2008

Does content knowledge matter in scoring teacher work samples?: A study of rater differences

Yana A. Cornish
University of Northern Iowa

DOES CONTENT KNOWLEDGE MATTER IN SCORING TEACHER WORK SAMPLES? A STUDY OF RATER DIFFERENCES

A Dissertation

Submitted

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

Approved:

_____

Dr. Robert Boody, Chair

_____

Dr. Victoria Robinson, Co-Chair

_____

Dr. Maria Basom, Committee Member

_____

Dr. William Callahan, Committee Member

_____

Dr. John Henning, Committee Member

Yana A. Cornish

University of Northern Iowa

May 2006

ACKNOWLEDGEMENTS

I am eternally grateful to my committee members and, especially, to my co-chairs, Dr. Robert Boody and Dr. Victoria Robinson, for their support, encouragement and for sharing their wisdom and expertise. Special recognition goes to Dr. Boody who was my conscience and encouragement in this process; his office doors were always open for me and he supported me all the way. Additionally, I am very thankful for the support of the dissertation efforts by the UNI College of Education and its' Dean who is also my committee member – Dr. William Callahan. It is hard to mention all the assistance that I have received from Dr. John Henning and Dr. Maria Basom, who were always ready to help me look critically at my work and guided me in my thinking and writing.

I could not have done this work without the support of my family – my husband, Garth, and my mother, Luba, who assumed my share of taking care of our three young children, all born during my dissertation work. I would not have succeeded without my family's support. I am especially thankful to my husband for always being there for me, continuing to offer his never ending love, for providing a sympathetic ear, and for encouraging me to go on. I am also thankful to my three babies, Maya, Lukas, and Evan, for being so good and cooperative and letting mama do her thing.

Special recognition and a deep thank you goes to my supervisor, Dr. Timothy O'Connor, for his understanding, flexibility and support of my academic work. I would also like to thank Dr. David Else and his office staff, for providing a quiet space for me to write and edit.

I wish to thank several people from the Renaissance Partnership for Improving Teacher Quality Project, especially, Dr. Roger Pankratz, Dr. Peter Denner, and Ms. Gaye Pearl, for all their assistance, support, encouragement, and for sharing their work with me. I am thankful to the Iowa teachers who responded to my call and agreed to participate in my study.

Lastly, thanks goes to many others who helped me along this long journey of working on my dissertation, for their support, a word of advice, a helping hand here and there, and sometimes just for their encouraging smile.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DOES CONTENT KNOWLEDGE MATTER IN SCORING TEACHER WORK SAMPLES? A STUDY OF RATER DIFFERENCES

An Abstract of a Dissertation

Submitted

in Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

Approved:

_____

Dr. Robert Boody, Committee Chair

_____

Dr. Sue Joseph, Dean of the Graduate College

Yana A. Cornish

University of Northern Iowa

December 2006

ABSTRACT

The study examines the role of rater characteristics in the assessment of teacher practice as presented in Renaissance Teacher Work Samples (RTWS). The study analyzed ratings of 10 teacher work samples submitted by teacher candidates at the University of Northern Iowa between Fall 2000 and Spring 2004. These teacher work samples, created in the area of Spanish language learning at 7-11 grade levels, were analyzed to determine the impact of content expertise, amount of teaching experience, and previous RTWS rating experience on reviewer's ratings. Three study questions form the foundation for the investigation:

1. Is there a significant difference between ratings assigned by raters with foreign language content experience and raters without foreign language content experience? Does this differ by sections of the teacher work sample?
2. What is the relationship between rater's overall teaching experience and his/her scoring of foreign language teacher work samples?
3. What is the impact of a rater's work sample scoring experience on his/her scoring of foreign language work samples?

In order to address these questions, the study used a causal-comparative research design. Dependent variables in this study were scores of work samples reported by 30 raters; independent variables were presence or absence of foreign language content expertise, as well as other demographic characteristics, such as (a) the amount of foreign language teaching experience, (b) amount of teaching experience, (c) experience with scoring work samples, (d) gender, and (e) level of education. The

study employed two instruments: (a) a demographic questionnaire and (b) a RTWS scoring rubric.

The investigator recruited 30 participants from various middle and high schools in Iowa. Sixteen of these participants were foreign language teachers, while the remaining fourteen were educators teaching in various content areas other than foreign language or ESL. Participants of the study were asked to participate in a Renaissance Teacher Work Sample training and scoring session, rating 10 Spanish work samples submitted by UNI teacher candidates.

The analysis of the demographic data revealed that participants varied greatly in almost all the areas of the questionnaire. The TWS data analysis contributed to the further understanding of the participants' rating process and outcomes, their scoring speed, and allowed to answer the questions of the study.

The findings of the study indicate that there is no statistically significant difference between the ratings of the Spanish teacher work samples reported by the participants in the study. Thus, the study did not find any statistically significant impact of the rater characteristics on scorers' perception of teacher practice as presented in the Spanish teacher work samples. The findings of the study support other validity and reliability studies of the RTWS methodology and instrumentation. Additionally, the study outlines several areas for further research.

CHAPTER I

INTRODUCTION

For years educators, policymakers, and the community at large have been discussing issues of quality in American education and the preparedness of new generations to join the professional world. This issue has become even more pressing since the employment market, developing with ever accelerating speed, has become more global and internationally competitive. Concerns regarding the quality of American education increased substantially after the publication of *A Nation at Risk* by the National Commission on Excellence in Education (1983) which warned that "a rising tide of mediocrity [in our schools] threatens our very future as a nation" (p. 5). Shortly after the release of the report, federal and state governments began a mission to fix America's schools, introducing and passing new laws focused on improving the quality of public education. One of the examples of federal efforts in this direction was *Goals 2000: Educate America* Act, signed into law by President Clinton on March 31, 1994 (U.S. Department of Education, 1994). The goal of the Act was "to encourage local community-based actions that meet pressing educational needs, help more students achieve to higher standards, increase parental participation, and improve teaching" (U.S. Department of Education, 2001). In 2001, the federal government stopped funding *Goals 2000* programs, and President Bush secured passage of the *No Child Left Behind* Act (NCLB; U.S. Department of Education, 2002a). The new law reflected current concerns regarding the quality of American education and provided a framework for

improving the performance of America's elementary and secondary schools and presenting all children with quality learning opportunities.

In spite of all the federal and local efforts to improve American education and the billions of dollars spent on these programs over the years, research suggests that student achievement, as measured by standardized tests, is experiencing very modest, if any, gains. As stated in the recent report, *National Assessment of Educational Progress* (National Center for Education Statistics, 2003), no significant changes in average reading scores for fourth-graders were detected when compared with average score data collected in 1992. According to Symonds (2001):

> Less than half of America's fourth-, eighth- and twelfth-graders read at a proficient or advanced level. For fourth-graders, the figure is only 32 percent, with black students faring the worst. Just 12 percent of them read at grade level. And by 12th grade, U.S. students score well below teenagers in almost every other developed country on math and science tests. (pp. 99-100)

Several major research studies (e.g., Darling-Hammond, 1997; National Commission on Teaching and America's Future, 1996; U.S. Department of Education, 1997) offer compelling evidence that teacher quality is the single most important factor that affects student achievement. Darling-Hammond argues that investing in high-quality teaching is one of the most important approaches to improving schools and raising student achievement. Therefore, quality of education is closely tied with teacher quality. Based largely on this belief, leading professional education organizations, since the mid-1980s, have been developing standards for specific content areas and for teacher practice in general.

In addition, guided by research, the NCLB Act (U.S. Department of Education, 2002a) emphasizes placing "highly qualified" teachers in the classrooms and calls for more accountability. Accountability is viewed as connecting individual schools and teachers to student performance and linking the quality of teacher preparedness to teacher preparation programs. The Act also called for implementation of subject area and professional education exams for teachers. With this new law, the nature of teacher licensing and certification is changing, introducing more rigorous, standards- and/or performance-based evaluation of teacher candidates before they are allowed to teach. In addition to more rigorous standards for teacher candidates, states are working on increasing quality of in-service teachers; for example, by promoting National Board of Professional Teaching Standards (NBPTS) certification. More and more in recent years, perception of quality of in-service teachers and "good teaching" is equated to and measured through increased student achievement.

In order to increase accountability regarding teacher quality in general and in teacher preparation, universities and state departments of education have been searching for effective tools to assess effectiveness of new teachers graduating from teacher preparation programs. The majority of states, 80 percent, have chosen to use paper-and-pencil tests, like PRAXIS II, to measure content and pedagogical knowledge of teacher candidates (McAllister, 2003). Research (Wilson, Floden, & Ferrini-Mundy, 2001) tells us that teachers need both content knowledge and knowledge of pedagogy, because "while an academic major guarantees that teachers know the subject, it does not

guarantee that they know how to teach that same subject to children" (Cross & Rigden, 2002, p. 25).

Pedagogy is a complex concept that refers to "the pedagogical (teaching) skills teachers use to impart the specialized knowledge/content of their subject area(s)" (National Board for Professional Teaching Standards [NBPTS], 1998). Effective teachers display a wide range of skills and abilities that lead to creating a learning environment where all students feel comfortable and are sure that they can succeed both academically and personally. Due to its complexity and ties to performance, content pedagogy is difficult to assess with a paper-and-pencil test; therefore, several states selected a different route and some employ Teacher Work Sample Methodology (TWSM). It is important to mention, that even in states that require paper-and-pencil tests, like PRAXIS II, or even more complex assessments, like PRAXIS III, which is portfolio/performance-based, many teacher preparation programs supplement them with portfolios and/or other performance assessments.

There is a variety of other approaches, most of which are either a "portfolio approach" or an "applied performance approach." TWSM is one of the latter, originally formulated at the Western Oregon University in the 1980s, that requires creation of teacher work samples (TWS) by teacher candidates to demonstrate "their professional skills including their ability to foster pupil learning" (Girod, 2002, p. xi). TWSM has been adapted and used by many teacher preparation programs in the nation (for example, the Renaissance Partnership for Improving Teacher Quality Project, n.d.a) to evaluate quality of teacher candidates. TWSM has also been employed by several state

departments of education as an assessment mechanism for first and second year practicing teachers (for example, Oklahoma and Oregon). The TWS Methodology is typically used as a basis for a teacher candidate evaluation system and as a way to ensure accountability and increase teacher quality in teacher preparation. Pankratz states that "the work sample methodology provides direct evidence of a teacher candidate's effect on student learning in a relatively short time period and clearly connects the elements of standard-based teaching and learning" (1999, p. 37). Schalock and Myton further expand on the TWSM connection of teaching and learning by stating that "teacher work sampling assesses the effectiveness of teachers close to their work... ...[and it is] a quality assurance system that holds student learning at its core" (2002, p. 11). Overall, teacher work sampling provides, with a greater degree of validity than traditional paper-and-pencil-based tests, information regarding teacher candidate's readiness to teach effectively focusing on improving student learning, making it a unique and effective assessment tool defining good teaching though improved student learning, that sets it apart from the NBPTS certification, Interstate New Teacher Assessment and Support Consortium (INTASC), and PRAXIS III (Henning & Robinson, 2004; Schalock, Schalock, & Myton, 1998; Girod, 2002).

Moreover, some scholars consider TWSM to be both a process and a product (Henning & Robinson, 2004); or, in other words, a vehicle for instruction as well as an approach to measurement (Girod, 2002). Girod also notes that TWSM "is a vehicle that helps perspective teachers learn to think about teaching in ways that are linked tightly and continuously to pupils' learning, to gain experience in teaching in this manner, and

to demonstrate effectiveness in doing so" (2002, p. 1). In teacher preparation and initial licensing, teacher work sampling can serve as:

1. A *model* for thinking about teaching and learning;
2. A *frame of reference* for designing and operating teacher preparation programs that systematically connect teaching and learning;
3. A *vehicle for practicing and obtaining feedback* on one's effectiveness as a teacher in fostering pupils' progress in learning (formative evaluation);
4. A *methodology for demonstration and documenting* one's effectiveness in fostering learning gains by pupils (summative evaluation), and
5. A *source of evidence to be used in recommending and granting* a license to teach. (Schalock & Myton, 2002, pp. 12-13)

University of Northern Iowa was one of eleven higher education institutions-partners in the Renaissance Partnership for Improving Teacher Quality Title II Consortium Grant located in California, Idaho, Iowa, Kentucky, Kansas, Michigan, Missouri, Pennsylvania, and Virginia (The Renaissance Partnership for Improving Teacher Quality Project, n.d.a). This five-year project, originated in 1999, has adapted the TWSM and developed its own version of the teacher work sample that included: (a) performance prompt, (b) teaching process standards, and (c) a scoring rubric. The Renaissance Project Teacher Work Sample (RTWS) is organized around seven teaching processes:

1. Contextual factors – description of the school and surrounding community that would include a demographic description of the group of students and any other relevant factors, which may be impacting student learning.
2. Learning goals – provides a list of challenging and appropriate learning goals to be addressed in the unit described in the work sample. These goals should be aligned with national, state, or local standards.
3. Assessment plan – contains multiple pre- and post-assessment measures that were employed in the unit for formative and summative assessments of student learning. These assessments should be aligned with unit learning goals.
4. Design for instruction – provides a summary of instructional methods used by the teacher candidate to help students meet learning goals for the unit. The instruction should take into consideration various student needs.
5. Instructional decision-making – this section describes formative assessment measures used by the teacher candidate to make changes to instruction based on student learning.
6. Analysis of student learning – contains analysis of student data collected by the teacher candidate. The teacher candidate is expected to comment on why individual students and groups of students were successful or less than successful in learning the material of the unit.
7. Self-evaluation and reflection – is a teacher candidate's reflection on the effectiveness of his/her teaching and attempts to improve student learning of all students. Candidates are expected to propose future activities that would be

effective in helping all students meet unit learning goals (The Renaissance Partnership for Improving Teacher Quality Project, n.d.a).

During the life of the project, over 2,500 teacher work samples have been submitted by teacher candidates from all the partner institutions combined. At the University of Northern Iowa alone 572 teacher work samples were submitted by teacher candidates from eight teaching centers. The University of Northern Iowa continues to use RTWS as an integral part of its teacher preparation program.

Teacher work samples are compiled by UNI teacher candidates during their student teaching experience and are about 20-25 pages in length. Each student teacher describes his/her activities during a period of 2-3 weeks following the Renaissance Teacher Work Sample Prompt and Rubrics. In addition, teacher candidates submit examples of assessments used during the course of instruction, accompanied by samples of student work, and conclude their TWS documentation with an extensive reflection. Later, these samples are scored by trained educators from UNI and K-12 schools in the Cedar Falls/Waterloo and surrounding areas using Renaissance Teacher Work Sample Rubrics. Each section of the work sample receives separate ratings following a three-point scale: 1 – standards were not met; 2 – standards were partially met; and 3 – standards were met. Additionally, each sample receives an overall score following the same three-point scale.

Because of the current focus on standards-based education and interest in a performance-based teacher assessment, the potential of teacher work sampling is great. Data generated by the work samples provides a variety of insights into knowledge and

skills of specific prospective teachers as well as contributes to the overall accountability of a teacher preparation program. Moreover, RTWS is used by a number of teacher preparation programs to make high-stake decisions regarding their teacher candidates. The authors of a number of research studies, examining various reliability and validity matters of RTWS, further voice their beliefs that any high-stake assessment should be thoroughly scrutinized:

> Institutions using performance assessments for high-stakes decisions are also faced with the challenges of showing the evidence derived from these assessments is valid and credible. As noted by Popham (1997), assessments used for high-stakes decisions such as program admission and certification or licensure must be accompanied by rigorous studies of the credibility of evidence including the validity of the assessment and the reliability of scoring decisions. (Denner, Salzman, & Harris, 2002, p. 2).

Although various aspects of TWS have been researched, many more remain to be studied. For example, the role of rater characteristics in assessment of teacher practice as documented in teacher work samples, specifically if teacher work samples are written in content-specific areas, e.g., foreign languages, is one of the areas that has not been studied. Such studies will contribute to the overall credibility of the assessment.

Statement of the Problem

Teacher Work Sample Methodology (TWSM) and its variations, like the Renaissance Partnership Teacher Work Sample (RTWS), are described as an applied performance approach that "links preinstructional planning, conduct of the instructional process, and subsequent reflection with a strong emphasis on assembling and analyzing data about student learning and growth" (Imig & Smith, 2002, pp. ix-x). When used as

a vehicle for instruction and assessment, TWSM "has been designed to portray the learning progress of pupils on outcomes desired and taught by a teacher over a sufficiently long period of time for appreciable progress in learning to occur" (Girod, 2002, p. 1). In order for the methodology to be used for matters of teacher candidate performance accountability and assessment of improved student learning, additional research needs to be carried out to address questions regarding procedures and factors that may have impact on scoring of teacher work samples, thus, addressing reliability and validity of the assessment.

Some aspects of TWSM have been studied in recent years. In addition to extensive field testing of the instruments and establishment of benchmarks, a number of research studies have been carried out to test content validity, score generalizability, quality of student learning assessment, and alignment with standards of the teacher work sampling (e.g., Denner, Norman, Salzman, Pankratz, & Evans, 2003; Denner, Salzman, & Bangert, 2001; Denner, Salzman, & Harris, 2002; McConney, Schalock, & Schalock, 1998; Salzman, Denner, Bangert, & Harris, 2001). The findings indicate direct correspondence between the targeted teaching behaviors and actual teaching practice, support the generalizability of the work sample scores and high dependability coefficients for panels of three or more raters, reveal positive correlation of TWS student assessments with ratings on an independent scale, and demonstrate close alignment with evaluation standards.

In spite of the studies carried out in the area of TWSM (e.g., McConney, Schalock, & Schalock, 1998) and RTSM (Denner et al., 2001; Denner, Norman,

Salzman, Pankratz, & Evans, 2003; Fredman, 2004; Salzman et al., 2001), some issues need further examination, especially regarding TWSM use with student teachers majoring in specific content areas, for instance, in the area of Foreign Language Learning. One of the issues still lacking empirical support deals with a role of rater characteristics, for example, content knowledge, amount of teaching experience, and previous TWS rating experience, in their assessment of teacher practice as defined by the Teacher Work Sample Methodology.

Purpose of the Study

This study analyzed ratings of 10 teacher work samples submitted by teacher candidates at the University of Northern Iowa between Fall 2000 and Spring 2004. These teacher work samples, created in the area of foreign language learning (Spanish) at 7-11 grade levels, were analyzed to determine the impact of content expertise, amount of teaching experience, and previous TWS rating experience on reviewer's rating. Three study questions, based on the statement of the problem, form the foundation for the investigation:

1. Is there a significant difference between ratings assigned by raters with foreign language content experience and raters without foreign language content experience? Does this differ by sections of the teacher work sample?
2. What is the relationship between rater's overall teaching experience and his/her scoring of foreign language teacher work samples?
3. What is the impact of a rater's work sample scoring experience on his/her scoring of foreign language work samples?

Methodology

In order to address these questions, the study used a causal-comparative research design. Dependent variables in this study are scores of work samples; independent variables are presence or absence of foreign language content expertise as well as other demographic characteristics, such as (a) the amount of foreign language teaching experience, (b) amount of teaching experience, (c) experience with scoring work samples, (d) gender, and (e) level of education.

The investigator recruited 30 participants for this study from various middle and high schools in the state of Iowa. Sixteen of these participants were foreign language teachers, while the remaining fourteen were educators teaching in various content areas other than foreign language teaching or ESL. Participants of the study were asked to participate in a Teacher Work Sample training session and later rate foreign language work samples submitted by UNI teacher candidates.

Definition of Terms

Candidate performance data: Information derived from assessments of teacher candidate proficiencies, in areas of teaching and effects on student learning, candidate knowledge, and dispositions. Candidate performance data may be derived from a wide variety of sources, such as projects, essays or tests demonstrating subject content mastery; and work samples as well as assessments, projects, reflections, clinical observations, and other evidence of pedagogical and professional teaching proficiencies (Indiana State University, n.d.).

Clinical faculty: School and higher education faculty responsible for instruction, supervision, and assessment of teacher candidates during field experience and student teaching (Indiana State University, n.d.).

Content and content area: The subject matter or discipline that teachers are being prepared to teach at the elementary, middle, and/or secondary levels. Content also refers to the professional field of study (e.g., special education, early childhood, foreign language, school psychology, school administration, etc.; Indiana State University, n.d.).

Content standards: represent what students should know and be able to do (Pritchard, 1996). "Content standards" describe what students will learn and teachers will teach within an academic discipline (Ravitch, 1995a, p. 12).

Foreign language instruction: a school subject which usually does not employ the foreign language as a medium of instruction, studied usually either for communication with foreigners who speak the language, or for reading printed materials in the language (Richards, Platt & Platt, 1993, pp. 142-143).

Iowa initial license: The initial license is issued to the graduates of approved teacher education programs. It is valid for two years, and it may be renewed for one additional two-year term. Any new graduate who has received a teaching contract is required to have this license. The license includes two parts: License area(s) and School setting(s). School settings are: Preschool, Elementary: Primary, Elementary: Intermediate, Middle School/Junior High, High School (Iowa State University, n.d.).

INTASC: The Interstate New Teacher Assessment and Support Consortium that has developed model performance-based standards and assessments for the licensure of teachers (Indiana State University, n.d.).

In-service teacher: Practicing K-12 educator.

Language acquisition: The process of learning language, usually in a subconscious manner as in learning one's native language. This process is often contrasted to "language learning," which refers to the conscious focus on knowledge and applying rules, as in a formal classroom situation (National Council for Accreditation of Teacher Education [NCATE], 2003, p. 91).

NBPTS: The National Board for Professional Teacher Standards, an organization of teachers and other educators, which has developed both standards and a system for assessing the performance of experienced teachers seeking national certification (NCATE, 2003, p. 91).

NCATE: National Council for Accreditation of Teacher Education is a professional accrediting organization that is recognized by the U.S Department of education as the accrediting body for colleges and universities that prepare teacher and other professional personnel for work in elementary and secondary schools (State University of New York College at Cortland, n.d.).

Opportunity-to-learn standards: represent the conditions and resources necessary to help students achieve the performance standards (Pritchard, 1996).

Oral Proficiency Interview (OPI): A standardized procedure for the global assessment of oral proficiency. It measures language production holistically by

identifying patterns of strength and weakness within the assessment criteria of function, contexts, and accuracy. The official OPI is administered by Language Testing International (LTI), a central testing service with has procedures in place for validating the ratings (NCATE, 2003, pp. 91-92).

Pedagogical content knowledge: The interaction of the subject matter and effective teaching strategies to help students learn the subject matter. It requires a thorough understanding of the content to teach it in multiple ways, drawing on the cultural backgrounds and prior knowledge and experiences of students (Indiana State University, n.d.).

Pedagogical knowledge: The general concepts, theories, and research about effective teaching, regardless of content areas (Indiana State University, n.d.).

Performance assessment: A comprehensive assessment through which candidates demonstrate their proficiencies in subject, professional, and pedagogical knowledge, skills, and dispositions, including their abilities to have positive effects on student learning (Indiana State University, n.d.).

Performance-based program: A professional preparation program that systematically gathers, analyzes, and uses data for self-improvement and candidate advisement, especially data that demonstrate candidate proficiencies, including positive effects on student learning (Indiana State University, n.d.).

Performance criteria: Descriptions or rubrics that specify qualities or levels of candidate proficiency that are used to evaluate candidate performance (Indiana State University, n.d.).

Performance standards: represent the levels of learning students should attain in relation to the content standards (Pritchard, 1996). "Performance standards" define the knowledge and proficiency requirements expected of students upon completion of specific levels of instruction (Levin, 1998, p. 4). These standards define levels of attainment and describe what kinds of performance characterize insufficient, sufficient, or outstanding achievement (Ravitch, 1995a, p. 12-13).

Portfolio: An accumulation of evidence illustrating individual skills, abilities, proficiencies, and performance, especially in relation to explicit standards and rubrics, used in the evaluation of one's competency as a teacher or in another professional school role. Contents might include end-of-course evaluations and tasks used for instructional or clinical experience purposes such as projects, journals, and observations by faculty, videos, or comments by cooperating teachers or internship supervisors, and samples of student work (NCATE, 2003, p. 92).

Pre-service teacher: – see teacher candidate.

Standards: Written expectations for meeting a specified level of performance (Indiana State University, n.d.).

Reliability: The extent to which a measurement instrument yields consistent, stable, and uniform results over repeated observations or measurements under the same conditions each time (Juvenile Justice Evaluation Center Online [JJECO], n.d.).

Renaissance Teacher Work Sample rater training: For all work sample raters, the training typically consists of two hours of a review of the teaching processes and standards targeted by the RTWS assessment, examination of the relationship between

the standards and the RTWS components, instruction on how to use the scoring rubrics to rate TWS performances, and anti-bias training (based on procedures described in Denner, Salzman, & Bangert, 2001; Denner, Norman, Salzman, Pankratz, & Evans, 2003).

Rubrics: Written and shared criteria for judging performance that indicate the qualities by which levels of performance can be differentiated, and that anchor judgments about the degree of success on a candidate assessment (Indiana State University, n.d.).

Teacher candidate: individual admitted to, or enrolled in, a program for the initial or advanced preparation of teachers (NCATE, 2003, p. 89).

Teacher Work Sample: the product [teacher candidates] develop to demonstrate a significant portion of their professional skills including their ability to foster pupil learning (Girod, 2002, p. xiii).

Teacher Work Sampling: the assessment strategies and materials associated with teacher work samples (Girod, 2002, p. xiv).

Teacher Work Sample Methodology: An applied performance approach that can be tailored to: (a) learning goals, (b) teaching style, (c) group & individual student needs, and (d) the context of the classroom, school, & community (Western Oregon University, n.d.).

Teacher Work Sample rater: Administrators, faculty members or teachers who participated in RTWS training and scored teacher work samples.

Teaching standards: written criteria for making judgments about progress toward the vision; they describe what teachers at all grade levels should understand and be able to do (The National Academies Press, n.d.). The teaching standards are the educational experiences teachers should provide inside and outside the school environment (Pritchard, 1996).

Validity of assessment results: The extent to which a measure accurately reflects the concept that it is intended to measure (International Foundation for Functional Gastrointestinal Disorders [IFFGD], n.d.).

CHAPTER II

REVIEW OF LITERATURE

The primary purpose of this study was to investigate the role of rater characteristics, such as content knowledge, amount of teaching experience, and previous Teacher Work Sample rating experience, in a rater's assessment of teacher practice. The study assesses teacher practice as presented in teacher work samples submitted by teacher candidates majoring in foreign language teaching at the University of Northern Iowa between fall 2000 and spring 2004. Chapter II provides a literature review of important issues related to the focus of the study. The review is organized into three major sections: (a) accountability in education, (b) assessment methodologies of teacher practice, and (c) foreign language teaching and learning in the United States.

A section on accountability in education offers a historical overview of the last two decades of efforts for increase in teacher accountability and teacher quality. A section on assessment methodologies of teacher practice includes an overview of several approaches used for assessment tools, with a focus on the Teacher Work Sample Methodology and its components, providing examples of how teacher work samples are used in teacher preparation programs and state efforts for increase in quality of teacher candidates and accountability of in-service teachers. A section on foreign language teaching and learning includes a brief historical overview of methods used to teach foreign languages in the Unites States, describes foreign language education in public schools, as well as discusses new developments in foreign language instruction and teacher preparation.

Accountability in Education

Recently matters of accountability and high quality instruction have become the focus of attention of educators, policymakers, and the general public. Individual institutions, professional organizations, and state and federal structures have developed a number of initiatives to increase student achievement and ensure accountability at all levels. The drive for increasing teacher quality that would result in improved student learning is not easy and many educators are working hard to meet federal, state, and in some cases, local goals to provide all children with quality teachers and opportunities to succeed in meeting rigorous academic standards.

The terms "accountability" and "standards" increasingly become the focus of educational practices at all levels. In education, standards are used as accountability measures to judge quality of teaching and learning. According to Pritchard (1996), the term "standard" has four general meanings in the field of education: (a) content standards, (b) performance standards, (c) opportunity-to-learn standards, and (d) teaching standards. Teaching standards are defined as written criteria for making judgments about progress toward professional competency; they describe what teachers at all grade levels should understand and be able to do (The National Academies Press, n.d.). In addition, teaching standards are defined as educational experiences teachers should provide inside and outside the school environment (Pritchard, 1996). In the recent years, the field of education has been experiencing a major shift towards standards-based instruction and increased accountability of teachers for student learning.

Historical Overview of Accountability Efforts

Since its publication in 1983, *A Nation at Risk: the Imperative for Educational Reform* (The National Commission on Excellence in Education) has been named as the driving force behind the modem standards movement in American education (e.g., Pritchard, 1996; Ravitch, 1995b; Shepard, 1993). This report called for public school educators and higher education teacher preparation programs to be held accountable for the quality of student learning, teachers, and teacher candidates. The authors of the report emphasized the need for change in American education. These are some changes proposed by the authors: moving away from "cafeteria-style curriculum" to more uniform programs, strengthening high school graduation requirements and raising college admission requirements, using school time more efficiently, and strengthening teacher preparation (The National Commission on Excellence in Education, 1983, pp. 18-31).

Publication of *A Nation at Risk* report became a catalyst in the movement to improve American public education and teacher preparation programs. Since its publication, educators and general public came to recognize the close connection between American educational system and the financial security and economic competitiveness of the nation and called for standards and accountability in the educational system in order to maintain high quality of education in public schools. The report reminded Americans how important education was to the U.S. international leadership in science, technology and trade stating that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our

very future as a nation and a people,... We have, in effect, been committing an act of unthinking, unilateral educational disarmament" (National Commission on Excellence in Education, 1983, p. 5). The report warned American public that due to the decline in the educational standards in nation's public schools, international competitors would be able to increase their presence and even become leaders in areas that U.S. historically was a champion in, such as business, manufacturing industry, science, and technology. Overall, according to Blosser (1989) and Shepard (1993), after the publication of *A Nation at Risk*, a major shift has occurred towards support for reform in American education.

Another major milestone towards accountability and standards in education was reached when a set of six broad goals, called *Goals 2000*, aimed at improving the educational standards in our nation, was formulated in 1989 (U.S. Department of Education, 1994). *Goals 2000* included, among several other things, a plan that by the year 2000 every child in America would meet rigorous academic standards.

Support for educational standards continued during the 1990s, and in 1994 the United States Congress ratified the *Goals 2000: Educate America Act.* This document endorsed the development of nation-wide educational standards and outlined national educational goals as a way of evaluating and advancing student achievement (U.S. Department of Education, 1994). More importantly, this act allocated resources to states to be used for development and implementation of educational improvements intended to assist all students in meeting rigorous academic standards.

In 2001 another milestone was reached when U.S. federal government stopped funding *Goals 2000* programs and a new initiative was proposed, the *No Child Left Behind Act* (NCLB), that President Bush signed into law on January 8, 2002 (U.S. Department of Education, 2002a). The new law was a sign of government will, but at the same time reflected current concerns regarding quality of American education. At the same time, the new Act provided a framework on how to continue improving the performance of America's elementary and secondary schools and present all U.S. children with quality learning opportunities. The NCLB Act also focused on increased accountability of the nation's schools and teachers for student achievement. The Act aimed to identify and use research-based strategies as effective teaching methods leading to increased academic achievement. According to the American Association of Colleges for Teacher Education's Education Policy Clearinghouse (2004), the NCLB changed "the federal government's role in kindergarten-through-grade-12 education by asking America's schools to describe their success in terms of what each student accomplishes." The act also contained the President's four basic education reform principles: (a) stronger accountability for results, (b) increased flexibility and local control, (c) expanded options for parents, and (d) an emphasis on teaching methods that have been proven to work.

Recognizing that every American family deserves public schools that work, *No Child Left Behind* pledged "highly qualified teachers" in every classroom by the 2005-06 school year. In its 2003 report *No Dreams Denied: a Pledge to America's Children*, the National Commission on Teaching and America's Future defined "highly qualified

teachers." This definition is based on both research on effective teaching and common sense. According to the report, the highly qualified teachers are those who:

- Possess a deep understanding of the subjects they teach;
- Evidence a firm understanding of how students learn;
- Demonstrate the teaching skills necessary to help all students achieve high standards;
- Create a positive learning environment;
- Use a variety of assessment strategies to diagnose and respond to individual learning needs;
- Demonstrate and integrate modern technology into the school curriculum to support student learning;
- Collaborate with colleagues, parents and community members, and other educators to improve student learning;
- Reflect on their practice to improve future teaching and student achievement;
- Pursue professional growth in both content and pedagogy; and
- Instill a passion for learning in their students.

Focus on Standards in Education

Since the 1990s, school accountability, content standards, and student achievement have become major concerns. Publication of reports like *A Nation at Risk*, followed by the Federal acts of the *Goals 2000* (U.S. Department of Education, 1994) and the *No Child Left Behind* (U.S. Department of Education, 2002a) increased pressure on policymakers and educators to propose concrete steps towards improvement of the American schools. Even before the *No Child Left Behind* Act, but even more so after its introduction, policymakers started to hold schools accountable for their pupils' learning. Some states have experimented with rewarding successful schools and punishing failing ones in an effort to ensure that all children get the quality education they deserve.

However, in spite of the states' efforts to increase school accountability, these improvements were not easy to implement in practice. In the late 1990s, *Quality Counts* (*Education Week*, 1999) reported that in a 50-state survey of state policies on accountability, the general indication was that states fell short of really holding schools accountable for their pupils' academic success or failure. The report discovered that the majority of states (48) assessed their pupils' knowledge, but only 36 states published annual report cards on individual schools. In addition, fewer than half (19) of the states publicly rated the academic performance of all state schools or at least identified low-performing ones.

With the introduction of the NCLB act, all states were required to collect scientific evidence and report on student achievement of all students in the state public schools. This data is supposed to communicate the true state of American public education to the general public. President Bush's plan of *No Child Left Behind* called for performance-based assessment, professional education and subject area examinations, and no out-of-content area teaching. The assumption was that quality teachers would provide better instruction resulting in increase in student learning.

The call for accountability did not stop with the reports on school performance or with the Title II Report Card, "requiring each state to report the pass rates on teacher assessments for all program completers from a state higher education institution teacher education program and a comparison between their institutions statewide" (Fredman, 2002, p. 3).

In efforts to provide further guidance for education reform efforts, some educators (e. g., Ravitch, 1995b) called for creation of national standards and national assessments, because "they are a way of establishing what needs to be taught and learned and whether progress is being made" (Ravitch, 1995b, p. 3). The educational historian and a former Assistant Secretary of Education, Diane Ravitch (1995a), comments on the demand for standards and accountability in American education:

> Americans ... expect strict standards to govern construction of buildings, bridges, highways, and tunnels; shoddy work put lives at risk. They expect stringent standards to protect their drinking water, the food they eat, and the air they breathe ....Standards are created and perfected because they improve the quality of life. (pp. 8-9).

Ravitch goes on to state that Americans hope that presence of standards will result in improved public education:

> Without content and performance standards, there is no way to determine objectively whether resources are deployed effectively. ...Standards can improve achievement by clearly defining what is to be taught and what kind of performance is expected (1995a, pp. 12-25).

Content Standards

The call for teaching standards and accountability in education is also driven by demands for content standards and systemic reform. Ravitch (1995a) defines content or curricular standards as descriptors of:

> ...what teachers are supposed to teach and students are expected to learn... Content standards should be specific enough to be readily understood by teachers, parents, students, and others. They should be clear enough so that teachers know what students are supposed to learn and can design lessons to help them learn what is expected. (p. 12)

In the late 1980s – early 1990s, authors like Smith, O'Day, Cohen and Spillane argued that school reform would require major systemic changes in various aspects of American education. Smith and O'Day (1991) made a strong case in their publications that such systemic efforts would require education officials to formulate core content standards and base all educational policies that were to follow on these standards. Furthermore, Smith and O'Day maintained that presence of clearly articulated content standards would influence instructional materials and assessments used in schools, as well as shape teacher preparation and professional development activities. According to Smith, O'Day and Cohen (1990), content standards would lead to teaching of more rigorous content, which in turn would require teachers to be involved in more challenging work than before. In this new educational environment emphasizing student performance, school administrators, teachers, and students themselves would have to assume new roles and responsibilities that would require higher levels of collaboration and participation of all parties involved (Cohen & Spillane, 1993; Smith & O'Day, 1991). Smith, O'Day, and Cohen (1990) went on to argue that there were also positive lessons to be learned from the current system:

> The first and central lesson is this: If exams are used to motivate students to be more serious about their studies, then examinations' content must be closely tied to the curriculum frameworks that are used to teach students (Smith et al., 1990, p. 41).

The calls for development of content standards were addressed in recent years, when many professional organizations and consortia developed national subject specific standards in such subject areas as arts, foreign language and English as a second

language, health and physical education, language arts, mathematics, social studies, science, career technical education, technology, and several others. Despite the positive publicity and substantial support for national and state-wide educational standards from policymakers, educators, and various professional organizations, some voices of caution continued to expressed negative opinions towards these efforts. In order to assess the progress of developing standards, research studies were needed.

Research on Educational Standards

National efforts to improve educational content standards in various subject areas, as well as standards for teaching, have led to standard-setting policies at the state level nation-wide. Research studies were conducted in order to assess the progress with the development of state standards. According to McLaughlin, Shepard, & O'Day, (1995), by the mid-1990s, the majority of states developed, or was in the process of developing, their variations of state-wide standards were designed to guide educational reforms in the states' local communities. Moreover, Finn, Petrilli, and Vanourek (1998) in their study on *The State of State Standards*, evaluated the state standards of all fifty states and the District of Columbia in five core academic subjects: English, geography, history, mathematics, and science. The study indicated that overall, some states did well in certain subjects, but the final conclusion was that most states still have a long way to go in meeting the demands for higher academic standards. Research on standards showed that many states struggle with formulating rigorous academic standards. In their summary of the individual states' "marks," Finn, Petrilli, and Vanourek (1998) reported that:

> In every subject, the number of states receiving "Ds" or "Fs" outnumbered those receiving "As" or "Bs." In English, only one state received an "A" while 12 received "Fs." In history, just one state received an "A" while 19 jurisdictions flunked, in geography, 3 states earned "As" and 18 failed. The numbers for mathematics were 3 and 16, and for science, 6 and 9 (p. 1).

Although the authors of the study concluded that, overall, the status of the state standards was less promising than they had expected, they also emphasized successes of individual states in establishing rigorous state standards in particular subject areas, for examples, California in mathematics, Colorado in geography, Indiana in science, Massachusetts in English, and Virginia in history. The authors went on to suggest that these states should serve as models to other states, thus further contributing to the success of the reform movement to standardize American education at a state level.

Research on Linking Teacher Work to Student Learning

In addition to calls for standards in education, other areas of educational process were being examined as important components of reforming American education, which would result in improved student achievement. According to Darling-Hammond and Rustique-Forrester (1997), given limitations of available resources and the increase in demand for student achievement to improve, it is important to understand key components of educational process that influence student learning. A growing body of research established a strong connection between teacher quality and academic success of their students. For example, a study of 900 school districts in the South (Ferguson & Ladd, 1996) provided empirical evidence of a strong correlation between teacher knowledge, as measured by licensing exam scores, master's degree, and amount of teaching experience, and academic achievement of their pupils. Several other studies

(Darling-Hammond, 2000; McRobbie, 2001; Sanders & Rivers, 1996) resulted in similar findings, reporting that the relationship between teacher knowledge and student achievement was pronounced in a variety of subject areas and settings, including low socio-economic schools (McRobbie, 2001). Moreover, several major publications (e.g., Darling-Hammond, 1997; National Commission on Teaching and America's Future, 2003; National Commission on Teaching and America's Future, 1996; U.S. Department of Education, 1997) have consistently indicated that teacher expertise is the single most important factor that affects student achievement. Several experts (Darling-Hammond & Rustique-Forrester, 1997; Ferguson, 1991; Greenwald, Hedges & Laine, 1996; Sanders & Rivers, 1996; Webster & Mendro, 1997; Webster, Mendro, Orsak, & Weerasinghe, 1998) argued that investing resources in teacher quality is one of the most important approaches to improving American public education and raising student achievement. Given the body of research on factors impacting student learning, pointing out the importance of teacher quality to the increase in student achievement, the following areas were proposed by the U.S. Department of Education (1998) as leading to high quality teaching: (a) recruit talented and diverse people, (b) improved teacher preparation, (c) raise licensing and certification standards, (d) improve induction of new teachers, (e) improve professional development, and (f) improve teacher accountability and incentives (U.S. Department of Education, 1998). Research also indicated that improvements in these areas should produce better trained and qualified teachers who are ready to meet higher professional standards and are capable of helping students to achieve higher academic standards.

As efforts to establish standards for student achievement and school accountability continue nationwide, teacher effectiveness in facilitating learning is an important component of these efforts. Studies (Ferguson, 1991; Greenwald, Hedges & Laine, 1996; Sanders & Rivers, 1996; Webster & Mendro, 1997; Webster et al., 1998) have shown that the improvement of teacher quality is an important step toward the overall improvement of American education. Some educators believe that one of the ways towards improvement of teacher quality is the National Board for Professional Teaching Standards certification. As stated by the former U.S. Secretary of Education, Richard Riley:

> We must recruit, support, and retain the most talented people in teaching. We must invest in high-quality professional development. We must require tougher licensing and certification standards for teachers, and increase dramatically the number of teachers who meet the demanding standards of the National Board for Professional Teaching Standards (in NBPTS, 1998, p.2).

National Board for Professional Teaching Standards (NBPTS)

Three years after the release of *A Nation at Risk*, in its 1986 report, entitled *Nation Prepared: Teachers for the 21st Century*, the Carnegie Task Force on Teaching as a Profession, recommended establishment of a national teacher certification program. In 1987, the National Board for Professional Teaching Standards (NBPTS) was established as a nonprofit organization. The National Board is governed by a 63 member board, consisting of directors, the majority of whom are classroom teachers (NBPTS, 2007).

According to NBPTS, its mission is "to establish high and rigorous standards for what accomplished teachers should know and be able to do, to operate a national

voluntary system to access and certify teachers who meet these standards, and to advance related education reforms for the purpose of improving student learning in American schools" (NBPTS, 2007). An important accomplishment of the NBPTS is the National Board Certification process, which is often considered a symbol of professional excellence (Baratz-Snowden, 1990, 1992; Harman, 2001). First awarded in 1994, this certification is based on advanced standards for experienced teachers (NBPTS, 2007) and signifies a teacher's knowledge and skills. NBPTS certification was developed by teachers, with teachers, and for teachers, setting rigorous standards for the profession, creating performance assessments based on those standards, and recognizing experienced teachers who meet the standards (NBPTS, 1998).

Since the beginning of the NBPTS certification in 1994, and as teachers came to recognize the value of achieving "master teacher" status, the number of certified teachers skyrocketed from just over 500 teachers in 1996 to 24,000 in 2002 (see Figure 1). In 2006 there were nearly 50,000 National Board Certified teachers nationally (NBPTS, 2006). Many school districts and states recognize the value of the certification and offer NBPTS-certified teachers extra pay and other benefits and professional assignments.

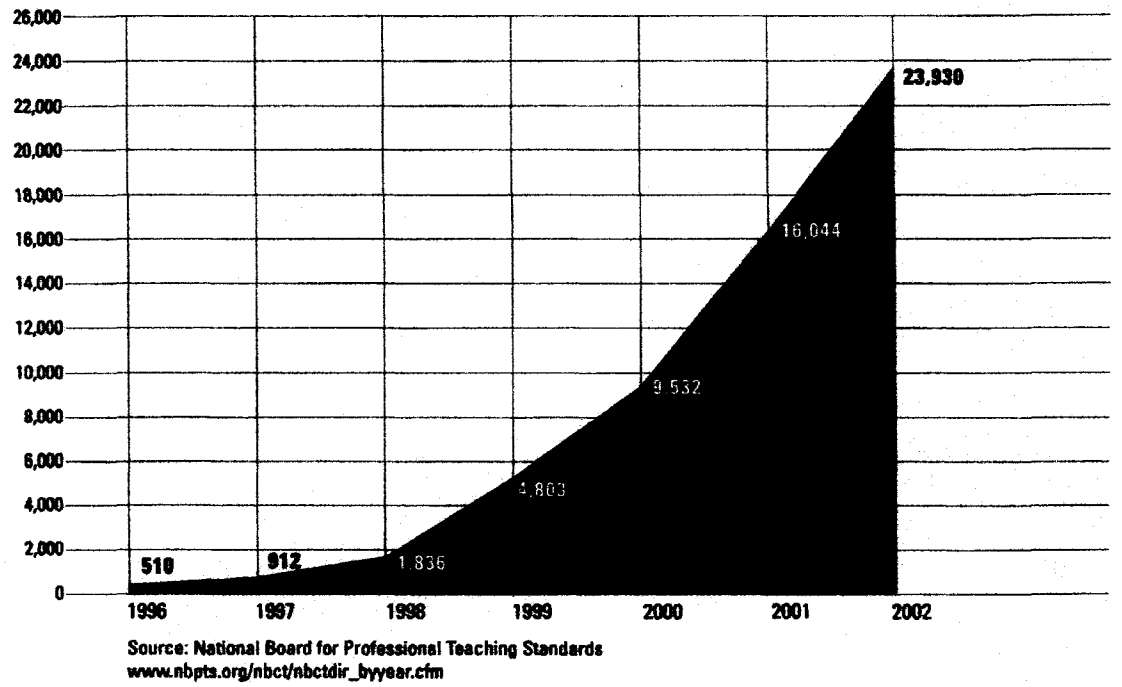

Figure 1. Number of National Board Certified Teachers 1996-2002 (National Board for Professional Teaching Standards, 2004).

Accountability of Teacher Preparation Programs

Although a substantial body of research indicates that quality teachers influence student performance, according to the U.S. Department of Education report on *Promising Practices*, "teacher education has long been considered weak among higher education degree programs, one that lacks high standards and strong contacts with the field" (1998). At the same time, teacher preparation is viewed as one of the important areas contributing to teacher quality, because the pre-service teachers of today will be the practicing teachers of tomorrow (U.S. Department of Education, 1998). It is

expected that the knowledge and skills acquired by a pre-service teacher will determine the effect that teacher will have on his or her students.

In addition, several authors (e.g., Jackson, 2006; Schackner & Lee, 2002; Selwyn, 2005/2006) reported increasing pressure from state and federal lawmakers in recent years to hold higher education teacher training programs accountable for the quality of their graduates. According to Jackson, "today, states need to assess not only the knowledge and skills of graduates of teacher preparation programs, but also the graduates' ability to improve student learning" (2006, p. 1).

As a result of this increasing pressure, teacher preparation institutions in many parts of the country have been faced with state or regional accountability initiatives. To guide teacher preparation programs, sets of national standards and/or assessments have been outlined by the Interstate New Teacher Assessment and Support Consortium (INTASC), National Board for Professional Teaching Standards (NBPTS), National Council for Accreditation of Teacher Education (NCATE), and the National Commission on Teaching and America's Future (NCTAF).

To emphasize the important role of teacher preparation programs and increase their accountability, the National Council for Accreditation of Teacher Education (NCATE) has set rigorous standards "and expect colleges to demonstrate that teacher candidates are gaining the knowledge, skills, and dispositions necessary to have a positive impact on P-12 student learning" (Mitchell, 2001, p. 4). Moreover, according to Fredman, "by 2003, NCATE accredited programs must provide evidence that their

education graduates, who are in their first year of teaching, are impacting student learning in their own classrooms" (2002, p. 3).

National Council for Accreditation of Teacher Education (NCATE)

Founded in 1954, NCATE is an accrediting non-profit nongovernmental coalition of more than thirty national associations representing the field of education. NCATE accredits higher education institutions with teacher preparation programs. It is governed by policy boards, which consist of teacher educators, teachers, state and local policymakers, and other education professionals. To many educators, NCATE accreditation indicates that a teacher preparation program "produces competent, caring, and qualified teachers and other professional school personnel who can help all students learn" (The Gale Group, 2007). NCATE accreditation is based on a set of standards and involves a self-study process conducted by the accreditation-seeking institution. NCATE standards highlight the following components of a quality teacher preparation program: (a) a coherent program of studies for each student rather than the typical hodgepodge, (b) a firm foundation in the liberal arts and teaching disciplines, (c) programs that prepare teachers for the higher content standards set for students, (d) programs that prepare teachers for classroom diversity and for new technologies, and (e) the use of performance-based standards rather than "seat time" in classes to determine the readiness of candidates to teach (U.S. Department of Education, 1998).

According to the Gale Group's *Encyclopedia of Education* (2007), NCATE standards have been adopted by 28 states as the state standards of teacher preparation,

with more than 600 teacher preparation institutions nation-wide being a part of the NCATE system in 2002.

Assessment Methodologies of Teacher Practice

In his 1999 sixth annual State of American Education Address speech, Mr. Riley called on states, school districts, and teacher preparation institutions, to concentrate on accountability and teacher quality and make necessary improvements in teacher recruitment, preparation, and professional development (U.S. Department of Education, 1999). Moreover, Mr. Riley viewed the National Board for Professional Teaching Standards as an organization contributing a great deal towards efforts of teacher accountability and improvement of teacher quality. In the last decade NBPTS has been a leader in establishing national teaching standards. However, many would agree that "the development of standards alone cannot ensure the success of school reform" (Holbein, 1998, p. 560).

Currently a variety of assessments, and their combinations, is being used in the Unites States to assess teacher quality of future and practicing teachers. These assessments range from paper-and-pencil tests, like PRAXIS II, to the NBPTS certification assessment, to several other unique systems developed by states and teacher preparation programs. Regardless of their format all these approaches strive to provide data for accounting for local efforts to prepare all students to meet rigorous academic standards.

PRAXIS I, II and III

The series of PRAXIS tests was developed by the Educational Testing Service, a non-profit organization employing over 2,500 people (ETS, 2006). According to Wakefield (2003), ETS administers over 20,000 various assessments in 180 countries of the world. These tests are specifically designed for teacher preparation programs or states to assess future teachers' readiness to enter the teaching profession.

Additionally, some professional organizations require one or both tests as part of their licensing process. Paper- or computer-based PRAXIS I test is intended to assess basic skills of reading, writing and math of teacher candidates. This test is typically taken earlier in the candidate's college career. The focus of the paper-and-pencil PRAXIS II test is a subject area content knowledge of a candidate, as well as "general and subject-specific teaching skills and knowledge" (ETS, 2006). This assessment comes in three variations: (a) subject assessment, (b) principles of learning and teaching, and (c) teaching foundations tests. PRAXIS II assessment usually takes between one and four hours to complete.

According to ETS website (2006), PRAXIS III is a classroom performance assessment test which judges the professional skills of new teacher in a classroom setting. ETS established guidelines for the use of this test and does not permit its use with practicing licensed teachers, especially for making employment decisions. The test includes three components: (a) direct classroom observation, (b) review of documentation prepared by the teacher, and (c) semi-structured interviews. According to ETS (2006), the PRAXIS III test "consists of a framework of knowledge and skills

for a beginning teacher that contains 19 assessment criteria in four interrelated domains: (a) organizing content knowledge for student learning (planning to teach); (b) creating an environment for student learning (classroom environment), (c) teaching for student learning (instruction), and (d) teacher professionalism (professional responsibilities). PRAXIS II assessment is delivered, scored and managed by individual states. Overall, PRAXIS I and PRAXIS II tests are employed by most state education agencies in the country. The results of the tests are used to make high-stake decisions pertaining to licensure of new teachers. PRAXIS III test is used as a requirement for the licensure or as professional development tool for practicing teachers in a number of states, e.g. Florida, Ohio, and Mississippi (ETS, 2006). PRAXIS series are frequently criticized for their bias discriminating against low-income and minority test takers. Some educators (e.g. Wakefield, 2003) recommend the use of PRAXIS tests along with other authentic assessments, GPA, and face-to-face interviews.

NBPTS Assessment

NBPTS has identified and outlined assessment for each of the 25 areas of teacher specialization. The assessment is based on the Standards for certification in each area, which were formulated by special Standards committees. According to the organization, "the Standards and the assessments for all certificate areas are based on five core propositions for accomplished teaching:

1. teachers are committed to students and their learning;
2. teachers know the subjects they teach and how to teach those subjects to students;
3. teachers are responsible for managing and monitoring student learning;

4. teachers think systematically about their practice and learn from experience; and
5. teachers are members of learning communities. (NBPTS, 2006)

According to the NBPTS website (NBPTS, 2006), the assessment process is two fold and consists of (a) Portfolio entries and (b) Assessment Center exercises. The certification process requires a total of four portfolio entries. Three of these entries suppose to highlight teacher's classroom practice and should be accompanied by samples of student work. The fourth portfolio segment should highlight accomplishments outside of classroom, involving parents and community at large. This segment also should demonstrate an impact of these activities on student academic progress.

The Assessment Center section of the certification is primarily focused on the *subject area content knowledge*. For instance, foreign language teachers are required to demonstrate oral and written proficiency in the target language in a series of four exercises, with the fifth exercise being devoted to the candidate's knowledge of language acquisition and the sixth exercise to the knowledge of how language works. The same section of the assessment is unique for other 25 specialization areas. For example, for an upper-level math teacher, the six exercises are focused on the following areas: (a) algebra, (b) calculus, (c) discrete math, (d) geometry, (e) statistics and data analysis, and (f) technology.

NBPTS certification assessment is often criticized for its lack of validity and reliability support (e.g. Cunningham & Stone, 2005; Kershner, 1999), as well as the high costs of certification process paid by the applicants. In his article Kershner points

out that the part of the assessment focused on teacher's subject knowledge is not very strong. Kershner also provides anecdotal evidence that teachers applying for NBPTS certification are rated not on their knowledge of content, but on "how well they can justify their teaching decision. In one example lauded by the board, the teacher explained that she gave a student an "A" in the name of self-esteem building – even though the student had several misspelled words on his paper" (1999).

On its website, NBPTS lists a number of recent studies that emerged as a response to voices of critics, questioning whether students of NBPTS certified teachers were doing better than students of non-certified teachers (Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Smith, Gordon, Colby, & Wang, 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004). All these studies, commissioned by NBPTS, report greater testing results and learning outcomes of students taught by NBPTS certified teachers. However, critics of NBPTS still raise a question whether these differences in effectiveness can be attributed to the certification process or whether these teachers had already been more effective prior to the certification process (e.g., Cunningham & Stone, 2005).

State-wide Assessment Models

As required by the *No Child Left Behind* Act (U.S. Department of Education, 2002a), all states were expected to institute accountability systems to serve as indicators of how well local schools achieve the established standards and meet benchmarks of NCLB. As a result, according to the report by the U.S. Department of Education, in 2002 every state across the nation have developed and implemented a school

accountability system designed to hold at least school districts responsible for student achievement based on established standards; every state had a process of identifying poor performing schools; and most state accountability systems were measuring student achievement as a primary indicator of educational progress.

Prior to the 1990s, state accountability systems have traditionally focused on limited school factors, such as class-size and staff qualifications. However, since mid-1990, some progress has been made when states started to incorporate the findings of the effective-schools research into their accountability systems, such as (a) staff development, (b) teacher evaluation, (c) principal leadership, (d) overall goal setting, and (e) student achievement (Sturm, 1995). In addition, since 2002, to comply with the *No Child Left Behind* Act, all states were required to produce reports on meeting their educational mission, as well as on their student achievement of educational standards. When reporting on student achievement, the states were required to give an account of accomplishments of various student sub-groups, based on their ethnicity, gender, socio-economic status and several other characteristics (U.S. Department of Education, 2002a). It is important to mention, that NCLB required all states to provide annual reports of student academic progress, closely focusing on two subject areas: reading and mathematics. These annual reports were expected to include the following components: (a) student achievement, (b) assessment rates, (c) graduation rate, (d) Adequate Yearly Progress (AYP) decisions, (e) school improvement, and (f) teacher quality (U.S. Department of Education, 2002b).

In addition to the abovementioned assessment models, a number of Southern states emphasized school accountability early on. According to Mark Musick (1998), the Chairman of the National Assessment Governing Board and the President of the Southern Regional Education Board, since 1988 several southern states took major steps towards establishing school accountability programs with an intent to make a difference in improving schools and student achievement. Each of these models incorporates a teacher evaluation component as a part of the overall school assessment. Under the leadership of the Southern Regional Education Board, several states (e.g., Tennessee, Texas, and Kentucky) have passed comprehensive K-12 schools accountability initiatives aimed at assessing and improving student academic progress. These state initiatives emphasized the importance of school accountability as a determining factor in improving and maintaining standards in education. Although unique, each of these approaches focuses on student achievement, and holds the teachers and schools accountable by measuring their students' learning.

Millman (1997), a professor of Educational Research at Cornell University, in his review of accountability systems developed by several southern states, wrote that any method which:

> ...evaluates teachers and schools with the hope of making them accountable should be fair to the teachers and the schools, should be comprehensive in terms of the types of learning objectives measured, should be competitive in relation to other methods of evaluating teachers and schools for an accountability purpose, and should not cause undesirable effects when used properly. (p. 243)

Below is a brief summary of the three of the Southern models: (a) the Dallas Value-Added Accountability System, (b) the Tennessee Value-Added Assessment

System, and (c) the Kentucky Instructional Results Improvement System. In addition, another accountability system with a focus on teacher quality, the Oregon Teacher Work Sample Methodology, is the assessment this study will examine in greater depth.

The Dallas Value-Added Accountability System

This school ranking system was first introduced in 1992, however, since then it has undergone some growth and now includes a process for identifying effective teachers. It is considered to be a "fair accountability system, based on variables in addition to norm-referenced test data...tying together district and campus improvement planning, principal and teacher evaluation, and school and teacher effectiveness" (Webster & Mendro, 1997, pp. 81-82). According to the authors, the ultimate measure of school effectiveness is based on pupils' test results, thus "a school with improving achievement results is held to be on the right track" (p. 83). Statistical analysis used by the system controls for demographic characteristics of the students as well as for the prior achievement. Under this system effective schools receive additional funding, spent on monetary awards to staff, while ineffective schools receive increased attention "ranging from additional services to replacing administrators to restructuring the school" (p. 88). In their article, analyzing the Dallas Value-Added Assessment system, Meng Thum and Bryk (1997) bring up a number of serious technical questions regarding its validity. Another criticism comes from the system's use of standardized test scores as a primary measure of student learning (Meng Thum & Bryk, 1997; Sykes, 1997). It is widely known that standardized tests do not capture all the complexity of cognitive processes (e.g., Gardner, 1991; Perkins, 1992). It is recommended that the

system supplement its standardized tests data with additional more authentic assessments. Furthermore, concerns have been expressed regarding the impact of standardized test data on teachers, when they are used as a primary measurement of teacher effectiveness.

The Tennessee Value-Added Assessment System

This system is unique in its attention to individual student academic performance. This system uses previous year data of individual students and compares it with the student's current achievement, calculating individual student academic gain. Student success is then attributed to the work of his/her teacher, school, and school district. This assessment model is based on works by McLean and Sanders (1984) and their conclusions that: (a) schools and teachers differed in their effect on student learning, (b) school and teacher effects seemed to be consistent across time, (c) teacher effects were not influenced by the location of the school, (d) teacher effects found statistically were highly correlated with subjective reports of supervisors, and (e) student gains were not correlated with previous achievement levels (Sanders & Horn, 1994, p. 300). Although currently the system employs results of standardized tests as a measure of student progress, Sanders & Horn state that other assessments, or their combination, can be used in this model. According to the standards set by the state legislature, all educational institutions in the state are expected to show "a mean gain for each academic subject [science, math, social studies, language arts, and reading] within each grade [3-8] greater than or equal to the national gain" (Ceperley & Reel, 1997, p. 136). Moreover, special provisions are made in this assessment system to maintain its

fairness, e.g., a minimum of three years of data is used to make decisions, educational institutions and teachers cannot be judged only on the outcomes of this assessment. One of the critiques of this assessment model is its lack of compensation or reprimanding policies for educators, based on their evaluation (Darlington, 1997). In addition, the assessment only incorporated those students with three-year-long profiles at the schools, thus leaving out a number of students in the state. Overall, data required by this model are rather complex to collect and maintain, since each child must have an extensive file maintained and annually updated with his/her detailed information available to allow for data comparisons.

The Kentucky Instructional Results Improvement System

With the passage of the Kentucky Education Reform in 1990, a new statewide school accountability system was established based on student academic progress. According to Kingston and Reidy, who worked for the Kentucky Department of Education, "the primary goal of the Kentucky's school-based accountability system is to motivate educators and the public to dramatically improve student learning" (1997, p. 191). The assessment system emphasizes evaluation of progress of *all* students, including those with disabilities, and holding schools, and all their certified staff, accountable for student learning. The system uses individual school's data as baseline for measuring academic progress of all students at each education institution. Student progress is assessed using cohorts of students, i.e., performance of current 3rd-graders was compared to the performance of 3rd-grades of the previous year.

One of the unique characteristics of this system is its use of performance-based test data along with several other indicators (i.e., attendance, retention, etc.) rather than standardized tests. The system uses financial compensation to reward schools with positive progress. Failing schools at first receive planning funds and assistance of a consultant to improve their performance, with prospects of a state takeover for those schools that are not able to improve. Moreover, in order to assist all schools in meeting accountability goals, the state devoted additional funds for professional development of educators. Critics of this approach (e.g., Stufflebeam, 1997) point out issues with reliability and validity of the assessments used in the model, and especially are disappointed with the absence of standardized test data.

It is important to mention that while each of the presented models have its strengths and offers a unique solution to the issue of school accountability at a state level, these models have their shortcomings and limitations. Jason Millman concludes the discussion of the abovementioned models and their relation to the body of research on teacher improvement by stating:

> On the one hand, one could argue that any information is valuable, including information on what students know and can do. On the other hand, merely describing the product (what students know and can do) provides scant information on what the teacher did or should have done to yield better results. Such an assessment is similar to the old-fashioned process-product research in which the explanatory goodies are kept in a mystical black box (1997, p. 247).

Teacher Work Sample Methodology

Oregon's Teacher Work Sampling

Since the late 1980s, the demand by policymakers and the public for increased student achievement and more school accountability for student learning has intensified. With a substantial body of research pointing out that teachers play a crucial role in student achievement, the state of Oregon approached this situation by developing its unique assessment system that focused on evaluation of teacher preparedness to influence student performance.

Educational reforms in Oregon. In 1987, the state of Oregon recognized the importance of quality teacher preparation for student achievement, which resulted in changes in state initial licensing. The state selected to move away from a teacher preparation program approval system to a new system that focuses on what an individual teacher candidate can do (Schalock & Myton, 1989; Schalock, 1998). A few years later, in 1991, a new school-reform bill was passed, which required all students to meet high academic standards (McConney, Schalock & Schalock, 1998; Schalock & Cowart, 1993). In its account of the Oregon reform efforts, *Education Week* (1997) reports that the goal of the new law was "to make the Oregon workforce the best educated and best prepared in America by the year 2000 and equal to any in the world by 2010" (Education Week, 1997). Additionally, as stated in the *Quality Counts '97* report (Education Week, 1997), in part due to the reform efforts mentioned above, "Oregon became the first state in the nation to win approval of its *Goals 2000* plan from the federal government in January 2005."

In 1991, in response to the new laws, the Oregon Department of Education along with school districts, were able to create a new educational environment in state schools (McConney, Schalock, & Schalock, 1998). According to the authors, this new educational environment was organized around content and performance standards that were closely aligned with assessments for benchmark grades 3, 5, 8, 10 and 12. The state then required its students to meet these performance standards in order for them to be admitted into state universities.

These innovative efforts have contributed to Oregon's grade of "A" in the report *Quality Counts 1997* (Education Week, 1997). The authors of the 1997 report state that "Oregon trailblazed one of the most ambitious school-reform plans in the nation in 1991. And with true pioneer grit, it has stayed the course, despite controversy and midcourse corrections" (1997). However, it is important to mention that in the most recent report *Quality Counts at 10: A Decade of Standards-Based Education* (Education Week, 2006), Oregon received a grade of "C+" for its standards and accountability, and "D" for its efforts to improve teacher quality. Overall, in the 2006 report the state scored below average in three of a total of four categories in which grades were assigned. According to the report:

> Oregon ranks near the bottom of the nation on indicators of teacher quality and posts mediocre scores in each specific area within this policy category. For example, the state does not require prospective teachers to have a major or equivalent coursework in the subjects they will teach to earn an initial license. In addition, the state does not fund or require professional development for teachers (Education Week, 2006).

Moreover, Oregon education reform efforts went beyond K-12 level and included teacher preparation programs. As documented by Schalock and Myton (1989) and Schalock, Schalock, Myton and Girod (1993), in its response to the state reform efforts, the Oregon Teacher Standards and Practices Commission (Oregon's teacher licensing agency) designed requirements for teacher preparation and licensure in the state that were in alignment with Oregon's new education model. For instance, in order to satisfy the initial licensing requirements, Oregon teacher candidates were expected to develop and submit for a review work samples illustrating their effectiveness in facilitating student learning.

Development of Oregon TWSM. Oregon institutions of higher education played an important part in the reform movement by providing research support to the state. One of the state institutions that provided a substantial contribution to the reform efforts, especially in the area of teacher assessment, was Western Oregon University. In response to the need for a teacher "performance-based assessment tool that can be used to not only measure teacher quality, but also to link students achievement to teacher quality" (Fredman, 2002, p. 4), the university engaged in the development of an assessment approach. After several years of extensive research, a team led by H. Del Schalock developed a performance-based approach to preparing and evaluating teachers. This new outcome-based assessment, called Oregon Teacher Work Sample Methodology (TWSM), was able to capture the spirit of the overall state reform efforts and was grounded in a context-dependent theory of teacher effectiveness (Schalock, Schalock, & Myton, 1999). The development of this assessment model led to the

establishment of a state-wide teacher preparation program of a new kind, the program with a focus on teacher candidates' ability to impact student learning, which corresponded well with Oregon education reforms.

What is Oregon TWSM. The Oregon Teacher Work Sample Methodology, a key component of the Western Oregon University teacher preparation program, is a performance-based assessment system that requires pre-service teachers to provide work samples demonstrating their proficiency in positively impacting student learning. In addition to providing a framework for what knowledge and skills teacher candidates should acquire during their studies, the Oregon TWSM also provides an insight into what pre-service teachers can actually do in a classroom setting.

TWSM is regarded by many in the field of education as an appropriate performance-based assessment instrument, which not only allows a pre-service teacher to showcase his/her professional knowledge and skills, but also gives the teacher candidate a framework for learning while completing the teacher work sample process. Some also emphasize the value of TWSM as an assessment tool promoting reflective skills of pre-service teachers (Pankratz, 1999). The original idea behind this assessment is that while working on a teacher work sample, future teachers think, learn, practice, and reflect upon their effectiveness as teachers in ways that align closely with standards-based education system. Oregon TWSM guides future teachers towards standards-based instruction by using the following ten steps:

1. Define the sample of teaching and learning to be described;
2. Identify the learning outcomes to be accomplished within the work to be sampled;

3. Prior to instruction, assess students' status with respect to the post-instruction outcomes to be accomplished;
4. Develop instruction and assessment plans that align with proposed learning outcomes and current status of students with respect to the proposed outcomes;
5. Describe the context in which teaching and learning are to occur;
6. Adapt the desired outcomes and related plans for instruction and assessment to accommodate all students and the demands of the teaching-learning context;
7. Implement a developmentally and contextually appropriate instructional plan.
8. Assess the post-instructional accomplishment of learners, and calculate each student's growth in learning;
9. Summarize and interpret the growth in learning achieved (or lack thereof) for the class as a whole and for selected groups with the class; and
10. Examine and reflect on student learning in light of the pre-instructional developmental levels of students, targeted learning outcomes, the context in which teaching and learning occurred, and personal professional effectiveness and development. (McConney, Schalock, & Schalock, 1998, p. 347)

The following elements of the methodology constitute the Oregon Teacher Work Sample methodology: (a) sample of work, (b) targets for learning, (c) measures of learning, (d) descriptors of process, (e) descriptors of context, (f) analyses of learning gains, and (g) reflection and next steps (Schalock, Schalock, McConney, Brodsky, & Myton, 2002, pp. 3-4).

Moreover, as stated by the director of the *Renaissance Partnership for Improving Teacher Quality* project, Roger Pankratz, "the work sample methodology provides direct evidence of a teacher candidate's effect on student learning in a relatively short time period and clearly connects the elements of standard-based teaching and learning" (1999, p. 37). Shalock and Myton also contributed to this thinking by stating that "teacher work sampling assesses the effectiveness of teachers close to their

work... ...[and it is] a quality assurance system that holds student learning at its core" (2002, p. 11). Girod further supports this position by stating that TWSM "is a vehicle that helps perspective teachers learn to think about teaching in ways that are linked tightly and continuously to pupils' learning, to gain experience in teaching in this manner, and to demonstrate effectiveness in doing so" (2002, p. 1). Finally, according to Schalock, Schalock, and Myton (1999) "TWSM is a quality assurance system that can assess what students learn, how well they are to learn it, the progress each student is making in his or her learning, and how each student who is not making the progress can be helped to do so" (p. 1.9).

Uses of Western Oregon TWSM. According to Ayres, Girod, McConney, Schalock, Schalock, and Wright (1996), Oregon TWSM is a methodology that was designed for a number of purposes: (a) teacher preparation and licensure, (b) teacher development and evaluation, and (c) research and program development. In spite of its original primary use as an assessment towards initial licensure, as time passed, the application of TWSM in Oregon has expanded, as stated by Schalock, Schalock, and Myton, the TWSM assessment system was used in the state as a "continued requirement... for initial licensure of teachers... and [a] recent addition as a requirement for continuing licensure" (1999, p. 1.6). In instances when TWS is used for teacher preparation and initial licensing, Schalock and Myton point out that teacher work sampling can serve as:

1. A *model* for thinking about teaching and learning;
2. A *frame of reference* for designing and operating teacher preparation programs that systematically connect teaching and learning;

3. A *vehicle for practicing and obtaining feedback* on one's effectiveness as a teacher in fostering pupils' progress in learning (formative evaluation);
4. A *methodology for demonstration and documenting* one's effectiveness in fostering learning gains by pupils (summative evaluation), and
5. A *source of evidence to be used in recommending and granting* a license to teach. (2002, pp. 12-13)

When Oregon TWSM is used with pre-service teachers, future teachers are asked to focus on instructional units to be taught over a period of 3-5 weeks. It is expected that each pre-service teacher will complete a total of two work samples during their student teaching experience (Schalock, Schalock, & Myton, 1998). The first work sample is produced with a substantial assistance of a university faculty, while the second sample is compiled independently by the future teacher. After its completion, the second teacher work sample is assessed, typically by the faculty supervising student teaching experience, and used as evidence of pre-service teacher professional readiness.

Overall, many in the field of education would agree that teacher work sampling provides, with a greater degree of certainty than paper-and-pencil tests, extensive information regarding a teacher candidate's readiness to be an effective educator. The sample also supplies materials for teacher preparation programs to see how capable their new teachers to focus on improving student learning. The overall informative capacity of this assessment tool makes it a unique and effective evaluation mechanism that defines good teaching though improved student learning and sets it aside from the NBPTS certification, INTASC, and PRAXIS III (Girod, 2002; Henning & Robinson, 2004; Schalock, Schalock, & Myton, 1998;).

TWSM and reflective practice. Experts on teacher work sampling (e.g., McConney, Schalock, & Schalock, 1998) argue that while preparing a work sample, pre-service/in-service teachers become engaged in the "reflective process" of designing activities which incorporate aspects that are known to impact student learning. Each of the components of the teacher work sample (e.g., learning environment, assessment) stimulates and guides its users to ask questions like:

1. What are the learning outcomes I want my students to accomplish?
2. What activities and instructional methodologies are appropriate or necessary for these students to achieve these outcomes?
3. What resources and how much time do I need to implement these activities and methodologies?
4. What assessment activities and methodologies are appropriate for these students and these outcomes when using these instructional methodologies?
5. How successful was I at helping each of my students achieve the learning?
6. What went right? What went wrong? Why? (McConney, Schalock, Schalock, 1998, p. 346)

In their collaborative work McConney and Ayres (1998) stress TWSM's potential for assisting pre-service/in-service teachers with establishing alignments between important components, such as instruction, assessment, and outcomes. Moreover, the TWS experience helps teachers in identifying, collecting, interpreting, and reflecting upon the evidence of student progress made toward meeting outlined instructional goals, thus connecting their teaching to learning of their students.

Western Oregon TWSM and student learning. According to Oregon TWSM experts (e.g., Ayres et al., 1996; McConney, Schalock, & Schalock, 1998), the methodology was designed to guide pre-service/in-service teachers in assessing student learning which occurred as a result of instructional activities the learners were engaged

in by the teacher. This assessment of student learning is typically done using pre- and post-tests or similar evaluation tools and comparing student outcomes on these tests. In many cases assessments are designed by pre-service/in-service teachers themselves. In the end, teachers are asked to reflect on the assessment data regarding student progress or "learning gain" and connect it to their teaching. This component of Oregon TWSM, helping teachers to establish connections between learning gain and teaching, adds "credibility" to the methodology and "meaning" to the process, as viewed by teachers and their evaluators, making work sampling an effective assessment and teaching tool to be used with future and current teachers alike (McConney, Schalock, & Schalock, 1998).

TWSM adaptations. Several examples in this section indicate that the TWS process can be adapted to fit a variety of learning goals and contexts. The flexibility of the TWSM approach also allows teacher preparation programs and future teachers to integrate a variety of unique teaching and learning standards. Some states that have adopted the original TWS methodology and instrumentation, which are closely aligned with the NCATE standards, modified the instrumentation to include their local state standards. For instance, according to Fredman (2002), the state of Oklahoma has successfully integrated their 15 teaching competencies, creating a new instrumentation version called OKTWS, which is closely meeting the unique needs of this state education system. The OKTWS is used in Oklahoma with first year teachers, along with a portfolio and a teacher observation, as a tool to assess the linkages between

teacher preparation and the impact of the teaching program graduate on classroom learning.

Another example of TWSM versatility is a use of its modified version to assess the impact of university seniors – reading tutors – on the reading skills of struggling K-12 readers (Cartwright & Blacklock, 2003). In addition, the study examined the influence of the modified TWS use on the dispositions of the tutors. The key modification of the TWS model in this case was in the focus of instruction and data collection on a single child, and not a group of learners. Overall, the study reported overwhelmingly positive effects of the TWSM on the reading skills of the struggling readers and the dispositions of the university students. The authors conclude that "the teacher work sample process provides a powerful method of closing the gap for struggling readers while documenting the learning of the candidates and their students" (Cartwright & Blacklock, 2003, p. 16). Furthermore, in stressing the success and wide application of the TWS approach, Cartwright and Blacklock state that "by demonstrating one aspect of a candidates' proficiency through their ability to effect student progress in reading, the institution is partially addressing a requirement of the accountability movement in ways that strengthen, not impede, student learning" (2003, pp. 17-18).

Kay Hegler (2003) reports on the use of TWSM with special education teacher candidates seeking their first teaching license. The author is pleased with the flexibility of the methodology allowing for the assessment of all eight special education outcomes. The author concludes that:

> The TWSM has been effective in assessing teacher candidate competence at the junior and senior-level in the special education licensure program. The methodology enables candidates to describe their impact on student learning. Faculty can summarize this data by type of outcome for the K-12 student and aggregate data for courses (p. 9).

Moreover, another example of Oregon TWSM adaptation is a multi-state Title II Grant Consortium, the Renaissance Partnership for Improving Teacher Quality, directed by Dr. Roger Pankratz. The project involved eleven teacher preparation institutions in ten different states and attempted to "improve the quality of their graduates and teachers in local partner schools by focusing attention on P-12 student learning" (Pankratz, 2004; The Renaissance Partnership for Improving Teacher Quality Project, n.d.a). The Consortium has adapted the original TWS to meet the needs of the multi-state project partners, calling the new version Renaissance Teacher Work Sample (RTWS).

According to Denner, Salzman, and Harris (2002), the Consortium greatly revised Western Oregon TWSM. This was done to "ensure that our teacher work sample assessment responds to the mandates for program accountability and to address the technical issues of validity and scoring reliability" (p. 3). These modifications included the following: (a) development of guidelines for the completion of samples, (b) scoring rubric closely aligned with standards and indicators, (c) establishment of benchmarked performances, (d) developing rater training, and (e) accumulating validity and reliability data on the assessment (Denner et al., 2002; Salzman et al., 2001). Especially, the Consortium has carried out extensive empirical work to minimize psychometric limitation of the Western Oregon TWSM in the studies on reliability and validity of the RTWS assessment (Denner, Pankratz, Norman, & Newsome, 2004).

Renaissance Partnership for Improving Teacher Quality

The Renaissance Partnership for Improving Teacher Quality Title II grant is a five-year project, originating in 1999 and "committed to a shift from focusing on the teaching process to focusing on learning results, and trying to connect teacher performance to student learning" (Robinson & Boody, 2003, p. 20). University of Northern Iowa is one of the eleven higher education institutions-partners in the Renaissance Partnership for Improving Teacher Quality. These teacher preparation institutions are located in California, Idaho, Iowa, Kentucky, Kansas, Michigan, Missouri, Pennsylvania, Tennessee, and Virginia (Pankratz, 2004; The Renaissance Partnership for Improving Teacher Quality Project, n.d.a). The Project has adapted the Oregon TWSM and developed its own version of the teacher work sample, which includes: (a) performance prompt, (b) teaching process standards, (c) scoring rubrics, and (d) scoring guide (The Renaissance Partnership for Improving Teacher Quality, 2002a, 2002b, 2002c). *The Renaissance Teacher Work Sample Prompt* is designed to guide teacher candidates in developing their work samples, providing them with criteria and rubrics they can use to self-evaluate their work in progress. Cooperating teachers and student teaching coordinators also use *the Prompt* to provide feedback to teacher candidates regarding their work samples. Later on, during the scoring process, raters use *the Prompt* and *Scoring Rubrics* to assess specific teaching processes within each teacher work sample.

In addition to its unique version of RTWS methodology and specific instrumentation (the *Prompt* and *Scoring Rubrics*), the Project has developed a large

number of other documentation to assist teacher preparation faculty, cooperating teachers, and teacher candidates in working together on documenting teacher candidate growth towards improving the learning of their students. For instance, *the Teacher Work Sample Scoring Guide* and *the Road Map for Locating Evidence* were also designed with this purpose in mind (The Renaissance Partnership for Improving Teacher Quality Project, 2002a, 2002c). These documents point out kind of evidence needed for each section of the sample. Moreover, the project's website offers nearly 50 examples of scored RTWS produced by student teachers at various partner schools. These RTWSs also have annotations to help users understand strengths and weaknesses of each exemplar (The Renaissance Partnership for Improving Teacher Quality Project, n.d.b).

The Renaissance Teacher Work Sample organization. The Renaissance Teacher Work Sample is a complicated instrument that "focuses very directly on things that make a difference in student learning, things that teacher candidates can improve" (Robinson & Boody, 2003, p. 20). The following are the seven teaching processes the Renaissance teacher work sample (RTWS) is organized around:

1. *Contextual factors* – description of the school and surrounding community that would include a demographic description of the group of students and any other relevant factors, that maybe impacting student learning.
2. *Learning goals* – provides a list of challenging and appropriate learning goals to be addressed in the unit described in the work sample. These goals should be aligned with national, state, or local standards.
3. *Assessment plan* – contains multiple pre- and post-assessment measures that were employed in the unit for formative and summative assessments of student learning. These assessments should be aligned with unit learning goals.

4. *Design for instruction* – provides a summary of instructional methods used by the teacher candidate to help students meet learning goals for the unit. The instruction should take into consideration various student needs.
5. *Instructional decision-making* – this section describes formative assessment measures used by the teacher candidate to make changes to instruction based on student learning.
6. *Analysis of student learning* – contains analysis of student data collected by the teacher candidate. The teacher candidate is expected to comment on why individual students and groups of students were successful or less than successful in learning the material of the unit.
7. *Self-evaluation and reflection* – is a teacher candidate's reflection on the effectiveness of his/her teaching and attempts to improve student learning of all students. Candidates are expected to propose future activities that would be effective in helping all students meet unit learning goals (The Renaissance Partnership for Improving Teacher Quality Project, n.d.a).

The process of teacher work sample creation is clearly described in paper-based manuals and web-based tutorials. Multiple examples of previous samples are available to pre-service teachers to review and learn from (The Renaissance Partnership for Improving Teacher Quality Project, n.d.b).

RTWS preparation and scoring. Typically the samples are compiled by teacher candidates in partner institutions during their student teaching experience. In about 20-25 pages, following the *RTWS Prompt and Rubrics*, each student teacher describes his/her activities during a period of 2-3 weeks, and includes examples of student work and assessments used, and provides reflections. Later on, these samples are scored by trained educators during specifically designed scoring sessions, using the RTWS Rubric (see Appendix B). Each sample receives an overall score on the following 3-point scale: 1 – standards were not met; 2 – standards were partially met; and 3 – standards were met.

During the 5-year life of the Project, since its beginning in 1999, over 3,000 work samples in total have been collected and scored by eleven partner institutions. In addition to the extensive field testing of the RTWS instruments (i.e., *Prompt and Scoring Rubric*), a number of research studies have been carried out to test content validity, score generalizability, quality of student learning assessment, and alignment with standards of RTWS (e.g., Denner, Norman, Salzman, & Pankratz, 2003; Denner, Norman, Salzman, Pankratz, & Evans, 2003; Denner, Pankratz, Norman, & Newsome, 2004; Denner, Salzman, & Bangert, 2001; Keese & Brown, 2003). Overall, the research findings of these studies indicate direct correspondence between the targeted teaching behaviors to actual teaching practice; support the generalizability of the work sample scores and high dependability coefficients for panels of three or more raters; reveal positive correlation of RTWS student assessments with ratings on an independent scale; and demonstrate close alignment with evaluation standards.

RTWS at the University of Northern Iowa. The University of Northern Iowa (UNI) is a partner institution in the Renaissance Partnership for Improving Teacher Quality Title II grant. UNI has a well established teacher preparation program graduating approximately 500-700 new teachers per year.

Starting in Fall 2000 and by the time this study was conducted in May of 2004, nearly 900 teacher work samples had been submitted by UNI teacher candidates from eight teaching centers. Each UNI sample receives a unique ID and is entered in a special RTWS database along with the information about the sample and the pre-service

teacher who compiled it. After that the name of the student teacher is removed from the sample itself and it is ready for scoring.

Scoring sessions are organized twice a year and attract over 100 educators each time. RTWS raters are recruited among UNI faculty and state's K-12 schools. Many K-12 scorers are cooperating teachers supervising UNI students during their student teaching. Overall, each scoring session lasts a total of about three-four hours and starts with a brief training, explaining the rating procedure and documentation used. Each rater progresses at his/her own speed. Several experienced session facilitators are present to answer questions and assist raters with their tasks. After the rating session is over, individual sample ratings are entered into the database and communicated to the individual pre-service teachers.

Currently, the University of Northern Iowa's College of Education continues to use the Renaissance Teacher Work Sample Methodology and assessment tools in its teacher preparation program. To date, over 2,000 teacher work samples have been collected and scored at UNI. The University of Northern Iowa uses RTWS for high-stake decision making, such as recommendation of its teacher candidates for initial licensure.

Using TWSM to Connect Teacher Work to Student Learning

TWSM as authentic assessment. Ayers et al. (1996) declared that since TWSM required a student teacher to demonstrate and document a 3- to 5-week period of work done in a classroom, it could be regarded as an extended, authentic performance instrument. The design of the TWS methodology assists and guides pre-service teachers

towards functioning as effective teachers usually do in their classrooms, as they perform duties related to instruction (Cotton, 1995; National Board for Professional Teaching Standards, 1998; Scriven, 1994, 1996). As such, TWSM and its variation - RTWSM, are increasingly recognized in teacher preparation as instruments for authentic assessment (e.g., Cartwright & Blacklock, 2003; Fredman, 2002; Girod, 2002; Hegler, 2003; Keese & Brown, 2003; Rudden, 2003).

Licensure use. Overall, the work sample uses are not limited to assessment, it can be used for improving instruction and learning, as well as an evaluation tool. When used for evaluation, it may be used for formative, summative, high-stakes, and not high-stakes evaluations. For licensure purposes, (R)TWSM can provide a direct measurement of the work performed by a student teacher, and the effects that work has on student learning, rather than relying on such proxy measures as grade point averages earned in a teacher education program or testing the student teachers' knowledge of content. There are anecdotal reports that new teachers even use their work samples at employment interviews to illustrate their teaching skills with concrete examples.

RTWSM is an outcome-based and content-dependent assessment tool. While it is not intended to be used as a *single* indicator of teacher candidate readiness and the quality of a teacher preparation program, it provides ample evidence on the candidate's ability to impact student learning. Moreover, RTWSM takes into consideration school and community factors, teacher's knowledge and skills, assessment procedures used, student competences and accomplishments, and a teacher's reflections on his/her effectiveness. All of these components together form the basis of the teacher work

sample methodology, making (R)TWS "an unusually complex applied performance assessment system that is embedded in a teacher's daily work" (Schalock & Myton, 2002, p. 8).

Linking teaching and learning. Although some may argue that Oregon model of TWSM is not capable of linking teaching practices to student outcomes (Darling-Hammond, 1998), a mounting number of studies indicate that in fact TWSM can empirically connect teacher performance and student learning.

For instance, Ayres et al. (1996) stated that at Western Oregon University, the design of teacher work sample as an extended, authentic performance task, which focuses on pupil learning and is reported by student teachers, makes the methodology *meaningful* for both the student teachers and the university supervisors who evaluate them. Schalock and Myton (2002) supported this thinking by stating that "a TWS connects teaching and learning through an informed interweaving of the seven interrelated core features that define the methodology" (p. 8).

The head of the *Renaissance Partnership for Improving Teacher Quality* project, Roger Pankratz, argued that RTWS methodology has a high potential for improving student learning opportunities (Pankratz, n.d.). Moreover, Keese and Brown (2003) supported the position that Renaissance TWSM also strongly connected teaching and learning by stating that it is "a method to document the effects of teacher performance on student learning outcomes. The work sample uses whatever form of assessment – authentic or standardized – the teacher develops to document increase or decrease in student learning" (p. 4). Cartwright and Blacklock (2003) contributed to the same idea

by stating that "through a teacher work sample process, candidates document their ability to diagnose needs, plan instruction, deliver instruction, and assess progress of a [learner]" (p. 7).

According to Ayres et al. (1996), TWSM has been evaluated on three fronts in order to demonstrate its efficacy in teacher preparation:

1. by comparing it to criteria for quality assessments laid down by experts in the field of assessment;
2. by conducting statistical analyses of how data are distributed, especially data that are related to the learning gains made by pupils taught using the methodology; and
3. by statistically analyzing the results obtained to determine whether TWSM measures and related variables explain student progress in learning (Ayres et al., 1996).

However, it is important to mention that concrete and specific evidence of teacher's impact on student learning may not be readily available in future teachers' work samples. For example, in their search for *specific examples of concrete evidence* of teachers' impact on student learning, as recorded in RTWS, Denner, Salzman, and Harris (2002) discovered that such data was difficult to locate in most TWSs examined in their study. Given this result, the authors conclude that:

> This finding has important implications because it points to a need to improve our guidelines and task prompts for producing teacher work samples. It also suggests that we may need to alter our teacher preparation program to better prepare our candidates to supply this data, if our TWS are to supply credible quantitative evidence for our candidates' impact on student learning (p. 24).

Research on Teacher Work Sampling

Oregon Teacher Work Sampling. In the late 1990s, Airasian (1997) described the Oregon Teacher Work Sample Methodology (TWSM) as a *developing* method

aimed at linking learning gains made by students to teacher performance. Since 1997, TWSM has been researched in great depth: the efficacy of the methodology was explored in a number of ways that include its use in teacher education as a measurement technique, a research topic, and a licensure tool (e.g., Girod, 2002; McConney, Schalock, & Schalock, 1998; Pratt, 2002; Rudden, 2003; Schalock, 1998).

Renaissance Partnership Teacher Work Sampling. In addition to field testing of the Oregon TWSM described above, extensive empirical studies were carried out on its variation – RTWS and its instruments. For instance, a number of research studies have been carried out to test content validity, score generalizability, quality of student learning assessment, alignment with standards of the teacher work sampling, and correlation of RTWS student assessments with ratings on an independent scale (e.g., Cartwright & Blacklock, 2003; Denner, Salzman, & Bangert, 2001; Denner, Salzman, Harris, 2002; Denner et al., 2003, 2003; Fredman, 2002; Hegler, 2003; Keese & Brown, 2003).

Validity of the assessment and instrumentation of RTWSM. The findings of several studies (e.g., Denner, Norman, Salzman, & Pankratz, 2003; Denner, Pankratz, Norman, & Newsome, 2004; Denner, Salzman, & Harris, 2002; Salzman, Denner, Bangert, & Harris, 2001) indicate that TWS assessment and instrumentation meets the elements of the Crocker's (1997) construct of *content representativeness*, which includes the three criteria: realism, criticality, and frequency. For instance, RTWS tasks, i.e., targeted teaching behaviors, are viewed to represent realistic classroom

experiences and correspond well with specific standards, like INTASC (Interstate New Teacher Assessment and Support Consortium, 1992).

Generalizability and RTWS. Several studies (e.g., Denner, Norman, Salzman, & Pankratz, 2003; Denner, Salzman, & Harris, 2002; Salzman, Denner, Bangert, & Harris, 2001) examined a matter of inter-rater reliability in scoring RTWSs. The importance of this research is stated in Denner, Pankratz, Norman, and Newsome (2004):

> It is important for scores on performance assessments to show a high degree of accuracy and consistency, if the scores are gong to be used for making high-stake decisions about the performance levels of your teacher candidates. Hence, the judgments of the raters must be in close agreement with one another. It is also important to show that the scores can be generalized beyond the particular tasks, the particular raters, and the particular occasion of assessment, if the scores are to be used to make general inferences about candidates' abilities to meet institutional and state teaching standards and their abilities to perform successfully as teachers. (p. 35)

These authors suggest using the Generalizability Theory (Shavelson & Webb, 1991) for the abovementioned purposes. First introduced by Cronbach and his colleagues in the 1970s (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), the basis of Generalizabiity theory (aka G theory) is

> the ability to determine multiple sources of error in measurement using analysis of variance (ANOVA) techniques. This yields a generalizability (reliability) coefficient that may include multiple error sources, unlike Classical Test Theory, and also avoids the requirement of parallel tests. Instead, generalizability theory relies on a less restrictive assumption by randomly drawing items from the same pool of possible items. (Measurement Experts, n.d.)

In contrast with Classical Test theory, which assumes that there is a single error, G theory presumes that there are multiple components of error each of which that can be estimated if data are collected correctly. Moreover, the value and uniqueness of G

theory is in its ability to separate error due to differences in measurement conditions. G theory reinterprets classical reliability theory as a theory regarding the adequacy with which one can generate from a sample of observations to a universe of observations from which it was randomly sampled.

When applied to the RTWSM research, several studies that applied G-theory (e.g., Denner, Norman, Salzman, & Pankratz, 2003; Denner, Salzman, & Harris, 2002; Salzman, Denner, Bangert, & Harris, 2001) reported high generalizability coefficients for panels of three or more trained raters. These findings indicate that RTWS scores can be used for high-stakes decision making regarding candidate performance on the teacher work sample.

Finally, in their *How To Manual for Teacher Educators Who Want to Collect, Use, and Report Valid and Reliable Performance Data on Teacher Candidates with a Link to P-12 Student Learning*, Denner, Pankratz, Norman, and Newsome (2004), describe a step-by-step process of conducting a generalizability study of performance data.

Other aspects of (R)TWS research. Some of the other aspects of TWSM that have been substantially researched include the advantages of TWSM, the measures used in the TWSM, and the methods used to link student learning to teacher performance. Schalock, Schalock, and Girod (1997), among others, stated a number of ways in how information obtained from TWSM can be used *confidently* and how it can add value to teacher preparation and licensure. For instance, the researchers state that a teacher work sample is a truly authentic work of a pre-service teacher that is rather complex and

demanding, assuring multifaceted representation of the teacher's professional knowledge and skills. In this respect, teacher work sampling allows for "quality assurance" in the licensing of teachers. Furthermore, Schalock et al., (1997) indicate that, in addition to bringing legitimacy to the teacher preparation and licensure process, TWSM has the potential for use as an *instructional tool.* Several separate studies serve as illustrations to this use of (R)TWSM (e.g., Cartwright & Blacklock, 2003; Fredman, 2002; Girod, 2002; Hegler, 2003; Keese & Brown, 2003; McConney et al., 1998; Pratt, 2002; Rudden, 2003; Schalock, 1998). Overall, teacher work sample methodology is designed to serve multiple purposes and audiences.

Several benefits and overall usefulness of teacher work sampling have been praised by one of the leaders of American education. In 1997, Darling-Hammond stated that the Oregon work sampling approach should be commended because teaching is highlighted in the context of the educational goals developed by the teacher, the context of the classroom, and student learning, which is measured in ways that try to link learning to the desired educational goals. In this respect, the teacher work sample can contribute to effective teacher evaluation and improvement. Furthermore, Darling-Hammond (1997) also pointed out the value of the TWSM approach to teacher assessment, which helps teachers to carefully evaluate practices, contexts, and outcomes, including the work done by pupil and teacher. These qualities give teachers the opportunity to reflect on their work in productive ways and develop habits of critical thinking and practice. In this respect, Darling-Hammond suggested that teacher work sampling resembles other assessment programs such as the National Board for

Professional Teaching Standards (NBPTS) certification and licensure testing, and the Interstate New Teachers Assessment and Support Consortium (INTASC), which use work sampling methodology to evaluate teachers based on their lesson plans, instructional practices, assessment of student work, feedback in the context of the samples of student work and progress made over a period of time.

The Renaissance Partnership for Improving Teacher Quality Project conducted extensive field testing of the RTWS methodology and instruments developed by the grant (i.e., *Prompt and Scoring Rubric*). The Project staff also researched extensively content validity, score generalizability, quality of student learning assessment, and alignment with standards of the teacher work sampling (e.g., Denner et al., 2001; Denner et al., 2003, 2003; Keese & Brown, 2003). The Project research findings indicate direct correspondence between the targeted teaching behaviors to actual teaching practice; support the generalizability of the work sample scores and high dependability coefficients for panels of three or more raters; reveal positive correlation of TWS student assessments with ratings on an independent scale; and demonstrate close alignment with evaluation standards.

However, as Darling-Hammond (1997) notes, further empirically-based work needs to be done to make the teacher work sampling a tool for formative evaluation of student teachers or even veteran teachers. The most important areas that need improvement are the measurement of outcomes, practice, and the methods used in evaluating the effects of the intended learning outcomes of TWSM.

Research on rater characteristics. The need for further research on the role of teacher demographic characteristics (e.g., professional training and years of teaching experience), beliefs, and other variables that impact their teaching and other professional activities, have been stressed for several decades. Currently, the body of research is rather limited. One of the studies (Tinker Sachs, Kong, Lo, & Lee, 1994) suggests that when teachers are employed as "judges" or "raters" of educational practice, their individual characteristics also play a major part in their decision making. Tinker Sachs, Kong, Lo, & Lee in their study of Hong Kong foreign language teachers reported that "low or high feelings about one's knowledge would affect how one feels about the degree of decision making one can make" (p. 183). The results of this study also indicate that teacher "qualification" or level of education impacts the way they teach. Additionally, the study points out that teaching experience plays a major part in how teachers define "good" teaching.

In research studies that involve educators in judging or rating the quality of a service (e.g., teaching) or product (e.g., teacher work sample or portfolio), researchers usually examine inter-rater reliability of raters to determine how reliable (or representative) the ratings are. Inter-rater reliability is also examined when a new assessment tool is being piloted to minimize any sources of error. These studies attempt to answer a question regarding necessary qualifications of a rater or rater characteristics, for example, content area expertise, teaching experience, level of teaching experience (K-12 vs. university level), experience with the assessment in question, as well as basic demographic characteristics (i.e., gender, age, etc.). In several studies focused on

Renaissance teacher work sampling, researchers were reporting high inter-rater reliability levels (e.g., Denner et al., 2003, 2003; Denner et al., 2001). Studies that involved portfolio assessments (e.g., Campbell, Melenyzer, Nettles, & Wyman, 2000; DeFina, 1992; Devlin-Scherer, 2003; Long & Stansbury, 1994; Wolf, 1991) and classroom observations (e.g., Burry, 1990; Evertson & Burry, 1988; Webb & Brown, 1969) reported low within- or between-rater reliability among subjects of the studies (i.e., portfolio reviewers or classroom observers).

However, it is important to mention that the majority of the existing RTWS studies examined a limited set of rater characteristics and their impact on scoring of TWSs. For example, Salzman, Denner, Bangert and Harris (2001) report that their study involved PK-12 teachers, a principal, and faculty members from a teacher education program and the arts and sciences as scorers of TWSs. In addition, the scorers varied in their amount of teaching experience, level of education, and gender. During their statistical analysis, the authors used Generalizability Theory (Shavelson & Webb, 1991) to calculate "total score dependability coefficients for absolute decisions." The study reported that the effect of individual raters was not statistically significant and reliable results can be achieved with as few as two raters scoring each TWS. Additional studies (Denner, Norman, Salzman, & Pankratz, 2003; Denner, Salzman, & Harris, 2002) replicated this analysis and recommended that a high level of inter-rater reliability can be achieved with panels of three or more raters, especially in cases when results are being used in high-stakes decision making.

Foreign Language Teaching and Learning in the United States

Foreign language teaching and learning has long been considered an important part of the public school curriculum. This notion has been reinforced in recent years with passage of the *No Child Left Behind* Act (U.S. Department of Education, 2002a) that listed foreign language study as one of the core subject areas that all children need to become proficient in. This section of the chapter will (a) briefly summarize the history of foreign language teaching; (b) discuss the importance of foreign languages for our country's political, economic, and commercial success; (c) describe the present situation with respect to foreign language instruction, focusing on types of programs employed and languages taught; (d) address national standards for foreign language learning and newly developed national standardized assessment of foreign language proficiency among high school-age students; and (e) present efforts to improve language instruction by preparing high quality foreign language teachers.

Historical Overview

In the early days of foreign language study in the United States, educators were more concerned with development of grammar knowledge and skills as well as literary knowledge in a foreign language, thus teaching "about" the language rather than the language itself. Speaking skills were considered to be "irrelevant" and "impractical" to the study of a foreign language (Coleman, 1929). The grammar-translation approach to language teaching, popular in the first half of the 1900s, was replaced by a new approach – the Audio-Lingual Method (ALM) – based on behaviorist principles in the late 1950s stimulated by the launching of Sputnik by the Soviet Union. For a number of

years the ALM continued to be used in various foreign language programs in the country, but by the 1980s, it was considered to be ineffective in developing proficiency in another language and was replaced by new approaches focused on communicative language teaching (Richards & Rodgers, 2001). Presently, the methodology of foreign language teaching is becoming more standards-based with an emphasis on developing foreign language proficiency in all its complexity, including knowledge and practical skills related to the language and culture.

The Importance of Foreign Language Education

In spite of the multilingual American heritage, the majority of modern Americans remain monolingual and, currently, only one out of three secondary school students studies a foreign language (Draper & Hicks, 2002). However, in a world that is becoming more global and interdependent, the need for foreign language skills is increasing rapidly. In addition to the linguistic outcomes of the foreign language study, learners also acquire various cultural knowledge and skills that are needed in order to function effectively in a multicultural society which requires at least some understanding of other cultures. As stated in the *National Standards for Foreign Language Learning*:

> The businessperson, the poet, the emergency room nurse, the diplomat, the scientist, and the teenage computer buff are representative Americans who play diverse roles in life, yet each could present a convincing rationale for the importance of studying a foreign language. Their reasons might range from the realistic to the idealistic, but one simple truth would give substance to them all: to relate in a meaningful way to another human being, one must be able to *communicate*. (American Council on the Teaching of Foreign Languages, 1996)

There are multiple reasons why people choose to study a foreign language. Some are looking for a challenging cognitive experience; others hope that knowledge of another language will help them find a rewarding career; some are simply interested in learning about other cultures; while others see it as a fulfillment of a graduation or college admission requirement. Regardless of their reasons, proficiency in more than one language and culture enables individuals to do the following:

- communicate with other people in other cultures in a variety of settings;
- look beyond their customary borders;
- develop insight into their own language and culture;
- act with greater awareness of self, of other cultures, and their own relationship to those cultures;
- gain direct access to additional bodies of knowledge; and
- participate more fully in the global community and marketplace (American Council on the Teaching of Foreign Languages, 1996).

In spite of all the benefits of foreign language study, in reality, multiple studies point out a serious lack of understanding of world affairs among American school children (cited in Benevento, 1985). Benevento continues by citing a study conducted by UNESCO that involved teenagers from nine countries. The results of the study were very alarming since "American students ranked next to last in comprehension of foreign cultures" (in Benevento, 1985, p. 10).

In the early 1980's, in his book *The Tongue-Tied American: Confronting the Foreign Language Crisis* (Simon, 1980), Congressman Paul Simon provided a convincing argument that knowledge of foreign languages in the United States is not simply a matter of interest, but is an issue of national security. He listed unfortunate events and serious dangers in the area of national diplomacy and commerce that took

place due to the lack of proficiency in foreign languages and knowledge of culture. Although his statistics are dated, Simon's arguments are still appropriate today. This point was demonstrated by security problems at U.S. military bases caused by the lack of trusted translators during recent military conflicts. Thus, insufficient foreign language skills in the United States continue to represent a national problem.

Importance of foreign language education summary

To summarize this section, by recognizing the importance of foreign language education to our society as a whole and placing more emphasis on language study in educational institutions at all levels, we will provide the citizens of the 21st century with a greater opportunity to acquire knowledge and skills for both academic and personal success. This in turn will provide for a more secure future for our nation.

Languages Taught

A wide variety of foreign language programs currently exists in the United States at elementary, middle, high school, and university levels. In addition to formal language programs at public and private institutions, there are many summer camps, exchange, and study abroad programs that offer learners more opportunities to study languages other than English.

It is also important to note some positive changes in foreign language teaching and learning. In recent years, the number of high school students that are enrolled in foreign language courses increased dramatically, in part due to the increased requirements for college admission (Draper & Hicks, 2002). American students currently study a much wider variety of languages than they have in the past. Although

Spanish, French, German, and Latin continue to be the most commonly taught languages, students around the nation are learning other languages as well. Programs in Arabic, Chinese, Italian, Japanese, Portuguese, and Russian continue to attract students.

Historically, the percentage of American high school students studying a foreign language reached an all-time high in 1910, when 49% were studying Latin and 34% learning modern foreign languages, for a combined total of 83% (Parker, 1957). Since the early 1900s, along with the increase in the nation's population, developmental changes occurred in American public education which resulted in a substantial increase in high school-age youth receiving secondary education. While the percentage of students studying foreign languages was higher in the 1900s, the number of students enrolled in foreign languages is now at an all-time high.

According to the American Council on the Teaching of Foreign Languages (ACTFL) *Foreign Language Enrollments in Public Secondary Schools Report* (Draper & Hicks, 2002), foreign languages are now studied by nearly seven million American students in public schools, primarily in grades 7-12, which represents 33.8% of total enrollment in these grades. These numbers show an increase in overall foreign language enrollment since the previous survey, carried out in 1994 (cited in Draper & Hicks, 1996), when about six million public school students were enrolled in foreign language classes (32.8% of total school enrollment in 7-12 grades). According to Rhodes and Branaman (1999), a 1997 survey of secondary enrollments conducted by the Center for Applied Linguistics (CAL) found that 51% of all U.S. high school students were enrolled in a foreign language that year. Moreover, John Watzke reports that "transcript

analysis of a national sample of graduated high school seniors by the National Center for Education Statistics found that 80.6% of 1998 graduates had enrolled in a foreign language course during their last four years in school" (2003, p. 213). The abovementioned numbers clearly indicate a substantial increase in foreign study in American schools since the 1980s, when only 19% of students in grades 7-12 were studying languages other than English (cited in Benevento, 1985).

Duration of foreign language study

In regards to the length of foreign language study, researchers seem to be contradicting one another. For instance, Watzke (2003) reports some positive changes: "like other academic subjects, foreign language experienced increases in advanced-level enrollments during the 1990s. From 1985 to 1994, the proportion of foreign language enrollments at the advanced level increased from 17.7% to 20.4%" (p. 213).

In their study, Draper and Hicks (2002) report that in spite of the presence of the National Standards for Foreign Language Education (ACTFL, 1996), calls to begin language study earlier in children's schooling, and apparent public interest and need for foreign language knowledge in the U.S. government, business, and industry, no changes in length of language study can be observed. Their 2002 report did not find any significant differences in length of foreign language study, with the majority of students still taking the language for only two years, usually not long enough to develop usable skills.

Study of individual languages

When study of individual languages is examined (see Figure 2), Draper and Hicks (2002) report that Spanish continues to remain the most commonly taught foreign language. Primarily for budgetary reasons, U.S. public schools continue to offer Spanish classes, while in some cases closing other foreign language programs. This would explain why the enrollments in Spanish programs have increased dramatically from 54% of the total foreign language enrollment in grades 7-12 in the 1980s (cited in Benevento, 1985) to almost 70% of all students taking foreign language classes in 2000 are being enrolled in Spanish (Draper & Hicks, 2002).

Figure 2. Foreign Language Enrollments as Percentage of Total Foreign Language Learning (Draper & Hicks, 2002).

As was mentioned above, during the same time period, between the 1980s and 2000, most other languages, except for Spanish, either had a reduction in student enrollments or remained relatively steady. French is the second most commonly taught language in the Unites States. However, enrollment in French has been decreasing from 18.3% in the 1982 survey to 12% in 2000 of all students taking foreign languages (Draper & Hicks, 2002). German continues to be the third most commonly taught language, followed by Latin, but their enrollments also decreased from 9% and 5% in 1982 to 4.8% and 2.7% in 2000 respectively. Overall, study of the Italian language is also down when compared with the results of the 1982 survey, from 2% in 1982 to 1.2% in 2000 of all students taking foreign language classes. But as reported by Draper and Hicks, "Italian was the one bright spot of the non-Spanish languages. Enrollments were up by 22,000 students, a 38% increase over the prior survey [1994], and the first measurable increase in the percentage of high school students studying Italian in 20 years" (2002, p. 1).

Overall, the importance of study of foreign languages and cultures has been affirmed by educators and policy makers (ACTFL, 1996; U.S. Department of Education, 1994). Historically, foreign language teachers played an important role in identifying weaknesses in the foreign language field and outlined the following problems in the early 1980s (as cited in Benevento, 1985):

1. inappropriate content, outdated materials, and ineffective methods;
2. inconsistent standards on measures of language proficiency;
3. weak teacher training programs;
4. limited development and dissemination of research; and

5. poor communication to students and the public in general about the importance of foreign language study.

Several of these problems were addressed by the national language organizations through development of national and state foreign language standards, performance guidelines for language learners, and performance assessment tools. However, a number of these problems still remain.

The National Standards for Foreign Language Learning

By the mid-1990, many national educational organizations have launched ambitious projects to define specific content standards in their respective subject areas in response to the initiatives passed in the *Goals 2000: Educate America* Act (U.S. Department of Education, 1994). Professional organizations and consortia in such subject areas as math, science, language arts, economics, foreign language, English as a Second Language (ESL), art, geography, history, health and physical education, civics and government, career technical education and several others (e.g., technology and early childhood education) have established or are in the process of establishing standards.

In the area of foreign language, the non-language-specific *National Standards for Foreign Language Learning: Preparing for the 21st Century* first appeared in 1996 (National Standards in Foreign Language Education Project). Moreover, the study of foreign languages was the seventh subject area to receive funding from *Goals 2000* federal initiative for development of national standards. These standards were the outcome of a national collaborative effort supported by four major national associations

for foreign language education: the American Council on the Teaching of Foreign Languages (ACTFL), the American Association of Teachers of French (AATF), the American Association of Teachers of German (AATG), and the American Association of Teachers of Spanish and Portuguese (AATSP). They were endorsed by several other organizations involved in the field of language teaching and learning, such as the American Association of Applied Linguistics (AAAL), the Chinese Language Association of Secondary-Elementary Schools (CLASS), the Modern Language Association (MLA), and state and regional associations of language teachers.

The foreign language standards present five goal areas that foreign language study should strive to encompass: communication, cultures, connections, comparisons, and communities, which are often referred to as "five C's" of foreign language study. Within each of these goal areas several content standards are specified representing a total of eleven individual content standards:

> Goal: Communication: communicate in languages other than English. Communication goals are considered the key of language study and should be carried out orally, in writing, and through reading of literature.
>
> > Content Standard 1.1: Students engage in conversations, provide and obtain information, express feelings and emotions, and exchange opinions.
> >
> > Content Standard 1.2: Students understand and interpret written and spoken language on a variety of topics.
> >
> > Content Standard 1.3: Students present information, concepts, and ideas to an audience of listeners or readers on a variety of topics.
>
> Goal: Cultures: gain knowledge and understanding of other cultures. Knowledge and understanding of another culture is also considered to be an important part of language proficiency and this proficiency cannot be achieved fully without understanding of the foreign culture.
>
> > Content Standard 2.1: Students demonstrate an understanding of the relationship between the practices and perspectives of the culture studied.

> Content Standard 2.2: Students demonstrate an understanding of the relationship between the products and perspectives of the culture studied.
>
> Goal: Connections: connect with other disciplines and acquire information. Language study offers learners unique opportunities to establish connections with other subject areas and information sources and learn from them.
>
> Content Standard 3.1: Students reinforce and further their knowledge of other disciplines through the foreign language.
>
> Content Standard 3.2: Students acquire information and recognize the distinctive viewpoints that are available only through the foreign language and its cultures.
>
> Goal: Comparisons: develop insight into the nature of language and culture. The study of foreign languages helps learners to increase their understanding of their own language and culture through establishing comparisons between their native language and heritage and the foreign culture(s) they study.
>
> Content Standard 4.1: Students demonstrate understanding of the nature of language through comparisons of the language studied and their own.
>
> Content Standard 4.2: Students demonstrate understanding of the concept of culture through comparisons of the cultures studied and their own.
>
> Goal: Communities: participate in multilingual communities at home and around the world. Through the study of foreign language and culture learners are able to become active participants in multilingual communities.
>
> Content Standard 5.1: Students use the language both within and beyond the school setting.
>
> Content Standard 5.2: Students show evidence of becoming lifelong learners by using the language for personal enjoyment and enrichment. (National Standards in Foreign Language Education Project, 1999, p. 9)

Since their first publication in 1996, the impact of the foreign language standards has been great. Shortly after formulating the generic standards, ACTFL, AATF, AATG, AATSP, and CLASS, joined by the American Association of Teachers of Italian (AATI), the American Council of Teachers of Russian (ACTR), the Association for Computational Linguistics (ACL), and the National Council of Japanese Language Teachers (ACJLT), rewrote the standards as the *National Standards for Foreign Language Education* and complemented them by nine language-specific standards for

Chinese, Classical Languages, French, German, Italian, Japanese, Portuguese, Russian, and Spanish (National Standards in Foreign Language Education Project, 1999). The language-specific standards were closely aligned with the non-language specific standards, but contained language-specific examples of learning scenarios and progress indicators, as well as offered lists of language- and culture-related classroom and bibliographic resources. The majority of the language-specific standards are focused on foreign language education at K-12 and post-secondary education levels (K-16). The standards are also designed to be used for assessment and follow the format of *Goals 2000*, specifically by indicating performance standard benchmarks for grades four, eight, and twelve. Several foreign languages (French, Japanese, German, Italian, Portuguese, and Spanish) have articulated additional benchmarks to include the undergraduate college years of language study. Each of the performance indicators appears in a form of a sample description of what learners should be able to do. For example, under communication content standard 1.1, the following growth can be observed during K-12 language study:

Sample Progress Indicators, Grade 4:

- Students give and follow simple instructions in order to participate in age-appropriate classroom and/or cultural activities.
- Students ask and answer questions about topics such as family, school events, and celebrations.
- Students share likes and dislikes with each other and the class.

Sample Progress Indicators, Grade 8:

- Students follow and give directions for participating in age-appropriate cultural activities and investigating the function of products of the foreign culture. They ask and respond to questions for clarification.

- Students exchange information about personal events, memorable experiences, and other school subjects with peer and/or members of the target cultures.
- Students compare, contrast, and express opinions and preferences about the information gathered regarding events, experiences and other school subjects.

Sample Progress Indicators, Grade 12:

- Students discuss, orally or in writing, current events that are of significance in the target culture or that are being studied in another subject.
- Students develop and propose solutions to issues and problems that are of concern to members of their own and the target cultures through group work.
- Students share their analyses and personal reactions to expository and literary texts with peers and/or speakers of the target language (National Standards in Foreign Language Education Project, 1999, p. 42-43).

Moreover, according to Watzke (2003):

> in terms of classroom practice, the goals and content standards are illustrated in a series of learning scenarios for each language. The learning scenarios provide a third-person account of how content standards are met in actual instructional activities or units. The scenarios describe the thematic topic, the setting and classroom activity, unit or project, the content standards targeted by the activity, and reflections on how each targeted content standard is met as well as additional instruction that might further meet these standards (p. 223).

In addition to the national generic and language-specific sets of foreign language standards, many states responded to the call for standards by formulating their own sets of foreign language standards, which were closely related to the national standards. According to a national survey (Rhodes & Branaman, 1999), by the late 1990s, 30 states either had their own foreign language standards or reported that they used the national standards. As reported by the National Assessment Governing Board (NAGB), about

70% of the states have standards that "reflect the national [foreign language] *Standards* entirely or to a great extent" (Kenyon, Farr, Mitchell, & Armengol, 2000, p. 9).

National Foreign Language Assessment

With publication of the original National Standards for Foreign Language Learning in 1996, language educators were informed *what* their students should do in a foreign language, however the issue of performance, or *how well* they can complete language tasks, was addressed in a different document titled *Proficiency Guidelines*, published by ACTFL first in 1986 and then revised in the late 1990s (ACTFL, 1999a). *Proficiency Guidelines* are defined as "global characterizations of integrated performance in each of four language skills: speaking, writing, reading, and listening. The ACTFL Guidelines are based in large part on the language skill level descriptions and adapted for use in academic environments" (Breiner-Sanders, Swender, & Terry, 2001, p. 1). The guidelines have been extensively tested, revised, and refined primarily through Oral Proficiency Interview programs that have been in place since the 1950s and involved foreign language teachers as interviewers since the 1980s. Originally developed primarily for assessing language abilities in post-secondary level students, "the ACTFL *Proficiency Guidelines* have become widely used in schools, colleges, teacher training institutions, and the private sector" (Kenyon et al., 2000, p. 10). In 1998, in order to further assist K-12 foreign language educators, ACTFL developed *Performance Guidelines for K-12 Learners* that specified performance levels that should be achieved by elementary, middle, and high school students in the foreign

language (ACTFL, 1999a). These guidelines are closely aligned with the national foreign language standards.

ACTFL continues to expand its expertise in applying the national foreign language standards and *Performance Guidelines* through their work on developing *Performance Assessment Units* to assess learners' competence across the standards. The K-12 foreign language *Performance Assessment Units* were used as a basis for development of the National Foreign Language Assessment of Educational Progress (FL NAEP) developed by the National Assessment Governing Board (NAGB) in collaboration with ACTFL and the American Institute for Research (AIR). The assessment was developed in response to the call of the United States Congress that emphasized the importance of foreign language study, along with English, math, science, and other subject areas, in the *Goals 2000 Act* (U.S. Department of Education, 1994). According to Kenyon et al., (2000), in 1997 NAGB included the development and administration of a foreign language assessment in its 10-year schedule, with specific plans to conduct actual assessment activities during the 2003/2004 school year. This was the first planned attempt of this kind to collect national data regarding foreign language performance students in U.S. public schools that would also have major implications for the future of foreign language education. In the years preceding the assessment, many language educators and researches were thrilled about the upcoming data collection. These high expectations were well summarized in a statement by Kenyon et al.:

> Now, for the first time, the United States will have a comprehensive national source of information on what its students know and can do in a language other than English. Developing the framework for this national assessment is a critical task that presents an unprecedented opportunity to foster national discussion and to build national consensus - within the foreign language community and across government, business, industry, and the general public - on the role of foreign language education in America's future. (2000, p. 3)

About the Spanish National Assessment of Educational Progress

According to the National Center for Educational Statistics (2005), in a 2003 pilot study, the assessment was administered to a representative sample of 12th graders across the nation and consisted of two stages. Stage one had its focus to describe student demographic characteristics as well as attitudes towards language study, experiences with foreign languages, and language abilities. Stage two of the assessment was suppose to focus on language performance of a national sample of 12th graders who studied Spanish in a variety of programs, attempting to examine the connection between length of study and language competence. According Kenyon et al., (2000):

> The Spanish NAEP is based on the framework for assessing communicative ability in languages other than English. In this framework, listening, speaking, reading, and writing skills will be assessed through authentic communication tasks that are called for in daily life, school, and work. Assessment tasks will reflect four interrelated goals that provide the basis for communication. These goals include the following:
>
> - gaining knowledge of other cultures;
> - connecting with other academic subject areas to acquire knowledge;
> - developing insights into the nature of language and culture through comparisons; and
> - participating in multilingual communities at home and around the world.
>
> Performances will be evaluated on how well the student understands (comprehension) and can be understood (comprehensibility). The criterion of comprehension/comprehensibility subsumes language knowledge, the appropriate use of communication strategies, and the application of cultural knowledge. The Spanish assessment will require demonstration of the following:

- listening and speaking in the interpersonal mode,
- listening in the interpretive mode,
- reading in the interpretive mode, and
- writing in the presentational mode. (pp. i-ii)

Currently, it is unclear what the results of the national foreign language assessment pilot study were. The report of the study was not publicized. Moreover, according to the National Center for Education Statistics (NCES) website, "on March 6, 2004, the National Assessment Governing Board postponed the planned 2004 administration of the 2004 Foreign Language National Assessment of Educational Progress" (NCES, 2005). Instead, national assessments conducted in 2003 through 2006 focused on the following subject areas: reading, mathematics, science, and U.S. history. Furthermore, according to the NAEP 2002-2012 schedule in the *Nation's Report Card: An Overview of NAEP*, a publication of the NCES, the foreign language assessment (12-grade only) will not take place until the year 2012 (Johnson, 2004).

Foreign Language Teacher Preparation

In recent years, many foreign language teachers are experiencing a number of challenges, such as (a) increasing enrollments, (b) diverse learners, (c) challenging standards, and (d) emphasis on technology. Curtain and Pescola (1994) suggested that foreign language teachers today "require a combination of competence and background that may be unprecedented in the preparation of language teachers" (p. 241). Teacher preparation in general, and of foreign language educators in particular, has gained more attention since the importance of language learning was stressed in the *Goals 2000* (U.S. Department of Education, 1994). It also became clear that the success of integration of

the *National Standards for Foreign Language Education* (ACTFL, 1999b) also depended upon the knowledge and skills of foreign language teachers. These events lead to close examination of quality of teacher preparation programs preparing foreign language teachers. Research studies (e.g., Schrier, 1993; Wolf & Riordan, 1991) indicate that many foreign language teacher education programs continue to use a traditional model, where a teacher candidate is expected to complete his/her foreign language and education courses, and then spend some time student teaching in a public school setting. Additionally, the majority of teacher preparation programs at universities are administered by either departments or colleges of education or by a department of modern/foreign languages. According to these studies, this arrangement may cause some problems or uncertainties regarding preparedness of foreign language teacher candidates. For example, in many cases there is no mechanism, other than grades, that would provide programs with information regarding foreign language competences of teacher candidates. In some cases, foreign language-specific methods courses may not be available to teacher candidates and they end up taking general methods courses. Finally, it is quite common for foreign language teacher candidates during their student teaching experience to be supervised by educators with expertise in areas other than foreign language teaching and learning. According to Glisan (2001), new foreign language teachers often leave teacher preparation programs unable to speak the foreign language well enough to teach effectively.

The foreign language profession also identifies a number of chronic problems that may interfere with successful second language acquisition of all learners, as well as

those interested in becoming language teachers. One of the problems is "students typically begin foreign language study in grade nine and continue for only two years" (National Standards in Foreign Language Education Project [NSFLEP], 1999, p. 17). Moreover, this delayed introduction to a foreign language in combination with a short exposure to language learning, produces "learners with skills limited to learned expressions and restrained interactions" (NSFLEP, 1999, p. 14).

Another obstacle recognized by NSFLEP is overall accessibility of language programs to all students, "foreign language programs have not traditionally accommodated all students" (NSFLEP, 1999, p. 98). Finally, NSFLEP states that there is a "lack of multiple entry points into foreign language programs that accommodate prior learning" (1999, p. 22-23). These problems indicate that foreign language professionals not only struggle with identifying content and teaching methodology, but also finding ways to make their subject accessible and challenging to all students.

However, it is expected that the introduction of new national foreign language standards for K-16 and calls for increase in teacher quality will have a significant impact on teacher preparation programs for language teachers. The National Commission on Teaching and America's Future in its report *No Dream Denied: a Pledge to America's Children* (National Commission on Teaching and America's Future, 2003) identified a set of steps to quality teacher preparation. If followed, these steps will provide a framework for preparing high quality beginning teachers in all subject areas, including foreign language instruction:

1. Careful recruitment and selection of teacher candidates,
2. Strong academic preparation for teaching,
3. Strong clinical practice to develop effective teaching skills,
4. Entry-level teaching support in residencies and mentored induction,
5. Modern learning technologies, and
6. Assessment of teacher preparation effectiveness (p. 20).

In response to the national concerns regarding quality of preparation of foreign language teachers, the National Foreign Language Standards Collaborative, in partnership with ACTFL, has developed a set of standards for teacher preparation programs for foreign language teachers that were approved by the National Council for Accreditation of Teacher Education (NCATE) on October 19, 2002 (ACTFL, 2002). NCATE is recognized by the U.S. Department of Education as a professional accrediting body for teacher preparation. The overall mission of NCATE is to determine whether colleges of education meet rigorous national standards in preparing future teachers for various content areas and grade levels. According to ACTFL (2002), it was planned that beginning in 2004, foreign language teacher preparation programs seeking NCATE accreditation would be required to base their program reports on the *ACTFL Standards for the Preparation of Foreign Language Teacher*s. Additionally, the NCATE standards are often used as a guide by state departments of education in their efforts to assess quality of teacher preparation programs in their states, and therefore, it should be expected that foreign language programs will be required to follow the new standards during state certification reviews as well. As Schrier (2002) points out:

> The purpose of the new standards for teacher preparation programs for foreign language teachers is "to serve as a catalyst to programs so that they in turn may

> prepare highly qualified teacher candidates for an educational system that increasingly needs a globally educated citizenry. (p. 14)

The new ACTFL/NCATE foreign language standards for teacher preparation programs require institutions of higher education to (a) provide evidence that their teacher candidates meet each of the standards, (b) put in place an accountability system assessing progress of individual teacher candidates at various stages in the program, (c) assess foreign language proficiency levels using the well-accepted *ACTFL Proficiency Guidelines* (ACTFL, 1999a), and (d) encourage collaboration between various colleges and departments in order to provide foreign language teacher candidates with quality experiences in foreign language, literature, culture, and pedagogy (Phillips & Glisan, 2002).

What's Needed?

Based on the current tendency to focus on standards-based education and emphasis on performance-based teacher assessment, the potential of teacher work sampling is great. Data generated by the work samples provides a variety of insights into the knowledge and skills of specific prospective teachers as well as contributes to the overall accountability of a teacher preparation program. Although various aspects of TWSM have been researched, many more remain to be studied. The TWSM still can be described as "a work in progress," which implies that there are questions that need to be answered if the methodology is to be used widely. For example, the role of rater characteristics in assessment of teacher practice, as documented in teacher work samples, still remains to be examined. This research area should study any possible

rater differences that may become evident when scoring teacher work samples written in content-specific areas (e.g., foreign languages, music, etc.).

The current study analyzes scores of Spanish language teacher work samples produced by student teachers at the University of Northern Iowa in order to explore the role of rater characteristics (such as content knowledge) in their assessment of teacher practice.

Summary

The review of literature for this study involved the investigation of how standards in education have impacted the educational system in America from the national and state levels down to the local school building. The last two decades witnessed an increase in the demand for the establishment of standards in education, a demand which has resulted in a corresponding increase in the demand for accountability in achieving those standards. States across the nation have responded to these demands by developing diverse forms of accountability instruments and procedures believed to be appropriate to their situations. One such instrument is the Oregon Teacher Work Sample Methodology, which because of its design (linking teacher effectiveness to student learning), has the potential for decision making that affects teachers. As the Teacher Work Sample Methodology evolves and further research on its appropriateness as an effective tool for high-stakes decisions about achievement of established educational standards is conducted, it could become a valuable tool for preparing and assessing both beginning and veteran teachers. In summary, this study seeks to determine how rater characteristics impact perceptions of teacher practice as presented

in Spanish language teacher work samples submitted by teacher candidates at the University of Northern Iowa.

CHAPTER III

METHODOLOGY

This study examined the role of individual rater characteristics in rater's assessment of pre-service teacher practice. By analyzing the contents of ten teacher work samples submitted by teacher candidates at the University of Northern Iowa, this chapter presents a description of the methods and procedures used in answering the questions of the study. The following three study questions formed the foundation for the investigation:

1. Is there a significant difference between ratings assigned by raters with foreign language *content experience* and raters without foreign language content experience? Does this differ by sections of the teacher work sample?
2. What is the relationship between the amount of rater's overall *teaching experience* and his/her scoring of foreign language teacher work samples?
3. What is the relationship between rater's work sample *scoring experience* on his/her scoring of Spanish language work samples?

In order to address these study questions, the study used a causal-comparative research design. As defined by Gall, Gall, and Borg (2003), this research design:

> is a type of nonexperimental investigation in which researchers seek to identify cause-and-effect relationships by forming groups of individuals in whom the independent variable is present or absent – or present at several levels – and then determining whether the groups differ on the dependent variable. (p. 296)

Moreover, according to Gall et al. (2003), this research design typically does not allow for making strong conclusions about cause-and-effect in question, but it is "useful

for initial exploratory investigations or in situations where it is impossible to manipulate the independent variable" (p. 295). Overall, one of the major benefits of this research design is that it allows studying cause-and-effect relationships where an experimental research is not possible, i.e., an experimental manipulation of the independent variable cannot be done.

Sampling

Subjects of the Study

The population of this study is Iowa educators-raters of teacher work samples. For the purposes of the study, the investigator recruited 30 participants from various middle and high schools in Iowa. Sixteen of these participants – members of the experimental group – were foreign language teachers, while the other fourteen – the comparison group – were educators with various content specialties other than foreign language teaching and ESL. Participants of the study were asked to participate in Teacher Work Sample scoring training and later rate foreign language work samples submitted by UNI teacher candidates.

Recruitment of Participants

The investigator recruited participants for this study by sending out a letter of invitation via e-mail to all middle and high school teachers at the schools located in the area mentioned above (Appendix C). Each participant received a thank you letter (Appendix D), a small stipend (Appendix E), and a certificate of appreciation (Appendix F) for his/her participation in the study.

Teacher Work Samples Used in the Study

The principal investigator in the study reviewed all Teacher Work Samples (TWS) submitted by pre-service teachers at the University of Northern Iowa between Fall 2000 and Spring 2004, a total of about 600 samples. Only a small number of available TWSs, about 16, were focused on foreign language units and thus were suitable for the proposed research questions. The majority of the available foreign language TWS dealt with Spanish language units and only a small number (one or two) were focused on German or French. This was expected, given the information regarding the overwhelming presence of Spanish language programs in the American public schools presented in Chapter II. Since the preparation of teachers of Spanish is the largest segment of the UNI foreign language teacher preparation program, and given a larger number of Spanish language teacher work samples available for the study, the researcher made a decision to use only Spanish TWS in this study. In order to keep the number of Teacher Work Samples reasonable for raters to score during their scoring session, it was decided to select a total of 10 Spanish TWS created by pre-service teachers during their student teaching experience.

Moreover, only two of the Spanish TWSs were based on units taught in elementary grades. Since the number of elementary-grade samples was very limited, the researcher decided against using elementary grades Spanish TWSs in the study. Therefore, all 10 TWSs used in the study focused on Spanish units taught at 7-11 grade levels (see Table 1). The majority of TWSs selected for the study were from high

school Spanish classes; while only one sample was of an exploratory Spanish unit taught in the seventh grade.

Table 1

Teacher Work Samples Used in the Study.

TWS number	Foreign Language Content Area	Grade Level	When TWS was submitted	Total Length in pages
#1	Spanish	9th grade	Spring 2002	64 pages
#2	Spanish	9th grade	Fall 2003	42 pages
#3	Spanish	9th grade	Fall 2002	64 pages
#4	Spanish	11th grade	Fall 2001	48 pages
#5	Spanish	10th grade	Fall 2001	53 pages
#6	Spanish	10th grade	Fall 2003	31 pages
#7	Spanish	9th grade	Spring 2004	28 pages
#8	Spanish	9-11th grades	Fall 2000	35 pages
#9	Spanish	9-11th grades	Spring 2001	68 pages
#10	Spanish	7th grade	Spring 2002	25 pages

Finally, as can be seen in the Table 1, the TWSs used in the study varied in length from 25 to 68 pages, with an average length of about 46 pages. Since the length

of the main part of the sample is regulated and is approximately 20 pages long, this variation in length can be attributed to the amount of attached evidence that teacher candidates included with their sample.

Instrumentation

Participants of the study were asked to take part in a five-hour-long Teacher Work Sample training and scoring session. About one-fifth of the event (approximately 50 minutes) consisted of a brief introduction to the Renaissance Partnership for Improving Teacher Quality project activities and updates, a description and examination of the work sample rubric and scoring guide, and instruction on how to use the scoring instrument (rubric) in rating teacher work samples. During the rest of the session (approximately 240 minutes) participants completed a short demographic questionnaire (see Appendix A) and scored ten Spanish teacher work samples selected for the study, using *the Scoring Rubric* (see Appendix B), designed specifically for scoring TWS by the Renaissance Partnership for Improving Teacher Quality Project. The Spanish teacher work samples used in the study were submitted by UNI teacher candidates between Fall 2000 and Spring 2004. At the beginning of the training and scoring session, each of the raters received a packet of materials along with a unique ID number, handed out at random. During the event, raters used their ID number on all the documentation they submitted: the demographic questionnaire and all ten scoring rubrics.

The Demographic Questionnaire

This questionnaire was designed by the researcher and included a total of 17 questions (see Appendix A). The purpose of the questionnaire was to collect demographic data about each of the raters who participated in the study. In addition to the questions regarding participants' gender, level of education, teaching level(s), and participants' content area, the questions also asked about participants' years of teaching experience, knowledge of world languages, previous TWS experience and several others. The items on the questionnaire were similar to other demographic instruments used in empirical studies (e.g., Clark, 1988; Fullan & Stiegelbauer, 1991; Richards, Tung, & Ng, 1992). The questionnaire was administered right after the Teacher Work Sample training and collected before the participants started to rate work samples.

The Renaissance Partnership for Improving Teacher Quality Scoring Rubric

The *Rubric* used to rate Spanish teacher work samples is the instrument commonly employed by the Renaissance Partnership for Improving Teacher Quality project to rate teacher work samples (see Appendix B). The *Scoring Rubric* is organized around seven main processes of the Renaissance Teacher Work Sample methodology: (a) contextual factors, (b) learning goals, (c) assessment plan, (d) design for instruction, (e) instructional decision making, (f) analysis of student learning, and (g) reflection and self-evaluation. Each rubric section is based on descriptions of key indicators. The number of indicators varies from process to process, ranging from three (in the decision making section) to six (in design for instruction section). All components of the rubric are scored on a three-point scale with a three standing for

"standard met," two – "standards partially met," and one – "standard not met." These indicators of the rubric are written in such a way that they can be easily applied to work samples submitted by students teaching at various grade levels and content areas. After individual areas within seven processes are scored, each process receives an overall process score using a three-point scale. The final step of the scoring is to assign an overall score on the three-point scale to the whole teacher work sample. After the initial development of the Renaissance Partnership for Improving Teacher Quality *Prompt* and *Scoring Rubric*, studies (e.g., Denner, Salzman, & Bangert, 2001, Denner et al., 2002; Denner, Norman, Salzman, & Pankratz, 2003; Denner, Norman, Salzman, Pankratz, & Evans, 2003) have been done to determine the amount of variance in the total scores due to rater differences and whether several raters can use the instruments with a high degree of consistency. This was done by calculating a correlation coefficient of inter-rater reliability using concepts from Generalizability Theory (Shavelson & Webb, 1991). The Generalizability Theory also offers formulas to calculate dependability coefficients of raters, in a way similar to the classical test theory's reliability coefficient. In Denner, Norman, Salzman, and Pankratz (2003), these formulas were used to calculate the minimum number of raters necessary for making "high-stakes decisions about absolute teaching performance level" (p. 34). The inter-rater reliability was reported to be high. The same study also uses Generalizability theory formulas to determine a minimum number of raters necessary to achieve a reliable inter-rater reliability to allow for generalizations and making high-stake decisions. The findings in the study by Denner, Norman, Salzman, and Pankratz

indicate that three or more raters with rating experience are needed for a "sufficient inter-rater agreement" (2003, p. 37). In the current study each teacher work sample was scored by all members from each (control and comparison) group (N=30).

In addition, a number of validity studies (e.g., Denner et al, 2001; Denner, Norman, Salzman, & Pankratz, 2003; Denner, Norman, Salzman, Pankratz, & Evans, 2003) were carried out to examine alignment between the TWS *Prompt* guidelines, the TWS standards, and the *Scoring Rubric* and to collect evidence in support of *content validity* of the TWS assessment. Data analysis in these studies indicated strong alignment among prompt tasks, standards, and the assessment rubric. Moreover, frequency, importance, authenticity, and representativeness findings of the study supported overall content representativeness of the TWS instrumentation. Specifically, frequency analysis indicated that "all of the targeted teaching behaviors were considered to have a high frequency in actual teaching practice" (Denner, Norman, Salzman, & Pankratz, 2003, p. 35). The analysis of importance of the teaching behaviors targeted in the assessment rubrics also indicated that it was focusing on behaviors that were important or very important. Authenticity analysis results also indicated that vast majority the TWS prompt tasks were authentic or very authentic teaching practices. Finally, representativeness analysis of the prompt tasks indicated that all the teacher work sample tasks "reflect and represent targeted standards" (Denner, Norman, Salzman, & Pankratz, 2003, p. 36).

There are two basic scoring approaches used by partner institutions in the Renaissance Partnership for Improving Teacher Quality project: *holistic* and *analytic*.

Based on multiple field testing and project research (Denner et al., 2001) each educator was expected to spend on average 13.5 minutes when scoring each teacher work sample holistically. However, UNI scoring sessions employ the *analytical approach* to scoring teacher work samples by assigning both total scores and subscores by process and indicators. This scoring process generally takes longer than the holistic approach. Based on UNI scoring records, raters spend about 20-25 minutes per work sample when scoring analytically. It was expected that the raters in the current study would spend similar amounts of time per teacher work sample since they were scoring them analytically.

Data Collection

The planned one-day training and scoring session took place on a Saturday in early May 2004 and lasted for about five hours. The event began with a short training session, approximately 30-45 minutes long, followed by a scoring session, lasting at least four hours with additional time for short breaks and a lunch. Study participants attended both the training and the scoring sessions of the event. To keep track of the scoring results and protect the identity of the participants of the study, each of the participants of the study was assigned a unique identification number that appeared on all the scoring sheets, demographic questionnaire, and other documentation completed by each rater.

The training briefly discussed the teacher work sampling process, guidelines used by teacher candidates to develop work samples, and provided a detailed overview of the scoring rubrics. Moreover, the raters were presented with *Assessor Guidelines*

that instruct scorers to "maintain the proper attitude towards performances" (The Renaissance Partnership for Improving Teacher Quality, 2002c). In addition, the session included a brief anti-bias segment in order to assist scorers in uncovering potential biases caused by personal perceptions of what "good" or "bad" teacher work sample should look like. At the end of the training component of the session, the raters were reminded to respect confidentiality of the teacher candidates. They were also shown how to search for evidence throughout the work sample using the *Road Map for Locating Evidence* created by the Renaissance Partnership for Improving Teacher Quality Project (2002a).

Following the instructions provided during the training session, the raters were asked to complete a short Demographic Survey (Appendix A). At the same time they received sets of ten Spanish teacher work samples each and scored all of them, assigning both total scores and scores by process and indicators using a Rubric designed by the Renaissance Partnership for Improving Teacher Quality Project (Appendix B). In addition, the participants were asked to note and indicate start and finish time on each of their scoring sheets. At the end of the scoring event, participants were asked to submit all the teacher work samples and scoring sheets to the researcher. The overall length of the scoring session varied from participant to participant. It took some scorers about three hours to complete rating ten work sample sets, while others spent over four-and-a-half hours rating the same ten teacher work samples. Finally, the subjects of the study received a thank you letter (Appendix D), a modest compensation of $200 each

(Appendix E), along with a certificate of appreciation (Appendix F) for their participation and time and efforts devoted to the scoring process.

Analysis of Data

The data collected for the study were first analyzed descriptively. There are several *independent variables* in this study, such as (a) presence or absence of world language content expertise, (b) the amount of teaching experience, (c) experience with scoring work samples, (d) gender, (e) level of education and several others. The *dependent variables* in the study are ratings of work samples, both total and broken down by process and indicators.

After collecting the data, statistical analysis of data was conducted. All data from the scoring sheets and demographic questionnaires were entered into the Statistical Package for the Social Sciences (SPSS) software program Version 10.0. The results of the data were tabulated by linking each item on the scoring sheet and demographic questionnaire to one or more of the research questions. All computational procedures were done using the SPSS software.

The first step in the data analysis was to conduct an exploratory data analysis and compute descriptive statistics for subgroups in the study. The subgroups were organized based on the participants' content area (foreign language vs. non-foreign language) and several other characteristics. The descriptive statistics included raw frequencies, group means, and standard deviations. The next step focused on examining statistical significance. To address the questions of the study, a comparison of data from the control and experimental groups was conducted using Analysis of Variance

(ANOVA). ANOVA is "a statistical procedure that compares the amount of between group variance in individuals' scores with the amount of within-group variance" (Gall, Gall, & Borg, 2003, p. 307). A regression analysis was used to assess the degree of relationship between scoring and additional rater characteristics used in the analysis, i.e., amount of world language teaching experience, teaching experience at a high school level, experience with scoring work samples, gender, level of education, and several others. Statistical tests were conducted at the .05 level of significance. Results were analyzed and conclusions drawn and described in the next chapter.

CHAPTER IV

RESULTS

In order to examine the role of teacher characteristics in their assessment of teacher practice, the first step in data analysis was to utilize descriptive statistics and data representations to arrive at descriptors of demographics, variables, and Teacher Work Sample (TWS) score distributions.

Descriptive Data

Demographic Data

Rater content area

Information concerning demographic variables was collected using the 17 question Demographic Survey (see Appendix A). The study involved 30 Iowa middle and high school teachers. Sixteen of the participating teachers were world language educators (experimental group), while the remaining teachers were teaching content areas other than languages or English as a Second Language (comparison group).

Gender

Overall, twelve males and eighteen females participated in the study. While only three of the language teachers were males, the non-language group contained nine males.

Education

In regards to the highest degree received, groups were very similar with about half of the participants reporting having MA degrees. The number of teachers with MA degrees directly correlated with number of years of teaching experience (See Table 2).

Teaching level

The language teachers group contained four teachers (25%) practicing at the *middle school level* (three of whom also taught at a high school level), while only one (7%) of the non-language teacher group member reported teaching middle school, as well as high school, classes.

Table 2

Highest Degree Received by Years of Teaching

1-5 Years of Teaching		6-20 Years of Teaching		21+ Years of Teaching	
BA	MA	BA	MA	BA	MA
9	1	5	5	3	7
Note: N = 30					

Languages taught

Eleven (70%) out of 16 foreign language teachers participating in the study reported that their primary teaching appointment was teaching Spanish, with the remaining 30 percent of participants teaching French. While only one teacher of Spanish also taught French, two-thirds of French teachers (three out of five) in the study reported having taught Spanish. Distribution of languages taught varied by the number of years of teaching experience that participants of the study reported (See Table 3).

Table 3

Languages Taught

1-5 Years of Teaching N = 5		6-20 Years of Teaching N = 7		21+ Years of Teaching N = 4	
Spanish	French	Spanish	French	Spanish	French
5	0	4	3	2	2
Note: N = 16					

Teaching experience

The participants in the study varied in regards to the amount of teaching experience they reported (see Table 4). The average amount of teaching experience for the language group was almost 14 years, while for non-language teachers it was nearly 17 years (see Table 5).

Table 4

Years of Teaching Experience

Years of Teaching Experience	1	2	3	4	5	6	10	11	14	15	20	21	22	28	29	30	33	36	42
Number of Participants	3	1	1	3	2	2	1	2	1	3	1	1	1	1	2	1	2	1	1
Approx. Percent	9	3	3	9	6	6	3	6	3	9	3	3	3	3	6	3	6	3	3
Note: N = 30																			

Table 5

Years of Teaching Experience by Type of Teacher

Language Teachers			Non-language Teachers		
Mean	SD	Median	Mean	SD	Median
13.94	11	11	16.64	13.74	15
Note: N = 30					

Knowledge of world languages

In regards to their knowledge of world languages, only four (30%) of the non-language teachers reported knowing a language other than English, with one teacher reporting knowing two foreign languages. Overall, non-language teacher indicated their knowledge of the following world languages: French, Latin, and Spanish. Knowledge of Spanish was reported by three (17%) of non-language teachers, with one teacher (6%) reporting his/her knowledge of French.

Five language teachers (31%) reported knowing two languages other than English (French and Spanish), while one teacher (6%) reported knowing three foreign languages (French, German, and Spanish).

Of four non-language teachers with the knowledge of a world language, all respondents reported rather lower levels of language proficiency: beginning (75%) or intermediate (25%) levels. Not surprisingly, all language teachers self-reported

possessing advanced or native-like proficiency in at least one language other than English.

Previous knowledge of Teacher Work Sample Methodology

Overall, twice as many non-language teachers (n= 10) reported having some previous knowledge of Teacher Work Sample Methodology than language teachers (n=5). Several of non-language teachers participating in the study were veterans of UNI TWS rating sessions. Since language teachers were specifically recruited for their participation in the study, most of them were new to this experience: data indicates that only about one third (n=5) of all language teachers (n=16) reported hearing about TWSM prior to their participation in the study. Similar analysis of responses of non-language teachers revealed that two thirds (n=10) of them (n=14) reported hearing about TWSM before taking part in the study.

Previous scoring experience

Overall, there were a total of seven participants with TWS scoring experience. Among the non-language group six out of fourteen teachers (43%) reported having scored TWS in the past, with three teachers having scored once, one teacher - three times, and two - four times. In the language teacher group only one participant reported having a one-time previous TWS scoring experience. Overall, participants of the study with more years of teaching experience tended to be more likely to have previous TWS scoring experience (See Table 6).

Table 6

Previous TWS Scoring Experience

1-5 Years of Teaching N = 10		6-20 Years of Teaching N = 10		21+ Years of Teaching N = 10	
Scored	Not scored	Scored	Not scored	Scored	Not scored
1	9	2	8	4	6
Note: N = 30					

Serving as a cooperating teacher

The majority of the non-language teachers (86%) reported serving as a cooperating teacher to a future teacher, while only 38% of language teachers reported having the same experience.

Serving as a cooperating teacher for a candidate with TWS

Nearly equal numbers of teachers in each group reported being a cooperating teacher to a student working on a Teacher Work Sample (two language teachers and three non-language teachers).

NBPTS certification

The participants of the study were similar in respect to their NBPTS certification status. None of the teachers reported being NBPTS certified.

Participation in scoring other high stake assessments

While two (14%) of the non-language teachers stated previous participation in other high stake assessments training and scoring, none of the participants in the language teacher group reported having similar experiences.

Teacher Work Sample Data

This section will describe overall descriptive analysis of the Teacher Work Sample scoring results. The results are organized in the following way: (a) time spent on rating teacher work samples and (b) ratings of teacher work samples.

Time Spent on Rating Teacher Work Samples

Average scoring speed

In respect to timing that it took participants in the study to score each of the ten Spanish Teacher Work Samples, participants varied greatly in their scoring speed of individual samples. Table 7 summarizes timing data for this aspect of the study. Overall, on average it took participants almost 22 minutes to score an individual teacher work sample. Based in the UNI's informal records of Teacher Work Sample ratings, this average timing corresponds well with the University of Northern Iowa scoring records that indicate that raters spend 20-25 minutes per TWS when scoring analytically.

Table 7

Individual Participant Scoring Time (in minutes)

ID#	TWS #1	TWS #2	TWS #3	TWS #4	TWS #5	TWS #6	TWS #7	TWS #8	TWS #9	TWS #10	Mean
P1	18	18	19	27	16	22	18	20	12	11	18.1
P2	19	22	22	24	19	19	18	15	19	13	19
P3	30	22	40	20	24	17	19	30	18	20	24
P4	20	26	21	19	15	17	14	22	22	15	19.1
P5	25	13	25	17	14	21	20	12	20	16	18.3
P6	42	28	40	31	29	29	28	28	19	27	30.1
P7	31	22	12	17	18	15	16	18	25	15	18.9
P8	19	24	24	28	22	32	20	22	23	17	23.1
P9	23	24	27	20	18	27	23	23	20	18	22.3
P10	15	32	30	24	20	9	30	10	10	11	19.1
P11	17	17	16	23	20	18	15	15	17	9	16.7
P12	22	28	27	28	33	26	23	25	15	18	24.5
P13	20	23	30	15	30	28	15	20	20	12	21.3
P14	25	35	34	20	20	18	19	21	19	19	23
P15	33	29	37	32	26	29	18	27	23	18	27.2
P16	19	25	15	20	19	23	19	14	22	18	19.4
P17	23	33	37	22	37	25	19	30	16	27	26.9
P18	19	23	25	20	21	15	15	22	14	15	18.9
P19	14	32	26	18	15	18	20	23	20	18	20.4
P20	23	23	37	16	17	20	20	25	15	18	21.4
P21	18	30	19	19	20	26	17	17	15	17	19.8

(table continues)

P22	15	21	30	25	**47**	26	30	25	19	17	25.5
P23	15	18	21	27	23	24	20	23	21	15	20.7
P24	24	25	30	20	20	25	20	26	28	25	24.3
P25	25	22	20	20	20	19	25	17	21	20	20.9
P26	24	26	29	20	36	20	22	27	23	18	24.5
P27	18	19	14	17	30	29	24	25	25	17	21.8
P28	24	15	17	17	18	18	16	18	22	17	18.2
P29	21	21	34	34	23	17	21	38	23	19	25.1
P30	17	23	17	27	24	29	17	21	25	15	21.5
Mean	**21.9**	23.9	**25.8**	**22.2**	**23.1**	22	20	21.9	**18.8**	18	**21.8**
Note: N of participants = 30. N of Teacher Work Samples = 10											

Individual scoring speed

Participants' individual timing averages when scoring samples varied from slightly less than 17 minutes to a little over 30 minutes per sample. It is important to note that scoring time of specific samples was quite different from participant to participant; the shortest time spent on scoring a sample was nine minutes and the longest was 47 minutes.

Scoring time of specific samples

Means of scoring times of specific samples did not vary greatly. The minimum was an average of 18 minutes for sample number 10 and a maximum time of approximately 26 minutes for sample number 3. In part, the amount of time spent on

scoring individual samples can be attributed to the length of the samples. For the most part, the RTWSs consisting of over 45 pages (samples 1, 3, 4, 5, and 9) on average took longer to score (see Table 5). One exception should be made to the previous statement: the longest RTWS – sample number nine with 68 pages – was scored relatively quickly, possibly due to its structure. This particular sample had a medium-size narrative component, while the bulk of it consisted of multiple attachments-examples of student work. The fact that the shortest RTWS – number 10 – took the participants the shortest time to score on average, supports the linkage between the length of the individual samples and the time it takes to score them. However, it is important to mention that the main part of all RTWSs has a fairly standardized length, about 20 pages, while the rest of the sample consists of some attachments used to illustrate points in the sample.

Scoring speed language teachers vs. non-language teachers

When average timing is compared between the sub-groups of participants, language teachers vs. non-language teachers, there is no statistically significant difference in their average timing ($t = .095$, $df = 299$, $p > .05$). On average, it took languages teachers 21.8 minutes ($SD = 6.28$) to score a sample, while non-langauge teachers spent 21.7 minutes ($SD = 5.91$).

Scoring speed and previous scoring experience

Moreover, when scoring time is examined for those with a previous scoring experience and those without such an experience, the difference in means is almost 3 minutes, which is statistically significant ($t = -3.67$, $df = 279$, $p < .001$). An average

timing for those with scoring experience is 19.2 minutes ($SD = 5.13$), while for the group without such experience it is 22.1 minutes ($SD = 5.98$).

Scoring speed and level of education

Highest level of education was related to scoring speed of the participants on the study. Scoring time of those with bachelor level of education (Mean = 22 minutes, $SD = 5.49$) was longer on average than of those with masters degrees (Mean = 20 minutes, $SD = 6.05$), ($t = -3.01$, $df = 279$, $p = .003$).

Rating of Teacher Work Samples

Every teacher participating in the study was asked to rate each of the ten Spanish teacher work samples using the RTWS rubric (see Appendix B). Participants assigned a score on the scale from one to three to each of the listed indicators, stating an overall process score and an overall rubric score, with one being "indicator NOT met," two – "indicator partially met," and three – "indicator met."

Overall Teacher Work Sample scores

As a group, participants of the study assigned relatively high overall scores to all ten Spanish Teacher Work Samples, with a mean of 2.6, median 3.00, and mode of 3. It is important to mention that TWS scoring is *criterion referenced*, not norm referenced, thus this "ceiling effect" is not necessarily indicative of a problem. TWS is a performance-based assessment and is meant to be competency oriented. Typically, future teachers receive guidance and other assistance that help them understand TWS process and be able to produce quality work sample. Top performance indicators of

TWS assessment are reachable and, currently, approximately 80% of future teachers submitting their work samples receive scores of "3".

As it was described earlier, the study ended up selecting and focusing on only Spanish Teacher Work Samples. Due to the uni-linguistic nature of the Teacher Work Samples used in the study, the questions of the study were revised to reflect this change.

Research Question 1: Is there a significant difference between ratings assigned to Spanish Teacher Work Samples by raters with foreign language *content experience* and raters without foreign language content experience? Does this differ by sections of the teacher work sample?

Overall, based on a t-test, this study did not discover any significant difference in overall ratings of Spanish Teacher Work Samples by sub-groups of teachers formed based on their content/subject area (see Table 8) at a .05 significance level ($t = .309$, $df = 28$, $p > .05$).

Table 8

Overall TWS Ratings by Type of Teacher

Type of Teacher	N of samples	Mean	Std. Deviation
Language Teacher	160	2.61	.549
Non-language teacher	140	2.58	.601
Note: N = 300			

To answer the second part of the first question a series of t-tests has been performed to study the difference between the ratings of Spanish Teacher Work Samples done by language teachers and non-language teachers. The results of the t-tests are presented in Table 9. T-tests for equality of means indicate that there is no significant statistical differences between the means of ratings of sections of TWS assigned by a group of language teachers and a group of non-language teachers.

Table 9

Ratings of Sections of TWS by Type of Teacher

Rubric Score	Mean		Std. Deviation		Std. Error Mean		Significance (2-tailed)
	LT	NLT	LT	NLT	LT	NLT	
Overall Rubric Score	2.61	2.58	.54	.6	.043	.051	.6
Context	2.44	2.49	.62	.64	.049	.054	.5
Learning	2.66	2.66	.54	.54	.043	.046	.9
Assessment	2.58	2.46	.56	.67	.045	.057	.1
Design	2.64	2.6	.50	.62	.04	.052	.5
Instruct	2.6	2.62	.59	.66	.047	.056	.7
Analysis	2.64	2.6	.54	.58	.043	.049	.5
Reflect	2.49	2.54	.62	.59	.049	.05	.5
Note: N = 300							

As indicated in Table 10, language teachers, as a group, seem to agree slightly more in their overall ratings of individual Spanish teacher work samples, while non-language teachers displayed slightly less of an agreement in assigning overall rubric score to each work sample. By "agreement" the researcher means that 70% or more participants in each group assigned the same rating to a sample. Therefore, language teachers agreed in six instances, while non-language teachers agreed in five instances when assigning overall rubric scores to the Spanish teacher work samples used in the study.

Interestingly enough, language teachers seemed to be more inclined to assign higher overall rubric scores than non-language teachers. For example, out of total 160 possible ratings, only five ratings (3.1%) of "Indicator NOT Met" were assigned by language teachers to three separate work samples; while non-language teachers assigned a total of eight ratings (5.7%) out of possible 140 ratings to the same category of "Indicator NOT Met." A total of five teacher work samples used in the study received such negative ratings by the non-language teachers.

Table 10

Overall TWS Rubric Score by Type of Teacher

Type of Teacher	Overall Rubric Score			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Number				
1	2	9	5	16
2		2	14	16
3		2	14	16
4		3	13	16
5		4	12	16
6		5	11	16
7		5	11	16
8	2	9	5	16
9		8	8	16
10	1	5	10	16
Total:	5	52	103	160
%	3.1%	32.5%	64.3%	100%
Non-language Teachers TWS Number				
1	2	6	6	14
2	1	3	10	14
3		2	12	14
4			14	14
5		4	10	14
6		2	12	14
7		6	8	14
8	3	8	3	14
9	1	7	6	14
10	1	5	8	14
Total:	8	43	89	140
%	5.7%	30.7%	63.5%	100%
Note: N = 30				

Overall scores for each of the seven processes

Table 11 summarizes the overall scores for each of the seven processes reported by all 30 participants of the study. As with the overall rubric scores, overall scores for each of the seven processes were also overwhelmingly positive (see Table 11). Less than seven percent of all processes have received "indicator NOT met" score, while the overwhelming majority of processes as described in the TWS used in the study were rated as "indicator met" (at least 54%).

Table 11

Descriptive Overall TWS Scores for Individual Processes for the Whole Group

	Indicator NOT Met - "1"		Indicator Partially Met – "2"		Indicator Met – "3"	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Context	22	7.3%	116	38.7%	162	54%
Learning	11	3.7%	80	26.7%	209	69.7%
Assessment	20	6.7%	103	34.3%	177	59%
Design	12	4%	89	29.7%	199	66.3%
Instruction	23	7.7%	71	23.7%	206	68.7%
Analysis	12	4%	90	30%	198	66%
Reflection	18	6%	110	36.7%	171	57.3%
Note: N = 30						

Moreover, Table 12 showcases results of analysis of frequencies of overall group scores for each of the seven processes of the scoring rubric for all ten foreign language teacher work samples used in the study. A closer examination of these means, ranging from 2.47 (Context) to 2.66 (Learning), also supports the earlier statement regarding overwhelmingly positive ratings of individual teacher work samples by the participants of the study.

Table 12

Overall TWS Rubric Processes Scores for the Whole Group

	Con-textual Factors	Learning Goals	Assess. Plan	Design for Instruct.	Instruct. Decision-Making	Analysis of Student Learning	Reflect. & Self-Eval.	Overall
Mean	2.47	2.66	2.52	2.62	2.61	2.62	2.51	2.57
Median	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Mode	3	3	3	3	3	3	3	3
Std. Dev.	.630	.546	.620	.562	.627	.563	.610	
Note: N = 30								

Ratings of individual processes by control and comparison groups

After analyzing group scores of the experimental and comparison groups in each of the processes of the Renaissance Teacher Work Sample *Scoring Rubric* (see

Appendix B), it seems that overall both groups rated each section of the rubric in a similar way. These findings are supported by the results of a t-test. However, it should be mentioned once more that analysis of Teacher Work Samples scoring data for each group of raters indicates some differences between the groups' ratings of individual processes. Languages teachers, as a group, seem to be less negative in assigning scores for individual processes of the Spanish teacher work samples used in the study. For example, as indicated in Tables 13 through 19, as a group, the languages teachers assigned less "indicator NOT met" ratings when assessing *overall process scores* for the following of the seven processes within the individual work samples, all dealing with classroom instruction: (3) Assessment plan (4% of ratings by language teachers vs. 10% of ratings by non-language teachers), (4) Design for instruction (1% of ratings by language teachers vs. 7% of ratings by non-language teachers), and (5) Instructional decision-making (6% of ratings by language teachers vs. 10% of ratings by non-language teachers). It is also important to mention that on the contrary to their earlier "more positive" rating of some of the processes, in the seventh and final area of the rubric - Reflection and self-evaluation - the language teachers, as a group, assigned slightly more failing scores, "indicator NOT met," to the Spanish teacher work samples used in the study than the non-foreign language teachers, 7% vs. 5% respectively.

Although the study did not find any statistically significant differences between ratings cast by foreign language teachers and their colleagues from other content areas, the author looked at some relative tendencies that indicated differences at the indicator level. These were included to generate more discussion and for further research.

Table 13

Overall TWS Process Scores for Contextual Factors by Type of Teacher

Type of Teacher	Contextual Factors			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Number				
1	3	6	7	16
2	1	7	8	16
3		7	9	16
4		2	14	16
5	1	8	7	16
6		6	10	16
7		7	9	16
8	2	8	6	16
9		6	10	16
10	4	10	2	16
Total:	11	67	82	160
%	7%	42%	51%	100%
Non-language Teachers TWS Number				
1	2	7	5	14
2	1	5	8	14
3		4	10	14
4		1	13	14
5		5	9	14
6		2	12	14
7		5	9	14
8	2	8	4	14
9	2	4	8	14
10	4	8	2	14
Total:	11	49	80	140
%	8%	35%	57%	100%
Note: N = 30				

Table 14

Overall TWS Process Scores for Learning Factors by Type of Teacher

Type of Teacher	Learning Goals			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Number				
1	2	10	4	16
2		4	12	16
3	1	1	14	16
4		1	15	16
5		2	14	16
6		1	15	16
7		4	12	16
8	1	9	6	16
9		3	13	16
10	2	7	7	16
Total:	6	42	112	160
%	4%	26%	70%	100%
Non-language Teachers TWS Number				
1	1	2	11	14
2	1	5	8	14
3		1	13	14
4			14	14
5		2	12	14
6		2	12	14
7		6	8	14
8		7	7	14
9	1	6	7	14
10	2	7	5	14
Total:	5	38	97	140
%	3.5%	27%	69.5%	100%
Note: N = 30				

Table 15

Overall TWS Process Scores for Assessment Plan by Type of Teacher

Type of Teacher	Assessment Plan			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Numbers				
1	4	9	3	16
2		3	13	16
3		2	14	16
4			16	16
5		7	9	16
6		4	12	16
7		8	8	16
8		10	6	16
9	1	7	8	16
10	1	6	9	16
Total:	6	56	98	160
%	4%	35%	61%	100%
Non-language Teachers TWS Numbers				
1	2	5	7	14
2	2		12	14
3		4	10	14
4		2	12	14
5		7	7	14
6		5	9	14
7	2	4	8	14
8	4	6	4	14
9	2	7	5	14
10	2	7	5	14
Total:	14	47	79	140
%	10%	33.5%	56.5%	100%
Note: N = 30				

Table 16

Overall TWS Process Scores for Design for Instruction by Type of Teacher

Type of Teacher	Design for Instruction			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Number				
1		9	7	16
2		4	12	16
3		2	14	16
4		5	11	16
5		5	11	16
6		6	10	16
7		6	10	16
8	2	5	9	16
9		8	8	16
10		3	13	16
Total:	2	53	105	160
%	1%	33%	66%	100%
Non-language Teachers TWS Number				
1	2	4	8	14
2	1	3	10	14
3		3	11	14
4		2	12	14
5		5	9	14
6	1	3	10	14
7	1	6	7	14
8	4	3	7	14
9	1	3	10	14
10		4	10	14
Total:	10	36	94	140
%	7%	26%	67%	100%
Note: N = 30				

Table 17

Overall TWS Process Scores for Instructional Decision-Making by Type of Teacher

Type of Teacher	Instructional Decision-Making			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Number				
1	6	8	2	16
2		1	15	16
3		2	14	16
4	1	4	11	16
5		5	11	16
6		5	11	16
7		3	13	16
8	1	8	7	16
9		6	10	16
10	1	4	11	16
Total:	9	46	105	160
%	6%	28%	66%	100%
Non-language Teachers TWS Number				
1	6	5	3	14
2		2	12	14
3		2	12	14
4		1	13	14
5		2	12	14
6		2	12	14
7	1	4	9	14
8	4	1	9	14
9		6	8	14
10	3		11	14
Total:	14	25	101	140
%	10%	18%	72%	100%
Note: N = 30				

Table 18

Overall TWS Process Scores for Analysis of Student Learning by Type of Teacher

Type of Teacher	Analysis of Student Learning			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Number				
1	3	10	3	16
2		2	14	16
3		2	14	16
4		6	10	16
5		1	15	16
6		4	12	16
7		6	10	16
8	1	9	6	16
9		6	10	16
10	1	2	13	16
Total:	5	48	107	160
%	3%	30%	67%	100%
Non-language Teachers TWS Number				
1	4	5	5	14
2		1	13	14
3		2	12	14
4		3	11	14
5		2	12	14
6		9	5	14
7		4	10	14
8	2	8	4	14
9	1	4	9	14
10		4	10	14
Total:	7	42	91	140
%	5%	30%	65%	100%
Note: N = 30				

Table 19

Overall TWS Process Scores for Reflection and Self-Evaluation by Type of Teacher

Type of Teacher	Reflection and Self-Evaluation			Total
	1 "Indicator Not Met"	2 "Indicator Partially Met"	3 "Indicator Met"	
Language Teachers TWS Number				
1	3	6	7	16
2		6	10	16
3		2	14	16
4		7	9	16
5	1	5	10	16
6		6	10	16
7	1	8	7	16
8	5	8	3	16
9		5	11	16
10	1	6	9	16
Total:	11	59	90	160
%	7%	37%	56%	100%
Non-language Teachers TWS Number				
1		10	4	14
2	2	3	9	14
3		3	11	14
4		5	9	14
5		4	10	14
6		1	13	14
7		8	6	14
8	3	9	2	14
9	1	5	8	14
10	1	3	10	14
Total:	7	51	82	140
%	5%	36.5%	58.5%	100%
Note: N = 30				

Summary of Research Question 1

In spite of some slight differences in ratings by the control and comparison groups of the Spanish Teacher Work Samples, there is no statistically significant difference in overall sample ratings by sub-groups of participants of the study formed based on their content/subject area (see Table 8) at a .05 significance level. Moreover, the study did not find any statistically significant differences in the ratings assigned to individual sections of the TWS rated by groups of language teachers and non-language teachers.

Research Question 2: What is the relationship between the amount of rater's *teaching experience* and his/her scoring of Spanish teacher work samples?

The participants of the study varied in the amount of their teaching experience overall and within the comparison and control groups (see Tables 4 and 5). To answer the second question of the study the overall scores and scores of individual TWS sections were correlated with years of teaching experience. The overall rubric score did not have any statistically significant correlation ($r = .004, p = .93$). The results of the statistical analysis by individual rubric section are presented in the table below (see Table 20). All but one correlation appear to be statistically insignificant. The only statistically significant correlation between the amount of teaching experience and ratings of TWS is in the Contextual factors section, however, it is not a large correlation ($r = .124, p = .031$). The researcher does not have a definite explanation as to why the Contextual Factors section stood out in the study. Perhaps, raters with more teaching experience were recognizing a set of certain characteristics in this TWS section, while

less-experienced scorers were looking for another set of items. The researcher was unable to establish if it was an interesting occurrence or a Type I error.

Table 20

Relationship Between Raters' Amount of Teaching Experience and Scoring of TWS

	Con-textual Factors	Learning Goals	Assess. Plan	Design for Instruct.	Instruct. Decision-Making	Analysis of Student Learning	Reflect. & Self-Eval.
Pearson Correlation	.124	.085	-.007	.103	-.024	-.073	-.008
Sig. (2-tailed)	.031	.142	.906	.075	.673	.207	.889
Note: N = 300							

Summary of Research Question 2

The study did not find any major statistically significant correlations between the years of teaching experience and scores of sections of TWS, as well as the overall rubric score. Only one relatively small correlation was found between the amount of teaching experience and ratings of the Contextual Factors section of the TWS.

Research Question 3: What is the relationship between rater's previous work sample *scoring experience* and his/her scoring of Spanish work samples?

A small number of participants of the study, seven teachers (23%), had previous scoring TWS experience. Four of these participants had scored one time prior to the data collection event, one participant had scored three times, and two had scored four

times before participating in the study. Table 21 summarizes the data analysis for this section. As one can see, results of the t-test show that there is no significant correlation between having any previous scoring experience and overall ratings, as well as ratings of individual sections, of the Spanish TWS in the study.

Table 21

Impact of the Previous TWS Scoring Experience

Rubric Scores	Mean		Std. Deviation		Std. Error Mean		Significance (2-tailed)
	Scored	Not score	Scored	Not scored	Scored	Not scored	
Overall Rubric Score	2.57	2.6	.57	.57	.069	.038	.67
Context	2.4	2.49	.66	.61	.08	.041	.31
Learning	2.66	2.66	.56	.54	.067	.036	.96
Assessment	2.57	2.51	.55	.63	.066	.042	.45
Design	2.67	2.61	.55	.56	.067	.037	.41
Instruct	2.57	2.62	.65	.62	.078	.041	.55
Analysis	2.63	2.62	.59	.55	.071	.037	.88
Reflect	2.5	2.52	.6	.61	.073	.04	.83
Note: N = 300							

It is important to mention that in their study of inter-rater reliability using Generalizability Theory, Denner, Salzman, and Harris (2002) came across a statistically significant rater effect due to some of the raters having less experience scoring TWSs. On average, these raters scored TWSs lower than other raters with more TWS experience. This effect was true in both rating scenarios using holistic and analytic scoring rubrics. In their study, the authors recommended that "only experienced raters should be used when making absolute decisions about candidates' levels of teaching performance using a holistic scoring rubric" (p. 22).

Summary of Research Question 3

The analysis of data pertaining to this section of the study indicates that there is no significant correlation between previous rating experience and scoring of Spanish TWS.

Chapter Summary

This chapter focused on the statistical analysis of the data collected during the course of the study. The study employed two instruments: (a) a demographic questionnaire and (b) a TWS scoring rubric. The short questionnaire was developed specifically for the study, while the TWS scoring rubric is a tool commonly used to score all Renaissance Teacher Work Samples. The analysis of the demographic data revealed that participants of the study varied greatly in almost all the areas of the questionnaire, except for being members of the NBPTS, which none of the participants had any experience with. The TWS data analysis contributed to the further

understanding of the participants' rating process and outcomes, their scoring speed, and allowed to answer the questions of the research.

Furthermore, the chapter reported data analysis for the three key questions of the study: (a) is there a significant difference between ratings of Spanish Teacher Work Samples assigned by raters with foreign language *content experience* and raters without foreign language content experience and does this differ by sections of the teacher work sample, (b) what is the relationship between the amount of a rater's overall *teaching experience* and his/her scoring of Spanish teacher work samples, and (c) what is the relationship between a rater's work sample *scoring experience* on his/her scoring of Spanish work samples.

Overall, the study did not find any statistically significant differences between the control and comparison groups, or relationships between the amount of teaching experience or previous scoring experience and the ratings of the Spanish TWS.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

American educators at all levels are constantly under increasing pressure to better prepare children and youth to succeed in the developing global society. Classroom teachers have been demonstrated to be central to pupils' academic success (e.g., Darling-Hammond, 2000). This evidence and pressure from parents, politicians, and a general public to improve American schools, has prompted a creation of more rigorous standards for teacher preparation programs to assess preparedness of their candidates as well as program effectiveness in preparing quality teachers.

There are several approaches and instruments used by teacher preparation institutions, school districts, and states to assess teacher preparation and teacher quality (Millman, 1997). One of these approaches, based on the Western Oregon Teacher Work Sample Methodology, is the Renaissance Partnership Teacher Work Sample (RTWS). This instrumentation is widely used by teacher preparation programs and is primarily utilized to inform future teachers and their teacher preparation programs on the readiness of the candidates to enter the teaching profession. Moreover, a growing number of programs are using TWS as a learning tool, helping their candidates develop their teaching skills. Additionally, according to the National Council of Accreditation of Teacher Education (NCATE), the TWSM meets their requirements for documenting impact of teacher candidates on student learning, and thus can be used by teacher preparation programs as a part of the accreditation process.

Since RTWS is a relatively new instrument, studies are being conducted to validate various matters related to its use. One of the assumptions of the RTWS is that any professional educator, after participating in a RTWS training, is capable of scoring teacher work samples in *any* content area using the associated rubric. This study chose to test this assumption by focusing on the specific content area of foreign langauge as a criterion for selection of TWS and formation of control and comparison groups.

Summary

The objective of this study was to determine how rater characteristics affect evaluation of teacher practice as presented in Spanish Teacher Work Samples (TWS) submitted by teacher candidates at the University of Northern Iowa. The study involved a total of 30 Iowa middle school and high school teachers who participated in a day-long training and scoring session. For the purposes of the study, the participants were divided into two groups, foreign language teachers (n = 16) and non-foreign language teachers (n = 14), in rating ten Spanish language TWS using the existing RTWS *Scoring Rubric*. The study also chose to examine some additional rater characteristics and their potential impact on scoring of TWS: various demographic characteristics such as length of teaching experience, previous scoring experience, and several others. The study employed a causal-comparative research design.

Discussion

Question 1: Is there a significant difference between ratings assigned by raters with foreign language content experience and raters without foreign language content experience? Does this differ by sections of the teacher work sample?

Independent t-tests between a group consisting of the foreign language scorers (n=16) and a group of non-foreign language scores (n=14) did not indicate any statistically significant differences at the .05 level. In other words, non-language teachers were as capable of assessing foreign language teacher practice, as defined and presented in Spanish TWSs, as their foreign language colleagues. This finding indicates that there is no need to assign foreign language content specialist to rating of teacher work samples compiled on units dealing with foreign language learning. Any other teacher would be quite competent in rating foreign language TWS using the associated assessment tools.

Question 2: What is the relationship between the amount of rater's overall teaching experience and his/her scoring of Spanish teacher work samples?

The study found no statistically significant relationships between the amount of raters' teaching experience and their scoring of Spanish TWSs. This finding further supports the notion of reliability of the RTWS methodology and instrumentation. Additionally, this finding suggests that the amount of teaching experience does not need to be a factor in selecting educators for TWS rating sessions.

Data analysis further indicated that this relationship between years of teaching and TWS scores was statistically significant for only one section of the TWS:

Contextual Factors. The researcher does not have a definite explanation as to why the Contextual Factors section stood out in the study, it could be an interesting occurrence or Type I error. Perhaps, raters with more teaching experience were recognizing a set of certain things in this TWS section, while those less-experienced scorers were looking for another set of items. Another possible explanation could be that it is a random case that exhibited itself as a result of multiple analyses ran with α nominally at .05. In any case, the statistical significance of this item was rather low ($r = .124$, $p = .031$).

Question 3: What is the relationship between a rater's work sample scoring experience on his/her scoring of Spanish teacher work samples?

In answering the question on the relationship between the raters' TWS scoring experience and the scoring of the Spanish TWSs, a series of t-tests was performed. The t-tests looked for any relationship between the scoring experience and overall rubric ratings, as well as ratings of individual sections of the TWS. Once again, the study found no statistically significant relationship between the previous scoring experience and rating of TWS at the .05 level. Thus, previous TWS scoring experience should not be used as a factor in selecting teachers to score work samples compiled by future teachers.

Scoring Time

Participants of the study varied greatly in their amount of time it took to score individual samples. The average time of 22 minutes per sample is similar to UNI existing timing records of previous scoring sessions.

Several other RTWS studies report a shorter average time spent by rates when scoring analytically: 13.5-14 minutes per sample on average (Denner et al., 2002; Salzman et al., 2001). However, it is important to mention several factors contributing to the difference in average scoring time between these studies and the research in hand. First, these studies used benchmarked RTWSs, i.e. samples selected by a group of experienced raters as proto-typical examples of work done corresponding with each of the four proficiency levels: beginning, developing, proficient, and exemplary. When non-benchmarked RTWSs were used, as in this study, the average rating time increased to 24 minutes, which is very similar to the findings of this study (Denner et al., 2002). Second, these studies (Denner et al., 2002; Salzman et al., 2001) used a modified version of the RTWS analytical rubric that did not require raters to score individual indicators, as the current rubric does; it only asked to assign scores to overall processes. In addition, the rubric used in this study also asked raters to assign an overall holistic score, which was not done in the abovementioned studies. The "shorter" version of the rubric used in other studies may also account for the difference in average timing between the earlier studies and the current one. Overall, research finds the average scoring time of around half hour per sample to be reasonable.

Even though no statistically significant differences were reported after data analysis to answer the main research questions of the study, the analysis of the descriptive data indicated differences in speed of scoring between the participants with previous scoring experience and those without any scoring experience. Experienced participants scored substantially faster than raters without any previous scoring

experience, by 3 minutes on average per work sample. The statistical analysis indicated that this difference was statistically significant at the .05 level.

In addition, data indicated that the level of education of scorers was related to their scoring speed. Participants with masters degrees were scoring faster by two minutes on average than their counterparts with bachelor degrees ($p = .011$).

Although the findings of the study indicated a lack of any statistically significant difference between the control and comparison groups, as well as a lack of impact of the raters' individual differences on their scoring of Spanish TWSs, nevertheless, these are important findings. The finding of no significant difference in Question 1 provides support that the foreign language content experts (foreign language teachers) and non-foreign language teachers in the study assessed Spanish teacher work samples in a reasonably similar way. A lack of statistical power would threaten the significance of this finding, but a power approximation analysis, carried out by the researcher, indicated that power was nearly .80 (α =.05, and positing a .25 point difference on rubric score), which is considered substantial (Cohen, 1988).

Some may argue that a study that finds no significant difference is relatively unimportant. However, due to the causal-comparative research design used in the study, which did not aim to examine an impact of an intervention, a finding of no significant difference is not a rejection of an intervention.

Implications of the Study

This study's findings support the reports of earlier studies (e.g., Denner, Salzman, & Bangert, 2001; Denner, Norman, Salzman, & Pankratz, 2003; Denner,

Norman, Salzman, Pankratz, & Evans, 2003; Salzman et al., 2001) that RTWS methodology and instrumentation are valid and reliable, and work well regardless of the TWS subject area and independent of individual rater differences (e.g. amount of teaching experience, subject area, previous rating experience). This is an important finding because now teacher educators, future and experienced teachers, and policymakers should gain a greater degree of confidence in using RTWS as an assessment tool measuring a teacher's ability to impact student classroom learning.

Several earlier studies (e.g., Denner et al., 2001; Denner, Norman, Salzman, & Pankratz, 2003; Denner, Norman, Salzman, Pankratz, & Evans, 2003; Salzman et al., 2001) examined validity of RTWS using criteria developed by Crocker (1997) for the content representativeness – consisting of frequency, criticality, necessity, and realism – of the teaching tasks in the instrumentation when compared with the actual teaching practice. These studies report that panels of expert raters observed moderate to high (high in most instances) levels of content representativeness of the RTWS prompt and scoring rubric. In addition, several studies (e.g., Denner, Norman, Salzman, & Pankratz, 2003; Denner, Salzman, & Harris, 2002) indicate direct correspondence of RTWS tasks and certain standards (e.g., INTASC). This information indicates a substantial degree of *validity* of the RTWS instrumentation use as a way to assess teacher competence. A close alignment of the RTWS assessment tasks with the vast majority of the Interstate New Teacher Assessment and Support Consortium (INTASC) standards is reported in several studies (e.g., Denner, Norman, Salzman, & Pankratz, 2003; Salzman et al., 2001). This alignment provides further evidence of two types of

validity of the RTWS instrumentation: content validity (alignment with national, state and institutional standards) and construct validity (alignment with the knowledge base on effective teaching), contributing to the greater degree of confidence regarding use of the assessment.

Several studies (e.g., Denner, Norman, Salzman, & Pankratz, 2003; Denner et al., 2002) investigated the amount of variance in the scores caused by individual rater differences and the generalizability potential of the scores across raters. The findings of these studies indicate that in order to achieve sufficient inter-rater *reliability* for high stakes decisions, three or more raters are needed to score each TWS.

The review of the body of literature and findings of this study also indicate that the RTWS is a *useful* assessment. It can be done in a reasonable and practical amount of time (about 20-25 minutes per sample) with high dependability coefficients for panels of three or more raters. This latter information is crucial for those intending on or currently using RTWS assessment for high-stake decision making, like granting initial licensure to teacher candidates or recommending first year teachers for permanent licensure.

Moreover, among several rater characteristics, the current study explored the role of previous rating experience in scoring TWSs. The findings of this study indicate that previous TWS scoring experience becomes useful to scorers at least in term of their rating speed. This finding can be used to beef-up or even restructure pre-scoring training session used at the University of Northern Iowa, to include a simulated scoring session. Such a session should give raters more competence and increase their level of

comfort in the real scoring that will follow. At a minimum, it should increase their rating speed, allowing rating a greater number of TWS or finishing ratings in a shorter amount of time.

Furthermore, the findings of the study can be useful in the selection process of RTWS raters. It is important to mention that even though the ratings of content teachers did not differ statistically from the non-content raters, it can be valuable to include content specialist in RTWS rating, especially those content specialists involved in the teacher preparation program. This hands-on experience of reading and evaluating teacher work samples compiled by student teachers in their program, and maybe even department, may provide some unique insights into the teacher preparation program and facilitate discussion regarding further program improvement.

Overall, the findings of this study contribute to the existing body of research on TWSM. The empirical evidence presented in the study, combined with the earlier studies and the review of literature, suggests that RTWS is a valid, reliable, and useful assessment tool, suitable for high-stake decision making regarding (a) future or current teachers' ability to positively impact student learning, as well as (b) accountability of a teacher preparation program. RTWS is a response of a Consortium of 11 teacher preparation institutions to the national calls for a development and implementation of an assessment system that would "yield defensible and credible evidence regarding candidates' ability to meet ...standards and impact PK-12 student learning" (Salzman et al., 2001, p. 3).

In conclusion, information provided by RTWS allows decision makers to assess teaching qualities of future and practicing teachers and their potential of making a positive impact on student learning. RTWS is one of the existing "applied performance approach" tools created to assess teacher quality. Armed with such tools, teacher preparation programs, school districts, and states have better chances in addressing the issue of teacher quality by requiring future and practicing teachers to demonstrate their impact on student learning, thus improving the quality of American public education.

Delimitations

1. A limited pool of foreign language teacher work samples (total of 16) available for the study with a limited selection of languages (Spanish or German)
2. A limited number of Spanish teacher work samples (10) produced by students at UNI were used in the study.
3. Work samples used in the study came only from student teachers of the University of Northern Iowa teacher preparation program.
4. Work samples in the study were based only on Spanish language units at 7-11 grades.
5. All raters were Iowa teachers.
6. The majority of foreign language teachers were teaching at a high school level, and only one had a teaching experience at both high school and university level.
7. Foreign language teachers who took part in the study were teachers of French and/or Spanish.

8. The study was based on a total of 30 subjects representing each of the groups as defined by demographic characteristics, therefore limiting its potential to generalize to a larger population.

Recommendations for Further Research

This study should be viewed as a beginning of many needed studies examining a variety of tools currently used in (a) teacher preparation to assess the quality of future teachers and (b) evaluating teacher quality of the practicing teachers in the field. This study only examined one of the tools; thus, studies of other assessment instruments should continue to be carried out.

The results of the current study echo other studies on RTWS (Cartwright & Blacklock, 2003; Denner et al., 2001; Denner, Norman, Salzman, & Pankratz, 2003), indicating that RTWS is a valid approach to assessing teacher preparedness and supporting the generalizability of the work sample scores for groups of three or more raters. However, it would be beneficial to replicate this study diversifying its participant pool and TWS pool. Since this study only involved participants-scorers from middle school and high school levels, it would be beneficial to carry out a study involving university level professors from teacher preparation programs. Due to a limited foreign language TWS pool at the University of Northern Iowa, only Spanish language samples were selected for the study. Future studies should attempt to select more linguistically diverse TWS pool to be used for research. Ideally, it would be interesting to use foreign language samples of less commonly taught languages, like Arabic, Chinese, or Russian.

Moreover, further research should focus on other content-specific areas of the curriculum, just as this study focused on foreign language teaching/learning. These subject areas, like business, information technology, music, etc., may create challenges for TWS raters, thus special training may be required prior to the scoring session.

Finally, it may be beneficial to look into the content and length of training sessions for raters conducted prior to a scoring session. As this study confirmed, previous rating experience helped scorers do their job faster, without losing quality. It can be assumed that thorough pre-scoring training, incorporating simulation activities, may be useful in training new raters. These training activities will allow rates become familiar or re-familiarize themselves with the instrumentation and give them an opportunity to clarify any uncertainties they may have about their role as a rater, which will most likely result in higher rating speed.

REFERENCES

Airasian, P. (1997). Oregon teacher work sample methodology: Potential and problems. In J. Millman (Ed.). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 46-52). Thousand Oaks, CA: Corwin Press.

American Association of Colleges for Teacher Education, Education Policy Clearinghouse. (2004). *No Child Left Behind* (NCLB) *research.* Retrieved March 12, 2004, from http://www.edpolicy.org/research/nclb/index.php

American Council on the Teaching of Foreign Languages. (1996). *National Standards for Foreign Language Education.* Retrieved on March 11, 2004, from http://actfl.org/

American Council on the Teaching of Foreign Languages. (1999a). *ACTFL Proficiency Guidelines.* Retrieved March 21, 2004, from http://www.actfl.org/

American Council on the Teaching of Foreign Languages. (1999b). *National Standards for Foreign Language Education.* Retrieved on March 11, 2004, from http://www.actfl.org/i4a/pages/index.cfm?pageid=3392

American Council on the Teaching of Foreign Languages. (2002). *ACTFL Standards for the Preparation of Foreign Language Teachers.* Retrieved March 20, 2004, from http://www.ncate.org/documents/ProgramStandards/actfl2002.pdf

Ayres, R., Girod, G., McConney, A., Schalock, M., Schalock, H., & Wright, D. (1996). A guide to teacher work sample methodology. In A. McConney (Ed.). *Connecting teacher work and student learning* (pp. 61-65). Monmouth, OR: Teacher Effectiveness Project, Teaching Research Division, Western Oregon University.

Baratz-Snowden, J. (1990). Research news and comments: The NBPTS begins its research and development program. *Educational Researcher, 19*(6), 19-24.

Baratz-Snowden, J. (1992). *National Board for Professional Teaching Standards – update.* (ERIC Document Reproduction Service No. ED351336)

Benevento, J. (1985). *Issues and innovations in Foreign Language Education.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Blosser, P. (1989). *The impact of educational reform on science education.* (ERIC Document Reproduction Service No. ED320764)

Breiner-Sanders, K., Swender, E., & Terry, R. (2001). *ACTFL revised proficiency guidelines – writing.* Retrieved March 6, 2004, from http://actfl.org/

Burry, J. (1990). *Validity and reliability of classroom observations: A paradox.* (ERIC Document reproduction Service No. ED320967)

Campbell, D., Melenyzer, B., Nettles, D., & Wyman, R. (2000). *Portfolio and performance assessment in teacher education.* Needham Heights, MA: Allyn & Bacon.

Carnegie Task Force on Teaching as a Profession (1986). *Nation Prepared: Teachers for the 21st Century.* Carnegie Forum on Education and the Economy, Hayattsville, MD. (ERIC Document Reproduction Service No. ED268120)

Cartwright, D., & Blacklock, K. (2003, January). *Teacher work samples and struggling readers: Impacting student performance and candidate dispositions.* Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, New Orleans, LA.

Cavalluzzo, L. (2004). *Is National Board Certification an effective signal of teacher quality?* Retrieved November 3, 2006, from http://www.nbpts.org/resources/research/browse_studies?ID=165

Ceperley, P., & Reel, K. (1997). The impetus for the Tennessee Value-Added Accountability System. In J. Millman (Ed.). *Grading Teachers, Grading Schools. Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, California: Corwin Press, Inc.

Clark, J. (1988). The relationship between teacher in-service and educational change. *ILEJ, 4*, 30-38.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, D., & Spillane, J. (1993). Policy and practice: The relationship between governance and instruction. In S. Fuhrman (Ed.), *Designing Coherent Education Policy: Improving the System.* San Francisco: Jossey-Bass.

Coleman, A. (1929). *The teaching of modern foreign languages in the United States.* New York: Macmillan.

Cotton, K. (1995). *Effective schooling practices: A research synthesis, 1995 update.* Retrieved March 21, 2004, from http://www.nwrel.org/scpd/esp95.html

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10,* 83-95.

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley.

Cross, C., & Rigden, D. (2002). Improving teacher quality. *American School Board Journal, 189*(4), 24-27.

Cunningham, G., & Stone, J. (2005). Value-added assessment of teacher quality as an alternative to the National Board for Professional Teaching Standards: What recent studies say. In R. Lissitz (Ed.). *Value added models in education: Theory and applications*. Maple Grove, MN: JAM Press. Retrieved November 3, 2006, from http://www.education-consumers.com/Cunningham-Stone.pdf

Curtain, H. & Pescola, C. (1994). *Languages and Children – Making the Match.* Reading, MA: Addison-Wesley Publishing Co.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching.* New York, N.Y: National Commission on Teaching and America's Future.

Darling-Hammond, L. (1998). *Standards for assessing teaching effectiveness are key.* Phi Delta Kappan, *79*(5), 471-472.

Darling-Hammond, L. (2000). *Solving the dilemmas of teacher supply, demand, and standards: How we can ensure a competent, caring, and qualified teacher for every child.* New York, NY: National Commission on Teaching and America's Future.

Darling-Hammond, L., & Rustique-Forrester, E. (1997). *Investing in quality teaching: State-level strategies*. Denver, CO: Education Commission of the States.

Darlington, R. (1997). The Tennessee Value-Added Assessment System: A challenge to familiar assessment methods. In J. Millman (Ed.). *Grading Teachers, Grading Schools. Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, California: Corwin Press, Inc.

DeFina, A. (1992). *Portfolio assessment: Getting started.* New York: Scholastic Books.

Denner, P., Norman, T., Salzman, S., & Pankratz, R. (2003, February). *Connecting teaching performance to student achievement: A generalizability and validity study of the Renaissance Teacher Work Sample assessment.* Paper presented at the Annual Meeting of the Association of Teacher Educators, Jacksonville, FL.

Denner, P., Norman, T., Salzman, S., Pankratz, R. & Evans, C. (2003). *The Renaissance Partnership Teacher Work Sample: Evidence Supporting Score Generalizability, Validity, and Quality of Student Learning Assessment.* Retrieved February 28, 2002, from http://fp.uni.edu/itq/Research/ATEFinalfromTony061203.pdf

Denner, P., Pankratz, R. S., Norman, A., & Newsome, J. (2004). *Building credibility into performance assessment and accountability systems for teacher preparation programs: A "how to" manual for teacher educators who want to collect, use and report valid and reliable performance data on teacher candidates with a link to P-12 student learning.* The Renaissance Partnership for Improving Teacher Quality, Western Kentucky University, Bowling Green, KY. Retrieved March 15, 2004, from http://fp.uni.edu/itq/PDF_files/Cred_Manual_Jan_30_2004_Working_Draft.pdf

Denner, P., Salzman, S., & Bangert, A. (2001). Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Education, 15*, 287-307.

Denner, P., Salzman, S., & Harris, L. (2002, February). *Teacher work sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning.* Paper presented at the 54th Annual Meeting of the American Association of Colleges for Teacher Education, New York, NY.

Devlin-Scherer, R. (2003). Preservice professional employment portfolios for middle schools. *Essays in Education. Department of Education at the University of South Carolina Aiken.* Retrieved November 15, 2006, from http://www.usca.edu/essays/vol62003/scherer.pdf

Draper, J., & Hicks, J. (1996). Foreign language enrollments in public secondary schools, Fall 1994. *Foreign Language Annals, 29,* 303-06.

Draper, J., & Hicks, J. (2002). *Foreign language enrollments in public secondary schools, Fall 2000, Summary report.* Retrieved March 6, 2004, from http://actfl.org/files/public/Enroll2000.pdf

Educational Testing Service. (2006). *Homepage.* Retrieved November 3, 2006, from www.ets.org

Education Week. (1997, January). *Quality counts 1997.* Retrieved November 3, 2004, from http://www.teachermag.net/sreports/qc97/

Education Week. (1999, January). *Quality counts 1999: Rewarding results, punishing failure.* Retrieved November 3, 2004, from http://www.edweek.org/rc/articles/2004/10/15/qc-archive.html

Education Week. (2006, January). *Quality Counts at 10: A Decade of Standards-Based Education.* Retrieved November 10, 2006, from http://www.edweek.org/ew/toc/2006/01/05/index.html

Evertson, C., & Burry, J. (1988). *Capturing classroom contex: The observation system as lens for assessment.* (ERIC Document Reproduction Service No. ED298109)

Ferguson, R. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal of Legislation, 28,* 465-498.

Ferguson, R., & Ladd, H. (1996). How and why money matters: An analysis of Alabama schools. In H. Ladd (Ed.). *Holding schools accountable* (pp. 265-298). Washington, DC: Brookings Institute.

Finn, C., Petrilli, M., & Vanourek, G. (1998). The State of state standards. Retrieved March 11, 2004, from http://www.fordhamfoundation.org/institute/publication/publication.cfm?id=25

Fredman, T. (2002, February). *The TWSM: An essential component in the assessment of teacher performance and student learning.* Paper presented at the 54th Annual Meeting of the American Association of Colleges for Teacher Education, New York, NY. (ERIC Document Reproduction Service No. ED464046)

Fredman, T. (2004). Teacher work sample methodology: Implementation and practical application in teacher preparation. *Action in Teacher Education, 6,* 3-11.

Fullan, M., & Stiegelbauer, S. (1991). *The meaning of educational change.* London: Cassell.

The Gale Group. (2007). *Encyclopedia of Education.* Retrieved April 3, 2007, from http://www.answers.com/topic/national-council-for-accreditation-of-teacher-education

Gall, M., Gall, J., & Borg, W. (2003). *Educational Research: An Introduction, 7th Ed.* Boston, MA: Allyn and Bacon.

Gardner, H. (1991). *The unschooled mind: How children think and how schools should teach.* New York: Basic Books.

Girod, D. (Ed.). (2002). *Connecting teaching and learning: A handbook for teacher educators on teacher work sample methodology.* Washington, D.C.: AACTE Publications.

Glisan, E. (2001). Reframing teacher education within the context of quality, standards, supply, and demand. In R. Lavine (Ed.). *Beyond the boundaries: Changing contexts in language learning* (pp. 165-200). Boston: McGraw Hill.

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Retrieved November 3, 2006, from http://www.nbpts.org/resources/research/browse_studies?ID=167

Greenwald, R., Hedges, L., & Laine, R. (1996, Autumn). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361-396.

Harman, A. (2001). *National Board for Professional Teaching Standards' national teacher certification.* (ERIC Document Reproduction Service No. FE460126)

Hegler, K. (2003, January). *Evaluating the use of teacher work samples to describe teacher candidate competence and PK-12 learning.* Paper presented at the 55th Annual Meeting of the American Association of Colleges for Teacher Education, New Orleans, LA.

Henning, J. & Robinson V. (2004). The teacher work sample: Implementing standards-based performance. *The Teacher Educator, 29*, 231-248.

Holbein, M. (1998). Will standards improve student achievement? *Education, 188*(4), 559-564.

Imig, D., & Smith, C. (2002). Foreword. In G. Girod, (Ed.). *Connecting teaching and learning: A handbook for teacher educators on teacher work sample methodology* (pp. ix-x). Washington, D.C.: AACTE Publications.

Indiana State University (n.d.). *Glossary of Education Terms.* Retrieved February 28, 2004, from http://soe.indstate.edu/ess/tchandGloss.htm

International Foundation for Functional Gastrointestinal Disorders (n.d.). *Glossary of Terms.* Retrieved November 20, 2006, from http://www.iffgd.org/GIDisorders/glossary.html

Interstate New Teacher Assessment and Support Consortium. (1992). *Model standards for beginning teacher licensing and development: A resource for state dialogue.* Washington, DC: Council of Chief State School Officers.

Iowa State University (n.d.). *Teaching Licensure.* Retrieved October 5, 2006, from http://www.teacher.hs.iastate.edu/teaching_licensure.php

Jackson, J. (2006). *Increasing accountability for teacher preparation programs.* Southern Regional Education Board. Retrieved March 28, 2007, from http://www.sreb.org/main/goals/Publications/06E18_Increasing_Accountability.pdf

Johnson, C. (2004). *Nation's report card: An overview of NAEP.* Washington, DC: National Center for Educational Statistics.

Juvenile Justice Evaluation Center Online (n.d.). *Glossary.* Retrieved November 20, 2006, from http://www.jrsa.org/jjec/resources/definitions.html

Keese, N., & Brown, T. (2003, August). *Student teacher input and Teacher Work Sample as part of a Teacher Education Unit Accountability System.* Paper presented at the Annual Meeting of the Association of Teacher Educators, Santa Fe, NM.

Kenyon, D., Farr, B., Mitchell, J., & Armengol, R. (2000). *Framework for the 2004 foreign language assessment of educational progress.* Retrieved March 6, 2004, from http://www.nagb.org/pubs/FinalFrameworkPrePubEdition1.pdf

Kershner, C. (1999). *The National Board for Professional Teaching Standards: National teacher certification in the works.* Retrieved November 3, 2006, from http://www.hslda.org/docs/nche/000000/00000031.asp

Kingston, N., & Reidy, E. (1997). Kentucky's accountability and assessment systems. In J. Millman (Ed.). *Grading Teachers, Grading Schools. Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, California: Corwin Press, Inc.

Levin, H. (1998). Educational performance standards and the economy. *Educational Researcher, 27*(4), 4-10.

Long, C. & Stansbury, K. (1994, December). Performance assessments for beginning teachers: Options and lessons. *Phi Delta Kappan, 76*(4), 318-322.

McAllister, P. (April, 2003). *The Praxis Series: Meeting the "Highly Qualified Teacher" Challenge*. Retrieved February 28, 2004, from http://www.ets.org/aboutets/testimony2.html

McConney, A., & Ayres, R. (1998). Assessing student teachers' assessments. *Journal of Teacher Education, 49*(2), 140-150.

McConney, A., Schalock, M., & Schalock, H. (1998). Focusing improvement and quality assurance: Work samples as authentic performance measures of prospective teachers' effectiveness. *Journal of Personnel Evaluation in Education, 11,* 343-363.

McLaughlin, M., Shepard, L., & O'Day, A. (1995). *Improving education through standards-based reform.* Standard, CA: National Academy of Education.

McLean, R., & Sanders, W. (1984). *Objective component of teacher evaluation: A feasibility study* (Working paper no. 199). Knoxville: University of Tennessee, College of Business Administration.

McRobbie, J. (2001). *Career-long teacher development: Policies that make sense.* San Francisco, CA: West Education.

Meng Thum, Y., & Bryk, A. (1997). Value-added productivity indicators: The Dallas system. In J. Millman (Ed.). *Grading Teachers, Grading Schools. Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, California: Corwin Press, Inc.

Measurement Experts. (n.d.) *Definitions.* Retrieved September 8, 2006, from www.measurementexperts.org/instrument/term_pocket_terms.asp

Millman, J. (Ed.). (1997). How do I judge thee? Let me count the ways. In J. Millman(Ed.). *Grading Teachers, Grading Schools. Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, California: Corwin Press, Inc.

Mitchell, A. (2001). Learnings from the NCATE pilot institutions. *NCATE: The Standard of Excellence in Teacher Preparation, 10*(1), 4-5.

Musick, M. (1998, July). Getting results: A fresh look at school accountability. Retrieved March 11, 2004, from http://www.sreb.org/main/highschools/accountability/gettingresults98.asp

National Academies Press. (n.d.). *Teaching Standards.* Retrieved February 28, 2004, from http://www.nap.edu/readingroom/books/nses/html/3.html

National Board for Professional Teaching Standards. (1998). *Early Childhood/Generalist Standards.* Washington, D.C.: Author. Retrieved March 10, 2004, from http://www.nbpts.org/

National Board for Professional Teaching Standards. (2004). *NBCTs by State.* Retrieved March 11, 2004, from http://www.nbpts.org/nbct/nbctdir_bystate.cfm

National Board for Professional Teaching Standards. (2006). *Milestones.* Retrieved November 3, 2006, from http://www.nbpts.org/about_us/background/milestones

National Board for Professional Teaching Standards. (2007). *History of NBPTS.* Retrieved March 29, 2007, from http://www.uni.edu/coe/nbpts/history.html

National Center for Education Statistics. (2003). *The National Assessment of Educational Progress.* Retrieved March 11, 2004, from http://nces.ed.gov/

National Center for Education Statistics. (2005). *The National Assessment of Educational Progress.* Retrieved November 18, 2006, from http://nces.ed.gov/

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative of educational reform* (Report No. 065-000-00177-2). Washington, D.C.: U.S. Government Printing Office.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future.* New York, N.Y.: N.Y. Author.

National Commission on Teaching and America's Future. (2003). *No dream denied: A pledge to America's children.* Retrieved March 9, 2004, from http://www.nctaf.org/dream/summary_report.pdf

National Council for Accreditation of Teacher Education. (2003). *Assessing Education Candidate Performance: A Look at Changing Practices.* Washington, D.C.: NCATE.

National Standards in Foreign Language Education Project. (1996). *Standards for Foreign language learning: Preparing for the 21st century*. Lawrence, KS: Allen Press.

National Standards in Foreign Language Education Project. (1999). *Standards for Foreign language learning in the 21st century*. Lawrence, KS: Allen Press.

Pankratz, R. (n.d.) *Why the Renaissance Teacher Work Sample strategy has a high potential to improve student learning opportunities.* Retrieved March 10, 2006, from http://cstl.semo.edu/rtwsm/article.htm

Pankratz, R. (1999, February). *Becoming accountable for the impact of graduates on students and schools: Making operational the shift from teaching to learning.* Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, Washington D.C. Retrieved February 28, 2004, from http://fp.uni.edu/itq/Paper_Publication/Becoming_Accountable.htm

Pankratz, R. (2004). *The power of partnerships in becoming accountable for the impact of teacher candidates on P-12 learning.* Retrieved May 12, 2005, from http://fp.uni.edu/itq/Paper_Publication/Chapter_5_The_Power_of_Partnerships_in_Becoming_Accountable_093004.pdf

Parker, W. (1957). *The national interest in foreign languages.* Washington, DC: U.S. National Commission for UNESCO, Department of State.

Perkins, D. (1992). *Smart schools: Better thinking and learning for every child.* New York: Free Press.

Phillips, J., & Glisan, E. (Eds.). (2002). *American Council on the Teaching of Foreign Languages, program Standards for the Preparation of Foreign Language Teachers*. New York, NY: ACTFL.

Pratt, E. (2002). Aligning mathematics teacher work sample content with selected NCTM standards: Implications for preservice teacher education. *Journal of Personnel Evaluation in Education, 16*(3), 175-190.

Pritchard, I. (1996). *Judging standards in standards-based reform.* Washington, DC: Council for Basic Education.

Ravitch, D. (1995a). *National Standards in American Education: A Citizen's Guide.* Washington, DC: The Brookings Institution.

Ravitch, D., Ed. (1995b). *Debating the Future of American Education: Do We Need National Standards and Assessment?* Washington, DC: The Brookings Institution.

Renaissance Partnership for Improving Teacher Quality Project. (n.d.a). *Project Overview.* Retrieved February 28, 2004, from http://fp.uni.edu/itq

Renaissance Partnership for Improving Teacher Quality Project. (n.d.b). *Scored TWS Exemplars produced by student teachers.* Retrieved February 28, 2004, from http://fp.uni.edu/itq/Scored_TWS/index.htm

Renaissance Partnership for Improving Teacher Quality Project. (2002a, January). *Teacher Work Sample road map for locating evidence.* Retrieved November 28, 2006, from http://fp.uni.edu/itq/PDF_files/roadmap.pdf

Renaissance Partnership for Improving Teacher Quality Project. (2002b, June). *Teacher Work Sample: Performance Prompt, Teaching Process Standards, and Scoring Rubrics.* Retrieved November 28, 2006, from http://fp.uni.edu/itq/PDF_files/June2002promptandrubric.pdf

Renaissance Partnership for Improving Teacher Quality Project. (2002c, June). *Teacher Work Sample Scoring Guide.* Retrieved November 28, 2006, from http://fp.uni.edu/itq/PDF_files/June2002ScoringGuide.pdf

Rhodes, N.C., & Branaman, L.E. (1999). *Foreign language instruction in the United States: A national survey of elementary and secondary schools.* Washington, DC and McHenry, IL: Center for Applied Linguistics and Delta Systems.

Richards, J., Platt, J., & Platt, H. (1993). *Longman Dictionary of Language Teaching and Applied Linguistics.* Essex: Longman.

Richards, J., & Rodgers, T. (2001). *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press.

Richards, J., Tung, P., & Ng, P. (1992). The culture of the English language teacher: A Hong Kong example. *RELC Journal, 23*(1), 81-102.

Robinson, V., & Boody, R. (2003). Doing instructional leadership through the Teacher Work Sample: A different way to play. *Iowa Educational Leadership, 5*(4), 19-20.

Rudden, J. (2003). The impact of teaching on learning: A case study of using a teacher work sample as a source of evidence of student learning. *Journal of Reading Education,* 28(3), 26-35.

Salzman, S., Denner, P., Bangert, A., & Harris, L. (2001, March). *Connecting teacher performance to the learning of all students: Ethical dimensions of shared responsibility.* Paper presented at the 53rd Annual Meeting of the American Association of Colleges for Teacher Education, Dallas, TX.

Sanders, W. & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel in Education, 8*, 299-311.

Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers of future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Schackner, B., & Lee, C. (2002, May 25). Teacher test results pressure universities. *The Pittsburg Post-Gazette.* Retrieved March 28, 2007, from http://www.post-gazette.com/localnews/20020525teachers0525p3.asp

Schalock, H., & Cowart, B. (1993). *Oregon's design for twenty-first century schools and its implication for teachers: A paradigm shift.* Monmouth: Division of Continuing Education, Western Oregon State College.

Schalock, H., & Myton, D. (1989). A new paradigm for teacher licensure: Oregon's demand for evidence of success in fostering learning. *Journal of Teacher Education, 39*(6), 8-16.

Schalock, H., & Myton, D. (2002). Connecting Teaching and learning: An Introduction to Teacher Work Sampling. In G. Girod (Ed.), *Connecting Teaching and Learning: A handbook for Teacher Educators on Teacher Work Sample Methodology* (pp. 5-31). Washington, DC: American Association of Colleges for Teacher Education Publications.

Schalock, H., Schalock M., & Girod, G. (1997). Teacher work sample methodology as used at Western Oregon University. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 15-45). Thousand Oaks, CA: Corwin Press, Inc.

Schalock, H., Schalock, M, McConney, A., Brodsky, M., & Myton, D. (2002). *Teacher work sampling: An evolving methodology for connecting teacher preparation and licensing to P-12 progress in learning.* Unpublished manuscript, University of Western Oregon.

Schalock, H., Schalock, M., & Myton, D. (1998). Effectiveness – along with quality – should be the focus. *Phi Delta Kappan, 79*(6), 468-470.

Schalock, H., Schalock, M., & Myton, D. (1999). Effective teachers please, Mr. Wise, quality is not enough. *Phi Delta Kappan, 80*(5), 1.6-1.9

Schalock, H., Schalock, M., & Myton, D., & Girod, G. (1993). Focusing on learning gains by pupils taught: A central feature of Oregon's outcome-based approach to the initial preparation and licensure of teachers. *Journal of Personnel Evaluation in Education, 7*(2), 135-158.

Schalock, M. (1998). Accountability, student learning, and the preparation and licensure of teachers: Oregon's Teacher Work Sample Methodology. *Journal of Personnel Evaluation in Education, 12*(3), 269-285.

Schrier, L. (1993). Prospects for the professionalization of foreign language teaching. In G. Guntermann (Ed), *Developing language teachers for a changing world.* pp.105-123. Lincolnwood, IL: National Textbook.

Schrier, L. (2002). The knowledge base supporting the Standards: Blending the past with the present. In J. Phillips & E. Glisan (Eds.). *American Council on the Teaching of Foreign Languages, program Standards for the Preparation of Foreign Language Teachers* (pp. 1-15). New York: ACTFL.

Scriven, M. (1994). Duties of the teacher. *Journal of Personnel Evaluation in Education, 8*(2), 151-184.

Scriven, M. (1996). Assessment in teacher education: Getting clear on the concept. *Teaching and Teacher Education, 12*(4), 443-450.

Selwyn, D. (2005/2006, Winter). Teacher quality: Teacher education left behind. *Rethinking Schools Online, 20*(2). Retrieved March 28, 2007, from http://www.rethinkingschools.org/archive/20_02/left202.shtml

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shepard, L. (1993). *Setting performance standards for student achievement.* Stanford CA: National Academy of Education, Stanford University.

Simon, P. (1980). *The Tongue-tied American: Confronting the Foreign Language Crisis.* New York: The Crossroad Publishing Company.

Smith, T., Gordon, B., Colby, S., & Wang, J. (2005). An examination of the relationship between depth of student learning and National Board Certification status. Retrieved November 3, 2006, from http://www.nbpts.org/resources/research/browse_studies?ID=167

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. Fuhrman, and B. Malen (Eds.), *The Politics of Curriculum and Testing, 1990 Yearbook of the Politics of Education Association.* London and Washington, DC: Falmer Press, 233-267.

Smith, M. S., O'Day, J., & Cohen, D. K. (1990, Winter). National curriculum American style: Can it be done? What might it look like? *American Educator, 10*, 40-47.

State University of New York College at Cortland. (n.d.). *SOPHE/AAHE Task Force on Accreditation Profession Survey: Definition Page.* Retrieved February 11, 2006, from http://www.cortland.edu/hlth/surveydefinitions.html

Stufflebeam, D. (1997). Overview and assessment of the Kentucky Instructional Results Information System. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, California: Corwin Press, Inc.

Sturm, H. (1995). *Public school accountability.* Background Paper 95-14. Nevada Legislative Counsel Bureau.

Sykes, G. (1997). On trial: The Dallas Value-Added Accountability System. In J. Millman (Ed.). *Grading Teachers, Grading Schools. Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, California: Corwin Press, Inc.

Symonds, W. (2001, September). A is for answers. *Reader's Digest,* 99-105.

Tinker Sachs, G., Kong, S., Lo, A., & Lee T. (1994). From task description to task enactment: Teachers' interpretation of language learning tasks. In Bird., N. (Ed.). *Language and Learning* (172-187). Paper presented at the Annual International Language in Education Conference, Hong Kong, China.

Unites States Department of Education. (1994). *Goals 2000: Educate America Act.* Retrieved March 11, 2006, from http://www.ed.gov/legislation/GOALS2000/TheAct/index.html

United States Department of Education. (1997). *Third International Mathematics and Science Study.* Retrieved March 11, 2004, from http://www.ed.gov/inits/TIMSS/

United States Department of Education. (1998). *Promising Practices: New Ways to Improve Teacher Quality.* Retrieved March 29, 2007, from http://www.ed.gov/pubs/PromPractice/title.html

United States Department of Education. (1999). *Riley Calls for Fundamental Changes to Improve American Education; Announces Steps to Address Accountability and Teacher Quality.* Retrieved March 11, 2006, from http://www.ed.gov/Speeches/02-1999/990216-a.html

United States Department of Education. (2001, October). *Misconceptions About the Goals 2000: Educate America Act.* Retrieved February 28, 2004, from http://www.ed.gov/G2K/myths.html

United States Department of Education. (2002a). *No Child Left Behind Act.* Retrieved February 28, 2004, from http://www.ed.gov/nclb/landing.jhtml?src=pb

United States Department of Education. (2002b). *On the Horizon: State Accountability Systems.* Retrieved on March 27, 2006, from http://www.ed.gov/admins/lead/account/stateacct/edlite-slide001.html

Vandevoort, L., Amrein-Beardsley, A., & Berliner, D. (2004). *National Board Certified teachers and their students' achievement.* Retrieved November 3, 2006, from http://www.nbpts.org/resources/research/browse_studies?ID=166

Wakefield, D. (2003). Screening teacher candidates: Problems with high-stakes testing. *The Educational Forum.* Retrieved November 3, 2006, from http://www.findarticles.com/p/articles/mi_qa4013/is_200307/ai_n9271541/pg_1

Watzke, J. (2003). *Lasting Change in Foreign Language Education: A historical case for change in national policy.* Portsmouth, NH: Greenwood Publishing Group.

Webb, J., & Brown, B. (1969, February). *Establishing reliability and validity estimates for systematic classroom observations.* Paper presented at the American Educational Research Association Meeting, Los Angeles, CA.

Webster, W., & Mendro, R. (1997). The Dallas value-added accountability system. In J. Millman (Ed.). *Grading Teachers. Grading Schools. Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks, CA: Corwin Press, Inc.

Webster, W., Mendro, R., Orsak, T., & Weerasinghe, D. (1998, April). *An application of hierarchical linear modeling to the estimation of school and teacher effect.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Western Oregon University. (n.d.). *Key Concepts.* Retrieved February 28, 2004, from http://www.wou.edu/education/worksample/twsm/page1.htm

Wilson, S., Floden, R., & Ferrini-Mundy, J. (2001, February). *Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations.* Seattle: Center for the Study of Teaching and Policy, University of Washington. Retrieved February 28, 2004, from http://depts.washington.edu/ctpmail/PDFs/TeacherPrep-WFFM-02-2001.pdf

Wolf, K. (1991). The schoolteacher's portfolio: Issues in design, implementation, and evaluation. *Phi Delta Kappan, 73*(2), 129-136.

Wolf, W., & Riordan, K. (1991). Foreign Language Teachers' Demographic Characteristics, In-service Training Needs, and Attitudes toward Teaching. *Foreign Language Annals, 24*(6), 471-478.

APPENDIX A

DEMOGRAPHIC QUESTIONNAIRE

Participant Survey ID number_____

1. **Gender:** ____Male ____Female

2. **Highest** degree received (**choose one**):
__AA __BA(general) __BA(teaching) __MA __Doctorate

3. **Now I teach at (check all that apply):**
__Elementary level __Middle School Level __High School Level

4. Years of teaching experience (state **total** number of years) _____

4a. If you are/were a world (foreign) language educator, state total number of years you have been teaching/taught in this field _____

5. Content area(s) I currently teach (**state all** content area(s)) _____

5a. If you are a world language teacher, which language(s) have you taught/do you teach (specify) _____

6. Do you know any world languages other than English? (**check one**) ____Yes ____No

6a. If yes, which world language(s) do you know? (specify) _____

6b. How would you rate your **proficiency** in each world language you know (**specify**):

Lang. 1: _____ __Beginner __Intermediate __Advanced __Native (-like)
Lang. 2: _____ __Beginner __Intermediate __Advanced __Native (-like)
Lang. 3: _____ __Beginner __Intermediate __Advanced __Native (-like)

7. Have you heard about Teacher Work Sampling prior to this event? (check one) ___Yes ___No

8. Have you participated in Teacher Work Sample (TWS) scoring before? (**check one**) ___Yes ___No

8a. If yes, how many times? (specify) _____

9.	Have you ever served as a cooperating teacher to any teacher candidate? (**check one**)	___Yes	___No
	9a. If yes, were you a cooperating teacher for a candidate with a Teacher Work Sample? (check one)	**___Yes**	**___No**
10.	Are you a NBPTS certified educator?	___Yes	___No
11.	**Have you ever received assessment training and scored high stake assessments (e.g., Advanced Placement Assessment, Oral Proficiency Interviews)?**	**___Yes**	**___No**

APPENDIX B

RENAISSANCE PARTNERSHIP FOR IMPROVING TEACHER QUALITY

PROJECT SCORING RUBRIC

The Renaissance Partnership for Improving Teacher Quality

Work Sample #: _____

Grade/Subject: _____

OVERALL Sample Score: _____

Rater ID #: _____

State EXACT TIME of Start:_____ **Finish:** _____

Contextual Factors Rubric ***Rubric SCORE:*** _____

TWS Standard: ***The teacher uses information about the learning/teaching context and student individual differences to set learning goals, plan instruction and assess learning.***

Rating → Indicator ↓	**1 Indicator Not Met**	**2 Indicator Partially Met**	**3 Indicator Met**	**Score**
Knowledge of Community, School and Classroom Factors	Teacher displays minimal, irrelevant, or biased knowledge of the characteristics of the community, school, and classroom.	Teacher displays some knowledge of the characteristics of the community, school, and classroom that may affect learning.	Teacher displays a comprehensive understanding of the characteristics of the community, school, and classroom that may affect learning.	
Knowledge of Characteristics of Students	Teacher displays minimal, stereotypical, or irrelevant knowledge of student differences (e.g. development, interests, culture, abilities/disabilities).	Teacher displays general knowledge of student differences (e.g., development, interests, culture, abilities/disabilities) that may affect learning.	Teacher displays general & specific understanding of student differences (e.g., development, interests, culture, abilities/disabilities) that may affect learning.	

Knowledge of Students' Varied Approaches to Learning	Teacher displays minimal, stereotypical, or irrelevant knowledge about the different ways students learn (e.g., learning styles, learning modalities).	Teacher displays general knowledge about the different ways students learn (e.g., learning styles, learning modalities).	Teacher displays general & specific understanding of the different ways students learn (e.g., learning styles, learning modalities) that may affect learning.	
Knowledge of Students' Skills And Prior Learning	Teacher displays little or irrelevant knowledge of students' skills and prior learning.	Teacher displays general knowledge of students' skills and prior learning that may affect learning.	Teacher displays general & specific understanding of students' skills and prior learning that may affect learning.	
Implications for Instructional Planning and Assessment	Teacher does not provide implications for instruction and assessment based on student individual differences and community, school, and classroom characteristics OR provides inappropriate implications.	Teacher provides general implications for instruction and assessment based on student individual differences and community, school, and classroom characteristics.	Teacher provides specific implications for instruction and assessment based on student individual differences and community, school, and classroom characteristics.	

Learning Goals Rubric ***Rubric Score*** ___

TWS Standard: ***The teacher sets significant, challenging, varied and appropriate learning goals.***

Rating → Indicator ↓	**1 Indicator Not Met**	**2 Indicator Partially Met**	**3 Indicator Met**	**Score**
Significance, Challenge and Variety	Goals reflect only one type or level of learning.	Goals reflect several types or levels of learning but lack significance or challenge.	Goals reflect several types or levels of learning and are significant and challenging.	
Clarity	Goals are not stated clearly and are activities rather than learning outcomes.	Some of the goals are clearly stated as learning outcomes.	Most of the goals are clearly stated as learning outcomes.	
Appropriaten. for Students	Goals are not appropriate for the development; pre-requisite knowledge, skills, experiences; or other student needs.	Some goals are appropriate for the development; pre-requisite knowledge, skills, experiences; and other student needs	Most goals are appropriate for the development; pre-requisite knowledge, skills, experiences; and other student needs.	
Alignment with National, State or Local Standards	Goals are not aligned with national, state or local standards.	Some goals are aligned with national, state or local standards.	Most of the goals are explicitly aligned with national, state or local standards.	

Assessment Plan Rubric *Rubric Score* ___

TWS Standard: ***The teacher uses multiple assessment modes and approaches aligned with learning goals to assess student learning before, during and after instruction.***

Rating → Indicator ↓	1 Indicator Not Met	2 Indicator Partially Met	3 Indicator Met	Score
Alignment with Learning Goals and Instruction	Content and methods of assessment lack congruence with learning goals or lack cognitive complexity.	Some of the learning goals are assessed through the assessment plan, but many are not congruent with learning goals in content and cognitive complexity.	Each of the learning goals is assessed through the assessment plan; assessments are congruent with the learning goals in content and cognitive complexity.	
Clarity of Criteria and Standards for Performance	The assessments contain no clear criteria for measuring student performance relative to the learning goals.	Assessment criteria have been developed, but they are not clear or are not explicitly linked to the learning goals.	Assessment criteria are clear and are explicitly linked to the learning goals.	
Multiple Modes and Approaches	The assessment plan includes only one assessment mode and does not assess students before, during, and after instruction.	The assessment plan includes multiple modes but all are either pencil/paper based (i.e. they are not performance assessments) and/or do not require the integration of knowledge, skills and reasoning ability.	The assessment plan includes multiple assessment modes (including performance assessments, lab reports, research projects, etc.) and assesses student performance throughout the instructional sequence.	
Technical Soundness	Assessments are not valid; scoring procedures are absent or inaccurate; items or prompts are poorly written; directions and procedures are confusing to students.	Assessments appear to have some validity. Some scoring procedures are explained; some items or prompts are clearly written; some directions and procedures are clear to students.	Assessments appear to be valid; scoring procedures are explained; most items or prompts are clearly written; directions and procedures are clear to students.	

Design for Instruction Rubric ***Rubric Score:*** ____

TWS Standard: ***The teacher designs instruction for specific learning goals, student characteristics and needs, and learning contexts.***

Rating → Indicator↓	1 Indicator Not Met	2 Indicator Partially Met	3 Indicator Met	Score
Alignment with Learning Goals	Few lessons are explicitly linked to learning goals. Few learning activities, assignments and resources are aligned with learning goals. Not all learning goals are covered in the design.	Most lessons are explicitly linked to learning goals. Most learning activities, assignments and resources are aligned with learning goals. Most learning goals are covered in the design.	All lessons are explicitly linked to learning goals. All learning activities, assignments and resources are aligned with learning goals. All learning goals are covered in the design.	
Accurate Representation of Content	Teacher's use of content appears to contain numerous inaccuracies. Content seems to be viewed more as isolated skills and facts rather than as part of a larger conceptual structure.	Teacher's use of content appears to be mostly accurate. Shows some awareness of the big ideas or structure of the discipline.	Teacher's use of content appears to be accurate. Focus of the content is congruent with the big ideas or structure of the discipline.	
Lesson and Unit Structure	The lessons within the unit are not logically organized organization (e.g., sequenced).	The lessons within the unit have some logical organization and appear to be somewhat useful in moving students toward achieving the learning goals.	All lessons within the unit are logically organized and appear to be useful in moving students toward achieving the learning goals.	
Use of a Variety of Instruction, Activities, Assignments and Resources	Little variety of instruction, activities, assignments, and resources. Heavy reliance on textbook or single resource (e.g., work sheets).	Some variety in instruction, activities, assignments, or resources but with limited contribution to learning.	Significant variety across instruction, activities, assignments, and/or resources. This variety makes a clear contribution to learning.	
Use of Contextual Information and Data to Select Appropriate and Relevant Activities, Assignments and Resources	Instruction has not been designed with reference to contextual factors and pre-assessment data. Activities and assignments do not appear productive and appropriate for each student.	Some instruction has been designed with reference to contextual factors and pre-assessment data. Some activities and assignments appear productive and appropriate for each student.	Most instruction has been designed with reference to contextual factors and pre-assessment data. Most activities and assignments appear productive and appropriate for each student.	
Use of Technology	Technology is inappropriately used OR teacher does not use technology, and no (or inappropriate) rationale is provided.	Teacher uses technology but it does not make a significant contribution to teaching and learning OR teacher provides limited rationale for not using technology.	Teacher integrates appropriate technology that makes a significant contribution to teaching and learning OR provides a strong rationale for not using technology.	

Instructional Decision-Making Rubric *Rubric Score* ____

TWS Standard: ***The teacher uses on-going analysis of student learning to make instructional decisions.***

Rating → Indicator ↓	1 Indicator Not Met	2 Indicator Partially Met	3 Indicator Met	Score
Sound Professional Practice	Many instructional decisions are inappropriate and not pedagogically sound.	Instructional decisions are mostly appropriate, but some decisions are not pedagogically sound.	Most instructional decisions are pedagogically sound (i.e., they are likely to lead to student learning).	
Modifications Based on Analysis of Student Learning	Teacher treats class as "one plan fits all" with no modifications.	Some modifications of the instructional plan are made to address individual student needs, but these are not based on the analysis of student learning, best practice, or contextual factors.	Appropriate modifications of the instructional plan are made to address individual student needs. These modifications are informed by the analysis of student learning/performance, best practice, or contextual factors. Include explanation of why the modifications would improve student progress.	
Congruence Between Modifications and Learning Goals	Modifications in instruction lack congruence with learning goals.	Modifications in instruction are somewhat congruent with learning goals.	Modifications in instruction are congruent with learning goals.	

Analysis of Student Learning Rubric *Rubric Score* ____

TWS Standard: ***The teacher uses assessment data to profile student learning and communicate information about student progress and achievement.***

Rating → Indicator ↓	1 Indicator Not Met	2 Indicator Partially Met	3 Indicator Met	Score
Clarity and Accuracy of Presentation	Presentation is not clear and accurate; it does not accurately reflect the data.	Presentation is understandable and contains few errors.	Presentation is easy to understand and contains no errors of representation.	
Alignment with Learning Goals	Analysis of student learning is not aligned with learning goals.	Analysis of student learning is partially aligned with learning goals and/or fails to provide a comprehensive profile of student learning relative to the goals for the whole class, subgroups, and two individuals.	Analysis is fully aligned with learning goals and provides a comprehensive profile of student learning for the whole class, subgroups, and two individuals.	
Interpretation of Data	Interpretation is inaccurate, and conclusions are missing or unsupported by data.	Interpretation is technically accurate, but conclusions are missing or not fully supported by data.	Interpretation is meaningful, and appropriate conclusions are drawn from the data.	

Evidence of Impact on Student Learning	Analysis of student learning fails to include evidence of impact on student learning in terms of numbers of students who achieved and made progress toward learning goals.	Analysis of student learning includes incomplete evidence of the impact on student learning in terms of numbers of students who achieved and made progress toward learning goals.	Analysis of student learning includes evidence of the impact on student learning in terms of number of students who achieved and made progress toward each learning goal.	

Reflection and Self-Evaluation Rubric ***Rubric Score*** ____

TWS Standard: ***The teacher analyzes the relationship between his or her instruction and student learning in order to improve teaching practice.***

Rating → Indicator ↓	1 Indicator Not Met	2 Indicator Partially Met	3 Indicator Met	Score
Interpretation of Student Learning	No evidence or reasons provided to support conclusions drawn in "Analysis of Student Learning" section.	Provides evidence but no (or simplistic, superficial) reasons or hypotheses to support conclusions drawn in "Analysis of Student Learning" section.	Uses evidence to support conclusions drawn in "Analysis of Student Learning" section. Explores multiple hypotheses for why some students did not meet earning goals.	
Insights on Effective Instruction and Assessment	Provides no rationale for why some activities or assessments were more successful than others.	Identifies successful and unsuccessful activities or assessments and superficially explores reasons for their success or lack thereof (no use of theory or research).	I Identifies successful and unsuccessful activities and assessments and provides plausible reasons (based on theory or research) for their success or lack thereof.	
Alignment Among Goals, Instruction and Assessment	Does not connect learning goals, instruction, and assessment results in the discussion of student learning and effective instruction and/or the connections are irrelevant or inaccurate.	Connects learning goals, instruction, and assessment results in the discussion of student learning and effective instruction, but misunderstandings or conceptual gaps are present.	Logically connects learning goals, instruction, and assessment results in the discussion of student learning and effective instruction.	
Implications for Future Teaching	Provides no ideas or inappropriate ideas for redesigning learning goals, instruction, and assessment.	Provides ideas for redesigning learning goals, instruction, and assessment but offers no rationale for why these changes would improve student learning.	Provides ideas for redesigning learning goals, instruction, and assessment and explains why these modifications would improve student learning.	
Implications for Professional Development	Provides no professional learning goals or goals that are not related to the insights and experiences described in this section.	Presents professional learning goals that are not strongly related to the insights and experiences described in this section and/or provides a vague plan for meeting the goals.	Presents a small number of professional learning goals that clearly emerge from the insights and experiences described in this section. Describes specific steps to meet these goals.	

APPENDIX C

INVITATION TO THE TRAINING AND SCORING EVENT

University of Northern Iowa

College of Education

When: Saturday, May 1st, 2004
8:00am-2:30pm
Where: Cedar Falls, Holiday Inn

Event Agenda:
8:00-9:00am – breakfast & scoring training
9:00-9:10am – break
9:10-12:00pm – scoring session
12:00-12:30pm – lunch
12:30-2:30pm – scoring session

Renaissance Partnership for Improving Teacher Quality Project

Dr. Victoria Robinson
UNI Project Director
University of Northern Iowa
College of Education
SEC 512
Cedar Falls, Iowa 50614-0604

Phone: 319-273-3070
Email: victoria.robinson@uni.edu

Project Website: http://fp.uni.edu/itq

[redacted]

Join us for UNI Training & Scoring Event

Dear World Language Educator (or Educator),

You are invited to participate in the fourth semi-annual Teacher Work Sample training and scoring session offered by UNI.

A continental breakfast and a brief scoring review will be provided prior to scoring of the Teacher Work Samples written by UNI students. Lunch will also be provided.

Teacher work samples are documents compiled by UNI teacher candidates during their student teaching where they record their instructional practices and impact on student learning.

You will receive:
- a stipend of **$300.00** for your scoring contributions, and
- a **professional development certificate** for your participation.

Seats are limited, **please RSVP**, no later than **April 17th**, to yana.cornish@uni.edu or by phone: 273-3064 if you are planning to attend the event. I will acknowledge your response with an e-mail or phone reply. Feel free to contact me if you have any questions regarding this event.

Hope to see you May 1st at the Cedar Falls Holiday Inn.

Victoria Robinson

UNI Project Director
Renaissance Partnership for Improving Teacher Quality

APPENDIX D

THANK YOU LETTER FOR THE PARTICIPANTS OF THE STUDY

05/01/04

Dear Scorer of Teacher Work Samples,

Thank you for taking your valuable time to participate in the fourth semi-annual Teacher Work Sample Training and Scoring Session on May 1st, 2004.

This scoring event was offered by the Renaissance Partnership for Improving Teacher Quality Title II Grant Project at the University of Northern Iowa. Without your help and expertise this aspect of teacher preparation would not have been possible. Through this scoring we were able to give valuable feedback to those entering the field of teaching and improve teacher quality for Iowa's children. Additionally, your participation contributed to the data collection efforts focused on validation of the scoring instrument.

I hope that your experience has been a positive one, and that in the future you would consider participating in this event again. Thank you for your participation and scoring contributions.

Dr. Victoria Robinson

UNI Project Coordinator
SEC 512
Phone: 319-273-3070
Email: victoria.robinson@uni.edu
Project Website: http://fp.uni.edu/itq

Ms. Yana Cornish

Technical Director
SEC 145
(319) 273-3064
yana.cornish@uni.edu

APPENDIX E

COMPENSATION FORM

Renaissance Partnership for Improving Teacher Quality Title II Grant

University of Northern Iowa

Thank you for participating in the Renaissance Teacher Work Sample scoring session in Cedar Falls on May 1, 2004. Your interest, contribution, and expertise are greatly appreciated. You will receive a $200.00 stipend for scoring and being a part of the research on the scoring process for University of Northern Iowa's Teacher Work Samples.

Rater ID _____

Please complete the following:

Name _____
First Middle Last

Social Security Number _____

Home Address _____

_____

Community and School Name _____

School or Home Phone Number _____

Grade/Subject Area _____

Email Address _____

Best practice is defined through student learning.

Dr. William Callahan, Interim Dean, College of Education • Dr. Victoria Robinson, Project Coordinator
Schindler Education Center 516 • Cedar Falls, Iowa 50614-0604 • Phone: 319-273-3070 • Fax: 319-273-5175 • http://fp.uni.edu/itq

APPENDIX F

CERTIFICATE OF APPRECIATION