 1 being the lowest score and 6 being the highest. Please follow the following categories in your rating of each submission and anecdotal report.

Submission # ______ of Portfolio # ______

Presentation of submission: 1 2 3 4 5 6
Completeness of submission: 1 2 3 4 5 6
Organization of submission: 1 2 3 4 5 6
Quality of submission: 1 2 3 4 5 6

Anecdotal report

Completeness of anecdotal report: 1 2 3 4 5 6
Quality of anecdotal report: 1 2 3 4 5 6
APPENDIX E

HOLISTIC RATING SCALE

Portfolio Number:__________

Criteria for Evaluation:

1. Content: ________ /10 (2 pts each)
   ______ a. Portfolio contained submissions of various types (library, instructional, and field work)
   ______ b. Each submission contained an anecdotal report of why that item reflects development on the part of the student, and the student’s self-evaluation of the item
   ______ c. Portfolio began with the welcome letter to inform the reader about the owner of the portfolio, the contents, and why they represent professional growth
   ______ d. Portfolio is well-organized, labeled, and easy to read
   ______ e. Established limits and guidelines for each item were followed

2. Quality: ________ /10 (2 pts each)
   ______ a. Portfolio was done with thoroughness and care; submissions show that owner intended to explore new aspects of teaching profession
   ______ b. Submissions were well-formatted, with evidence of careful preparation and substantial thought
   ______ c. Submissions were well-detailed reflecting continuous activity on them over a period of time
d. Evidence of editing and appropriate use of academic style

e. Evidence of careful reflection on personal growth by the student in the anecdotal report

3. Professional Development: ________ /5 (2.5 pts each)

   a. Evidence that owner attempted to extend professional network
   b. Evidence that owner shows an awareness/appreciation of resources available

Total: ________ /25

Comments:
APPENDIX F

PERFORMANCE TASKS

1. Videotaped game/learning activity: (individual assignment)  5% of grade
   Develop or borrow (with appropriate reference) a game/activity to do with class. Provide
   the following on a handout.
   a. State type of class and grade level (i.e., immersion, grade 5)
   b. Learning objectives: how activity reinforces basic concepts of the curriculum;
      specific L2 outcomes; thinking skills involved (Bloom’s taxonomy)
   c. Give language functions practiced with examples
   d. Describe the game/activity
   e. Provide a script for the teacher’s role in introducing and conducting the
      game/activity; demonstrate how the teacher surrounds the activity with language to
      provide comprehensible input
   f. Ensure a communicative orientation to the game
   g. Provide photocopy of reduced visuals

2. Lesson plan (individual or pair assignment)  10% of grade
   Develop a 30-minute lesson plan. Pay particular attention to the match of your activities to
   the stage of cognitive development of the students. Be sure to include some small group or
   pair activity. Prepare materials and visuals. Include a means of checking for
   comprehension. Must be a different grade and program model from game/activity above.
   a. State type of class and grade.
   b. State objectives
   c. Identify prerequisite learning
   d. Identify new subject matter
   e. State procedures, giving sequence and description for each activity
   f. Identify language functions with examples
g. Materials and visuals: bring photocopies of reduced visuals; list realia or props

h. Describe how you would subsequently evaluate learning.

3. Reactions to readings (individual assignment) 10% of grade

Three times during the quarter, prepare a typed, one-page, single-spaced reaction statement based on the course readings. Include your “big learnings”, “ahas”, remaining questions, etc.

4. Content-based FLES or ESL unit (learning group assignment) 25% of grade

Develop a list of six to ten activities on a selected theme. There should be an experiential component (e.g., arts and crafts, physical activity, music, motion, science, etc.) and a linguistic component (e.g., language structures, vocabulary). Describe each activity briefly and indicate the skills and concepts that would be developed (from core curriculum of grade level and from thinking skills. Plan to present three of the activities to the class.

   a. Plan 6-10 activities through thematic webbing. All activities briefly described.
      Three activities fully described.
   b. Indicate language structures and vocab used in activities.
   c. State concepts from core curriculum used in web.
   d. State linguistic outcomes for each activity.
   e. State thinking skills necessary for each activity.
   f. Describe experiential activity for at least one activity.
   g. Prepare materials, visuals for class presentation.

During presentation, you will briefly describe web. Each group member will then have five minutes to explain each activity in detail.
APPENDIX G

FINAL EXAM

Throughout this quarter, we have seen how communicative language use guides curriculum development in the field of second language acquisition. In order to determine whether students can use language for meaningful development, we have also discussed alternative assessments such as observation, portfolios, and performance-based assessments. During the course of this final exam, you will demonstrate your understanding of the field of childhood second language acquisition by performing the task delineated below.

Task: Individually, in partners, or in groups, write a rationale for FLES or ESL to be presented to the joint curriculum council/school board of a fictitious town.

In your rationale, briefly explain the program model you wish to support based on the community you envision. Include why the programs of the 60s failed and how current programs are more consonant with second language acquisition research. By way of explanation, citations, and examples, lead the school board/council to adopt your program.

Hints: In order to prepare this rationale, you will have to review the objectives on the course syllabus and use citations from the class lectures and the readings. You must include the following concepts somewhere in your rationale: Bloom’s taxonomy of thinking skills, Halliday’s language functions, grouping strategies, whole language, cognitive abilities/developmentally appropriate instruction, content-enriched/based instruction, assessment strategies, and second language acquisition research terms.

The exam must not exceed ten double-spaced type written pages. Be concise but thorough and choose citations carefully.

20% of the final grade
APPENDIX H

PORTFOLIO INTERVIEW QUESTIONS

1. What was your overall reaction to the use of portfolio assessment in the methods course?

2. What abilities did you feel the portfolio measured in terms of your knowledge?

3. Did you feel that the portfolio allowed you to demonstrate knowledge that was qualitatively different from the knowledge used in answering the question on the final exam?

4. How would you describe your portfolio in terms of what it says about you? Does it accurately demonstrate your abilities as a foreign language teacher?

5. What were some of the things you really liked about the portfolio? What were some of the things you were not pleased with?

6. Most people answered on the survey that they would be willing to use portfolios with future students, how would you implement this tool into your classrooms? Do you actually see yourself using portfolios at some level with students?