An Examination of the Relationship Between Depth of Student Learning and National Board Certification Status

Article · January 2005

CITATIONS
16

READS
183

4 authors, including:

Tracy Smith
Appalachian State University
12 PUBLICATIONS **249** CITATIONS
SEE PROFILE

Susan Colby
Appalachian State University
13 PUBLICATIONS **132** CITATIONS
SEE PROFILE

Jianjun Wang
California State University, Bakersfield
68 PUBLICATIONS **318** CITATIONS
SEE PROFILE

An Examination of the Relationship Between Depth of Student Learning and National Board Certification Status

Tracy W. Smith, Ph.D.
Appalachian State University

Belita Gordon, Ph.D.
Test Scoring and Reporting, University of Georgia

Susan A. Colby, Ed.D.
Appalachian State University

Jianjun Wang, Ph.D.
California State University, Bakersfield

Office for Research on Teaching
Appalachian State University

April 2005

This project is funded in part with grants from the U.S. Department of Education and the National Science Foundation. Through September 2005, NBPTS has been appropriated federal funds of $149.1 million, of which $136.7 million was expended. Such amount represents approximately 34 percent of the National Board Certification project. Approximately $261 million (66 percent) of the project's cost was financed by non-federal sources.
The contents of this publication were developed under a grant from the U.S. Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal government.

ABSTRACT

The purpose of this study was to examine the impact of National Board Certified Teachers (NBCTs) on student achievement (depth of student learning), compared to teachers who attempted, but did not receive National Board Certification. Participants were recruited from across the United States in four certificate areas. A total of 64 teachers from 17 states participated in the study. Thirty-five (55%) of the participants had achieved National Board Certification, and 29 (45%) had attempted but had not achieved National Board Certification. The overall findings from this study indicated that the relationship between student learning outcomes and teacher certification status was highly statistically significant on six of the seven student outcomes measures in favor of the NBCTs. The comparative teaching practices dimension of the study, also statistically significant, suggested that NBCTs fostered deeper understanding in their instructional design and classroom assignments.

Table of Contents

List of Tables

List of Figures

ABOUT THE AUTHORS

Tracy W. Smith, Ph.D., Principal Investigator, Executive Project Director, and English Language Arts Specialist

Dr. Smith is an assistant professor of Curriculum and Instruction at Appalachian State University. She received her doctorate in Curriculum and Teaching and was the 2000 recipient of the American Association for Colleges of Teacher Education Dissertation Award. At Appalachian State University, she received the Outstanding Faculty (student recognition) and College of Education Teaching (peer recognition) Awards in 2005.

Dr. Smith has taught English/language arts at the middle and high school levels and has been a school district coordinator for gifted education and middle grades education. She has been a consultant in the areas of writing instruction, gifted education, teacher leadership, and program evaluation. Her primary research interests include assessment of teacher quality, specialized middle level teacher preparation, and young adolescent advocacy.

Belita Gordon, Ph.D., Senior Researcher and Performance Assessment Specialist

Dr. Gordon has directed Georgia's statewide writing assessment program for 15 years. Her work has included the development of assessments at four grade levels, the creation of rater training and monitoring programs, and the publication of both instructional materials and research articles. The writing program is housed at the University of Georgia, where she serves as a Public Service Associate. She led the development of Gwinnett County's science and social studies high school graduation (Gateway) performance assessments, a ten-year project. She has assisted five other states as they developed large-scale writing assessment programs. Her work with the National Board includes commissioned papers for the initial Technical Analysis Group, observations of training for three certificate areas used to refine rater training for these and other

certificates, and scoring validation studies for four certificate areas. Her primary research interests are prompt variables, score resolution methods, and the influence of gender on the quality of student writing.

Susan A. Colby, Ed.D., Senior Researcher and Elementary Education Specialist

Dr. Colby is an assistant professor and elementary education program coordinator at Appalachian State University. Dr. Colby received her doctorate in Educational Leadership and was the recipient of the 2002 Edgar Morphet Dissertation Award from the National Council of Professors of Educational Administration. In 2004, she received the Outstanding Faculty Award at Appalachian State University. Her research interests include teacher development and teacher evaluation for preservice and practicing teachers.

Dr. Colby has served in the public schools as an elementary school principal, Title I Coordinator, clinical professor, and teacher. During her tenure in the public schools, she designed programs for the professional development of teachers; coordinated district efforts to develop performance-based assessments; taught courses in literacy and assessment; and wrote curricula for K-5 reading.

Jianjun Wang, Ph.D., Chief Psychometrician

Dr. Wang is a professor of educational research at California State University, Bakersfield. He has been a recipient of research grant awards from the Spencer Foundation, the U.S. Department of Education, and the AERA Research Grants Program. Besides a Ph.D. in education, he holds two MS degrees, one in physics and one in statistics. He also passed the second Ph.D. Qualifying Examination in statistics, and served on the Editorial Board of the *Journal of Modern Applied Statistical Methods*. In 2000, he was selected as a fellow of the

National Center for Education Statistics. Dr. Wang participated in multiple phases of this investigation, including the population definition, sample design, data analysis, and result reporting.

ACKNOWLEDGEMENTS

Such a comprehensive study would never be possible without the collaborative and creative efforts of many people. The authors want to acknowledge and commend the people who made this study possible. At each stage of this two-and-a-half year project, committed educational professionals generously gave their time and energy to ensure the completion of each component of the study. In an effort to recognize their contributions, we wish to thank the following persons.

Through his previous research examining the relationships between teacher expertise and student learning, Dr. John A. Hattie, University of Auckland, provided knowledge and inspiration for this project. He also supported the principal investigator in her understanding of the SOLO Taxonomy and offered advice and guidance in the research design and instrument development.

Tess Riedl, Program Assistant to the study, coordinated all the day-to-day administrative details of the grant. Her roles included correspondence with consultants, content specialists, teacher participants, trainers, scorers, and telephone recruiters; keeping the grant budget; preparing the mailings of the pilot and operational phases of the study; and maintaining databases and web pages.

Dr. Michael S. Hale, curriculum consultant for Vital Source Technologies, Inc., was Senior Researcher for the design and pilot phases of this study. He worked with all research team members in the development and pilot testing of the data collection instruments and procedures. He also served as a reviewer of the final report manuscript.

Dr. J. Steve Oliver, Professor of Science Education at the University of Georgia, was a prominent presence throughout the study. He provided expertise in science education, served as

our AYA/SCI content specialist, assisted with benchmarking and scorer training, and reviewed our final report documents. He and his research colleagues at the University of Georgia (Peg Graham and Nicholas Oppong) also provided feedback about the design of the study. Their important questions concerning the SOLO Taxonomy and rubrics kept us focused on critical issues of the study design, instrumentation, and audiences of interest.

National Board Certified Teachers Vickie Morefield and Linda Taylor provided support and expertise at multiple points during the study. Both are experienced classroom teachers and teacher-researchers. In this study, they coordinated benchmark identification and revision of scorer training materials. They were always willing to do whatever was needed to help with the design and implementation of the study.

Scarlet Davis, NBCT and doctoral student at Appalachian State University, assisted with the background research for the pilot phase of the study. She also developed the initial drafts of the participant eligibility website and on-line questionnaire.

Jody Holleman, doctoral student at Appalachian State University, assisted with research for preliminary literature reviews and provided feedback on materials and forms for the pilot study.

Jim Moody, former graduate assistant at Appalachian State University, provided technology expertise and continued revisions of our recruitment webpage and survey. In addition, he compiled demographic and statistical data for the pilot study.

David Lussier, NBCT, Advisor to the President, and Director of Research at the National Board for Professional Teaching Standards, coordinated and facilitated our access to NBPTS candidate data, while maintaining rigorous standards of confidentiality and professionalism.

Dr. Barbara Blackburn, Winthrop University, assisted in the identification and coordination of a team of telephone recruitment callers. She also assisted in the telephone recruitment herself because she recognized the need for expeditious recruitment.

Pádraig Michael Mann assisted by tallying and compiling data. He also spent countless hours photocopying case studies, organizing materials for the scoring teams, and transporting materials to the scoring site.

Appalachian State University staff personnel, Mary Beth McGee, Justin Cervero, Bryan Johnson, and Terry McClannon, all technology specialists, provided information and instructions to solve technology issues. Justin Cervero and Bryan Johnson also added valuable assistance to the study by developing web-based databases and other innovative tools we needed to collect and analyze data. They provided a method to obtain real time data from scorers working around the region. They also taught our scoring trainers how to use the custom data entry tools as well as how to monitor the progress made in scoring. Following the completion of data collection, they developed tools to aggregate and present the collected data for analysis.

Steve Cramer, Candace Langford, Jeremy Granade, and Karla Lefevre in the Writing Assessment Center of Test Scoring and Reporting Services at the University of Georgia provided critical analysis and expertise in their review of the writing rubrics and scoring of the writing samples.

Dr. Robert Johnson from the Educational Psychology Department of the University of South Carolina, used his breadth of expertise and experience (from the public school level to the development of language arts performance assessments) to analyze the writing assessment outcome data. The information he provided contributed to our plan for future in-depth study of the writing samples produced in this study.

Our telephone protocol recruitment team consisted of teachers from North Carolina and South Carolina who spent evenings on the phone contacting those teachers who responded to our survey and recruiting them to further participate in our study. Their commitment to the study was evident as they endured late nights, wrong numbers, and the frustration of not being able to reach people.

The scoring team consisted of teachers from North Carolina, Virginia, and Georgia, most of whom were National Board Certified Teachers or doctoral students. These outstanding professionals came to Boone, NC or Athens, GA, for a 2-5 day training session, and then scored the participants' data during summer vacation time in order to provide evaluation materials in a timely manner. Some of them also assisted with the training of the scorers, development of training materials, and identification of benchmark cases.

Finally, we wish to thank the classroom teachers who participated in the study. In order to preserve the confidentiality of the participants as well as their students, the teachers are unnamed in this report. More than anyone else, however, their participation and involvement in the study made this work possible.

EXECUTIVE SUMMARY

The number of National Board Certified Teachers (NBCTs) has increased dramatically within the last few years. Currently, over 40,000 teachers have achieved NBCT status. As of December 2004, all 50 states and approximately 544 school districts had implemented policies and regulations to recruit, reward, and retain NBCTs (http://www.nbpts.org/about/state.cfm). With those incentives have come calls for evidence that those teachers certified by the National Board for Professional Teaching Standards (NBPTS) are truly at the top of their profession. More specifically, there has been a call for evidence that students taught by NBCTs benefit from superior instruction.

The purpose of this study was to examine the impact of NBCTs on student achievement (depth of student learning), compared to teachers who attempted, but did not receive National Board Certification. Participants were recruited from across the United States in four certificate areas. A total of 64 teachers from 17 states participated in the study. Thirty-five (55%) of the participants had achieved National Board Certification, and 29 (45%) had attempted but had not achieved National Board certification. The findings provided information about the impact National Board Certified Teachers are having on the depth of student learning in classrooms across the country.

Two major research questions were addressed in this study:

- *Comparative Teaching Outcomes*. Do students taught by National Board Certified teachers produce deeper responses (to class assignments and standardized writing assessments) than students of teachers who attempted National Board Certification but were not Certified?

- *Comparative Teaching Practices*. Do National Board Certified teachers develop instruction and structure class assignments designed to produce deeper responses than teachers who attempted National Board Certification but were not Certified?

In regards to the first research question, student work samples collected in the context of teacher participants' regularly planned curriculum were collected and analyzed. These work samples included the responses of six randomly-selected students on all work produced during the course of the unit. The preponderance of evidence from the six students' work was used as a summary representation of the student outcomes in the teacher's class. The depth of student outcomes was scored separately from the depth of teachers' instructional aims and design. Based on the evaluation of the student work samples, the findings suggested that the student outcomes in most of the teachers' classrooms, regardless of certification status, were at the surface level (78%). However, students of NBCTs were almost twice as likely to achieve deeper learning outcomes (Certified: 29%; Non-Certified: 14%). While the difference between the student work samples of NBCTs and those who had attempted but did not achieve certification was not statistically significant, the descriptive analysis of this data has implications for understanding the complexity of the teaching and learning relationship.

As an additional source of evidence related to students' learning outcomes, a standardized writing assessment was administered to 377 students of teachers in the MC/Gen and EA/ELA certificate areas. Each writing sample was scored holistically and analytically. The analytic writing features included controlling idea, organizational structure, elaboration, voice, and sentence formation. In each of the six analyses (holistic and five writing features), results were statistically significant in favor of the NBCTs. In other words, the students of NBCTs outperformed the students of their non-Certified counterparts in all areas of writing assessed.

In regards to the second research question, teachers' instructional aims were assessed through qualitative and quantitative analyses of work samples submitted based on a unit of instruction. Specifically, data were gathered from stated instructional goals, descriptions of

related unit lessons, and copies of instructional materials and student responses to assignments. The findings indicated that a majority of the teachers (64%) aimed instruction and assignments toward surface learning outcomes. However, the NBCTs were more than two times as likely to aim instruction at deeper learning outcomes (Certified: 49%, or 17 of 35; Non-Certified: 21%, 6 of 29). There was a statistically significant difference between the aims related to the depth of student learning of NBCTs and those who had attempted, but did not receive certification. NBCTs more often intended to foster deeper student understanding.

The overall findings from this study indicated that the relationship between student learning outcomes and teacher certification status was highly statistically significant on six of the seven student outcomes measures. The comparative teaching practices dimension of the study, also statistically significant, suggested that NBCTs fostered deeper understanding in their instructional design and classroom assignments. This study contributes to the growing body of evidence that the relationship between the quality of instruction and the status of National Board certification (Certified vs. non-Certified) is indeed significant.

INTRODUCTION

With the bipartisan passage of the *No Child Left Behind Act of 2001* (NCLB), quality teaching became a political issue. The title – No Child Left Behind – implies that the spirit of the law is to guarantee success for every public school child in America. NCLB states that "Every child deserves highly qualified teachers," and it requires "states to have a highly qualified teacher in every public school classroom by the end of the 2005-2006 school year" (http://www.whitehouse.gov/infocus/education/teachers/quality_teachers.html). Success for every child in America is a noble, moral, straightforward goal; however, the means to accomplish this goal are not nearly so simple or clear-cut. So, while most would agree with the spirit of the law, it remains to be seen if we are willing to allocate the human and financial resources to make this goal a reality for all of America's children.

One clear message in the language of NCLB is that teacher quality is critical to the academic success of students. Public attention and national discussion about teacher quality have intensified as each state has struggled with the decision of how to define "highly qualified teacher." Some states and policymakers have opted for efficient, economical, expeditious methods for identifying, classifying, and multiplying a highly-qualified teaching force. Such solutions often embrace the importance of content knowledge, while minimizing the importance of pedagogical knowledge and skills.

No one would argue that every child deserves a quality teacher, and NCLB acknowledges that in an era of increasing standards and accountability in education, teacher quality will be more important than ever. One early model for identifying quality teachers was the National Board for Professional Teaching Standards. The National Board for Professional Teaching Standards was created in 1987 after the Carnegie Forum on Education and the Economy's Task

Force on Teaching as a Profession released *A Nation Prepared: Teachers for the 21st Century* (May 16, 1986).

The Carnegie Task Force report, *A Nation Prepared,* proposed that the key to the success of the American educational system was to create a profession equal to the task of preparing students for a changing American and global society. This new profession of well-educated teachers would be prepared to assume new powers and responsibilities to redesign schools for the future. The Task Force urged the teaching profession to set the standards and certify teachers who meet those standards and called for the formation of the National Board for Professional Teaching Standards.

For over a decade now, the National Board for Professional Teaching Standards has implemented its system of advanced certification to identify accomplished teachers. While some have been critical of the National Board's system of advanced certification because of its costs, many believe this system holds promise because it acknowledges the complexity of quality teaching. Many recent approaches to identify highly qualified teachers have focused primarily (and sometimes exclusively) on a teacher's subject matter knowledge. In contrast, Berry (2004) explains that the "National Board's assessments not only measure teachers' content knowledge, but also stipulate that candidates compile several samples of student work from different points in the school year and then require them to explain how they assessed this work, how they developed specific interventions, and how they documented student improvements in subsequent assignments" (pp. 9-10).

While the National Board's system for identifying quality teachers has achieved some recognition as a viable model for identifying quality teachers, many have questioned whether the

assessment system distinguishes teachers who are more accomplished at improving student learning.

The Purpose of this Study

The purpose of this study was to examine the impact of National Board Certified Teachers (NBCTs) on student achievement (depth of student learning), relative to teachers who had not gained National Board Certification. The findings provide information about the impact National Board Certified Teachers are having on the depth of student learning in classrooms across the country.

While the main focus of this study was to assess the validity of certification decisions of the National Board for Professional Teaching Standards, the authors anticipate that the design, instrumentation, and methodologies employed here will contribute to the research literature about the relationship between teaching and learning. We hope that the information contained herein will be not be limited in its application to only distinguishing teachers who are already the best at fostering deeper student learning. Rather, we hope that the model and empirical data will be used to provide the support all teachers need to foster deeper student learning outcomes.

Overview

The present investigation uses data collected during the course of teacher-participants regular classroom instruction as well as a standardized measure of writing to evaluate and compare the teaching practices and teaching outcomes of National Board Certified Teachers (NBCTs) to those of teachers who have attempted but did not achieve certification. More specifically, this study addresses the following questions:

- *Comparative Teaching Outcomes.* Do students taught by National Board Certified teachers produce deeper responses (to class assignments and standardized writing assessments) than students of teachers who attempted National Board Certification but were not Certified?

- *Comparative Teaching Practices.* Do National Board Certified teachers develop instruction and class assignments designed to produce deeper student responses than teachers who attempted National Board Certification but were not Certified?

Sixty-four teachers from 17 different states participated in this study. All participants had attempted Certification in one of four certificate areas: Middle Childhood/Generalist, Early Adolescence/English Language Arts, Adolescence Young Adulthood/Science, or Adolescence Young Adulthood/Social Studies-History. Thirty-five (55%) of the participants had achieved National Board Certification, and 29 (45%) had attempted but had not achieved National Board Certification.

Outline of the Report

In the remaining chapters of this report, we will address the following:

Chapter 2. *Review of the Relevant Literature*

Chapter 3. *Developing a Model for Examining the Relationship Between Depth of Student Learning and Teacher National Board Certification Status*

Chapter 4. *Methodology*

Chapter 5. *Results*

Chapter 6. *Discussion and Conclusions*

REVIEW OF THE RELEVANT LITERATURE

Most people can identify the best teacher they have ever had. If we ask each person to identify that teacher and explain why that teacher was good, even the best, we would likely get a description of that teacher's qualities, perhaps compared to other teachers: "That teacher cared about me personally" or "That teacher made me want to learn." And if we asked more and more people, we would likely have a variety of qualities of good teachers. Among these responses, it is doubtful that many people would say, "That teacher helped me obtain a good score on my standardized test." Yet, when we consider how teachers are evaluated by policy-makers and the general public collectively, standardized test scores seem to be the factor that is valued most.

Students and parents intuitively know what scholars are now able to prove (Sanders & Horn, 1998), the effects of a good teacher are long-lasting, and, conversely, so are the effects of a poor teacher. Numerous authors and researchers have presented compelling evidence that teacher effectiveness is related to student learning (Cohen, 2003; Darling-Hammond, 2000, 2003; National Commission on Teaching and America's Future [NCTAF], 2003; Stronge & Hindman, 2003; Wilson, Floden, & Ferrini-Mundy, 2001). While curriculum, class size, funding, family and community involvement, and many other factors contribute to school improvement and student achievement (Cawelti, 1999), many experts have concluded that the single most influential school-based factor contributing to school improvement and student achievement is the teacher (Darling-Hammond & Loewenberg-Ball, 1997; NCTAF, 1996, 2003; Sanders & Horn, 1998; Stronge & Tucker, 2000). The NCTAF report stated, "The bipartisan passage of the *No Child Left Behind Act of 2001* was a clear expression of national will. Recognizing that every American family deserves public schools that work, *No Child Left Behind* pledges highly

qualified teachers in every classroom by the 2005-06 school year" (p. 4). But, perhaps before we can fulfill this pledge, we must also agree on what a highly-qualified teacher is.

This chapter provides a review of the relevant literature related to the current study focusing on teacher expertise and National Board Certification. Five main topics are discussed in an attempt to provide the necessary research context for this study. The topics presented in this chapter include expertise in teaching, the relationship between teacher quality and student learning, measuring quality teaching by examining student learning, surface and deep learning, and NBPTS as a model for accomplished teaching.

Expertise in Teaching

Although it is increasingly clear that the nation has reached consensus that high quality teaching is the most valuable resource a community can provide to its young people, what is not as clear is how to define and assess high quality teaching. Models of expert teaching now exist, as do promising programs that help to identify quality teachers. One such program, the National Board for Professional Teaching Standards (NBPTS), acknowledges the idea of expertise in teaching and certifies teachers who demonstrate that they are accomplished teachers.

In recent years, several researchers have developed models of expert teaching and teachers (Berliner, 2004; Hattie, 2002; Hattie, Clinton, Thompson, & Schmitt-Davis, 1996; Smith, 2004; Stronge, 2002). While many current models identify highly qualified teachers based primarily on the assessment of content knowledge or a review of students' test scores, the models described here examine a wealth of attributes related to quality teaching. Most often, these attributes emerge from empirically based research studies, meta-analyses, and reviews of research and literature. The development of these models of teacher expertise indicate an interest in defining expert teaching and identifying expert teachers. Additionally, such models can inform

educators and policy makers about important skills and dispositions to foster in teacher preparation and professional development. While each model uses distinct language to describe the attributes of expert teaching, taken together, these models inform the emerging definition of quality teachers.

Berliner (2004) examined research-based propositions about the nature of expertise in teaching, specifically, expertise in pedagogy. By integrating research on expert teaching with understandings of expertise in fields outside of education, Berliner discussed the nature of expertise, outlined six policy implications, and described theories that contribute to the knowledge base on expert teachers.

Based on a synthesis of meta-analyses related to student outcomes and an extensive review of the literature related to domain-specific expertise, Hattie et al. (1996) identified a set of four major attributes and eighteen specific dimensions of teaching that can be used to discriminate between expert and novice, or expert and experienced teachers. The four major attributes included (a) extensive, accessible content knowledge; (b) pedagogical knowledge that transforms essential aspects of the subject matter to connect with students' ways of understanding; (c) affective attributes including a respect for learners and a passion for teaching; and, (d) attention to student outcomes including motivation, self-efficacy, challenge, and achievement gains. The Hattie et al. model was empirically tested in a validity study of the NBPTS assessment system (Bond, Smith, Baker, & Hattie, 2000).

Smith (2004) analyzed and interpreted three expert teachers' behaviors and verbal responses to examine the common attributes of these expert teachers. After examining the multiple data sources including surveys, transcripts of lessons, researcher notes, e-mail correspondence and teacher and student interviews, Smith proposed that a summary

representation of expertise existed by identifying six central tendencies that these teachers shared. Expert teachers had a sense of confidence in themselves and in their profession, demonstrated through practice their emphasis on classrooms as communities, emphasized the importance of developing relationships with students, demonstrated a student-centered approach to instruction, made contributions to the teaching profession through leadership and service, and showed evidence of mastery in their content areas.

Stronge (2002) developed a list of qualities of effective teachers based on research results across several decades. While Stronge outlined a comprehensive set of categories that encompassed many characteristics of effective teachers, he succinctly summarized these categories into three overarching attributes: The effective teacher recognizes complexity, communicates clearly and serves conscientiously. To illustrate these attributes, Stronge stated that effective teachers understand the intricacies involved in the teaching and learning process; successfully navigate the complexity of classroom life; clearly articulate expectations, encouragement, and content knowledge; and are dedicated to students, the profession and to their own learning.

One characteristic of expert teaching that was common to each model was a focus on student learning. Expert teachers used a student-centered instructional approach, employed flexible and diverse strategies, monitored student performance consistently, and understood that pedagogical expertise was situated in an understanding of students as individuals and learners (Berliner, 2004, Hattie et al., 1996, Smith, 2004; Stronge, 2002). The NBPTS Standards also reflect this focus on student learning. Teachers seeking certification must show evidence that they are committed to students and their learning and are responsible for managing and monitoring student learning. The NBPTS standards emphasize that accomplished teachers are

conscious of their learning expectations for students and what strategies they will use to help students accomplish these goals. NBCTs demonstrate their ability to set appropriate goals for student learning, articulate the connections between the goals and experience of students, analyze classroom interactions, examine student progress, and reflect on their practice (http://www.nbpts.org/standards/ncbert.cfm).

A second characteristic of expert teachers that is particularly relevant to the current investigation involves expert teachers' deep understanding of content. While expert teachers have a thorough understanding of domain specific knowledge, they also understand that knowledge is contextually bound (Berliner, 2004; Hattie et al., 1996). Content knowledge and instruction are inextricably related to students' needs and students' learning. According to the National Board, NBPTs possess deep subject knowledge combined with the ability to teach content in ways that help students learn. Certified teachers show evidence that they know the subjects they teach and how to teach those subjects to students. The National Board offers 27 certificate fields to honor the differences in content knowledge and knowledge of student developmental levels required for each field. During NB candidacy, teachers respond to six content specific exercises related to their field. Candidates are expected to have content and pedagogical knowledge across the full range of the selected certificate area.

The models of teaching expertise also emphasize the importance of effective classroom management. Expert teachers are proactive in their response to student behavior and have the ability to move beyond formal, static rules and procedures to a more adaptive approach (Stronge, 2002). In addition, expert teachers maximize student learning through the use of effective management of student behavior. Certainly, increases in student learning are more likely when students are working on learning-related tasks. Based on the same premise of expert teaching,

teachers pursuing NBC are asked to describe how they interact with students, how they create a climate in the classroom conducive to learning, and the how they engage in student learning.

Expert teachers can also be distinguished from experienced and novice teachers based on their professional characteristics. Expert teachers reflect on practice in a manner that facilitates professional growth and increased student learning. In addition, expert teachers make contributions to the teaching profession through leadership and service (Smith, 2004). Both of these qualities are evident in the NB Standards. One of the NBPTS Standards states, "Teachers think systematically about their practice and learn from experience." In addition, NBC aims to "reshape" the teaching profession by certifying expert teachers that make quality contributions to the profession. NBCTs are asked to document their work outside of the classroom to show evidence of facilitating growth for teachers, creating networks with the larger community, and promoting student learning.

Years of experience is also discussed in the models of teaching expertise. While not all experienced teachers are experts (Smith, 2004), expert teachers have crafted their knowledge over a period of five years or more. Experience is a prerequisite for learning domain-specific, contextualized knowledge that contributes to expertise in pedagogy (Berliner, 2004). The NBPTS acknowledges the importance of experience in its assessment design and eligibility requirements. Only experienced teachers (three or more years) are eligible to apply for certification.

Although characteristics of expert teachers can be identified and analyzed independently, it is important to consider the interrelationships among the characteristics (Smith, 2004; Stronge, 2002). Expert teachers demonstrate high levels of proficiency in a variety of areas, and, it is the integration of many attributes that lead to effectiveness in the classroom. This interrelationship is

also recognized in the National Board's performance assessments that measure teachers' performance against "high and rigorous" content-specific standards. The NB asserts that the process required for certification is unique in that "it assesses not only the knowledge teachers possess but also the actual demonstration of their skills and professional judgment applied daily in the classroom" (www.nbpts.org/standards/nbcert.cfm). While the National Board Standards were informed and developed by committees of teachers and other experts, empirical evidence supports the standards articulated and valued by NBPTS (Bond et al., 2000; Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Vandevoort, Amrein-Beardsley, & Berliner, 2004).

The Relationship between Teacher Effectiveness and Student Learning

Berliner (2004) asserted that the positive relationship between expert teaching and students' performance has only recently been supported with empirical evidence. Findings from large-scale research studies connecting teacher effectiveness and student learning indicated that teachers influence student performance (Sanders & Horn, 1998; Wang, Haertel, & Walberg, 1993; Wright, Horn, & Sanders, 1997). Wang et al. (1993) conducted a study to determine the factors that influence learning based on data gathered from 61 research experts, 91 meta-analyses, and 179 chapter and narrative reviews. To identify and estimate the influence of educational, psychological, and social factors on learning over 11,000 relationships were examined. Relationships among variables such as classroom management, metacognitive, cognitive, parental support, student and teacher interaction, school culture, curriculum, state and district policies, and student demographics were among the many variables examined. Three methods of analyses were conducted (content analyses, expert ratings and meta-analyses) to quantify the importance and consistency of variables that influence student learning. Findings indicated that proximal variables (e.g. psychological, instructional, and home environment)

exerted more influence than distal variables (e.g. demographic, policy, and organizational). Wang et al. concluded that the actions of students, teachers, and parents mattered most to student learning; policies had limited effect compared to the day-to-day efforts of the people most involved in students' lives.

From a state study using value-added methodology in Tennessee, Sanders and Horn (1998) found that the major determinant of student progress is the effectiveness of the teacher. In their study, factors with little influence on student progress included race, socioeconomic status, class size and classroom heterogeneity. Also using data from student achievement scores in Tennessee, Wright et al. (1997) conducted thirty separate analyses based on academic gain. After controlling for factors such as heterogeneity, student achievement level, and class size, the results indicated that the teacher (highly significant in all analyses) and the prior achievement level for the student were the most important variables influencing student gain. Effective teachers appeared to be effective with students of all achievement levels, regardless of the level of heterogeneity in their classrooms. Although critics have identified limitations of using value-added methodology (Cochran-Smith, 2004; Hattie & Clinton, 2001; Kupermints, 2003), especially for high-stakes decision making, the results of much of the value-added research do provide evidence that a positive relationship exists between teacher effectiveness and student learning.

Evidence of the long-term interest in research related to the relationship between teachers' instruction and the accomplishment of students is also documented in Floden's work (2002). In his review of the literature on this subject, Floden cited a long line of research studies with often conflicting views on the relationship between teaching and learning. He reported, "Many researchers, dismayed by the ways their findings were used, expressed doubts about the

search for generalizable connections between teaching and learning" (p. 3). The issue of whether a generalizable connection exists between teaching and learning becomes particularly important when used to inform two important national discussions. The first discussion is related to the importance of initial, specialized teacher preparation programs and the relationship between a teacher's preparation and the learning of that teacher's students. Studies have been and are being conducted that are yielding data to support and oppose the student outcome value of teacher preparation (Floden, 2002). Findings from one study suggested there is a significant correlation between teacher quality and certification status. Using data from a 50-state survey of policies, state case study analyses, the 1993-1994 Schools and Staffing Surveys, and the National Assessment of Educational Progress (NAEP), Darling-Hammond (1999) found that teacher credentials such as licensure status and degree in the field to be taught were very significantly and positively correlated with student outcomes. The qualitative and quantitative findings from this study suggested that policy investments in the quality of teacher preparation may be related to student achievement. In contrast, Goldhaber and Brewer (2000) concluded from a study analyzing data from a nationally-representative survey of about 24,000 eighth-grade students that there is little rigorous evidence connecting teacher licensure to student achievement. Based on their mixed results examining the relationship between student achievement in math and science and teacher credentials, they argued that the important policy question of whether imposing more rigorous standards in teacher licensure would lead to student achievement had not yet been definitively answered.

A second national discussion related to the teaching-learning connection involves the advanced certification system of the National Board for Professional Teaching Standards (NBPTS) and the degree to which it represents an important reform opportunity for the teaching

profession. Again, this issue is frequently debated, and NBPTS has both opponents (Finn & Wilcox, 1999; Holland, 2002; Podgursky, 2001; Thirunarayanan, 2004) and advocates (Berliner, 2004, Berry, 2004, Darling-Hammond, 1996; Darling-Hammond & Rustique-Forrester, 1997). The question of whether NBCTs and their non-Certified counterparts can be distinguished from each other based on the quality of their students' learning has become a most compelling question for policy makers, teachers, teacher educators, school administrators, and the business community. As more and more teachers become certified through the National Board process, the demand increases for evidence that students taught by NBCTs benefit from superior instruction.

Measuring Quality Teaching by Examining Student Learning

A major difficulty in the efforts to examine the relationship between teacher quality and student learning is how to measure teacher effectiveness by examining student learning in appropriate, fair and valid ways. In the 2003 Annual Report to Congress informing the public of the state of teacher quality in America, US Secretary of Education Rod Paige acknowledged that research has consistently shown that individual teachers contribute to student achievement. However, he indicated that the identification of teacher effectiveness has been reduced to a single factor: student achievement. Vandevoort, Amrein-Beardsley, & Berliner (2004) suggested that Paige's comments echo a "well-rooted" American view, one in which teachers, unlike other professionals, are publicly scrutinized and often evaluated based solely on the outcomes of those they serve, particularly through the use of standardized achievement tests.

In the educational field, using student test scores for high-stakes decisions related to the evaluation of students, teachers and schools is often viewed as problematic. While many scholars believe that student progress should be a factor in evaluating teacher effectiveness (Mendro,

1998; Millman & H. D. Schalock, 1997; Popham, 1997, 1998; Sanders & Horn, 1998), using student achievement data exclusively, specifically standardized achievement scores, is a controversial subject (Darling-Hammond, 1997, 1998; Millman & H. D. Schalock, 1997; Popham, 1998; Webster, 1995). Critics cite a variety of reasons why we should move beyond relying primarily on student test score data when evaluating student learning. The National Research Council (2001) proposed three limitations of large-scale standardized assessments: the inability of test score data to capture complex knowledge effectively and identify critical differences in students' levels of understanding, the inability of test score data to improve teaching and learning, and the inability of test score data to assess growth over time.

Other experts have advocated for an approach to assessment that honors the use of multiple data sources in evaluating student progress (Linn, 2000) and aligns with quality standards for performance (Thompson, 2001). Hattie and Jaeger (1998) argued for an approach to assessment that acknowledges the interplay between assessment, learning, and feedback:

> …assessment needs to be an integral part of a model of teaching and learning if it is to change from its present status as an adjunct to 'see' if learning has occurred, to a new status of being part of the teaching and learning process" (p. 111).

Darling-Hammond (1997) focused on two questions when examining assessment systems linked to student learning: Does the assessment system really measure the quality of schooling or teaching? and, Does the assessment system improve teaching and learning? Darling-Hammond (1997) stated that accountability policies that focus on higher student standards should ensure that teachers have the knowledge and skills necessary to more effectively understand and demonstrate student learning. This challenge requires improved methods for examining the work

of teaching and schooling and improved strategies for linking student learning to teacher effectiveness in ways that require more than a focus on student test scores.

An alternative to using student test score data as the primary indication of teacher effectiveness is the practice of providing teachers with the skills necessary to demonstrate student learning by gathering, interpreting and analyzing data. In this approach, teachers analyze a variety of data to document and reflect on their own work in relation to student progress. One system, the Oregon Teacher Work Sample Methodology (TWSM), assesses teacher quality in relation to artifacts presented in the course of teaching and learning. These artifacts focus on student learning gains (Airisian, 1997; H. D. Schalock, M. D. Schalock, & Girod, 1997). During the initial stages of the TWSM, concerns were raised related to the lack of standards against which to evaluate teaching practices, quality of the test items as constructed by teachers, the inferences made from the work samples, and the developing methodology (Airisian, 1997; Darling-Hammond, 1998). However, recent findings based on the TWSM are encouraging (Stronge & Tucker, 2000). Denner, Salzman, & Bangert (2001) found that work samples could provide valid and credible evidence linking teacher performance and student learning. In the current study, work samples from 132 practicing and prospective teachers were examined to assess the validity and generalizability of using such samples in assessing teachers' abilities to meet teaching standards and impact student learning. The results suggested initial support for teacher work sample assessment as a way to provide evidence connecting teaching performance to student learning.

While the TWSM may have limitations making it difficult to use for high-stakes decision making, the NBPTS assessments which also gather data from work sampling methods differ significantly in ways that make them more appropriate for high-stakes decisions (Darling-

Hammond, 1998). Darling-Hammond outlined these key aspects: (a) the assessment tasks are substantially standardized, (b) evaluations are based on clear standards of practice developed by expert teachers, (c) the examinations include both on-demand performance tasks and samples derived from teachers' work, (d) the scoring systems are highly developed and have been validated and tested for reliability, and (e) a careful process of standard setting has been validated and tested. The National Board has developed performance-based assessments to measure teaching practice against high and rigorous standards. Consistent with teacher work sample methodology, the NB assessment system asks candidates to offer direct evidence of their work, including student samples, and an analytical reflective commentary. Teachers are required to systematically analyze student work, and the quality of their teaching (http://www.nbpts.org/standards/nbcert.cfm).

Surface and Deep Learning

Assessing expert teaching based on student learning through teacher work samples requires a sophisticated understanding of the quality of student learning. Evidence has shown that teachers can adopt a surface or deep approach to teaching, which has consequential effects on what and how students learn (Boulton-Lewis, Dart & Brownlee, 1995; Boulton-Lewis, Smith, McCrindle, Burnett, & Campbell, 2001). Marton and colleagues analyzed conceptions of learning and formulated two major levels of learning: surface and deep (Marton & Säljö, 1984; Marton, Dall'Alba, & Beaty, 1993). Other researchers have developed similar, sometimes more specific, descriptors to describe levels of understanding (Pask, 1988: Svensson, 1984; Biggs & Collis, 1987, 1991). According to Marton's framework, a surface approach involves minimum engagement with the task, typically a focus on memorization or applying procedures that do not involve reflection, and usually an intention to gain a passing grade. In contrast, a deep approach

to learning involves an intention to understand and impose meaning. The student focuses on relationships between various aspects of the content, formulates hypotheses or beliefs about the structure of the problem or concept, and relates more to obtaining an intrinsic interest in learning and understanding. High-quality learning outcomes are associated with deep approaches whereas low-quality outcomes are associated with surface approaches (Biggs, 1987; Entwistle, 1988, 2001; Harper & Kember, 1989; Marton & Säljö, 1984).

Helpful in the assessment of deep and surface learning is the model created by Biggs and Collis (1979, 1982, 1991) referred to as the SOLO Taxonomy. The SOLO (Structure of the Observed Learning Outcome) system describes the levels of abstraction observed in student response. Biggs and Collis argued that while quantitative aspects of evaluating learning are well understood and applied, qualitative aspects of evaluating student work are less often researched and used in classrooms. The SOLO Taxonomy was designed with the understanding that qualitative evaluation is both feasible and helpful and proceeds in a hierarchy of levels of increasing structural complexity. The taxonomy is structured into five major levels, with transitional responses sometimes identifiable between levels: (1) pre-structural; (2) uni-structural; (3) multi-structural; (4) relational; and (5) extended abstract. These hierarchical levels reflect the quality of learning of a particular episode or task. The levels of the SOLO Taxonomy are described in more detail in Chapter 3 of this report.

Not only has the SOLO Taxonomy developed by Biggs and Collis (1979/1982) been used widely in education (Bouton-Lewis & Gillian, 1995; Boulton-Lewis, Gillian, & Wiliss, 1996; Chick, 1998; Chinn, 2002; Lam & Foong, 1996; McAlpine, 1996; Pegg & Davey, 1989), it has also been used in research studies examining surface and deep outcomes in a variety of contexts such as health sciences (Scholten, Ingrid, Keeves, John, Lawson, & Michael, 2002), counseling

(Burnett, 1999), engineering (Carew & Mitchell, 2002), and program evaluation (Dziuban, Cornett, Moskal, & Gyori, 2000).

Chan, Tsui, Chan, & Hong (2002) conducted a study comparing the application of three taxonomies measuring students' cognitive learning outcomes. In this study, responses from graduate students' term papers were analyzed using Bloom's Taxonomy (Bloom, Engelhart, Furst, Hill, & Drathwohl, 1956), the SOLO Taxonomy (Biggs & Collis, 1982) and a reflective thinking instrument designed to assess the level of critical thinking and reflection in written assignments (Kember, Jones, Loke, McKay, Sinclair, Tse, Webb, Wong, F., Wong, M., & Yeung, 1999). In the Chan et al. study a modified version of the SOLO Taxonomy was used in an effort to reduce ambiguity in scoring and increase assessment reliability by including sub-levels between each of the levels. The findings from this study validated the use of the SOLO Taxonomy in a variety of contexts: those involving a variety of subjects, ability levels and assignments. Findings also suggested that the conceptual ambiguity of SOLO can be improved by adding sub-levels to the scale as averaging rater scores does not accurately reflect the level of a student's cognitive attainment.

In the present study, as well as in the Chan et al. (2002) study, the SOLO Taxonomy was augmented with sub-levels to distinguish between the hierarchical levels. Consistent with the NBPTS evaluation system, this study focused on the depth and quality of the examples provided (http://www.nbpts.org/candidates/guide/1_assmnt.html).

In an attempt to understand the nature of cognitive processes at the highest level of formal thinking, Chick (1998) used the SOLO Taxonomy to examine the stages of mathematical cognition of a mathematics researcher by analyzing the data she collected as a graduate student. Chick stated that while both undergraduate and graduate students operated in a formal mode,

there was a difference between the two levels of formal functioning, formal-1 and formal-2. One significant difference Chick observed was between creating (formal-2) and understanding knowledge (formal-1). A second distinction between the two levels was observed in responses. Chick found it difficult to assess formal-2 cognition in student responses to prompted questions. While both formal-1 and formal-2 modes can produce relational responses on the SOLO Taxonomy, satisfactory performance at the formal-2 level, a criterion qualifying the individual as a "researcher," was evidenced most often with the ability to produce relational responses. Chick concluded that outcomes indicative of formal-2 cognition can be evaluated using the SOLO Taxonomy and that the levels of the taxonomy reflect the worthiness of the results, just as the SOLO Taxonomy in concrete-symbolic and early formal modes has been applied successfully.

Boulton-Lewis, Wilss & Mutch (1996) also used the SOLO Taxonomy to analyze student learning. In their study, the content of written statements from 40 teachers enrolled in a graduate course was categorized by structural organization according to the SOLO model. Their findings indicated that 80% of student responses fit the multistructural level indicating that students need help in structuring the content of their learning to reach a relational or abstract level. They advocated that students be provided opportunities to distinguish between models written at different SOLO levels and that students write and rewrite material individually and in groups until a relational level is met.

Highly regarded in the present study is the ability to distinguish between deep and surface learning. Teacher work samples were analyzed to identify whether NBCTs differed from their non-certified counterparts in their intent and attempts to facilitate depth in student learning.

In addition, two measures of achievement were collected from the students of the teacher-participants to determine depth of student learning. First, all student work associated with the

specified unit was collected from six randomly-selected students in each class and analyzed to determine students' depth of understanding of the unit concepts. Also, students of teachers representing the MC/Gen and EA/ELA certificates responded to a standardized writing assessment. Because writing is frequently viewed and used as a universal measure of student achievement, the research team felt that writing would be an appropriate measure to include in this investigation. Prior to designing the writing assessment, the research team engaged in multiple discussions around the question *What is depth of knowledge of writing*? Interestingly, neither the pedagogical nor the writing assessment literature provided a straightforward answer. Had we wondered how students learn to write and the struggles they encounter, the literature offered many ideas. Had we needed to know who receives the lowest test scores or which genres are hardest so that we could design a "rigorous instrument," this too could have been found in a literature review. Previous application of SOLO to written products was scant. Biggs and Collis (1992) used the SOLO levels to rate short paragraph-length creative writing samples. What the rich composition literature did provide, however, was the lens of rhetorical analysis and discourse acts.

Writers who move beyond surface responses, too often the product of prescriptive formulas, approach and complete any given writing task as a series of rhetorical questions. "What do I know about this subject?" "Is it enough to get me started?" "Where can I learn more?" "What do I know about my audience – their biases and predispositions, their likelihood to respond to logic and/or appeals to emotions?" "What does my audience know about my subject?" "What do I want my readers to know when they finish reading my piece?" "Do I want to affect their feelings – why?" "How can I make this happen?" "Who else has written about this subject?" "How did they treat it?" "Should I do the same?" "Now that I've done some

brainstorming, am I ready to start writing?" "Or, is this one of those days or one of those topics where I just have to start writing before I can back up and analyze?" "What's that word, the *mot juste*, I'm struggling to find so that my reader will know precisely what I mean?" "What can I compare these numbers to so my reader grasps how enormous this problem is?" This series of questions is only an illustration of rhetorical problem-solving, not a complete list. Writing is so much more than the imitation of models and the memorization of rules. The production of a writing sample requires global skills: generating ideas, organizing them, making connections within the text and with the reader, and adopting a stance suited to the subject, the task, and the audience (Atwell, 2002; Fletcher, 2001; Graves, 2000; Larson, 1992; Murray, 2002; Ray, 2001).

NBPTS: A Promising Model of Teaching Expertise

According to the organization's own description, NBC measures a teacher's practice against high and rigorous standards. The NBPTS is anchored in the belief that the single most important action this country can take to improve schools and student learning is to strengthen teaching. The NBC assessment process includes an extensive series of performance-based assessments that includes teaching portfolios, student work samples, videotapes and thorough analyses of the candidates' classroom teaching and student learning. Teachers also complete a series of written exercises that probe the depth of their subject-matter knowledge and their understanding of how to teach those subjects to their students. Teachers who have participated in the National Board Certification process have overwhelmingly stated it is the most powerful professional development experience of their careers. They suggest the experience changes them as professionals and that through the process they deepen their content knowledge and develop, master, and reflect on new approaches to working with their students (Darling-Hammond, 1998; NBPTS, http://www.nbpts.org/about/index.cfm).

The rationale behind using a performance-based system as the foundation for the certification process is stated by NBPTS:

> Teaching is at the heart of education, so one of the most important actions the nation can take to improve education is to strengthen the teaching profession. National Board Certification concentrates education reform in the classroom- where teaching and learning takes place (http://www.nbpts.org/standards/nbcert.cfm, para. 1).

Two important features help to define the National Board Certification philosophy: The certification process is based on high and rigorous standards that articulate what teachers should know and be able to do, and, performance-based assessments are utilized to measure teaching practice in relation to these standards. The standards are developed by committees of teachers and other experts and reflect the core propositions; identify the specific knowledge, skills and dispositions that support accomplished practice; show how a teacher's professional judgment is observable in actions; and describe how the standards come to life in different settings. Reviews are conducted both internally and externally during a public comment period, later to be revised and adopted for publication.

Numerous empirical studies have now been conducted in an attempt to validate the National Board Certification process as a model that accurately identifies superior teaching. Studies have been conducted examining the psychometric quality of the National Board for Professional Teaching Standard's Assessments (Jaeger, 1998; Myford & Engelhard, 2001). Jaeger discussed the findings of a Technical Analysis Group (TAG) established by the National Board with the responsibility for conducting research on the measurement quality of the assessment process. The TAG focused on four areas: validating the Board's assessments, characterizing the reliability of the Board's assessments, establishing standards of performance

for awarding certification, and investigating the presence and degree of adverse impact and bias in the assessments.

In a comprehensive validity study examining certification status in relation to dimensions of teaching expertise, Bond et al. (2000) found that NBCTs out-performed their non- Certified counterparts on 13 dimensions of teaching expertise with 11 of those 13 dimensions being statistically significant. The 13 dimensions of teaching expertise emerged from a comprehensive review of the research and scholarly literature on expert/novice comparisons (Hattie, 2002; Hattie & Clinton, 2001; Hattie, Clinton, Thompson, & Schmitt-Davis, 1996). The data sources were gathered from 65 teachers and included instructional objectives and lesson plans from a unit, observational records from all 65 teachers' classrooms, and scripted interviews of the teachers and their students. The study also examined samples of student classroom work produced in response to teacher-developed assignments as part of the instructional unit. An analysis of student work samples revealed that students taught by NBCTs showed a greater depth of understanding than students of teachers who attempted, but did not gain National Board Certification. Although these findings were important, they provided only a beginning to the research that is needed, with a small sample of teachers in two certificate areas.

Goldhaber and Anthony (2004) used a value-added research design and annual test scores from North Carolina students in grades three, four, and five for three academic years to determine the effect of NBCTs on student achievement. In all, over 600,000 student records in reading and math were linked to teacher records over the same time period yielding pre-test and post-test scores. Their findings indicated that NBCTs appear to be more effective than their non-certified counterparts. Students of NBCTs improved an average of seven percent more on their year-end math and reading tests than students whose teachers attempted but did not gain

certification. This performance differential was more pronounced for younger and lower-income students. The researchers qualified their findings by explaining that the "NBPTS effect" differs significantly when grade level and student type are considered.

Vandevoort et al. (2004) examined the relationship between National Board Certification and student achievement in grades 2-6 in 35 classrooms from 14 Arizona school districts as measured by the Stanford Achievement Test. A comparison was made between the adjusted gain scores of students of NBCTs and those of non-NBCTs. Findings from the achievement test scores indicated that on the average, students of NBCTs made over 1.3 months greater gain per year in reading and 1.4 months greater gain per year in math than the students of non-NBCTs. When generalizing across years and subjects, students of NBCTs averaged over 1.2 months greater gain than students placed with non-NBCTs. In three-fourths of the 48 comparisons the students of NBCTs outperformed their counterparts. Data were also gathered from questionnaires given to NBCTs and surveys completed by their respective principals. Classrooms of the NBCTs appeared to be heterogeneously grouped from both the teachers' and principals' perspectives. Findings from the data collected from the principals' survey indicated that about 85% perceived the NBCTs to be one of their best teachers and 90% believed that NBCTs were contributing to improvement in teacher quality. After considering the findings from a variety of studies including their own study examining the effects of NBCTs on student achievement, Vandevoort et al. concluded, "The preponderance of the evidence suggests that the students of NBCTs achieve more" (p. 36). Further, they contend, "the weight of the current evidence suggests that the NBPTS conducts a certification program that works as intended" (p. 37).

In a recent study examining the relationship between certification status and student learning gains, Cavalluzzo (2004), found that Miami-Dade math teachers who had achieved

National Board Certification helped their students achieve larger gains in testing than those who did not achieve certification. The data were gathered from 9th and 10th grade math students in a large urban school district from the years 1999-2000 to 2002-2003. The findings indicated that students made larger gains if their teacher was National Board Certified, and smaller gains if their teacher did not achieve certification or withdrew from the NBC process. In all, 7 of 9 indicators of teacher quality included in the analyses resulted in a positive, statistically significant gain in student achievement. Cavalluzzo concluded that National Board Certification proved to be an effective signal of teacher quality and a valid discriminator among applicants. Suggested from these findings is that school systems may use NBC to target pay increases to teachers of the highest quality.

Other studies have raised questions about National Board Certification focusing on two distinct issues: adverse impact on candidates from specific groups, and the relationship between expertise in teaching and NBCTs (Goldhaber, Perry, & Anthony, 2003; Ladson-Billings & Darling-Hammond, 2000; Pool, Ellett, Schiavone, & Carey-Lewis, 2001; Stone, 2002). Two recent studies examined the issue of adverse impact with respect to race and gender (Goldhaber, Perry, & Anthony, 2003; Ladson-Billings & Darling-Hammond, 2000). Ladson-Billings and Darling-Hammond examined the practice of successful urban teachers in relation to the NBPTS and Interstate New Teacher Assessment and Support Consortium (INTASC) assessments. Based on a review of the literature, practices of successful urban teachers were identified. Successful urban teachers focus on social relationships, focus on the whole child, understand students' cultural backgrounds, and are able to connect classroom content with student experiences. Further, culturally responsive teachers understand that the curriculum does not always benefit urban, poor children of color and effective teachers make demands for academic student success

for all students rather than lowering expectations for high-risk students. With these notions in mind about successful urban teachers, Ladson-Billings and Darling-Hammond examined the extent to which these aspects of teaching are well-represented and well-measured by NBPTS and INTASC. They concluded that the characteristics of successful urban teachers are not well represented in the current NBPTS EA/ELA assessment and suggested that while changes made to the assessment system in 1996-1997 may improve pass rates, there still may well be an adverse impact for urban teachers of color.

Goldhaber et al. (2003) found similar results. After examining the NBPTS applicant sample in the state of North Carolina for the years 1997-2000, large differences were noted between successful and unsuccessful candidates. Disparities existed between NBPTS certified and non-certified applicants by race and gender with African-American and male teachers less likely to be certified.

Regarding the relationship between expertise in teaching and NBCTs, Pool et al. (2001) examined the variation among the professional practices of six NBCTs using systematic classroom observations, individual teacher interviews, and focus group interviews with administrators and colleagues. Findings from the data revealed that considerable variability exists in the quality of teaching and learning in the daily practices of the six certified teachers. Two teachers were judged as exemplary, two average, and two fairly ineffective. Pool et al. found that teachers who valued the philosophy of the NBPTS maintained a higher quality teaching and learning environment. In contrast, teachers who cited monetary gain as the primary reason for seeking certification demonstrated more difficulties with elements of effective teaching. They concluded that the results of the study call into question monetary incentives for

receiving NBPTS certification and warrant further study of daily teaching practices of certified teachers.

Stone (2002) studied the "NBCT effect" by examining whether 16 of 40 NBCTs in Tennessee were exceptionally effective in producing objectively measured student achievement gains. Data for this study were gathered from the Tennessee Value Added Assessment System (TVAAS) for the year 2000. In this study, exceptional teaching was defined by the state for accountability purposes as teaching that increases student achievement equal to 115% annual gain in three core subjects. The findings indicated that the 16 NBCTs could not be considered exceptionally effective in terms of ability to increase student achievement. None of the teachers met the standard in one or more of the required subjects and all failed to meet the standard for three consecutive years. One of the 16 came close to qualifying with exceptional teaching in two of three areas for two of the three years. Stone concluded that bonuses based on NBPTS Certification should be suspended until it can be established that certification delivers what it promises. Stone also concluded that the certification process is not serving the teacher quality aims of public policy focusing on accountability for student achievement. While critics have questioned the methodology and conclusions (see Vandevoort et al., 2004, for a discussion), Stone raises important issues about certification-dependent teacher bonuses and incentives as well as the student benefits related to NBC.

These studies examining the practices and outcomes of NBCTs contribute to the growing knowledge base regarding the relationship between NBPTS Certification status and student learning. While each study may have individual and even acknowledged limitations, each adds to the empirical research base assessing the validity and even feasibility of the NBPTS Certification system. However, examining and addressing these limitations in future studies is critical as

policy makers, practitioners and educators make decisions related to practice and policy. Despite the concerns and questions that are raised in regards to these studies (Finn & Dunne, 1999; Holland, 2002; Podgursky, 2001a), there now exists a significant body of research that provides evidence that there is a positive relationship between National Board Certification and higher levels of student learning (Bond et al., 2000; Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Vandevoort et al., 2004) .

The present study provides a logical follow-up to the Bond, et al. (2000) study and a complementary study to the line of inquiry related to the relationship between teacher National Board Certification and student achievement. However, this study departs from many of the previous studies in significant ways. The Bond et al. study examined National Board Certified teachers based on a model of expert teaching (Hattie, 2002; Hattie & Clinton, 2001; Hattie et al., 1996). Subsequent studies (Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Vandevoort et al., 2004) examined the impact of NBCTs on student learning by examining test scores due to the current accessibility of test score data and sophisticated processes to control for multiple variables. Although the current study uses performance-based assessments to draw conclusions about student learning, it complements the earlier studies related to student learning based on student test scores (Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Vandevoort et al., 2004) by examining a broader range of content disciplines. For example, it examined instruction and student learning for high school students as well as elementary and middle grades students. The present study is similar to the line of inquiry begun in the Bond et al. study which focused on using performance-based assessments; however, in this design, more certificate areas are being examined, and student learning outcomes are assessed based on the SOLO Taxonomy distinguishing between deep and surface learning. The present study examined the link between

Certification status and student learning by analyzing work samples, an avenue less explored when comparing NBCT with their non-certified counterparts. In this study, student work samples produced in the course of a teacher's regularly planned and implemented instruction were examined to determine the depth of observed student learning outcomes.

In an article discussing authentic assessment of teaching in context, Darling-Hammond and Snyder (2000) discussed the complexities of assessing teachers in a time when teachers need to demonstrate a more sophisticated understanding of the effects of context and learner variability on teaching and learning. Assessment systems such as NBPTS provide a model for authentic assessment of teaching by accounting for issues of context when evaluating teacher performance. Darling-Hammond and Snyder suggested a framework for defining authentic assessment of teaching based on findings from the emerging research related to this topic.

First, assessments sample the actual knowledge, skills, and dispositions desired of teachers as they are used in teaching and learning contexts, rather than relying on more remote proxies. Darling-Hammond and Snyder (2000) offered an analogy about such assessments: "If one wants to assess a performance skill like swimming, for example, it is useful to have the swimmer in the water by some point" (A framework for defining authentic assessment of teaching section, para. 1). In the field of education, even observation of the teacher can be considered somewhat of a remote proxy as one cannot visibly see many aspects of teaching such as planning, work with families, and work with colleagues. Assessment tools such as interviews, teacher reflections and analyses, and other artifacts that represent practice may better meet the needs of authentic assessment. In the current study, this aspect of authentic assessment is evident in the research design because participants submitted descriptions of their student and teaching

contexts, profiles of their lessons, and student work samples. All data were from an authentic teaching situation, a unit in the teacher's regularly planned curriculum.

Darling-Hammond and Snyder (2000) also indicated that effective teacher assessments require the integration of multiple kinds of knowledge and skill as they are used in practice. They suggested that assessments that mirror teaching by seeking to integrate knowledge related to content, assessment, and pedagogy better represent the tasks teachers actually perform. For the current study, teachers submitted work samples from a self-selected teaching unit. To develop this teaching unit, teachers incorporated a variety of understandings about students as well as their expertise in content, pedagogy and assessment. In addition, teacher-participants in two certificate areas submitted student responses to a standardized writing assessment.

In *The Neglected "R": The Need for a Writing Revolution,* by the National Commission on Writing in America's Schools and Colleges, commission chair C. Peter McGrath and vice chair Arlene Ackerman (2003) argued for the educational value of writing:

> Writing extends far beyond mastering grammar and punctuation. The ability to diagram a sentence does not make a good writer. There are many students capable of identifying every part of speech who are barely able to produce a piece of prose. While exercises in descriptive, creative, and narrative writing help develop students' skills, writing is best understood as a complex intellectual activity that requires students to stretch their minds, sharpen their analytical capabilities, and make valid and accurate distinctions. As a nation, we can barely begin to imagine how powerful K-16 education might be if writing were put in its proper focus. Facility with writing opens students up to the pleasure of exercising their minds in ways that drilling on facts, details, and information never will. (pp. 13 - 14)

Writing plays a crucial role in learning. Further, as the commission recognizes, writing assessment must go beyond multiple-choice testing. A standardized writing assessment was added to the collection of student outcomes data for this study partially because of the clearly valuable role that writing plays in learning. It was also added in response to recommendations from the previous validation study by Bond, Smith, Baker, and Hattie (2000). The writing assessment added a constant, i.e., a controlled source of student performances to the varied student works that were produced as part of normal classroom instruction. We followed the "on-demand" writing model used by as many as 43 states to routinely test composition skills. On-demand is a timed writing to an assigned topic (Baldwin, 2004; CCSSO, 2002). The writing samples are evaluated using rubrics that identify well-defined and highly teachable skills (Popham, 2004).

Third, with effective teacher assessments, multiple sources of evidence are collected over time and in diverse contexts. Darling-Hammond and Snyder (2002) indicated that sound decisions take into account evidence based on adequate samples of thinking and behavior and relevant information such as the context for learning, goals for student learning, and information about the students. For this study, teachers were asked to discuss class/student information, their approach to and beliefs about teaching, description of the unit, a description of each lesson, and an overview of the student work samples. This data helped scorers to determine the intentions of the teacher as related to the performance of students.

The fourth aspect of authentic teacher assessment in context is that the evidence is evaluated by individuals with relevant expertise against criteria that matter for performance. Darling-Hammond and Snyder (2000) suggested that evaluators should demonstrate expertise in the subject area they are assessing, and expectations are clearly outlined based on standards for

performance. In this study, the content specialists, trainers, and scorers were experienced classroom teachers. All evaluators had experience in the specified content discipline and most where NBCTs themselves. If the purpose of teaching is to improve and deepen student understanding, certainly the criteria in this study (based on the SOLO Taxonomy) represent criteria that matter for performance in the field.

This review of the literature provides the research base for an examination of comparative teaching outcomes (student performance) and comparative teaching practices (aims and goals of instruction) assessed in relation to the SOLO Taxonomy. Attributes of expert teachers were identified, the relationship between teacher effectiveness and student learning was discussed, issues related to the evaluation of teaching based on student learning were examined, and the SOLO Taxonomy was presented as a method for evaluating deep and surface outcomes for both teaching outcomes and teaching practices. Additionally, NBPTS was highlighted as a certification process for accomplished teachers that acknowledges the benefits of authentic assessment of teachers that has the potential to "transform teaching quality over time" (Berry, 2004, p. 8)

DEVELOPING A MODEL FOR EXAMINING THE RELATIONSHIP BETWEEN DEPTH OF STUDENT LEARNING AND TEACHER NATIONAL BOARD CERTIFICATION STATUS

The present study is based on research related to the complex relationship between teaching and student learning. Chapter 2 provided a traditional review of the literature on research studies and theoretical examinations related to this relationship. Topics in the literature review included expertise in teaching, the relationship between teacher quality and student learning, measuring quality teaching by examining student learning outcomes, surface and deeper learning, and National Board Certification as a promising model of teaching expertise.

This chapter provides more detailed definitions and descriptions of the models and theoretical frameworks that influenced the research design. Beginning with a discussion about limitations of current assessments, this chapter describes a model for developing and evaluating assessments that are sensitive to the ways that students represent knowledge and develop competence. This model, the Assessment Triangle (National Research Council, 2001), provided a framework for the development of the assessments used in the present investigation.

The quality of learning concerns all educators. Most think they know quality when they see it, but most also find it difficult to define, particularly in terms understandable to students. Many recent studies have examined teacher quality by analyzing student performance on standardized test scores (Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Sanders & Horn, 1998; Stone, 2002; Vandevoort, Amrein-Beardsley, & Berliner, 2004; Wright, Horn, & Sanders, 1997) In addition, policymakers often use the data from large-scale assessments to make decisions regarding teacher compensation and performance as well as student progress and promotion (Cavalluzzo, 2004; Cochran-Smith, 2004; Darling-Hammond, 1999). However, the results (individual or aggregated) of such large-scale assessments provide very limited information

about how students organize knowledge and represent information or about how instruction might be changed to improve student learning (Darling-Hammond, 1997). In contrast to previous studies that relied on traditional, large-scale assessments to examine the complexity of the teaching-learning relationship, the present study examined teachers' instructional goals and materials as well as students' responses and work samples to determine the aim (implicit and explicit) of instruction and the depth of the student learning.

Limitations of Current Assessments

The National Research Council (2001) described limitations of many current, large-scale assessments. The first limitation is that many assessments do not capture the kinds of complex knowledge and skills that are emphasized in contemporary standards and deemed essential for success in an information-based economy and world. Many of these assessments, for example, do not examine students' organization of knowledge, their problem representations, their use of strategies, or their self-monitoring skills. Grant Wiggins writes, "The simplest way to sum up the potential harm of our current tests is to say that we are not preparing students for real, 'messy' uses of knowledge in context – the 'doing' of a subject" (1993). In addition, many of these assessments are not useful for improving teaching and learning – a critical goal in education reform. Most current, large-scale tests provide only limited information that educators can use to determine why students do not perform well or to modify the conditions of instruction in ways likely to improve student learning. Often the data generated from these tests indicate only general information about a student's performance relative to his or her peers (e.g., 35th percentile). Sometimes the results indicate that a student has performed poorly in a particular discipline (e.g., below grade level in math). Such results do not reveal if a student is using misguided strategies, advancing toward competence, or persisting with a partial understanding.

The National Research Council (NRC) also suggested that many current assessments provide only "snapshots" of achievement at particular points in time. They do not capture the progression of students' conceptual understanding over time. Even when students' growth is considered and calculated in an assessment model, many times the concepts that are assessed at the testing points are not conceptually congruent. For example, students may be assessed on their understanding of earth science in seventh grade, and they may be assessed on their understanding of physical science in eighth grade. Even though both assessments are labeled "science," they do not measure students' growth in understanding particular concepts. Therefore, true growth in understanding is difficult to infer.

The Assessment Triangle

The NRC report (2001) stated that assessments, no matter what their purpose, share certain common principles. One is that "assessment is always a process of reasoning from evidence" (p. 2). By its very nature then, assessment is imprecise to some degree – depending on the extent of the evidence and the reliability of the judgment. Assessments, therefore, are only estimates of what a person knows and can do. The NRC (2001) outlined three key elements that influence any assessment. These foundational elements, comprising what they introduced as the "assessment triangle," include a model of how students represent knowledge and develop competence; tasks or subjects that allow observation of students' performance; and an interpretation method for drawing inferences from the performance evidence obtained. The assessment triangle, therefore, is depicted with **cognition**, **observation**, and **interpretation** at the three corners. These three elements must be explicitly connected and designed as a coordinated whole. Otherwise, the meaningfulness of the inferences drawn from the assessment will be compromised.

The **cognition** corner of the assessment triangle refers to a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain (e.g., long division). In any particular assessment application, a theory of learning is needed to identify the set of knowledge and skills that is important to measure for the tasks at hand. The NRC explained that it is best when assessments are based on scientifically credible models that account for the typical ways students represent knowledge and develop expertise.

The **observation** corner of the assessment triangle represents a description or set of specifications for assessment tasks that will reveal illuminating responses. Every assessment is based on a set of beliefs about the kinds of tasks or situations that will prompt respondents to say, do, or create something that demonstrates important knowledge and skills.

Finally, the **interpretation** corner encompasses all the methods and tools used to reason from fallible observations. Every assessment is based on certain assumptions and models for interpreting the evidence collected from observations. In the context of large-scale assessment, the interpretation method is usually a statistical model. In the context of classroom assessment, the interpretation is often made less formally by the teacher, and is usually based on an intuitive or qualitative model rather than a statistical one.

The rest of this chapter provides descriptions of the three corners/constructs of the assessment triangle, with particular attention to how they were operationalized for this study.

Application of the Assessment Triangle to This Study

The cognition, observation, and interpretation dimensions of the assessment triangle informed the design of this study. These constructs will be used to describe and provide a rationale for the various decisions that were made in the design and implementation of this study.

Cognition

The NRC recommended that a model of cognition and learning serve as the cornerstone of the assessment-design process and that this model should be based on the best available understanding of how students represent knowledge and develop competence in the domain. The model of cognition that most influenced the design of this study was the cognitive perspective. Though the cognitive perspective is not well-represented in many traditional assessments, it has influenced several recent innovations in the design and use of educational assessments (National Research Council, 2001).

The NRC (2001) provided a summary and overview of the cognitive perspective as follows:

> Cognitive theories focus on how people develop structures of knowledge, including the concepts associated with a subject matter discipline (or domain of knowledge) and procedures for reasoning and solving problems. The field of cognitive psychology has focused on how knowledge is encoded, stored, organized in complex networks, and retrieved, and how different types of internal representations are created as people learn about a domain (NRC, 1999). One major tenet of cognitive theory is that learners actively construct their understanding by trying to connect new information with their prior knowledge.
>
> In cognitive theory, knowing means more than the accumulation of factual information and routine procedures; it means being able to integrate knowledge, skills, and procedures in ways that are useful for interpreting situations and solving problems. Thus, instruction should not emphasize basic information and skills as ends in themselves, but as resources for more meaningful activities.

> …cognitive theory also emphasizes what type of knowledge someone has. An important purpose of assessment is not only to determine what people know, but also to assess how, when, and whether they use what they know. This information is difficult to capture in traditional tests, which typically focus on how many items examinees answer correctly or incorrectly, with no information being provided about how they derive those answers or how well they understand the underlying concepts. Assessment of cognitive structures generally requires more complex tasks that reveal information about thinking patterns, reasoning strategies, and growth in understanding over time (pp. 62-63).

This study examined teacher assignments and assessments to determine whether they were structured to elicit information about students' depth of understanding and ways of knowing. In addition, student work samples were analyzed to determine whether students were, in fact, using what they learned in meaningful ways. For example, the work samples produced by students were analyzed to determine if students were connecting the facts in a unit of study in such a way that they understood the underlying subject domain concepts.

Observation

The observation dimension of the assessment triangle refers to the specifications for assessment tasks that will elicit illuminating responses or data. For this study, the design called for an examination of the relationship between teaching and learning in the context in which both occurred: the teacher's classroom. The teachers' regularly-planned curriculum and instruction and their students' responses were the specific tasks chosen to analyze for this research. The data that were collected, therefore, included teachers' instructions and assignments to students as well as students' responses and resulting work samples. A standardized writing assessment was also

administered to all students in the EA/ELA and MC/Gen classrooms. Detailed descriptions of these instruments will be provided in the Methodology Chapter.

Interpretation

The interpretation corner of the assessment triangle consists of the methods and tools used to draw conclusions about the observations. In the context of this study, interpretation methods and tools incorporated both qualitative and quantitative models. The SOLO Taxonomy was used to evaluate teacher data, including teachers' responses to questions about their practice and instructional design, resources and materials used in instruction, and written and oral instructions given to students. Students' work samples were also evaluated using the SOLO Taxonomy. The writing assessment was evaluated in two ways: the SOLO Taxonomy was used to obtain a holistic score, and an analytic writing features rubric was used to evaluate the writing on five traditional dimensions of writing quality: controlling idea, organizational structure, elaboration, voice, and sentence formation.

The SOLO rubric was applied by expert teachers and content specialists in each discipline. The scoring procedures are discussed in more detail in the Methodology Chapter.

SOLO Taxonomy. Consistent with the cognitive perspective, Biggs and Collis (1982) posited that the evaluation of thought, from childhood to adulthood, gives an important clue about quality. That clue is structural organization, which discriminates well-learned from poorly-learned material in a way not unlike that in which mature thought is distinguishable from immature thought.

The research and subsequent models of Biggs and Collis (1982, 1991) were influenced by the validated conceptions of learning styles research begun in Sweden in the mid-1970s by Ference Marton and Roger Säljö (1976). These researchers analyzed conceptions of learning and

formulated two major levels of learning: surface and deep (Marton, Dall'Alba, & Beaty, 1993; Marton & Säljö, 1976, 1984). A surface approach involved minimum engagement with the task, typically a focus on memorization or applying procedures that did not involve reflection, and usually an intention to gain a passing grade. In contrast, a deep approach to learning involved an intention to understand and impose meaning. When a deep approach was applied, the student focused on relationships between various aspects of the content, formulated hypotheses or beliefs about the structure of the problem, and related more to obtaining an intrinsic interest in learning and understanding. High-quality learning outcomes were associated with deep approaches whereas low-quality outcomes were associated with surface approaches (see Biggs, 1987; Entwistle, 1988, 2001; Harper & Kember, 1989; Marton & Säljö, 1984). Several researchers have determined that teachers can also adopt a surface or deep approach to teaching, which has consequential effects on what and how students learn (Boulton-Lewis, 1994; Boulton-Lewis, Dart & Brownlee, 1995; Boulton-Lewis, Smith, McCrindle, Burnett, & Campbell, 2001; Boulton-Lewis, Wilss, & Mutch, 1996; Campbell, Smith, Boulton-Lewis, Brownlee, Burnett, Carrington, & Purdie, 2001).

After studying the organization of responses from hundreds of students, from elementary through high school and college levels, in such subjects as history, mathematics, creative writing, reading, geography, and foreign languages, Biggs and Collis (1982) determined that a similar structure emerged in all cases. This structure of the observed learning outcome forms the basis of the SOLO Taxonomy, which may be applied to evaluate learning quality in a wide variety of school and college situations, in most subject areas. There have been a few previous attempts to evaluate quality, the most notable being the Bloom Taxonomy (1956). That taxonomy, however, has been used mostly to develop questions and items, not to evaluate open-ended responses to

existing questions and item types. The SOLO Taxonomy can be used to assess quality retrospectively in an objective and systematic way and is understandable by both teacher and student. It may be used as an instructional as well as an evaluative tool (Biggs & Collis, 1991; Hattie & Purdie, 1998).

The SOLO Taxonomy was derived from a study of outcomes in a variety of academic content areas and can be used to evaluate the quality of student responses. Since it was introduced in the early 1980s, the Taxonomy has been widely-used and modified for educational practice and research purposes. Levins (1995) suggested that "the newer developments have allowed for a greater utilisation of the Taxonomy in different educational practice and research environments, while, at the same time, not negating the initial formulation" (p. 1).

Biggs and Collis hypothesized that as the depth of student learning increases, the work students produce as evidence of their learning displays similar stages of increasing structural complexity. In their research related to student outcomes in a variety of content disciplines, student responses became quantitatively different first, as the amount of detail in the students' responses increased. Next, responses became qualitatively different as detail became integrated into a structural pattern.

Biggs and Collis (1991) explained that, in their progression from incompetence to expertise, learners displayed a consistent sequence, or learning cycle. This cycle is repeated at each mode of representation. The SOLO Taxonomy includes the five basic levels: prestructural, unistructural, multistructural, relational, and extended abstract. While traditional Piagetian theory asserts that cognitive development proceeds in discrete stages, with uneven performance across stages viewed as rare, the neo-Piagetian model of Biggs and Collis is more context-dependent and provides room for individual variation across levels relative to subject domains and tasks.

For this reason, the stages in the SOLO Taxonomy can be viewed as a continuum from pre-understanding to surface understanding to deeper[1] understanding. The continuum in Table 3.1 suggests that a surface understanding often precedes a deeper understanding.

Table 3.1 *The SOLO Taxonomy Level Continuum*

Pre-Understanding	**Prestructural**	The task is engaged, but the learner is distracted or misled by an irrelevant aspect or detail.
Surface	**Unistructural**	The learner focuses on the relevant domain and picks up one aspect to attend to.
	Multistructural	The learner picks up more and more relevant or correct features, but does not integrate them.
Deeper	**Relational**	The learner now integrates the parts with each other, so that the whole has a coherent structure and meaning.
	Extended Abstract	That coherent whole is generalized to a higher level of abstraction. The learner now generalizes the structure to take in new and more abstract features, representing a new and higher mode of operation.

In his most recent book, Biggs (1999) represents the SOLO Taxonomy graphically this way:

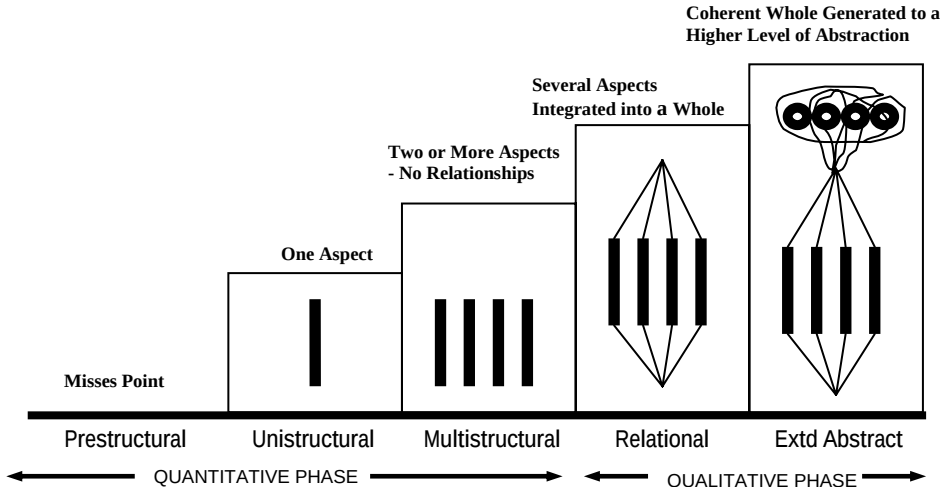

Figure 3.1 Graphic Representation of the SOLO Taxonomy

[1] While Biggs and Collis use the two categories of *surface* and *deep* to distinguish two levels of depth of learning, we prefer *surface* and *deeper* because this language more clearly indicates a learning cycle, or continuum.

The first level of the Taxonomy represents a lack of understanding. At this level, the response indicates that learning is inappropriate or irrelevant to the task in question. The next two levels (unistructural and multistructural) of the taxonomy correspond to surface learning, and the latter two (relational and extended abstract) correspond to deeper learning.

Expert teachers are more likely to lead students to deeper learning rather than surface learning (Hattie, 1998). These teachers structure lessons to allow the opportunity for deeper processing, set tasks that encourage the development of deeper processing, and provide feedback and challenge for students to attain deeper processing. A further advantage and unique distinction of the SOLO model is that it can be used to reliably code classroom lessons and assignments as well as the resulting student work produced as a result of those assignments (Hattie & Purdie, 1998).

To establish whether teacher assignments and student outcomes in each case were surface or deeper, teams of accomplished teachers were trained to use the Structure of the Observed Learning Outcome (SOLO) Taxonomy related to their specific content area. Descriptions of these evaluation procedures are included in the Methodology Chapter of this report.

Marzano's New Taxonomy of Educational Objectives. Although the SOLO Taxonomy provided the conceptual framework for evaluating teacher tasks and student work samples, the work of Robert Marzano and his colleagues (Marzano 1992, 2001; Marzano, Brandt, Hughes, Jones, Presseisen, & Rankin, 1998; Marzano, Pickering, Arredondo, Blackburn, Brandt, & Moffett, 1997) on the dimensions of learning, particularly the domains of information and mental procedures, were used to clarify and contemporize the SOLO Taxonomy. Several of our scorer training documents drew heavily from *Designing a New Taxonomy of Educational Objectives* (Marzano, 2001).

Based on four components of cognitive behavior and perceived limitations of Bloom's Taxonomy, Marzano developed a new taxonomy of educational objectives. Marzano suggests that knowledge can be organized into three general categories: information, mental procedures, and psychomotor procedures. For most school academic tasks, students exercise knowledge in the information and mental procedures domains; therefore, these domains were applied to the SOLO Taxonomy as evaluators made judgments about the aim of instruction and the depth of student learning.

Marzano's explanation of the domains of information and mental procedures is consistent with the levels in the SOLO Taxonomy model. Marzano organizes seven types of information that students need and use into two broader categories: details and organizing ideas. He also distinguishes skills from processes in the Domain of Mental Procedures. The research team for this study represented the hierarchical structure of the both domains from highest level of understanding to lowest level of understanding in Figure 3.2.

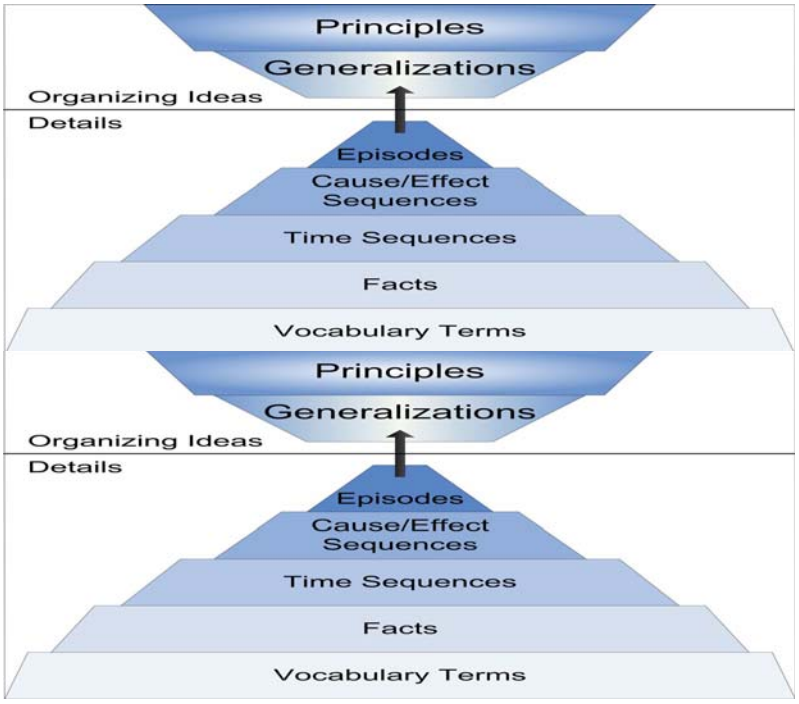

Figure 3.2 Marzano's Domains of Information and Mental Processes

Though the details dimension of Marzano's model can include somewhat sophisticated information, the focus is on reproduction and recollection of information. At the organizing ideas level, however, the focus is on drawing conclusions and applying and integrating details to form generalizations and articulate principles.

While the organizing ideas and details dimensions of Marzano's hierarchy are consistent with the quantitative (surface) and qualitative (deep) aspects of the SOLO Taxonomy, Marzano's explanations of each type of information help to clarify further the distinction between recall and reproduction of information characteristic of surface learning and the transformation and application of information characteristic of deeper learning.

Also critical to an examination of school-related tasks, assignments, and learning, is the consideration of the domain of mental procedures. Marzano distinguishes the information domain from the domain of mental procedures by explaining that the domain of information typically describes the "what" of human knowledge while procedural knowledge typically describes the "how-to." In this study, the domain of mental procedures became particularly important as assignments and responses related to procedures. Examples included such tasks as composing an essay and conducting a science experiment. These two tasks require students to use information and procedures to complete a task. Marzano separates the domain of mental procedures into two categories: skills and processes. The skills category represents the lower-level procedures that require students to apply one or more single rules, and algorithms, or a tactics. The processes category includes macroprocedures. Macroprocedures can include a diversity of possible products or outcomes and involve the execution of many interrelated subprocedures. Again, this hierarchy of mental procedures is consistent with the SOLO

Taxonomy with the skills dimension corresponding to the surface levels of performance and the macroprocedures corresponding to the deeper levels of performance.

Perhaps one of the greatest values of this study is that it provides a promising model for accomplishing a critical aim of assessment: improving student understanding and performance. Wiggins (1998) suggested that "the aim of assessment is primarily to *educate and improve* student performance, not merely to *audit* it" (italics in original, p. 7). The model proposed here has potential to inform teachers, teacher educators, researchers, and other education stakeholders about how to understand and improve student performance.

METHODOLOGY

While the nature of educational research requires that researchers be flexible and responsive to changing conditions, quality educational research also requires careful planning and forethought. This chapter provides details related to the design and methodology of the present study. Descriptions of the purpose, participants, procedures, materials, analysis, and standards of quality and verification are included.

This study could most accurately be described as a mixed method study. A mixed method approach was appropriate for this study intended to examine the complex relationship between teaching quality and student learning. Detailed descriptions support and exemplify the numerical values assigned to the teaching practices and student outcomes. Besides descriptions of various classroom assignments, activities, and learning outcomes related to the depth of learning, this validation study of the NBPTS system of advanced certification also examined representative information to facilitate generalization of the research findings across the nation. Thus, statistical inference is employed as part of the quantitative method to complement the in-depth inquiry of teaching effectiveness in a wide range of school settings.

Purpose

The purpose of this study was to examine the teaching practices of National Board Certified teachers (NBCTs) and their impact on student achievement (depth of student learning), relative to the practices and impact of teachers who have not gained National Board Certification (NBC). In this study, the research design and data collection were developed to examine two major dimensions of teaching performance: comparative teaching practices and comparative teaching outcomes. Two major research questions, aligned to these dimensions, were addressed in this study:

- *Comparative Teaching Outcomes.* Do students taught by National Board Certified teachers produce deeper responses (to class assignments and standardized writing assessments) than students of teachers who attempted National Board Certification but were not Certified?

- *Comparative Teaching Practices.* Do National Board Certified teachers develop instruction and class assignments designed to produce deeper student responses than teachers who attempted National Board Certification but were not Certified?

As a validation study, these research questions were used to evaluate the relationship among teachers' instructional aims and design, students' subsequent learning, and the National Board's vision of accomplished practice. The research questions and validation study issues are shown in Table 4.1.

Table 4.1 *Dimensions of Teacher Performance, Research Questions, and Validation Study Issues*

Dimension of Teacher Performance	**Research Question**	**Corresponding Validation Issue**
Comparative Teaching Practices	Do National Board Certified teachers develop instruction and class assignments designed to produce deeper student responses than teachers who attempted National Board Certification but were not Certified?	To what extent is the National Board's vision of accomplished practice, as articulated in the Standards documents and as instantiated in its assessments, consonant/consistent with teachers' instructional design and expectations?
Comparative Teaching Outcomes	Do students taught by National Board Certified teachers produce deeper responses (to class assignments and standardized writing assessments) than students of teachers who attempted National Board Certification but were not Certified?	To what extent is the National Board's vision of accomplished practice, as articulated in the Standards documents and as instantiated in its assessments, consonant/consistent with students' learning outcomes in the classroom?

The research crosswalks represented as Tables 4.2 and 4.3 were developed by the research team to guide instrument development, data collection, triangulation of data sources, and data analysis.

The qualitative results are limited to a descriptive response of the appropriate level of the SOLO Taxonomy.

Table 4.2 *Comparative Teaching Practices Research Crosswalk*

Comparative Teaching Practices			
Questions Guiding Data Collection and Analysis	**Data Sources**	**Analysis**	
		Qualitative	**Quantitative**
What *types* of assignments are given or required by NBCTs and non-NBCTs?	▪ Unit Context Responses ▪ Profile of Lessons ▪ Profile of Student Work Samples ▪ Student Work Samples	▪ Content analysis and evaluation using the Teacher SOLO Taxonomy ▪ Development of illustrative exemplars	▪ Parametric statistical testing on the SOLO score difference between Certified and non-Certified teachers ▪ Parametric analysis of the National Board scores between candidates who have shown surface instruction and those who demonstrated deep instruction ▪ Non-parametric analysis on the association between the Certification outcome (Certified vs. non-Certified) and the depth of instruction ▪ Reliability checking on the SOLO scoring
What assignments are given or required by NBCTs and non-NBCTs?	▪ Profile of Lessons ▪ Profile of Student Work Samples ▪ Student Work Samples	▪ Content analysis and evaluation using the Teacher SOLO Taxonomy ▪ Development of illustrative exemplars	
What is the intent of teachers' instruction relative to depth of student response?	▪ Unit Context Responses ▪ Profile of Lessons	▪ Content analysis and evaluation using the Teacher SOLO Taxonomy ▪ Development of illustrative exemplars	

Table 4.3 *Comparative Teaching Outcomes Research Crosswalk*

Comparative Teaching Outcomes			
Questions Guiding Data Collection and Analysis	**Data Sources**	**Analysis**	
		Qualitative	**Quantitative**
What level of responses do students produce in response to teacher assignments?	▪ Work Samples collected from six randomly selected students in each classroom	▪ Analysis of expert scores on SOLO-based rubrics ▪ Development of exemplars	▪ Multilevel analysis of the student outcomes to partition the variances at student and teacher levels ▪ Discriminant function analysis to test the degree of separation between Certified and non-Certified teachers ▪ Factor analysis of student writing samples to identify a latent variable of student writing performance ▪ Post hoc analyses of the work setting difference between Certified and non-Certified teachers ▪ Triangulation of the results of student writing and teacher certification status with assessment of the student work samples using both parametric and non-parametric methods ▪ Inter-rater reliability checking on the holistic SOLO ratings, as well as the analytic ratings from the writing rubric
What level of responses do students produce in response to an external writing prompt?	▪ Responses to standardized writing prompts (generated by our research team) for each student in the MC/Gen and EA/ELA classrooms	▪ Analysis of holistic rating using SOLO-based rubric ▪ Analysis of analytic scores on five writing features ▪ Development of exemplars	

The National Board for Professional Teaching Standards (NBPTS) system of advanced certification, including its portfolio assessment and performance scoring system, represents an extraordinarily complex and ambitious initiative in education. In a previous validation study of NBC, Bond, Smith, Baker, and Hattie (2000) developed some innovative approaches by

incorporating quality of student learning outcomes and collection of additional validation data by trained observers. Although the previous study was confined by a small sample of teachers in two certification areas, its empirical findings provided valuable references to consider in designing the sampling framework, collecting unbiased information, and completing the profound and extensive task of data analysis. In this regard, the research methodology demonstrates characteristics of both confirmatory and exploratory inquiries. Whenever pertinent, the existing knowledge from the previous study was utilized to make the inquiry more feasible and informative than a complete random exploration. Meanwhile, new methods have been developed to explore broader issues not covered in the Bond et al. project.

Participants

Teacher-participants were recruited from across the United States in four certificate areas. A total of 64 teachers from 17 states participated in the study. All participants had attempted Certification in one of the four certificates identified. Thirty-five (55%) of the participants had achieved National Board Certification, and 29 (45%) had attempted but had not achieved National Board Certification.

Procedures

The data gathering process can be examined on two dimensions that assemble a list of subjects and variables, respectively. To facilitate the generalization of the results, a proper mechanism of randomization needed to be introduced in the subject sampling process to support probabilistic inference of the population parameters across the nation. On the variable dimension, qualitative and quantitative measures were articulated to triangulate different perspectives of teaching effectiveness at various levels of the U.S. education system. A variety of qualitative and

quantitative data and analyses were used in this study. Thus, quality control issues largely hinge on details of the sample design and variable measurement.

Sample Design and Procedures

To detect potential differences between National Board Certified teachers and non-Certified teachers, if they existed, representative samples were drawn from candidates for the National Board assessment in four of the available certificate areas: one generalist certificate, Middle Childhood/Generalist, and three subject-specific certificates, Early Adolescence/ English Language Arts (EA/ELA), Adolescence/Young Adulthood Science (AYA/Science), and AYA Social Studies-History (AYA/SS-H). These certificates represent three of the developmental levels of the available certificates. Using the database provided by NBPTS in December 2003, the research team identified the population size for each certification area. The candidate population is represented in Table 4.4.

Table 4.4 *Candidates by Subject Area*

Certificate	Total number candidates
MC/Gen	5181
EA/ELA	1518
AYA/Science	1242
AYA/SS-H	1129

This candidate population represented the candidates in the database for all years from 1996 through 2003. Although some of the certificates were available prior to 1996, changes made to the NBPTS assessment design and scoring system make it impractical to compare scores for teachers who were candidates prior to 1996. To minimize the possibility of detecting unwarranted differences, re-takers (that is, teachers who made a second or subsequent attempt to gain certification after an initial failed attempt) were omitted from the database prior to sampling. Two withdrawal cases in the AYA/SS-H track were considered as non-participants,

and thus, dropped from the population list. When designating sample sizes in each of the *Certified* and *non-Certified* categories, consideration was given to differences in the sample attrition rate evidenced in the pilot study which was conducted Fall 2003. Due to the subject expansion in this investigation, results from the pilot study were used to reconfirm the response rate among the four certificate areas. The response rates for the pilot study are included in Table 4.5.

Table 4.5 *Attrition Data for Each Subject Area*

Certificate	**Number Certified**	**Number responded**	**Number non-Certified**	**Number Responded**
MC/Gen	25	12	25	7
EA/ELA	26	10	24	6
AYA/Science	25	7	25	5
AYA/SS-H	24	11	26	10
Total	100	40	100	28

Besides differences among the subject areas, the response rate was consistently higher for the Certified group in each of the four certificate domains. Accordingly, the sample sizes were carefully configured as shown in Table 4.6 to support a balanced comparison between the *Certified* and *non-Certified* groups on various measures of teaching effectiveness.

Table 4.6 *Sample by Certificate and Certification Status*

Certificate	**Total number**	**Number Certified**	**Number non-Certified**
MC/Gen	150	55	95
EA/ELA	180	70	110
AYA/Science	250	105	145
AYA/SS-H	125	50	75
TOTAL	705	280	425

To facilitate probabilistic inference from the sample statistics to population parameters, random numbers were generated from the SAS computer software to draw participants from the Certified and non-Certified categories. Whereas the National Board certification maintains the

same criteria for all states, differences in the certification incentives may have caused an uneven spread of candidates across states to pursue certification. In addition, those who have been certified may choose to move to a state that has better incentives. The policy impact is reflected by widely-varied sampling points for each state. To ensure a balanced representation across different geographic regions, telephone area codes were sorted, and the geographic representation was checked for a match between the sample and population lists.

In addition, the original population data from the National Board included candidate gender, race, and teaching experience information. As a criterion-referenced assessment, criteria for NBC are not gender/race specific. Although no stratification is needed on dimensions of *gender* or *race* at the sampling stage, the demographic information was documented in the population database so that sampling weight could be created to disentangle the results over different demographic categories through a post-stratification process (see Allen, Carlson, & Zelenak, 1999). The post-stratification analyses were conducted on those contextual factors to facilitate various policy analyses in the future.

Procedures for Determining Participant Eligibility

To maximize efficiency and reduce unnecessary communication with ineligible teachers (e.g., retired, no longer teaching, not interested in this research), the research team established an initial contact with potential participants that included a letter that briefly described our research and requested that recipients complete an eligibility survey. The survey included questions about recipients' current teaching context and interest in participating in a research study. Potential participants drawn from the sample could complete the survey in either of two formats:

- a paper format (included with the letter), or

- an electronic (web-based) format that they could access from their home or school computer.

NBPTS provided information to enable researchers at the Office for Research on Teaching (ORT) to develop databases with potential participant contact information. Based on the sampling results, 705 surveys were mailed to potential participants.

Recruitment Procedures

As potential participants met the eligibility criteria and expressed willingness to participate (or at least be contacted to obtain additional information about the project), the research team responded quickly by making a telephone contact. In an effort to minimize participant mortality, the research team established telephone recruitment procedures and protocols. The principal investigator and senior researcher developed a telephone recruitment training program. Practicing teachers, graduate students at Winthrop University in South Carolina, and members of the scoring team for the Bond et al. study (2000) were trained to make recruitment calls. The telephone recruitment protocol materials are included as Appendix A.

To manage the recruitment of potential participants, a telephone recruitment webpage and database were developed. Using this webpage and database, a member of the research team at the Office for Research on Teaching could assign a telephone recruitment team member to make a telephone contact to the potential participant to confirm eligibility, explain briefly the purpose and procedures of the study, and ask for a verbal agreement to participate.

If the potential participant continued to express an interest in the research, the scope of his or her involvement was described in more detail using the recruitment protocol, and a verbal agreement to participate was secured during the telephone call. A best mailing address for the participant was recorded by the telephone recruitment team member so that agreement and

honorarium forms and subsequent data collection materials could be mailed to the teacher participant.

Procedures for Securing Participant Agreement

Once a teacher agreed to participate in the study, the telephone recruitment team member forwarded the relevant data (e.g., mailing address, current teaching situation) to a member of the research team at the Office for Research on Teaching. The administrative assistant prepared an agreement packet for each potential participant. This packet included a letter describing the study, two copies of an agreement form, an honorarium, a parent consent/student assent form, and an information sheet for the school principal. The agreement packet is represented in this report as Appendix B. When participants returned agreement and honorarium forms, the research assistants for the study prepared and sent the data collection materials.

Teacher and Student Work Sample Data Collection Design and Procedures

Once agreement and honorarium forms were returned to the ORT, participants were sent a box that included all the materials they would need to collect data for the research project. Each box sent to participants included the following materials:

- a cover letter reiterating the components of the study,
- directions for collecting the student work,
- one Unit Context form with return envelope,
- directions for randomly selecting six students for research study,
- twenty Profile of Instruction forms with a return envelope,
- twenty Profile of Student Work Samples forms,
- fifteen Tyvek® envelopes,
- mailing tape, and

- one postage-paid return label.

The MC/Gen and EA/ELA participants also received the following additional materials for the writing assessment component of the study:

- one copy of the teacher directions for the writing assessment,
- thirty student writing folders, each containing a writing prompt, and
- one teacher participant form.

Forms and data collection materials are included in this report as Appendix C.

These materials comprised all supplies teachers needed to collect data related to their self-selected unit of instruction. For the purposes of the study, the research team defined a unit of instruction as at least five related lessons focused on a single topic, theme, or problem. The forms ask teachers to explain the context of the unit/lesson, the teacher's purpose for the various lessons, and other details related to the lesson, assignment, and student work samples. These completed data provided information related to the comparative teaching practices research question in this study.

Each teacher participant submitted work samples from six randomly-selected students in one class. Teachers were provided directions for randomly selecting six students for this study. They were asked to use a copy of their class roster and the sampling matrix provided. On the sampling matrix, the teachers were to use the row that corresponded to the number of students in the class. Then, they were to choose the students based on the matrix and their number on the class roster. If the student designated was not available (e.g., no parent/guardian permission, withdrawn from school), teachers were to go to the next available student on the roster. These work samples provided data related to the comparative teaching outcomes research question in this study. Table 4.7 is a copy of the sampling matrix provided to teachers.

Table 4.7 *Sampling Matrix for Random Selection of Six Students*

Number of Students in Class	Student A	Student B	Student C	Student D	Student E	Student F
10	1	4	9	10	3	6
11	7	6	9	10	8	11
12	3	7	4	8	10	6
13	8	6	7	10	12	4
14	8	14	7	12	6	9
15	1	2	10	7	9	5
16	1	16	3	11	9	12
17	14	10	11	17	16	9
18	10	8	1	12	7	4
19	6	8	3	18	5	17
20	15	12	8	10	5	3
21	6	17	14	1	20	18
22	1	5	8	17	4	2
23	15	17	21	10	23	20
24	2	9	3	4	5	8
25	1	11	4	3	16	24
26	17	11	21	7	6	19
27	23	24	13	16	7	4
28	24	17	19	15	1	27
29	23	22	6	13	14	1
30	1	17	10	13	23	22
31	2	11	15	22	20	5
32	28	14	27	5	19	3
33	11	32	3	7	33	21
34	32	20	21	30	12	9
35	24	27	16	22	5	11

Procedures for Managing Data

As the teachers returned completed materials to the ORT, an entry was created in the data log, and the contents of the boxes were marked with the participant's identification number, devised for the purpose of this study. ORT administrative personnel purged all student identifiers from student work samples and teacher responses before any external reviewers or content experts saw the materials. Teacher names and school/community identifiers were also removed from the materials. Unused forms and envelopes were removed from the boxes.

Development and Implementation of a Standardized Writing Assessment

Teachers in the MC/Gen and EA/ELA certificate areas were asked to administer the writing assessment based on the NBPTS descriptions of these certificates that the teacher candidate should have primary responsibility for students' writing instruction. Teachers in the other two certificate areas are not typically responsible for delivering writing instruction to students so they were not included in the writing portion of the study.

Several questions informed the development of the writing assessment component:

- What are the characteristics of state assessments for the participants in this study?
- What features of writing are typically included in standardized scoring guidelines?
- Does the standardized context allow students to produce both surface and deeper writing samples or does the on-demand context restrict their responses?
- What is "depth of knowledge" of writing in the context and content of school-based writing?
- How can writing tasks be designed to elicit deeper responses?
- What do surface responses, produced in a standardized context, look like?
- What do deeper responses, produced in a standardized context, look like?

Procedures for Developing the Writing Tasks and Administration Materials

State writing assessments are a reality for many public school classrooms and certainly for candidates included in this study. The debate over the positive or negative impact of such testing is beyond the scope of this discussion (Hillocks, 2002; Huot, 2002; Nunally, 1991; Paris & McEvoy, 2000; Thomas, 2004). However, in order to minimize the potential for advantaging or disadvantaging any segment of the student population on the basis of their familiarity and practice with their state's writing test, a review of all state assessments was completed prior to developing writing test materials for this study. Four characteristics of these tests were collected for the grade levels of students in the Middle Childhood/Generalist and Early

Adolescence/English Language arts certificates: the type of writing students were asked to produce, whether students wrote to an assigned topic or had a choice of topics, the length of time students had to complete their writing, and whether the writing samples were scored holistically or analytically. The study began with participants from almost all states. Because writing samples were only submitted from nine states, the relevant portion of the survey is summarized in Tables 4.8, 4.9, and 4.10.

Table 4.8 *State Department of Education Writing Assessment Programs for Participants Completing the Study (Elementary)*

State and Grade Level	Form of Writing Assessed at the Elementary Level	Administration Time	Topic Choice	Type of Scoring
Colorado 3,4,5	Narrative Expository Informative	3 days, 50 minutes	no	Analytic
Florida 4	Narrative Expository	45 minutes	no	Holistic
Kansas 5	Narrative Creative Expository	4 class periods	no	Holistic
Kentucky 4	Narrative Persuasive	90 minutes	yes	Holistic
Maryland	Narrative Expository	2 sessions, 3 hours maximum	no	Analytic
Massachusetts 4	Narrative	2 sessions, 45 minutes	no	Analytic
North Carolina 4	Narrative	50 minutes	no	Holistic
Ohio 4, 6	Narrative Expository	<2½ hours (2 Samples)	no	Analytic
Wisconsin 4	Narrative Expository Persuasive	not available	no	Holistic

Table 4.9 State Department of Education Writing Assessment Programs for Participants Completing the Study (Middle Grades)

State and Grade Level	Form of Writing Assessed at Middle School Level (7^{th} & 8^{th})	Administration Time	Topic Choice	Type of Scoring
Florida 8	Persuasive Expository	45 minutes	no	Holistic, 6 pt.
Kentucky 7	Narrative Persuasive Expository Descriptive	90 minutes Multiple samples	yes	Holistic
Kansas 8	Narrative Expository Creative	2-4 class periods	no	Analytic, 6 trait
Colorado 7	Narrative Expository Descriptive	50 minutes	no	Analytic, 4 pt.
Maryland (Functional Writing Test) 8	Narrative Expository	2 sessions, 3 hours max. 2 samples	no	Analytic, 4 pt.
North Carolina 7	Argument	75 minutes	no	Analytic: Content = 4pt Conventions= 2pt
Massachusetts 7	Expository	2-45 minutes sessions	no	Holistic: 2 forms Topic: 0-6 Conventions: 0-4
Wisconsin 8	Narrative Persuasive Expository	30 minutes	no	Holistic, 6pt+ 3 pt. Convention Rubric

Table 4.10 *State Department of Education Writing Assessment Programs for Participants Completing the Study (High School)*

State and Grade Level	Form of Writing Assessed at High School Level (9th and up)	Administration Time	Topic Choice	Type of Scoring
Colorado 9	Narrative Expository Descriptive	50 minutes	no	Analytic, 4pt.
Florida 10	Persuasive	45 minutes	no	Holistic
North Carolina	Informational	75 minutes	no	Analytic: Content = 4pt Conventions= 2pt

Development of Writing Prompts

While it was important to consider the context of statewide testing programs, it was more important to create an evaluative tool and data source consistent with the theoretical underpinnings of the study. The greater concern thus was to create a stimulus (writing prompt) that would allow student-writers to demonstrate a depth of knowledge of writing. A review of the prompt-development literature identified three primary factors affecting writing performance: student knowledge of the content or subject matter (Benton, Corkill, Sharp, Downey, & Khramtsova, 1995; Gradwohl & Schumacher, 1989; Hilgers, 1982; Langer, 1984); student knowledge of the forms of writing (Engelhard, Gordon & Gabrielson, 1992; Goldberg, Roswell, & Michaels, 1998; Larson, 1971; Ruth & Murphy, 1998; Wesley, 2000); and student knowledge of macroprocedures as they relate to writing (Camp, 1993; Fletcher, 1993; Larson, 1992; Marzano, 2001).

Writing prompts were developed by the writing assessment director and the principal investigator, both experienced language arts educators. Initially, to accommodate the variability

of the types of writing tested in state assessments, prompts were developed to allow students a choice of form (such as narrative, expository, or persuasive writing). Drafts of these "choice" prompts were shared with several classroom teachers who were asked to predict how their students would respond. When most of the teachers indicated that their students were unfamiliar with a choice of form embedded within an assigned topic and would want to ask questions before they could start writing, further development was suspended. The novelty of such a stimulus might interfere with students' ability to write.

Two types of writing, Informative and Persuasive, were finally selected for prompt development as they were the most likely to elicit responses at all levels of the SOLO Taxonomy. These types are familiar to American educators through the National Assessment of Education Progress (1998) and abroad (Glasswell, Parr, & Aikman, 2001; Wilkinson, 1980). The two prompt writers, when thinking through typical student responses to their initial set of over 30 prompts that included narrative and descriptive writing, found that these hypothetical responses would result in, at best, surface Multistructural writing. Each prompt included four elements: a topic familiar to students, the purpose of the piece of writing, a plausible audience, and a format. Format (such as a letter) was included to make the writing task as realistic as possible but it was not an element in the evaluation of the writing samples. In order to ensure that students would be familiar with the content of the prompt, 18 broad topics such as *animals, inventions, food,* and *school rules* were reviewed by teachers and two experts in large-scale writing assessment to ascertain content familiarity. Four topics were selected, and five prompts were developed for the 2003 pilot study. Three of these were retained for the operational study.

Development of Test Administration Procedures and Materials

The writing process is the result of a paradigm shift from teaching students the "modes of writing" which require students to imitate these forms as produced by adult, professional writers, to teaching them the recursive nature of writing. This process writing approach includes pre-writing, writing, and revising as a series of processes that are repeated in the production of a final draft. While process writing is exemplary as an instructional model (Atwell, 2002; Caulkins, 1994; Freeman, 2003; Graves, 2000; Murray, 2002; Portalupi & Fletcher, 2001; Ray, 2001), it is only approximated in standardized test conditions. To minimize loss of time for classroom instruction, students were given 45 minutes to produce a single draft. To remind them of the writing process, the total time was divided into 10 minutes to plan and prewrite, 30 minutes to draft and revise, and 5 minutes to proofread. These times were flexible in the actual administration. Additional process cues were presented on the Writing Topic Page in the form of Hints about "Getting Ready," "Writing for a Reader," and "Polishing your Writing." Test materials, including prompts, student test materials, and teacher directions are included in Appendix D. A sample Writing Topic Page from the final study is included as Figure 4.1.

Sample Writing Topic Page

Membership in the History Hall of Fame

Each person in your class is going to choose an important person for the History Hall of Fame. Write a letter to the selection committee that explains why the person you have chosen should become a member of the History Hall of Fame. Think of famous people in history that you have studied, read about, or seen in educational films. You can write about a person who has made important contributions to sports, the arts, politics, or taking care of others.

Think about WHY THE PERSON SHOULD BE SELECTED FOR THE HISTORY HALL OF FAME. Write to explain WHY THE PERSON SHOULD BE SELECTED FOR THE HISTORY HALL OF FAME.

Hints

Getting Ready

- Use the planning space provided in the Writing Folder.
- Think about the writing topic.
- Brainstorm or prewrite.
- If you need more space to plan your writing, ask the teacher for more paper.

Writing for a Reader

- Think about your audience. Remember that you are writing to explain to the selection committee why your person should be chosen for the History Hall of Fame.
- Give reasons and examples to help explain.
- Think about the order of your ideas.

Polishing Your Writing

- Read your paper and make any needed changes.
- Add any missing information or words.
- Make sure you have written complete sentences.
- Check your spelling and punctuation.

Figure 4.1 Sample Writing Topic Page

Data Collection for the Writing Samples

Student writing assessment responses were collected to assess the following indicators of writing performance:

1) A holistic score using the SOLO Taxonomy Rubric;

2) An analytic score for controlling idea using the SOLO-based Writing Rubric;

3) An analytic score for organization using the SOLO-based Writing Rubric;

4) An analytic score for elaboration using the SOLO-based Writing Rubric;

5) An analytic score for voice using the SOLO-based Writing Rubric; and

6) An analytic score for sentence formation using the SOLO-based Writing Rubric.

Three prompts were included in the operational study. Prompts were spiraled within classroom, with each classroom receiving two different prompts, and half the students within an individual classroom responding to one of the two prompts. One prompt was repeated across both certificates. Students in the Middle Childhood/Generalist classrooms were asked to write persuasively about a "most important animal" of their choice. Students in the Early Adolescence/English Language Arts classrooms were asked to write persuasively about "an invention that has improved or harmed people's lives." Students in both classrooms were asked to write an explanation of why a person they selected should be admitted to a "history hall of fame." Each student received and responded to a single prompt. Given the nature of the particular Informative prompt (The history hall of fame), students tended to produce persuasive responses.

The Scoring Operation for Work Samples

Preparation for Training and Scoring of Work Samples

Prior to scorer training, multiple scoring tools were developed to assist trainers and scorers in their understanding of this performance assessment. The teacher scoring tools included a teacher rubric, a Teacher Evidence Recording Form (T/ERF), and a teacher scoring pathway with guiding questions. The student scoring tools included a student rubric, a Student Evidence Recording Form (S/ERF), and a student scoring pathway with guiding questions. These scoring tools are included as Appendix E.

Rubric Development. Based on the literature related to student learning generally and the SOLO Taxonomy specifically, the research team developed a series of scoring rubrics for evaluating teacher aims and instructional design as well as several aspects of student learning and achievement. Beginning with the SOLO rubrics developed and used in the Bond et al. study

(2000), modifications were made to accommodate the data collected in the present investigation. Teacher and student work sample data from the pilot study were used to continue the revision process.

With the student and teacher work sample rubrics, content specialists closely examined "cases" to determine if these rubrics could be used without making each rubric specific to the content areas. The content specialist team determined that the wording on the rubrics was sufficient to score student and teacher work samples, but that a necessary condition was that they be applied by expert scorers (in the particular disciplines).

Benchmarking. During the benchmarking process, the teacher and student work samples were used to clarify the language of the scoring rubrics, especially those indicators that separated "surface" performances from "deeper" performances. Clarity was essential to the success of the operational scoring.

Using the final versions of the rubrics, benchmark performances for the teacher and student dimensions were identified. Benchmarking of cases began in May 2004. Content specialists began screening and carefully reading cases for potential benchmark cases that would further define evaluation points on the rubric for each content area. Benchmark cases were identified for the unistructural, multistructural, and relational levels in a variety of content areas and for the teacher and student dimensions. No cases were found during the benchmarking sessions that reached the extended abstract level.

Benchmarks represent a variety of unit topics and content areas. Because the content area units in ELA, Science, and SS-H provided a particular content to examine and evaluate, cases in these certificate areas were used as benchmarks. A team of content specialists and experienced scorers selected the benchmark cases.

The benchmark cases for the work sample evaluations are shown in 4.11 (1 = Prestructural, 2 = Unistructural, 3 = Multistructural, 4 = Relational, 5 = Extended Abstract).

Table 4.11 *Benchmark Cases for Work Sample Evaluations*

Certificate	**Teacher Dimension Benchmarks**					**Student Dimension Benchmarks**				
	1	2	3	4	5	1	2	3	4	5
EA/ELA		010106								
AYA/Science			040198				040009	040179	040033	
AYA/SS-H			110007; 110079				110002		110079	

Development of Evidence Recording Forms. In the same way that the rubrics from the Bond et al. study (2000) provided a starting point for modifications on the rubrics for the current study, the Evidence Recording Forms (ERFs) from the previous study were also used to inform the version of the form that was used for this study. Changes were made to accommodate the data collection instruments for the present investigation. An electronic version of the evidence recording form was also created. The same color-coding system that was used for the data collection instruments was used in the design of the electronic version. Therefore, when scorers accessed and used the electronic form, the color of each webform matched the color of the appropriate data collection instrument. For example, the Profile of Lesson forms that teacher-participants used was light blue so the electronic webform where scorers recorded evidence related to the Profiles of the Lesson was also light blue. This consistent color-coding system contributed to the ease of use and consistency of evidence recorded.

Development of Scoring Pathways. To assist scorers in their efforts to read and record, and evaluate evidence from the extensive and varied data sources for each case, the scoring consultant, and content specialists, trainers, developed dimension-specific scoring pathways with guiding questions. In other words, while scorers read all materials for each case and

dimension, the *teacher* scoring pathway served as a reminder of the most critical data sources for the teacher dimension of the evaluation. The emphasis for scoring the teacher dimension was on the teacher-generated responses: the Unit Context, the Profiles of Lessons, and the Profiles of Student Work Samples. Although scorers read these same materials to become familiar with the context of the instruction when scoring the student dimension, the *student* scoring pathway focused scorers' attention on the work samples produced by students. The Scoring Pathways, with the Guiding Questions are included in Appendix E. Guiding questions were used to focus and re-focus scorers' attention on the critical evidence related to the appropriate dimension. Guiding questions (Figure 4.2) were developed for both the teacher and the student scoring dimensions, and both the reading and synthesizing portions of the scoring path.

Student Version

Guiding questions to be used as you *read* the evidence:

- What did students produce?
- What instructional materials does the student use to develop responses?
- What is the pattern of response for the task or sequence of tasks?
  (Finds details, uses details, uses details to develop new understandings)

Guiding questions to be used as you *synthesize* the evidence:

- Do the student work samples represent deep levels of understanding?
- To what extent?

Teacher Version

Guiding questions to be used as you *read* the evidence:

- What task(s) did the teacher give the students?
  - What do the task(s) **allow** the students to do?
  - What do the task(s) **require** the students to do?
- What does the teacher do to set up the task(s)?
- What role does the teacher play in the teaching and learning process?

Guiding questions to be used as you *synthesize* the evidence:

- Does the teacher design tasks or a sequence of tasks that could elicit deeper student outcomes?
- To what extent?

Figure 4.2. *Guiding Questions from the Scoring Pathways*

Participants in Scoring Operation for Work Samples

Content Specialists. At the beginning of this study, content specialists with expertise related to each of the four certificate areas were identified to work as a team on the development of data collection instruments and procedures. These content specialists have doctoral degrees in education or the specified content area. In addition, all have K-12 public school experience as well as university teaching and research experience.

Trainers. Two content-area trainers were identified and recruited for each of the certificate areas in the study. These trainers were selected based on their experience and expertise in research, teaching, and scoring. In addition, the trainers worked in pairs so that their content areas would complement each other. For example, one of the AYA Social Studies trainers specialized in American government, politics, and history while the other specialized in World geography and history.

After the initial training session, trainers worked as scorers; therefore, their demographic data are included in the description of scorers in the next section.

Scorers. All members of the scoring team (including trainers) were experienced classroom teachers. The scoring team consisted mostly of classroom teachers, many National Board Certified; however, some NBCTs who served as scorers had taken other types of positions. One was in a district administrative position, one was in a Reading Specialist position, one was a literacy specialist, and four were doctoral students in the field of education. The team members ranged in years of teaching experience from 2 to 33 years, with a mean of 14.6 years. Two-thirds (66.7%) had advanced degrees in Education. Table 4.12 shows the highest degree earned and gender composition of the scoring team.

Table 4.12 *Degree and Gender Composition of Scoring Team*

Degree	Male	Female	Total
BA/BS	9	24	33
MA/MS	6*	12	18
ED.S.	2	0	2
Ed.D/Ph.D.	1*	1*	2
Total	18	37	55

*does not include degrees in progress

Scorers had a variety of teaching backgrounds and experiences ranging from elementary school through college level, and including content areas such as language arts, social studies, reading and writing, Reading Recovery, gifted education, technology, all areas of high school social studies and history, and all areas of high school science. Eighty-five percent of the scorers were National Board Certified.

The scoring team included members and officers of a variety of national professional organizations, including the following: National Education Association, National Middle School Association, International Reading Association, American Association of Physics Teachers, National Science Teachers Association, Association for Supervision and Curriculum Development, National Association of Black School Educators, Pi Lambda Theta, and Delta Kappa Gamma. On the state level they were members of the North Carolina Association of Educators, North Carolina Middle School Association, North Carolina Council for the Social Studies, North Carolina Science Teachers Association, Georgia Association of Educators, and Georgia Science Teachers Association. They have made a combined 42 presentations at national, regional, and state education conferences, and have conducted 43 workshops. One third of the scorers has been involved with Institutions of Higher Education or has taught courses at that level. Forty-five of the scoring team members have served on state department of public instruction committees (e.g., curriculum revision committee, textbook adoption committee), have been involved in writing End-of-Grade test items, reviewing the exam items, or reading test

items; have been certified writing trainers; or have been curriculum trainers. Many held multiple positions over the years. Twenty-seven percent of the scoring team members were professional development providers for teachers. For example, nine scorers participated in or were certified trainers at the North Carolina Center for the Advancement of Teaching, and seven were participants or trainers at the North Carolina Teacher Academy.

Team members have been recipients of many teaching honors, including sixteen teachers who have received Teacher of the Year Awards (several more than once). Many of the team members have been cooperating teachers, half have served as mentors to pre-service or in-service teachers, and seven have served as chairpersons of their departments. Almost half have received some kind of grant. Three teachers have received Bright Ideas grants from the Blue Ridge Electric Membership Corporation. One AYA/SS-H teacher has received grants through the Western Carolina Foundation and the Constitutional Rights Foundation. Two AYA/SCI teachers received a National Science Foundation grant through the University of Georgia, and another received a North Carolina Biotechnology Center grant. Other designations among the scoring team members included a Charlotte World Affairs Council Scholar, a Jaycees Young Educator, a Washington Mutual Fellow, a Toyota-Most Inspired Teacher, and several have received Time Warner Star Teacher Awards. Six scorers have published articles in magazines or have co-authored professional papers. Three scorers were Certified NBPTS mentor/trainers.

Our team of scorers lived and taught in the states of North Carolina, Virginia, and Georgia. Scorers and trainers were recruited from diverse sections of the state. Of our North Carolina scoring team, one taught in Ashe County, two in Avery County, and one in Buncombe County, representing our Northwestern Mountain counties. Four scorers taught in the counties of Guilford, Forsyth, and Davidson, representing our Western Piedmont region. Six scorers taught

in Caldwell and Catawba Counties, representing the Southern Piedmont region. One scorer was a gifted education specialist from Virginia. Eight teachers lived and taught in the Charlotte-Mecklenburg area of the state. Two teachers represented the Eastern Coastal area of Pender County. Half of our AYA/SCI scoring team, seven teachers and educators, were from the Athens, Georgia area. Finally, two scorers were retired teachers.

The gender and ethnic composition of the scoring team is shown in Table 4.13. While minorities are not well-represented, this under-representation is also reflected in the overall teacher population generally and in our study sample particularly (though not by design).

Table 4.13 *Demographic Composition of Scoring Team*

RACE	**SEX**		
	Female	Male	Total
Asian	2	0	2
African American	1	0	1
Indian	1	0	1
White	20	9	29
Total	24	9	33

Training of Trainers

In this study, all trainers and scorers were experienced expert teachers. Based on their experience, credentials, and expertise, two scorers in each scorer group were designated as trainers. The content specialists and two classroom teachers who were experienced, complex performance assessment raters provided training for the trainers in each content area. The train-the-trainer session included the following topics:

- Overview of the study,
- Discussion of data collection procedures and instruments,
- Discussion of SOLO Taxonomy rubrics and Marzano's Taxonomy of learning objectives,
- Discussion and application of reading for evidence,

- An independent reading of a case,
- Discussion related to sources of scoring error,
- Close examination of a teacher benchmark case,
- Training related to electronic data entry and the Evidence Recording Form,
- Practice reading and recording evidence,
- Close examination of a student benchmark case,
- Debriefing about the surface and deeper outcomes constructs,
- Discussion of preponderance of evidence,
- Certificate-specific case review to identify training cases,
- Discussion of the purpose and procedures for scoring augmentation, and
- Planning for training scorers in certificate area.

On-site contact time for the train-the-trainer session was approximately 20 hours. After this session, trainers worked independently and in pairs for approximately 15-18 hours to prepare materials for their certificate-specific scorer training sessions.

Training of Scorers

Scorer training sessions were content specific such that scorers for each certificate were trained separately and specifically for the appropriate content areas. Because of the high rate of participation among AYA/Science teachers, we conducted two sessions at separate sites for AYA/Science. Though each session was slightly different, depending on the subject area and the participants' needs, common training elements included the following: an overview of the study, a close examination of the teacher and student rubrics (based on the SOLO Taxonomy), a presentation related to the Domain of Information and Mental Procedures (Marzano), a discussion of sources of scoring error, a discussion related to surface and deeper instructional

aims and outcomes, a close examination of multistructural and relational cases for the certificate, discussions and exercises about classes and significance of evidence, and a "live" scoring session with rater pairs.

The scoring teams were trained in two phases. The first phase consisted of an initial one-day training workshop during which scorers received training related to the design of the study, the data collection materials and protocols, and general scoring procedures. The training team led a discussion about ethical issues related to scoring and how to address personal biases that might affect scoring judgments. The scorers received in-depth training about how to consider evidence contained in the data and arrive at a scoring decision. During this phase of the training, scorers received an executive summary of the study, a handout based on a PowerPoint presentation overview of the study, the teacher dimension rubric, and the student dimension rubric. The training team led an exercise in which scorers examined a known-score benchmark case and the appropriate scoring rubric to refine their understanding of the dimension and to practice making judgments about the various kinds of evidence discussed.

For the second phase of the training, the scorers examined evidence related to the second dimension (teacher or student) in a different benchmark case. For this exercise, scorers did not know the score that had been assigned to the case. After "scoring" this case, they used the data sources, evidence recording forms, and scoring pathways to discuss the differences between the teacher aim and student outcomes dimensions. Next, raters practiced using the scoring web page. They entered data from the practice cases onto the web page, so that they could understand the format of the design and the details related to submitting their scored cases. Finally, the scorers examined and evaluated cases "live" (in other words, these cases had not been previously scored) and independently, though they were assigned the same case as another scorer in the session.

During this portion of training, each scorer assigned a value to a case and waited for the paired scorer to complete his/her evaluation. Cases were photocopied so that scorers could work on a single case simultaneously. Once both evaluations were complete, the content specialist and principal investigator discussed the evidence with the pair to respond to questions related to evidence, the rubric, the dimension, or the technical aspect of entering the data. If the rater pair disagreed on the case, the content specialist and principal investigator discussed the details of the case with each scorer individually. The goal of this discussion was to understand the nature of the evidence relative to the scoring rubrics, not to bias the scorer. In all cases, this discussion resulted in agreement between the rater pair. Once these cases were resolved, scorers were given the remainder of their assigned cases, these cases were logged, and they could leave the scoring site.

Raters left the training site with the cases (or copies of cases) they were to score as well as paper copies of materials they would need (e.g. Scoring Manual, Evidence Recording Forms, sticky notes). Raters scored the cases in remote locations (usually their own homes) and submitted the evidence and scores electronically using a secure website designed specifically for this project. A summary of the operational training and scoring activities is shown in Table 4.14.

Table 4.14 *Summary of Scoring Activities*

Activity	**Dates**	**Location**	**Participants**
Identification of SOLO Level Exemplars	May 21-26	Boone, NC	Principal Investigator, Scoring Consultant, MC/Gen Content Specialist
Benchmarking	June 8-11; June 16-18	Boone, NC; Charlotte, NC; Kernersville, NC	Representative Content Specialists, Scoring Consultant, Principal Investigator
Design of scoring augmentation	June 16-18	Boone, NC; Athens, GA	Scoring Consultant, Principal Investigator
Train-the Trainer Scoring Session	June 18-20	Boone, NC	All Content Specialists, all trainers

MC/Gen Scorer Training	June 21-22	Boone, NC	Principal Investigator, MC/Gen Content Specialist, MC/Gen trainers, MC/Gen scorers
MC/Gen Scoring	June 22-25	Boone, NC	MC/Gen Content Specialist, MC/Gen trainers, MC/Gen scorers
EA/ELA Scorer Training	June 21-22	Boone, NC	Principal Investigator, EA/ELA trainers, EA/ELA scorers
AYA-SS/H Scorer Training	June 25-26	Boone, NC	Principal Investigator, SS/H Trainers, SS/H scorers
AYA/Science Scorer Training (Site I)	June 28-29	Boone, NC	Principal Investigator, NC Science Trainers, NC Science scorers
AYA/Science Scorer Training (Site II)	July 1-2	Athens, GA	Principal Investigator, AYA/Science Content Specialist, GA Science Trainers, GA Science scorers
Scoring in all content areas	July – August	Various sites in NC and GA	All scorers
Scoring resolutions; third reads	September – October	Various sites in NC and GA	Principal investigator; all scorers

Scoring Logistics

Procedures. Scorers who successfully completed training were assigned cases to be scored. Each scoring team was divided into pairs of raters for each case and dimension. A specialized software program and database were used to assign cases randomly to scorers, maximize rater pairs, and prevent the same scorer from evaluating a single case for both the teacher and student dimensions. Actual scoring took place off-site, typically at the scorer's residence.

Scorers reviewed all case material related to the specified dimension. They typically made extensive notes related to the relationship between the evidence and the scoring rubric and arrived at a judgment about whether the evidence represented surface or deeper instructional aims and student outcomes. Next, scorers designated each case as representative of one of the following more specific levels in the SOLO Taxonomy Rubric: *Prestructural, Unistructural,*

Multistructural, Relational, or Extended Abstract. Finally, the scorers designated the case as a solid, high, or low instance of the SOLO level. This last step has been referred to in the literature as scoring augmentation (Penny, Johnson, & Gordon, 2000). The decision to use scoring augmentation was both theoretical and pragmatic, evolving out of our understanding of the nature of the SOLO Taxonomy, the work sample and writing assessment, data and the scoring process itself. The SOLO Taxonomy represents a continuum, rather than five discrete levels neatly separated by five impermeable lines, as they appear in the rubric. Some performances contain consistent, confirmatory evidence across all the sources and types of evidence; others contain evidence that, looked at in isolation, fits into a higher or lower level on the taxonomy. Such "mixed" cases do not fit neatly into one of the levels. The work samples contain evidence from several sources that is complex, dense, and at times uneven. Even with the short one-to-two page single source writing samples, the evidence can be uneven. Many performances, whether lengthy or brief, with single or multiple sources of evidence, fit cleanly into a single SOLO level. For those that do not, augmentation provides the scorers with a way to designate that this case contains features of a higher or lower SOLO level.

Raters were supplied with paper copies of the ERF. Many raters completed a paper version of the ERF as they scored the case or immediately after, using their notes. Then, they entered their evidence and judgments onto a website designed specifically for this purpose.

Because MC/Gen cases were not used as benchmarks, the MC/Gen scoring team stayed on-site to score their cases during the scoring operation so that they could discuss case-related content questions with the principal investigator and the MC/Gen content specialist. Since the sample for this certificate was especially small, the cases were reserved for evaluation by the scoring team. It should also be noted that members of this certificate scoring team were most

experienced in the scoring process as most of them evaluated work samples in the Bond et al. study (2000).

Monitoring and Resolutions

The scoring manager monitored scores as they were submitted for rater agreement. If dichotomous scores for a given case did not agree, the scoring manager reviewed the scoring rubric and the evidence in the case with each scorer to determine whether there was any misunderstanding about the rubric or misinterpretation of the evidence. The scoring manager was careful not to sway the judgments of the scorers, but only to monitor and clarify their understanding of the rubric for that dimension and how to interpret the evidence they found. This procedure usually resolved any disagreement between scorers, but if the resolution did not occur, a third member of the scoring team for that certificate area was assigned to read the case and submit a third ERF. After team members had submitted scores for the first several cases, and the scoring manager determined that the level of rater calibration was satisfactory, the scoring manager continued to monitor scores for rater agreement but did not contact rater pairs who disagreed. Further disagreements were assigned to a third rater who submitted another ERF. Throughout scoring, the scoring manager continued to monitor assigned scores for individual raters who might require additional calibration. Table 4.15 represents a summary of approximate person hours spent on training and scoring.

Table 4.15 *Hours Spent on Training and Scoring*

Activity	**Number of participants**	**Hours (or hours per unit)**	**Total**
Benchmarking	6	20 hours	120
Train-the-Trainer	17	20 hours on-site; 15 hours off-site	595
Scorer Training	42	13 hours	546

Scoring of Cases	(all cases = 65) x 4 raters (2 for teacher; 2 for student)	2.5 average hours per case	650
Third reads to resolve scores	4	2.5	10
Total			1921

Score Assignment. For each dimension, (e.g., teacher aim, student outcome) a final dichotomous score (surface or deeper) and a more descriptive score level (prestructural, unistructural, multistructural, relational, extended abstract) were assigned. The final dichotomous score was derived in one of two ways. If both raters assigned the same "score," that was the final score. If they assigned different scores (one surface, one deeper), the case was sent to a third, independent reader. The third reader's score was matched with one of the first two. In both instances, exact agreement between two independent raters was required: Raters 1 and 2 or Raters 1 and 3 or Raters 2 and 3.

It should be noted that as scorers evaluated the student outcome dimension, the scorers carefully reviewed each of the student artifacts submitted and provided one overall score about the depth of response from the group of six randomly-selected students. This score was to represent the depth of response from the class. Since the teacher is actually the unit of analysis in this study, this score served as a summary representation of the outcomes in the teacher's class.

The final evaluation that scorers made was related to the augmented scoring value. That is, the scorer decided if each case was a solid, high, or low instance of the descriptive score level. The Evidence Recording Forms included in Appendix E demonstrate the order of the raters' scoring decisions. After the scoring operation was complete, the nominal scores were converted to numerical values. Table 4.16 represents the numeric values for the scores.

Table 4.16 *Numeric values for SOLO scores of Work Samples*

	Prestructural	**Unistructural**	**Multistructural**	**Relational**	**Extended Abstract**
Solid	1	2	3	4	5
High	1.33	2.33	3.33	4.33	5.33
Low	.67	1.67	2.67	3.67	4.67

After these numeric conversions were made, the final score was calculated by averaging the scores assigned by the two scorers who agreed. That is, the scores from Scorer 1 and Scorer 2 were averaged if those scorers provided the same dichotomous score (surface or deeper). If they disagreed on the dichotomous score, the score of Scorer 3 was averaged with the original scorer who provided a dichotomous score that matched that of Scorer 3.

Management of the Work Sample Scoring Operation

Data base development. A consultant was contracted to design and implement a secure database solution for the purpose of managing the scoring operation. The consultant coded procedures for input of data, including issues such as maximizing rater pairs, assigning cases to scorers randomly, excluding raters from scoring for both teacher and student dimensions on the same case, and locking evidence recording forms from editing once they were submitted.

Web page development. A consultant was contracted to design a customized web page application capable of retrieving data from remote locations and displaying it to the content specialists, trainers, and scorers. The design of the web page matched the design of the data collection instruments used by research participants (including color and language), so that scorers could easily navigate the page as they entered evidence for each case's score report.

The Scoring Operation for Writing Samples

Preparation for Training and Scoring of the Writing Assessments

Prior to rater training three scoring tools were developed and refined as they were applied: the SOLO Taxonomy rubric, a SOLO-based analytic scoring rubric, and anchor training papers.

Rubric Development

The initial phase of rubric development centered on the SOLO Taxonomy developed in the earlier validation study (Bond, et al., 2001). Global skills addressed in the research literature as characteristics of effective writing guided our development of the writing prompts and rubrics to measure depth of knowledge of writing.

Standardization of the writing rubric. The writing assessment provided a standardized student outcomes measure, making it important to blend pertinent aspects of the SOLO Taxonomy, Marzano's declarative and procedural knowledge; and standardized assessment practices for evaluating writing. Scoring rubrics were collected from a number of sources: state departments of education, the National Assessment of Educational Progress (1998), commercial publishing and testing companies, professional journals, professional texts (particularly Arter and McTighe, 2001) and technical reports (Glasswell, Parr & Aikman, 2001).

Common features or traits of effective writing were found across all these varied sources. Both holistic and analytic rubrics expressed standards for evaluating the following traits: response to the writing task, purpose, use and quality of ideas or evidence, organization, voice, sentence fluency, and conventions. All but the latter were incorporated into the SOLO-based Writing Assessment Rubrics developed for this study. Conventions could not be described in the SOLO framework beyond the surface level, perhaps because as Camp (1993) and Schuster

(2003) suggest, grammar, usage, and mechanics are the subskills of writing. While important in a polished, edited piece of writing, conventions could not be defined at all five SOLO levels.

Keeping in mind the depth-of-knowledge lens, the analytic writing features were combined into two clusters. Complexity of Ideas included Controlling Idea, Organizational Structure, and Elaboration of Ideas. Complexity of Craft included Voice and Sentence Fluency. While writing assessment practitioners would quickly recognize the analytic features (if not the complexity clusters) and also much in the descriptions of the levels from Prestructural to Extended Abstract, the SOLO influence is apparent in the rubrics designed for this study. For example, the deeper level of Controlling Idea is assigned only when that the controlling idea is a unifying principle, generalization, or theme. The writer must go beyond the immediate context set up in the writing prompt. (Rubrics are in Appendix F).

Anchor papers. Writing assessment trainers use student writing samples to illustrate the abstractions in the rubric. The anchors provide concrete examples of "how much, qualitatively and quantitatively, is enough." for each of the performance levels. Scoring rubrics are not meant to stand alone. Rubrics and writing samples are anchored to each other, typically in a benchmarking or range finding process. Because this was a research study and not an operational assessment that would be repeated annually, the range finding process was informal. Anchors were selected for each level and feature in the holistic and the analytic rubrics. They were deliberately chosen from the two certificate areas and several grade levels so that scorers would learn to apply the rubrics in the same fashion to papers that "looked different." The exception, as in the work samples, was Extended Abstract, as we did not locate examples of this level in range finding. Examples of some of the anchor papers are included in the Exemplar Writing Samples section of the report.

Scoring of the Writing Samples

Participants. Given the relative simplicity of scoring the student writing samples, only five people played the various roles of trainer and scorers. The principal investigator worked with the writing assessment director to determine the holistic SOLO ratings, using papers from the pilot and operational study. The writing assessment director selected the papers to be used during training. Three experienced, expert raters from the full-time staff of Test Scoring and Reporting Services (TSARS) of the University of Georgia completed the training and rating process. One of the scorers has worked for five years while the other two have worked at TSARS for over ten years and manage the writing assessment center for five scoring projects annually. TSARS itself has over fifteen years of experience in the development and scoring of writing tests at three grade levels and in the development and training for high school science and social studies essay assessments. The writing assessment director, who directs the writing assessment division of TSARS, served as the trainer and supervised resolutions for the pilot and operational studies.

Training of scorers. After a brief overview of the study, scorers were first trained to apply the holistic SOLO rubric so that they would shift from familiar writing rubrics to the new conceptual framework. They were informed that we did not and would not know the certification status of any of the teacher participants. For this portion of the study, each individual writing sample was to be evaluated, not the class set of papers. In addition to reading and discussing the rationale for the SOLO levels of the prescored student papers, training included a discussion of how these performance standards differed from the decisions and procedures for state assessments. In statewide assessments, reporting as much information as possible about a single sample is often a goal. This goal means that students are allowed great latitude in how they

respond to a prompt. In this study however, if a student did not follow the task, the writing sample received a Prestructural rating. For example, in the task to convince classmates about an important animal, papers that identified an important animal followed only by instructions for animal care received a Prestructural rating. There responses would have received higher ratings in some state writing programs.

The first training session introduced scorers to four of the five SOLO levels with 12 anchor papers from two of the three prompts. The Extended Abstract level was presented by explaining what a high Relational paper would have contained to move it to the next level. The second phase of training on the five feature analytic writing rubric introduced twenty additional papers from all three prompts and both younger and older writers. The twenty-paper set was constructed to provide additional practice in the holistic rubric and to introduce the analytic rubric. These papers covered all but the Extended Abstract level and varying degrees of quality within levels (such as solid, low, or high Multistructural). Once the scorers had rated and discussed the first ten papers, the next papers were read in sets of five and rated independently. Ratings and the rationale for them were discussed in a group setting. The training process took 12 hours.

Scoring Logistics

Procedures. The student writing samples were kept in teacher sets. Scorers were given photocopies of these papers, not the student test booklets that identified the student. Scorers rated the papers with no knowledge of teacher, certification status, grade level, or state. Papers were kept in the original order (as the teacher submitted them), which meant that sometimes papers alternated topics and sometimes several papers in a row would be on the same prompt topic before switching to the second one and then back again to the first.

Each student paper was read independently by two of the scorers. Pairings were rotated so that some class sets were read by scorers 1 and 2, some by 1 and 3, and some by 2 and 3. Sets were assigned to ensure that scorers read papers from both the MC/Gen and EA/ELA candidates. Each scorer's accuracy was checked through back-reads, with either a sampling of papers within a class set or the entire set rated by the writing assessment director. Training was ongoing, with the writing assessment director providing corrective feedback through discussion with the scorer. Scoring was completed in five days.

Scorers first completed the holistic rating for a paper and then the five feature analytic ratings. Augmentation, as described in the scoring of work samples, was used to indicate scores that did not fit solidly in the rubric levels (Penny, et al., 2000a, 2000b). Working with highly-experienced scorers made it possible for them to indicate the numerical value of the augmented score during the rating process rather than conversions applied once the rating process was complete. It was not until the end of the study that we learned that the system for assigning numerical values to writing samples differed from the scale applied to the student work samples. The relative value is the same. Augmented values for the writing assessment are included in Table 4.17.

Table 4.17 *Augmented Values for Writing Sample Scores*

	Low	**Solid**	**High**
Prestructural	0	0	0
Unistructural	.75	1.00	1.25
Multistructural	1.25	2.00	2.25
Relational	2.75	3.00	3.00
Extended Abstract	3.75	4.00	4.25

Monitoring and resolutions. Agreement was defined as less than a .50 difference between the first two scores. Resolutions were completed using the discussion method (Johnson,

Penny & Gordon, 2000; Johnson, Penny & Gordon, in press). This method was selected so that the resolution process served the dual purpose of additional training. In the few instances that discussion did not lead to resolution, the writing assessment director provided a third reading. The third score replaced either the first or second score.

Description of Data Analyses

In a comparative study, detection of the difference between two groups, if it exists, relies on unbiased data collection and proper statistical analyses. Data collection and analyses in this study were designed to assess Certified and non-Certified teachers on two major dimensions of teacher quality and performance: comparative teaching practices and comparative teaching outcomes. In this study, a randomization mechanism was adopted to ensure a balanced representation of Certified and non-Certified participants. The probabilistic sampling also included consideration of categorical factors, such as gender and race, through a post-stratification procedure to ensure a proportion match between the sample and the population on those dimensions. Besides the National Board assessment scores and the related demographic factors, empirical data were gathered from multiple sources to reflect school characteristics, teaching quality, and student performance. The school variables provide a contextual description of the instructional setting. The teacher and student data were collected to assess the instructional quality, as evidenced by qualitative and quantitative information for research triangulation. The instructional quality indexes were articulated with the contextual factors, as well as the National Board Certification outcome. A multilevel analysis of student learning outcomes justified the need for having the contextual examination.

The National Board Certification outcome is dichotomous, Certified or non-Certified. Whereas this categorical measure directly represents the Certification decision, scores of the

participants were also analyzed to assess the performance on an interval scale. The interval scale is suitable for parametric data analyses, and the categorical data can be handled through non-parametric methods. Intuitively, the score analysis seems more attractive because of the stronger power associated with parametric methods to detect significant difference. On the other hand, validation of the National Board Certification is not confined on the checking of assessment scores. Validation also hinges on a proper setting of the threshold level that defines the dichotomous Certified and non-Certified categories. In this regard, non-parametric methods should also be considered because the categorical measure presents a more comprehensive measure of the Certification outcome.

A comprehensive data analysis plan including both qualitative and quantitative methods was developed and implemented for this study. While similar to the research crosswalk, this plan addresses the method of analysis for each research question. Table 4.18 represents the data analysis methods.

Table 4.18 *Data Analysis Methods*

Research Question	**Data sources**	**Qualitative Method(s) of Analysis**	**Quantitative Method(s) of Analysis**
Do students taught by National Board Certified teachers produce deeper responses than students of teachers who attempted National Board Certification but were not Certified?	• Student Work Samples • Standardized Writing Assessment (MC/Gen and EA/ELA participants)	• Content analysis and evaluation using the Student SOLO Taxonomy Rubric • Development of illustrative student exemplars o Work samples o Writing samples	• Multilevel analysis of the student outcomes to partition the variances at student and teacher levels • Discriminant function analysis to test the degree of separation between Certified and non-Certified teachers • Factor analysis of student writing samples to identify

			a latent variable of student writing performance • Post hoc analyses of the work setting difference between Certified and non-Certified teachers • Triangulation of the results of student writing and teacher certification status with assessment of the student work samples using both parametric and non-parametric methods. • Inter-rater reliability checking on the holistic Student SOLO ratings, as well as the holistic and analytic values from Student Writing Assessment
Do National Board Certified teachers structure assignments designed to produce deeper responses than teachers who attempted National Board Certification but were not Certified?	• Unit Context Responses • Profiles of Instruction • Profiles of Student Work Samples • Student Work Samples	• Content analysis and evaluation using the Teacher SOLO Taxonomy Rubric • Development of illustrative teacher exemplar	• Parametric statistical testing on the SOLO score difference between Certified and non-Certified teachers • Parametric analysis of the National Board scores between candidates who have shown surface instruction and those who demonstrated deep instruction • Non-parametric analysis on the association between

			the Certification outcome (Certified vs. non-Certified) and the depth of instruction • Reliability checking on the SOLO scoring

Though the qualitative analysis of this study will continue for months and perhaps years, the completed data analyses of the work samples includes content analyses of each case on the two dimensions of comparative teaching practices and comparative teaching outcomes. After the raters used content analyses to evaluate the cases, the quantitative analyses ensued. Categorical evaluations (e.g., surface or deeper; prestructural, unistructural, multistructural, relational or extended abstract) of the cases were translated to numerical values for the quantitative phase of the data analysis. A similar process was used to assign nominal and then numerical values for measures of writing performance. Inter-rater reliability indexes for the writing scores were checked through correlation analyses.

To compare the effectiveness of Certified and non-Certified teachers in a classroom setting, student learning outcomes were analyzed to partition their variances at the class and student levels. The SAS and SPSS software packages were employed in multilevel analyses, factor analyses, and discriminant function analyses of these multiple learning outcomes. The research design not only fits the natural structure of the existing school setting in which students are nested within classes, but also facilitates assessment of the need to include contextual variables at the teacher or classroom level in disentangling explanation of the learning outcomes. The contextual teaching factors investigated in this project included characteristics of school locales, resources, and sizes.

In addition, the teaching effectiveness was linked to the type of student enrollments. To characterize the background of students served by participants, quantitative measures were constructed to reflect an *overall social economic status* (SES) and *race diversity* (RD) of the school population, and parametric statistical methods, including the *t*-test and/or *F*-test, were employed to examine the diversity indexes of the student population taught by Certified and non-Certified teachers. Meanwhile, teachers were examined on their instructional quality as reflected in the SOLO scores and augmented SOLO scores. Gender differences in those teaching outcomes was subjected to statistical testing. More importantly, an independent sample *t*-test was used to detect differences in the National Board certification scores among teachers who were categorized in an either "deep" or "surface" category from the SOLO assessment.

These parametric tests were grounded on a comparison of average scores on an interval or ratio scale. Whereas the parametric approach stronger statistical power to detect significant difference, the validation study should be focused on the dichotomous certification outcomes, i.e., "Certified" or "non-Certified," which is on a nominal scale. Thus, a more straightforward approach is to examine the categorical data using a non-parametric method. In contrast, a disadvantage from the non-parametric analysis is its lack of sensitivity to detect statistical differences in comparison to its parametric counterpart. Due to the pros and cons on each side, both parametric and non-parametric methods (such as the chi-square test) were employed in this investigation to triangulate research findings important to the National Board certification. Design of the data analyses is articulated as an indispensable component of the overall quantitative research strategy delineated in Figure 4.3 below.

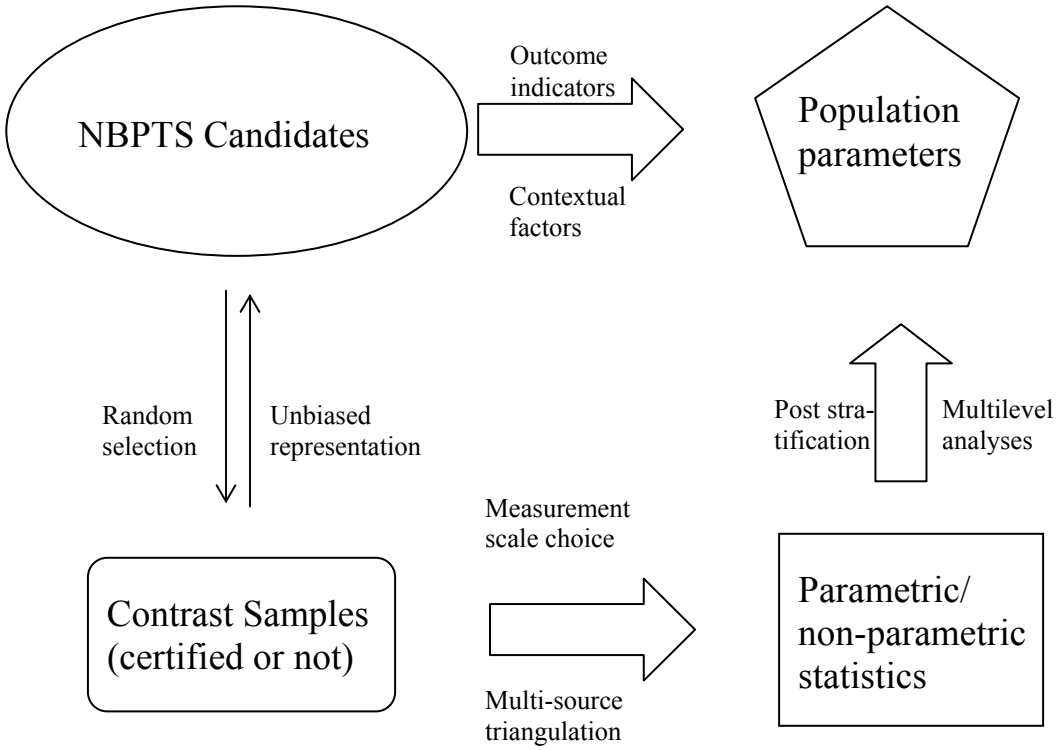

Figure 4.3 *Overview of the Project Components Methodology for Statistical Inference*

RESULTS

The question of whether National Board Certified Teachers (NBCTs) and their non-Certified counterparts can be distinguished from each other based on the quality of their teaching and the learning of their students has become a most compelling question for policy makers, teachers, teacher educators, school administrators, and the business community. This chapter describes the results from analyses of teacher planning documents and assignments and student work samples produced in the course of teacher-participants' regular curriculum, as well as the outcomes from a standardized writing assessment administered in the MC/Gen and EA/ELA teacher participants' classrooms.

The National Board for Professional Teaching Standards (NBPTS) delineates a system of advanced, voluntary certification for K-12 teachers. In this validation study, various and multiple sources of data were collected and analyzed to address the following questions:

- *Comparative Teaching Outcomes.* Do students taught by National Board Certified teachers produce deeper responses (to class assignments and standardized writing assessments) than students of teachers who attempted National Board Certification but were not Certified?

- *Comparative Teaching Practices.* Do National Board Certified teachers develop instruction and class assignments designed to produce deeper student responses than teachers who attempted National Board Certification but were not Certified?

The statistical investigation focused on an examination of differences between Certified and non-Certified teachers in their instructional designs and students' responses. Teachers who have sought National Board certification in four different certificate areas were randomly selected and then recruited to participate. The Certification status is designated into Certified and non-Certified categories, which, at the time of the Certification decision, involved a comparison between the assessment score and a threshold level (cut-off score). Given the importance of the Certification decision, parametric and non-parametric data analyses were conducted on the

assessment scores and the certification status to triangulate differences in the teaching practices and outcomes of Certified and non-Certified teachers. Before the findings from these analyses are discussed, it is important to understand the results of the recruitment and participation methods employed.

This chapter is organized into three major sections. Section I provides information about recruitment efforts and participation. Section II provides results of the analyses related to Comparative Teaching Outcomes. Section III provides results of the analyses related to Comparative Teaching Outcomes.

Section I

Recruitment and Participation

Recruitment Results

The original research design called for two hundred participants – fifty in each of the four certificate areas; however, we did not secure two hundred participants for the operational phase of the study. During the recruitment phase of the study, the team attempted over two hundred phone calls to teachers for recruitment purposes. Initially, 202 teachers verbally agreed to participate in the study. Written descriptions of the requirements and participation agreement packets were sent to these teachers. Although 159 teachers returned agreement packets, participant mortality rates were high. Teachers cited lack of time, problems with administrator permission to participate, professional and family commitments, lack of response or consent from students' parents, and the time-consuming requirements of participation as reasons for non-completion. Reminder emails were sent to the 136 participants who agreed to participate but had not returned materials or contacted us to indicate that they would not complete the study. Table 5.1 shows information related to recruitment efforts and participation in the study.

Table 5.1. *Recruitment Efforts and Participation*

	Surveys Mailed	Returned Surveys	Eligible	Recruited	Agreed to Participate	Boxes Sent	Dropped Out	Expected Return	Materials Returned
EA/ELA	178	65	43	41	31	30	7	23	11
MC/Gen	156	56	30	28	19	19	3	16	9
AYA/Science	252	125	92	87	74	72	4	68	34
AYA/SS-H	125	53	48	46	35	34	5	29	10
Total	711	299	213	202	159	155	19	136	64

Characteristics of the Participants

Certification Status

Participants were recruited from across the United States in four certificate areas. A total of 64 teachers from 17 states participated in the study. All participants had attempted Certification in one of the four certificates identified. Thirty-five (55%) of the participants had achieved National Board Certification, and 29 (45%) had attempted but had not achieved National Board Certification. Table 5.2 shows the number and percentage of participants representing each certificate area, including their Certification status.

Table 5.2 *Participants by Certificate and Certification Status*

Certificate	Total Number Participants	Number Certified	% Certified	Number non-Certified	% non-Certified
MC/Gen	9	5	55.6%	4	44.4%
EA/ELA	11	6	54.5%	5	45.5%
AYA/Science	34	18	52.9%	16	47.1%
AYA/SS-H	10	6	60.0%	4	40.0%

Race, Sex, and Certification Status

Examining information about race, sex, and Certification status related to participation is interesting. Table 5.3 shows the race, sex, and Certification status of participants by certificate area. Though the sample drawn included 98 non-whites (approximately 13.8% of the sample) and 68 people for whom race was not known (approximately 10% of the sample), all operational participants for whom race information was known were white. We cannot assume that non-white participation was random; however, we do not have data to determine why only white teachers completed the requirements for the study. We will return briefly to this issue in the Discussion and Conclusions Chapter.

Table 5.3 *Participation by Race, Sex, Certification Status*

	Race		**Sex**		**Certification**			**Certification Percentages**	
Certificate	White	Unknown	Male	Female	Certified	non-Certified	Totals	Certified	non-Certified
EA/ELA	11	0	2	9	6	5	11	54.5%	45.5%
MC/Gen	9	0	2	7	5	4	9	55.6%	44.4%
AYA/Science	34	0	12	22	18	16	34	52.9%	47.1%
AYA/SS-H	6	4	3	7	6	4	10	60.0%	40.0%
Total	60	4	19	45	35	29	64	54.7%	45.3%

The distribution of sex and Certification was more varied. Among various demographic variables of teachers, gender was a factor examined in this investigation because the sampling process assured a proper match between the sample and the population through a process of post-stratification. At the teacher level, gender differences were examined on the National Board assessment scores and the SOLO scores from this validation study to describe a deep or surface approach to instruction. The statistical analysis shows no significant gender difference in the National Board score [$t(61) = 0.52$, $p = 0.60 > 0.05$], the SOLO score [$t(62) = 0.48$, $p = 0.63 > 0.05$], or the augmented SOLO score [$t(62) = 0.55$, $p = 0.58 > 0.05$].

The *t*-test is part of the parametric statistical method that is more sensitive to detect significant differences, and thus, becomes a more preferred choice than its non-parametric counterpart. Parametric methods require assessment of the outcome measures, such as the National Board scores, on an interval or ratio scale. Whereas the scores are represented as continuous data, "Certified" or "non-Certified" is a dichotomous decision from the National Board assessment. Validation of the National Board Certification directly hinges on the final decision that involves both National Board scores and the threshold line that converts the continuous scale into a dichotomous decision. Therefore, a more straightforward approach is to examine the categorical data using a non-parametric method.

Results of the non-parametric data analysis are based on examination of a series of contingency tables. In this study, categorical data can be presented in the following table of frequencies from classifications of the gender and Certification status.

Table 5.4 *Frequencies of Teachers Across Gender and Certification Status*

	Female	**Male**
Certified	22	13
Non-Certified	22	7

A χ^2 test shows a non-significant association between gender and Certification status: χ^2 (1) = 1.25, p = 0.26 > 0.05.

The following table contains a frequency distribution on gender and the SOLO instruction (deep vs. surface) dimensions.

Table 5.5 *Frequencies of Teachers Across Gender and SOLO Categorization*

	Female	**Male**
Deep	17	6
Surface	27	14

A χ^2 test shows a non-significant association between those dimensions: χ^2 (1) = 0.45, p = 0.50 > 0.05.

Section II

Comparative Teaching Outcomes: Students' Depth of Understanding

Two data sources provided evidence for assessing students' depth of understanding. Student work samples were scored based on the Student SOLO Rubric. Additionally, writing assessments were scored holistically with the Student SOLO Rubric and for feature analysis with guidelines based on the SOLO Taxonomy. Scorers had no knowledge about the participants' Certification status. Accordingly, independent collection of the needed student and teacher data was essential to this investigation.

Student Work Samples

One measure of the comparative teaching outcomes involved an examination of student work samples from six-randomly selected students in each teacher's classroom. The preponderance of evidence from the six students' work was used as a summary representation of the outcomes in the teacher's class. Teachers were asked to collect all unit-related work from these six students to submit with their materials. For this measure, two independent content-expert raters assessed the depth of student understanding of unit concepts. The depth of student outcomes was scored separately from the depth of teachers' instructional aims and design (to be discussed later).

Quantitative Analysis of Student Work Samples

From the evaluation of student work samples, we learned that the outcomes in most (78%) of the teachers' classrooms, regardless of Certification status, were at the surface level. However, students of NBCTs were almost twice as likely to achieve deeper learning outcomes

(Certified: 29%; Non-Certified: 14%). Table 5.6 shows the number and percentage of Certified and non-Certified teachers whose collection of student work samples were scored as "surface" and "deep."

Table 5.6 *Student Classroom Learning Outcomes and Teacher Certification Status*

	Surface	**Deep**
Certified	25 (71%)	10 (29%)
non-Certified	25 (86%)	4 (14%)
Total	50 (78%)	14 (22%)

While students in classrooms with Certified teachers demonstrated deeper responses more often than students in classrooms with non-Certified teachers, results of the statistical analysis show no statistically significant association between the depth of student understanding and the teacher certification status: χ^2 (1) = 2.03, p = 0.15 > 0.05. In addition, the student level of understanding (i.e., deep vs. surface) is not significantly linked to teachers' National Board assessment scores: t (61) = 1.42, p = 0.16 > 0.05.

The depth of student understanding was scaled on a SOLO and an augmented SOLO scale. The correlation coefficient of 0.19 was insignificant (p = 0.13 > 0.05) between the Certification score and the SOLO score. The correlation coefficient between the augmented SOLO score and the Certification was 0.15, which was not significant (p = 0.25 > 0.05) either.

Qualitative Analysis of Student Work Samples

While student outcomes were not found to be statistically significant, the descriptive data are very useful to inform educators about the depth of student learning achieved over the course of the instructional unit. The raters observed that in many cases students "could not possibly" achieve deeper outcomes because the instructional design did not foster such outcomes.

To illustrate the nuances of the comparative teaching practices and comparative teaching outcomes evaluations, the research team developed a series of exemplars that provide descriptions of the teacher practices and student performances as well as corresponding commentary to describe how the data fit into the evaluation category. For the comparative teaching outcomes exemplar, Case 010133 (Adventure with Will Hobbs) was profiled because it provides a model of multistructural student outcomes. Though the exemplar is lengthy, it is helpful in communicating the "surface" constructs; therefore, it is included in its entirety here.

Comparative teaching outcomes exemplar. The following information and descriptions are provided to clarify the SOLO Taxonomy evaluations related to Comparative Teaching Outcomes. The primary data source for the evaluation focusing on the depth of student learning was the student work samples. When evaluating a case for depth of student learning, scorers *focused* on the student work samples and used the Unit Context responses, Profile of Instruction information, and Profile of Work Samples information to interpret student work samples (e.g., what assignment was given, what resources were provided). Descriptions and commentary of all data sources are included in the exemplar here because they provide information about the context of the teaching in which the student learning took place. It is important to note that scorers did not know the certification status of the teachers as they were assigning scores to the cases.

Case 010133: Adventure with Will Hobbs
Data Source: Unit Context

Description of teacher practice	**Commentary related to SOLO evaluation**
"Adventure with Will Hobbs" is the title of the unit designed by Ms. Harrison (a pseudonym). Per the instructions, she described her eighth grade language arts class based on their learning-related characteristics. The students in	

her class "range from being identified as talented and gifted to learning disabled and everywhere in between." Ms. Harrison reported that reading abilities vary from second grade to college level. A special education teacher teaches with Ms. Harrison "when she can," or pulls those students from class for extra help.

As part of the research design and data collection request, Ms. Harrison described her major beliefs about teaching. She reported that every student in a class is a member of the team and can make a valuable contribution. The team works to "improve weaknesses and build strengths." The main influences on instructional planning were the state-mandated curriculum and the interests of students.

In describing the unit of instruction, Ms. Harrison indicated that heterogeneously grouped "book clubs" would read 5 different novels written by Will Hobbs, write discussion questions, work with vocabulary, make personal connections to the reading, and complete other activities related to their respective novels. She stated that the unit would span three weeks. At the end of the unit, the group members will mix with other groups to find common themes and lessons within the writing of Will Hobbs. Ms. Harrison listed eight goals for the unit. These goals included the following:

- defining unknown words through context clues and author's use of comparison, contrast, and cause and effect,
- determining the meanings and pronunciations of unknown words by using dictionaries, thesauruses, glossaries, technology and textual features such as footnotes or sidebars,
- responding to literal, inferential, evaluative, and synthesizing questions to demonstrate comprehension of grade-appropriate texts,
- applying reading comprehension strategies to include making

Ms. Harrison's stated goals that require students to define, determine, and answer questions could limit student outcomes to recall of information, even if students are required to recall more than one detail. For example, Ms. Harrison wanted students to "define unknown words through context clues," answer various levels of questions, and identify examples of foreshadowing and flashback. These goals correspond to the multistructural level of response on the Student SOLO Rubric: "Student response focuses on multiple details of the tasks or sequence of tasks serially without relating them."

In addition, Ms. Harrison indicated that students would identify and explain universal themes across different works by the same author. Although the aim of this goal seems consistent with the relational level of the SOLO Taxonomy where "deep or transforming levels reflect an understanding gained by relating to the task in a way that links up with existing knowledge or that is personally meaningful," a scorer evaluating this case would need to examine the students' work samples to determine if this integration

predictions, comparing and contrasting, recalling and summarizing, making inferences and drawing conclusions, • identifying and explaining various types of characters and how their interactions and conflicts affect the plot, • comparing and contrasting different points of view, • identifying and explaining universal themes across different works by the same author, and • identifying examples of foreshadowing and flashback in a literary text.	of information into a more complex combination is supported by students' subsequent responses.

Data Sources: Profile of Instruction, Profile of Student Work Samples, and Student Work Samples

Description of teacher practices	**Description of student work samples**	**Commentary related to SOLO evaluation**
Ms. Harrison submitted "Profile of Instruction" forms for seventeen lessons in this unit. During the course of the unit, Ms. Harrison indicated that students read their novels and completed assignments in "workshop" format. She also reported using lecture, direct instruction, and assessments. The materials used in this unit of instruction were teacher-made handouts. <u>Day 1</u> On the first day of this unit, Ms. Harrison's students discussed the types/genres of literature covered during junior high. Ms. Harrison introduced the works of Will Hobbs. She then read summaries from the five books to be used in this unit. Afterwards, students were asked to select their book preferences.	Since Ms. Harrison was introducing the novels, no student work samples were provided for this day of instruction.	This activity provided students a brief summary of each book and allowed students to decide which book they would like to read. In terms of learning, Ms. Harrison stated that her students would "learn a bit about the genre of adventure stories." Here, scorers would look for evidence to support the study of genre in student work samples.

Day 2 On day two, Ms. Harrison reported that students were assigned to their book clubs and given a procedure page along with the recording sheets they were to complete as they read. Students were then instructed to begin reading. Ms. Harrison stated that her students would learn new vocabulary, create discussion questions, and learn about their books. Students were provided teacher-made handouts that provided space for students to record three questions and three vocabulary words from each chapter. One section of the handout provided students with space to copy a passage from the book and tell why they selected that passage. The bottom portion of the handout provided space for an illustration of the passage. A third handout asked students to tell where they made a personal connection to the novel and explain why.	No student work samples were included for this day of instruction.	While no student work samples were included for Day 2, the work done in class was used by students to complete subsequent assignments. Work completed on this day may be evidenced in data collection materials provided on a later date. Scorers would look for evidence that students have used the reading strategies listed as a goal in the Unit Context to make meaningful connections with the text.
Day 3 On the third day of instruction, Ms. Harrison asked students to begin the reading assignment, work on their questions, and work on vocabulary. As a result, she expected them to increase their reading skills, to learn new vocabulary, and to learn cooperative skills within their groups.	No student work samples were included for this day of instruction.	For days when no work samples are included, scorers would determine if work completed is evidenced in data collection materials provided on a later date. Here, scorers would note that the questions students were asked to write and their work on vocabulary were consistent with the surface level.
Day 4 On the fourth day of instruction, Ms. Harrison reported giving the students an opportunity to ask	No student work samples were included for this day of instruction.	

about events in their book. Then, students were directed to read, write questions, and work on vocabulary. As a result of the instruction, Ms. Harrison expected students to "read, write questions, identify and define vocabulary, ask clarifying questions and work cooperatively." She expected students to learn "better reading skills, to learn new vocabulary and to learn cooperative skills within their group." <u>Day 5</u> On the fifth day of instruction, Ms. Harrison stated that students were given the exercise and note sheets for a point of view lesson. Students were to write definitions for a variety of terms such as narrator, point of view, first person point of view and third person point of view (to include omniscient third person point of view and limited third person point of view). Definitions for these words as well as key understandings related to point of view were to be found in their literature books. On the note sheets students recorded the point of view used in the Will Hobbs selection as well as the point of view identified in five teacher-selected poems. They were asked to justify their selection for each. Students were instructed to work within their book clubs to complete the exercises by determining the best definitions and answers for each question on the note sheets. As a result of the instruction, Ms. Harrison expected students to work	Student work samples included the definition of terms and the application of those terms to their novel. In the final work samples, students showed evidence of identifying the point of view for five poems they read and justifying their answers.	Scorers would note that work samples elicited a simple recall of information. Even though students determined the point of view for the novel they were reading, students used the definitions as discrete rather than interconnected pieces of information. Students were asked to rewrite two or three sentences from their novels using a different point of view. Again, scorers would note that students did not attempt to make sophisticated connections, such as how point of view alters the reader's experience. Finally, students were asked to identify the point of view for each of the poems read and justify their answer. Student responses focused on identifying the point of view of each poem separately. The outcomes were consistent with the Multistructural level of the Student SOLO Rubric: "Students use details and/or skills and can make decisions based on details. However, need for closure produces decisions

cooperatively helping those who had difficulties. Students were to discuss material and arrive at a consensus. Ms. Harrison expected students to learn the definitions of vocabulary terms and be able to recognize examples of each. In addition, students were expected to learn the point of view of their particular novel. <u>Day 6</u> Students participated in their first book discussion on the sixth day of instruction. Ms. Harrison gave students a number to indicate their order of participation. After the group's discussion, students were to complete a worksheet focusing on flashback, foreshadowing, suspense, and prediction. Ms. Harrison directed students to "ask each other questions over the reading, discuss the answers, share vocabulary words and definitions, read their favorite passage aloud, share their illustrations, and share a personal connection" to the text. Students were asked to define foreshadowing, flashbacks, and suspense and to record examples of each from their novels. In addition, students were directed to make predictions about the story. Ms. Harrison stated that students were expected to understand their story to this point and to learn the definitions of foreshadowing, flashback, and suspense. Ms. Harrison also expected students to identify these techniques in literature.	Student work samples included a worksheet packet with three assignments. In the first section, students developed questions and vocabulary words for each chapter of the text. In the second section, students copied a passage from the text to share with the group, justified why they chose the passage, and drew an illustration about the passage. In addition, they identified and explained a personal connection they made with the story. In the third section of the worksheet packet, students wrote definitions for foreshadowing, flashback, and suspense and located examples from text.	that use several discrete pieces of information without interrelating them." Students were asked to explain where they made a personal connection with the story. Students named the event and told why they connected to this event. Scorers would note that while students made connections with the text to their daily life, these connections did not involve "a sophisticated and/or novel understanding" which could lead to a relational level (Student SOLO Rubric). In these work samples, students developed statements about how an incident in the story reminded them of an incident in their own life. This information was not used in a broader conceptual context. Finally, students were asked to complete an activity focusing on literary terms and the skill of predicting. Responses from these work samples demonstrated an understanding of the connection between point of view, the literary terms studied (flashback, foreshadow, suspense), and adventure as a genre of literature. Student responses indicated a

		multistructural level of performance because students focused on multiple details serially without relating them (Student SOLO Rubric) as they selected examples of flashback, foreshadow, and suspense from their novels.
<u>Day 7</u> Ms. Harrison reported that students were given a quiz to complete. Students wrote definitions for a variety of terms related to point of view, identified the point of view of a passage and justified their choice, and answered approximately 10 comprehension questions related to their novel. Students were then instructed to begin the next reading assignment for their novels. As a result of the instruction, Ms. Harrison expected students to provide her with information regarding their understanding of their novels.	No student work samples were included for this day of instruction.	If student responses had been available, scorers would determine if students were beginning to develop sophisticated connections and understandings related to the ideas and concepts in the novel.
<u>Day 8</u> Ms. Harrison instructed students to use class time to work on a reading assignment based on their novel. Students were directed to write and discuss questions related to various chapters; identify and define vocabulary; select and illustrate a favorite passage; and make a personal connection to the story. Students worked as a group, in pairs, or individually and were asked to help each other with challenging material. As a result of the instruction, Ms. Harrison expected students to "improve their reading skills, to learn new	No student work samples were included for this day of instruction.	Ms. Harrison's stated goals seem to focus more on skills than conceptual understandings. It may be difficult for students to achieve deeper understanding if the teacher's instructional design does not foster higher level outcomes.

vocabulary and improve their cooperative skills." <u>Day 9</u> Ms. Harrison reported that students were given the same directions as in the previous lesson. As students were working, she met individually with students to check their progress and answer questions. Students were asked to bring all of their materials to the table for their progress check. (Class time was shortened due to standardized testing.) <u>Day 10</u> Ms. Harrison reviewed the purpose of book club meetings and explained how to use a documentation sheet. Ms. Harrison reported that because groups had gotten off-task and had trouble staying focused, each group would choose a secretary to record their discussion on the documentation sheet. Groups would then submit their record sheet at the end of class. For this instructional session, students were expected to complete four activities: (a) ask each other questions, discuss their answers and make a judgment on the quality of questions written by the group members, (b) discuss vocabulary terms and determine the most difficult words from the passage, (c) share their favorite passages and illustrations and determine a high quality illustration, and, (d) share personal connections with the story and identify a "quality" connection. As a result of the	No student work samples were included for this day of instruction. No student work samples were included for this day of instruction.	

instruction, Ms. Harrison expected students to have a complete understanding of their novel up until this point, to learn what makes a good discussion question, and to learn at least five difficult vocabulary terms and their meanings. <u>Day 11</u> Ms. Harrison instructed students to write quality definitions for the terms conflict, internal conflict, external conflict, and resolution. She provided handouts for this activity. Students worked alone or together to define the terms, but worked alone when writing paragraphs applying these literary techniques to their novels. Students retained notes (definitions) to study for an upcoming test. Ms. Harrison expected students to construct quality definitions and then apply those terms by writing paragraphs about the conflicts from their books. Students were expected to learn the definitions of a variety of terms and how those terms applied to their novels.	Each student work sample included two to three written paragraphs identifying and explaining the various conflicts from their novels. The first paragraph in each response identified an internal conflict from the novel and explained why the conflict was internal. The second paragraph identified an external conflict and explained why the conflict was external. The third paragraph identified another form of external conflict (man vs. man, nature, or self) and explained why the conflict was man vs. man, nature, or self.	Student responses demonstrated the use of working memory (the newly acquired definitions for terms related to conflict) to make decisions about their book. Responses used several details from their novel, but the facts were not interrelated. This is consistent with the multistructural level of the SOLO Taxonomy. Connections to the universality of conflict to create tension in literature may have moved these responses into the relational level. However, in these work samples, students merely categorized details from their novels into predetermined conflict categories.
<u>Day 12</u> Ms. Harrison reported that a quiz was given during this class period. Students were instructed to define various terms focusing on point of view and conflict and "enhance, revise, edit, and add at least one quote" to a paragraph written during the previous day's lesson. Students were instructed to work quietly on an assignment	Student work samples included the book club quiz. More specifically, they included the definitions for the following literary terms: point of view, first person, third person limited, third person omniscient, internal	For the first portion of the quiz, student work samples suggested the reproduction of memorized definitions. Identifying the internal and external conflict of a novel is consistent with the multistructural level of the SOLO Taxonomy because students were applying knowledge gained from

following the quiz. Ms. Harrison expected students to demonstrate their knowledge of important literary terms and to create a cohesive paragraph during this instructional session. As a result of instruction, she hoped that students would learn what terms they needed to review and to determine details that would support their topic sentences.	conflict, external conflict, and resolution.	instruction and making decisions based on this information. However, there was no student attempt to interrelate this information with other knowledge gained over the course of the unit. The second portion of the quiz required students to select a paragraph from the previous day's work to revise and produce a final copy. Student work samples included a quotation that supported the conflict they had identified from their novels. Scorers would note that student work samples "reflected a relationship in terms of a few limited details" as students identified a conflict, labeled the conflict (man vs. man, nature, self), and selected a quotation to support the identified conflict. This is consistent with the multistructural level of the SOLO Taxonomy as students were looking at the relationship of limited and independent details. At the relational level of the SOLO Taxonomy students would begin to generalize within the context of the novel or within the context of Will Hobbs's works.
<u>Day 13</u> On day thirteen of this unit, Ms. Harrison instructed students to work on their third and final reading assignment (handouts provided). Students were allowed to choose their own "work configurations." Ms. Harrison expected "students to read, write discussion questions, identify and define vocabulary, select a	No student work samples were included for this day of instruction.	

favorite passage to share and illustrate, and make a personal connection with the story." Students were expected to learn new vocabulary and improve reading and cooperative learning skills. <u>Day 14</u> The topic for the fourteenth day of this unit was "Character, Mood, Tone." Ms. Harrison provided handouts with definitions and graphic organizers to help students examine flat, round, dynamic, and static characters. The resources also provided students with the definitions for key terms such as mood and tone. Students then answered questions in relation to this material. For this lesson, students worked in cooperative groups to define terms and apply those terms to the characters in their particular novels. Ms. Harrison noted students had "dealt with this material" three times this year. This indicated that students would be using prior knowledge. <u>Day 15</u> The topic for this day of instruction was "Book Club Meetings/Theme." Ms. Harrison stated that students who read the same book title met together for a teacher-directed discussion. Afterwards, students were asked to define theme and record four themes from their book (handouts provided). Ms. Harrison expected students to participate in the large group discussion by either posing questions or answering questions.	No student work samples were included for this day of instruction. Student work samples included a handout with the definition of theme and a list of four themes from their respective novels.	Student work samples showed that students could correctly define a "theme" and list four themes from their novels. Scorers would note that students used rote memorization to give a definition for theme. Definitions were almost exactly alike across student samples. Memorized responses are consistent with surface or reproducing levels of student outcomes on the SOLO Taxonomy. The serial listing of

In addition, students were expected to recognize themes in their respective novels.		themes from the novel evidenced in the student work does not demonstrate personal meaning or connection with any existing knowledge. Students did not organize information into a coherent whole as would occur at the relational level. An integration of point of view, character development, conflict, and literary devices (flashback, foreshadow, suspense) to generate a coherent whole using theme as the concluding purpose would move the student samples to the relational level of student outcomes as would a discussion about Will Hobbs's use of the devises across novels.
<u>Day 16</u> Ms. Harrison reported that students were struggling with point of view. She conducted a whole class review of the concept before handing out a note sheet to use as a reference for point of view. Ms. Harrison instructed students to prepare for an essay test. Students were told that no books would be allowed, but materials they had prepared would be permissible. Ms. Harrison provided students with the three possible essay questions prior to the test day and indicated that a number would be drawn to determine which test question they would answer. Ms. Harrison mentioned that her outcomes for the day were to have students participate in the lecture on point of view and begin preparing for the test. She expected them to learn the concept of point of view.	No student work samples were included for this day of instruction.	

Day 17 Ms. Harrison reported that students turned in their novels before testing. Students selected a number indicating which prompt they would answer. The following prompts were included: • Discuss the use of conflict in your novel. Discuss the three types of conflict and provide one example of each. Discuss each example thoroughly by explaining the events that developed the conflict and how it was resolved. Provide quotes from the book to support your examples. • Discuss the use of foreshadowing, flashbacks and suspense in your novel. Discuss one example of each. Discuss each example thoroughly by explaining the technique and how it enhanced the story. Provide quotes for the book to support your examples. • Discuss the development of your main character in the book. Explain the development in terms of round, flat, static, dynamic. Provide quotes from the story to support your assessments of the character. Discuss the characters choices, personality and changes throughout the story. Ms. Harrison expected students to demonstrate their understanding of the essay question and their novels. She expected students to identify what they had learned from their novel.	The student work samples for this day of instruction included essays focusing on one of the three provided prompts.	While two of the six students in this sample connected some of the details from the novel, including quotations, to the author's technique, the preponderance of the essays exhibited only a recall of information with no connection to the broader concept of literary technique. One student wrote the following: "Gabe also had a man versus man conflict with Johnny and Raymond about leaving camp. He thought it was best to leave camp and find the nearest town, but Johnny thought they could spend the winter as long as they kept in their spot. Gabe disagreed: 'Spend the winter back here? I don't think so. I'd rather take my chances on the river. I want you to come with me, Raymond, but if you stay, no hard feelings.' His conflict was resolved by convincing Raymond and Johnny to come with him down the Nahanni. This trouble Gabe had with the other characters was a man versus man conflict. " In this paragraph the student accurately identified the conflict and identified a passage from the novel that highlighted this conflict. However, each paragraph of the essay is distinct and does not relate to others. In other words, the essay, in its entirety, does not support a broader generalization.

Day 18 The final day of instruction for this unit was titled "Theme." Ms. Harrison noted that students were given worksheets on theme and instructed to complete an exercise by lumping similar themes together. She expected students to examine all five works (book titles) to identify the universal themes in Hobb's work. In addition, students were to analyze their findings and explain these in writing. Ms. Harrison expected students to learn about universal themes and why they are included repeatedly in an author's work.	Student work samples for this day of instruction included completed graphic organizers which included a list of themes for each novel (from a provided list), definitions to key terminology, and 1-2 paragraph explanations of what students had learned about Will Hobbs as an author.	Students relied on memorization to provide this list of themes for each novel. Students used this information to write on the following topic: "Use your completed graphic organizer to write a paragraph or to explain what you have learned about Will Hobbs as an author. Ideas: What themes does he like to include in his work? Why do you think he feels these ideas are important enough to be included in multiple works?" Student work samples listed a variety of themes as portrayed in Hobbs's work and identified a general category to which a specific theme belonged. According to Marzano's Domain of Information (2001) this is a detail skill, which is consistent with surface or reproducing levels of student outcomes as stated by the SOLO Taxonomy. An example of this can be seen in the following student's work sample: "Will Hobbs uses the universal themes adventure, friendship, bravery, and determination. I think he uses these themes often to show the struggles people go through. Will Hobbs uses adventure as a universal theme because where there is adventure there is always suspense and suspense tends to get the reader more interested. Will Hobbs uses bravery to show how courageous these characters were and determination to show how the characters always kept moving. He uses friendship as a

		universal theme to show that friendship was a reward to some characters. Will Hobbs used these universal themes."

Evaluation Summary

The SOLO Taxonomy rating for the collection of student work samples in this case was Multistructural Solid. This case was evaluated by two separate expert scorers before it was labeled. The rater's rationales for the final score assignment were as follows:

- The majority of the work samples showed mostly recalling and reproducing, with multistructural elements. While two students approached the relational level in the essay writing, the preponderance of evidence shows only a recall of information with no broader connections.
- Many student activities involved definition and identification. Although some activities allowed for the possibility of some deeper levels of understanding, in practice that did not happen to any significant degree. When student responses involved multiple concepts, there was very little evidence of students connecting concepts and drawing larger conclusions. There was very little evidence of any meaningfully personal connection with the material that led to a sophisticated and/or novel understanding.

Had the students used the definitions and concepts learned during this unit to generalize about how the writer's style is influenced by the genre and its intended audience, students might have been evaluated as relational. However, students did not use the information gained at the lowest level of learning (vocabulary) to build more sophisticated understandings and applications. Almost all student samples reflected decisions based on details and a need for closure. Though all assignments were related to Will Hobbs' texts, each assignment was separate and distinct. Students did not integrate new knowledge with existing knowledge.

Writing Assessment

An additional source for collecting evidence related to students' learning outcomes was a standardized writing assessment. This assessment was developed by Test Scoring and Reporting Services at the University of Georgia. Based on the NBPTS descriptions of criteria for teachers pursuing Certification in the MC/Gen and EA/ELA areas, the teacher should have primary responsibility for students' writing instruction. Therefore, in this study, teacher participants in the MC/Gen and EA/ELA certificate areas administered this assessment. Teachers in the other

two certificate areas are not typically responsible for delivering the primary writing instruction to students.

Quantitative Analysis of Writing Assessment

Although students of NBCTs did not demonstrate a statistically significantly deeper understanding of classroom curriculum in their *classroom work* (student work samples), student writing samples were gathered to cross-examine the findings. More specifically, the writing samples were assessed on the following dimensions:

- A holistic score using the SOLO Taxonomy Rubric;
- An analytic score for controlling idea using the SOLO-based Writing Rubric;
- An analytic score for organization using the SOLO-based Writing Rubric;
- An analytic score for elaboration of ideas using the SOLO-based Writing Rubric;
- An analytic score for voice using the SOLO-based Writing Rubric; and
- An analytic score for sentence formation using the SOLO-based Writing Rubric.

Data collected during the pilot phase of the study informed the operational administration and analysis of the writing assessment. To ensure a reliable assessment on the writing dimensions, a pilot assessment was conducted in 2003. In the pilot, writing samples were submitted from five teachers and 99 students representing the EA/ELA and MC/Gen certificate areas. This initial phase of the study was conducted to determine whether or not students would produce samples at both surface and deeper levels of the SOLO Taxonomy. If they did not go beyond multistructural responses, we would not have confidence that either the writing prompts or the on-demand context would elicit evidence of depth of learning. The pilot writing samples were read holistically, using the SOLO Taxonomy Rubric. This evaluative phase of the development process was completed by the principal investigator and the scoring consultant.

While students did not produce Extended Abstract responses, they did produce prestructural, unistructural, multistructural, and relational responses. The relational texts were found in the writing samples from one seventh grade and a ninth grade classes in the EA/ELA certificate area. They were also found in the responses from one of the sixth grade classes in the MC/Gen certificate area. Unistructural and Multistructural Level samples were also found from these same classrooms. Only a single Prestructural writing sample was produced in the pilot phase.

After some minor modifications, the operational study was completed in 2004. For each of the writing assessment measures in both the pilot and operational study, two raters were assigned to score the student responses independently, without knowledge of the teacher's Certification status. Consistency of the scoring was assessed by an inter-rater reliability index (Table 5.7).

Table 5.7 *Inter-rater Reliability for Student Writing Assessment Outcomes in 2003 (Pilot) and 2004 (Operational)*

		Analytic Writing Features				
Year	**Holistic SOLO (V1)**	**Controlling Idea (V2)**	**Organization (V3)**	**Elaboration (V4)**	**Voice (V5)**	**Sentence Formation (V6)**
2003	.94	.95	.98	.96	.99	.98
2004	.96	.97	.96	.97	.98	.98

Given the fairly high reliability indices, an average outcome between the two raters on each dimension was employed to assess the deep or surface responses at the student level.

In the operational administration and analysis of the writing assessment, 18 teachers submitted 377 writing assessment responses. Nine Certified and nine non-Certified teachers from the two certificate areas provided the writing samples. Table 5.8 contains the number of students from each classroom who completed writing responses as well as the certificate area and

Certification status of the teachers. The Certified or non-Certified status was not known by scorers at the time the writing responses were rated in either the pilot or the operational study.

Table 5.8 *Certificate, Certification Status, and Number of Responses for Writing Assessments*

Teacher Code	Certificate	Certification Status	Number of Student Responses
030066	MC/Gen	Not Certified	24
030070	MC/Gen	Not Certified	19
030074	MC/Gen	Certified	25
030079	MC/Gen	Certified	23
030093	MC/Gen	Not Certified	25
030107	MC/Gen	Not Certified	12
030113	MC/Gen	Certified	06
030131	MC/Gen	Certified	22
030138	MC/Gen	Certified	20
010003	EA/ELA	Not Certified	21
010019	EA/ELA	Certified	30
010022	EA/ELA	Certified	22
010076	EA/ELA	Not Certified	31
010085	EA/ELA	Not Certified	16
010086	EA/ELA	Certified	20
010106	EA/ELA	Not Certified	19
010133	EA/ELA	Not Certified	18
010142	EA/ELA	Certified	24
Total			377

Multiple *t*-tests, informative as they are, tell only part of the story. To the extent that two variables are correlated, it is to be expected that if two groups differ on one of the variables, the chances are increased that they differ on the other as well. Under these circumstances, an analysis that considers all of the dimensions simultaneously is desirable. Discriminant function analysis (DFA) is a multivariate technique that seeks to linearly combine the information contained in the entire set of variables under study in such as way as to maximally distinguish between groups of interest. In other words, DFA allows one to examine the pattern of characteristics that distinguish one group from the other. To the extent that individuals are correctly classified, the inference that the groups differ on the dimensions taken as a whole is

supported. In this validation study, DFA has been adopted to test the degree of separation between Certified and non-Certified teachers, and the statistical testing shows a significant separation between the Certified and non-Certified teachers [$\chi^2(6) = 22.38$, $p = 0.001 < 0.05$]. This result suggests the appropriateness of using these writing assessment outcomes in differentiating instructional effectiveness between Certified and non-Certified teachers. Accordingly, these writing indicators can be employed to articulate the effectiveness of instruction with teachers' Certification status.

The linkage among the writing assessment outcomes was confirmed through a correlation analysis of a total of 377 student records. The correlation matrix is listed below. Among the six assessment outcome variables, all those correlation coefficients are highly significant ($p < .0001$).

Table 5.9 *Correlation Coefficients for Writing Assessment Variables*

	V1	V2	V3	V4	V5
V2	0.94076				
V3	0.81179	0.83819			
V4	0.81663	0.84944	0.92624		
V5	0.82308	0.83192	0.86354	0.88412	
V6	0.79290	0.82886	0.87677	0.88262	0.90589

Based on the highly correlated writing assessment data, factor analyses have been conducted to identify a latent variable of the writing performance that accounts for a large portion of the assessment information from these six indicators. Eigenvalues from the factor analysis are the same as the variance accounted for by each of the latent factors (Sharma, 1996), and in this study, the six latent factors have their eigenvalues listed in Figure 5.1.

	Eigenvalue	Difference	Proportion	Cumulative
1	5.31053310	4.99987777	0.8851	0.8851
2	0.31065532	0.14998395	0.0518	0.9369
3	0.16067137	0.06970932	0.0268	0.9636
4	0.09096205	0.01632663	0.0152	0.9788
5	0.07463542	0.02209269	0.0124	0.9912
6	0.05254273		0.0088	1.0000

Figure 5.1 *Eigenvalues of the Correlation Matrix*

The results show that the first latent factor, depth of knowledge of writing, accounts for more than 88% of the variance in the six indicators. The squared multiple correlation (SMC) of these indicators with this latent factor is 0.98. According to Sharma (1996), "Squared multiple correlation simply represents the extent to which the variables or indicators are a good measure of a given construct" (p. 120). The high SMC value supports the use of these six indicators to identify the latent variable of student writing performance. To facilitate interpretation of the latent variable, the factor pattern has been presented below.

	Factor1
V1	0.90111
V2	0.92726
V3	0.93492
V4	0.94599
V5	0.93049
V6	0.93106

Figure 5.2 *Factor Pattern of Writing Features*

Because all the indicators have high loadings on the first factor, it is pertinent to use this latent factor to represent student writing skills across all six dimensions. An independent sample *t*-test suggests that students taught by teachers with the National Board Certification demonstrated significantly higher writing performance [$t(375) = 4.90$, $p < 0.0001$]. This result is reconfirmed by multiple *t*-tests on each of the six writing assessment dimensions (Table 5.10).

Table 5.10 *Statistical Testing on Each of the Writing Assessment Dimensions*

Writing Assessment Dimension	**df**	**T**	**p**
V1: Holistic SOLO/Depth of Knowledge of Writing	375	4.49	< 0.0001
V2: Controlling Idea	375	4.40	< 0.0001
V3: Organization	375	4.04	< 0.0001
V4: Elaboration of Ideas	375	4.53	< 0.0001
V5: Voice	375	5.47	< 0.0001
V6: Sentence Formation	375	4.62	< 0.0001

Multilevel Analysis. Under the regular school structure, students are nested within classrooms taught by different teachers. Because of exposure to similar learning opportunities, one may speculate that students who shared the same teacher are more or less similar in these learning outcomes than those taught by different teachers. To disentangle this hierarchical data structure, a multilevel analysis was needed to partition variances of the writing performance between student and teacher levels. The SAS PROC MIXED routine was employed for the multilevel analysis, and Table 5.11 lists the findings from the variance partition.

Table 5.11 *Variance Partition from Multilevel Data Analyses*

Level	**Depth of Knowledge of Writing**
Teacher	2.04
Student	3.09

The results indicate that more than 40% [2.04/(2.04+3.09)] of the variance in student writing performance is distributed at the teacher level. Besides reconfirming the important role teachers play in the student learning processes, this finding suggests the need for incorporating more factors at the teacher level to help explain the variation of student performance. Since the achieved sample of 377 students was nested within a total of 18 teachers**,** the teacher sample is

not large enough to estimate more parameters from the hierarchical structure, and thus, contextual factors must be examined separately using a more simple method of statistical inference.

Contextual Factors of the Teaching Setting

Because the variance partition between Certified and non-Certified teachers on the variable related to depth of knowledge of writing was substantial, it was necessary to continue with a systematic examination of the contextual factors of the participants' teaching situation to determine the effect of teaching context on student outcomes.

In addition to the professional teaching performance assessed through the National Board Certification, this study examined the instructional context in which the teaching occurred with a random sample of 64 teachers who applied for the National Board certification. The school setting represents a primary platform for teachers to deliver their instruction. However, the effectiveness of instructional strategies cannot be substantiated without referring to characteristics of the specific student population. Both school and student affiliations were considered in the analyses of instructional factors at the teacher level.

Type of class. Participants provided information related to the type or level of class for which they submitted data. Based on teachers' descriptions and the research team's knowledge about grouping in public schools, we created three categories for class type. These include Gifted/AP, Advanced, and General/Basic. We identified classes as Gifted/AP when students had to meet specific placement criteria to be placed in the class. When teachers identified the students as gifted, Regents, IB, honors, or Advanced Placement, we used the Gifted/AP category for this class. Three participants described their classes as Advanced. The final category was General. In this category, we included classes that the teachers identified as general or college

prep. Table 5.12 shows how many of each class type were included in the study as well as the Certification status of the teacher for those classes.

Table 5.12 *Participants' Class Type and Certification Status*

Class Type	**Total**	**Number Certified**	**% Certified**	**Number non-Certified**	**% non-Certified**
Gifted/AP	18	11	61%	7	39%
Advanced	3	2	67%	1	33%
General	43	22	51%	21	49%

To facilitate a chi-square analysis, advanced and general classes were combined into a single row to eliminate the second row that has small frequencies. The combined row includes participants whose students did not meet specific placement criteria for an AP, gifted, or honors class. The chi-square test shows an insignificant association between the Certification status (Certified vs. non-Certified) and the class type (Gifted/AP vs. Non-Gifted/AP), i.e., $\chi^2(1) = 0.42$, $p = 0.52 > 0.05$.

School size. Furthermore, a total number of teachers and students was used to indicate the school size. The statistical analysis did not reveal a significant difference in the National Board assessment scores between Certified and non-Certified teachers in terms of the size of schools in which they taught: $t(61) = 0.98$, $p = 0.33 > 0.05$. Census data and the NBPTS database were used to categorize the teaching context in terms of locale type, race ratio of school population, district per pupil expenditure, and percentage of students in the school who were receiving free and reduced lunch. The descriptive statistics on these variables indicated that teachers who had attempted Certification, no matter what the certification outcome, taught in similar contexts.

Student population. To characterize the student populations served by Certified and non-Certified teachers, overall social economic status (SES) and race diversity (RD) were examined in this investigation. The SES indicator at the school level was assessed by the percentage of

students participating in the free or reduced lunch program. There was no significant difference between Certified and non-Certified teachers in terms of this SES indicator: $t(61) = 0.98$, $p = 0.33 > 0.05$.

The RD indicator was computed from the number of the non-white students divided by the number of white students. The statistical testing failed to find a significant difference in the student race ratio for teachers between Certified and non-Certified categories: $t(62) = 1.03$, $p = 0.31 > 0.05$.

Per pupil expenditure. In addition, per pupil expenditure was checked in those school districts from which teachers were sampled to contrast the Certified vs. non-Certified categories. An independent sample *t*-test showed no significant difference in the district resources between Certified and non-Certified teachers: $t(61) = 1.02$, $p = 0.31 > 0.05$.

Locale type. The research team also collected information related to school and district context. Schools are characterized by their service regions, resources, and sizes. In this investigation, service regions were categorized by a total of eight locale types (Table 5.13). The locale code is a variable that the National Center for Educational Statistics has created for general description, sampling, and other statistical purposes. It is based on the location of school buildings, and in some cases may not reflect the entire attendance area or residences of students enrolled. The designation of each school's "locale" is based on its geographic location and population attributes such as density (http://www.opi.state.mt.us/PDF/RuralED/LocaleCodeMeth.pdf).

Table 5.13 *School Locale Types*

Locale	Label
1	large central city
2	mid-size central city
3	urban fringe large city
4	urban fringe mid-size city
5	large town
6	small town
7	rural outside CBSA
8	rural inside CBSA

For each locale type, data were gathered on two outcome measures: (1) percentage of all teachers who participated in the National Board assessment and received the Certification; (2) percentage of teachers who participated and failed to achieve Certification. Because the pair of measures is linked to the same locale type, a related sample *t*-test can be used, and the results suggest no significant differences in the locale percentage between Certified and non-Certified teachers: $t(7) = 0.83$, $p = 0.43 > 0.05$.

The insignificant differences on various dimensions of the teaching context seem to suggest that teachers who have attempted National Board Certification, regardless of Certification status, are employed in similar teaching settings. This study does not provide sufficient information to compare the context of teachers who have attempted NBC and those who have not because all teachers who participated in this study have attempted NBC. However, because all participants in this study have completed the process of attempting Certification, comparisons between the Certified and non-Certified may be more rigorous than comparisons made between Certified teachers and the general teaching population. The statistical analysis from the current investigation suggests a similar working context for Certified and non-Certified teachers.

Qualitative Analysis of Student Writing Assessment

Student samples are provided to illustrate how the SOLO Taxonomy rubric and the five feature writing rubric were applied. Both Surface and Deeper levels are presented here. The exemplars were produced in response to the task to *nominate an individual for a history hall of fame*, the prompt that was used in both the MC/Gen and EA/ELA classrooms. Two sets of commentary are provided for each paper included here. The first commentary for each paper is related to the Holistic SOLO Evaluation. The second commentary is related to the Writing Features that were assessed.

Unistructural Writing Exemplar
Case: Middle Childhood Generalist 33488

Student Writing Sample George Washington Carver	**Commentary Related to SOLO Evaluation**
There are loads of people in the History Hall of Fame, but I think there is room for one more. I think George Washington Carver should be in the History Hall of Fame. Why, you ask? George Carver has contributed a lot to the modern world. Mr. Carver invented the peanut. The peanut has been very useful and still is. In school cafeterias and many homes across the world, there is peanut butter. If not for peanuts then we wouldn't have peanut butter. If not for peanut butter then we wouldn't have PBJ's. I don't know about you, but I love PBJ's. If I could go back in time, I would thank Mr. Carver for inventing the peanut. There are also places like Logan's that give you a bucket of peanuts, and you just eat, then throw the shells on the floor. This makes eating at this restaurant more fun. We wouldn't have the peanuts to throw if not for Mr. Carver. We also have candy like Nutter Butters, Peanut Butter Crackers, and Payday's that all have peanuts. I love theses candies, so thanks Mr. Carver.	The writer's response to the task (explaining to the Selection Committee why a certain individual should become a member of the History Hall of Fame) is simplistic. The line of reasoning is limited to "because he gave us peanuts to eat." The writer does not include the value of peanuts as a food product (inexpensive, readily available, easy to store, source of protein, even picky eaters will eat peanuts, available in low fat and low sodium varieties, goes with other foods such as fruits and breads. Further, the writer does not provide information about the actual discovery (when, under what circumstances) or of Carver's other contributions. The reader learns little about Carver's contributions and much about the writer's enjoyment of peanuts. The response is Surface Unistructural Solid.

Overall, I think Mr. Carver should be added to the History Hall of Fame for his contributions to modern food and life.	
Student Writing Sample George Washington Carver	**Commentary Related to Writing Evaluation**
There are loads of people in the History Hall of Fame, but I think there is room for one more. I think George Washington Carver should be in the History Hall of Fame. Why, you ask? George Carver has contributed a lot to the modern world. Mr. Carver invented the peanut. The peanut has been very useful and still is. In school cafeterias and many homes across the world, there is peanut butter. If not for peanuts then we wouldn't have peanut butter. If not for peanut butter then we wouldn't have PBJ's. I don't know about you, but I love PBJ's. If I could go back in time, I would thank Mr. Carver for inventing the peanut. There are also places like Logan's that give you a bucket of peanuts, and you just eat, then throw the shells on the floor. This makes eating at this restaurant more fun. We wouldn't have the peanuts to throw if not for Mr. Carver. We also have candy like Nutter Butters, Peanut Butter Crackers, and Payday's that all have peanuts. I love theses candies, so thanks Mr. Carver. Overall, I think Mr. Carver should be added to the History Hall of Fame for his contributions to modern food and life.	<u>Controlling Idea</u> The writer's controlling idea, that George Washington Carver's contribution of the peanut to modern life makes him worthy of being in the History Hall of Fame, is developed through a single line of reasoning – his "invention" of the peanut means that we have peanuts for food. The broader concept, that "modern life" itself is therefore better, is introduced but neither developed nor apparently understood by the writer. The controlling idea is Surface Unistructural Solid. <u>Organization</u> The writer's placement of details is random, moving from peanuts in school and home to back in history then to peanuts in to a restaurant and from there to candy. If the organizational plan within the body paragraphs is spatial, it is inconsistent. Organization is Surface Unistructural Solid. <u>Elaboration of Ideas</u> Starting narrowly (I like peanut food products) results in limited support: we have peanut butter sandwiches, it's fun to throw peanut shells, and peanut butter candies contain peanuts. Word choice, limited to repetition of peanuts and peanut food products and the abundance of vague words such as "useful," "eat," "food," and "life" does little to develop the writer's topic. Elaboration is Surface Unistructural Low. <u>Voice</u> The writer begins by addressing the reader and anticipating their questioning his nomination, but then seems to lose sight of the Selection Committee as the audience as he thanks Mr. Carver for the peanut butter jelly sandwich and again for peanut "candies." The writer's

	enthusiasm for peanuts is not balanced by the Selection Committee's need for a more formal, less conversational tone. Sentences, though limited in quantity, begin in a variety of ways. Voice is Surface Multistructural Solid. <u>Sentence Formation</u> Sentences are generally clear though neither sophisticated nor controlled. Occasional awkwardness ("In school cafeterias…..there is peanut butter.") stands out in a text with few sentences. Sentence Formation is Surface Multistructural Solid.

Multistructural Writing Exemplar
Case: Early Adolescence/English Language Arts 010076

Student Writing Sample: J.R.R. Tolkein	**Commentary Related to SOLO Evaluation**
The choice is quite simple. The person who should be inducted into the History Hall of Fame is the great author J. R. R. Tolkein. Among all others, Tolkein stands out as an important person in history. Tolkein's tales of adventure include The Hobbit, The Lord of the Rings trilogy, and many other well known books. They have entertained generations of fans for many years. The Lord of the Rings was in fact so popular, that it was recently made into a movie by film director Peter Jackson. Nominated for many awards including "Best Motion Picture", the three movies topped the box office for weeks on end. Not only have readers been enticed by the tales of fantasy for pleasure, many schools across the nation now use The Lord of the Rings in their study of literature. The adventures mentioned in his books have been declared classics, and are known for the things learned through their analyzation. Tolkein is arguably the best author of all time. Having captivated readers through all these years, he is most definetely an important person in literature as well as history. As you can see, the late J.R.R. Tolkein is easily the best choice for the next member of	The writer's response to the assigned task, to explain to the selection committee why a certain individual should become a member of the History Hall of Fame, demonstrates a surface level of understanding. The writer's focus on the popular success of Tolkein's fantasies confines the explanation to a discussion of success at the box office and in school. Discussion of the films is more about director Peter Jackson's achievements than Tolkein's. Two serial supporting points, entertainment value and the educational value result in a Surface Multistructural Solid evaluation. The points are linked to the task but not to each other.

the History Hall of Fame.	
Student Writing Sample: J.R.R. Tolkein	**Commentary Related to Writing Evaluation**
The choice is quite simple. The person who should be inducted into the History Hall of Fame is the great author J. R. R. Tolkein. Among all others, Tolkein stands out as an important person in history. Tolkein's tales of adventure include The Hobbit, The Lord of the Rings trilogy, and many other well known books. They have entertained generations of fans for many years. The Lord of the Rings was in fact so popular, that it was recently made into a movie by film director Peter Jackson. Nominated for many awards including "Best Motion Picture", the three movies topped the box office for weeks on end. Not only have readers been enticed by the tales of fantasy for pleasure, many schools across the nation now use The Lord of the Rings in their study of literature. The adventures mentioned in his books have been declared classics, and are known for the things learned through their analyzation. Tolkein is arguably the best author of all time. Having captivated readers through all these years, he is most definetely an important person in literature as well as history. As you can see, the late J.R.R. Tolkein is easily the best choice for the next member of the History Hall of Fame.	<u>Controlling Idea</u> The explicit controlling idea, that J.R.R. Tolkein is an excellent choice for the History Hall of Fame is established throughout the piece. It is, in fact, repeated as a reason itself. The writer's general opinion does not however, evolve into a deeper unifying theme such as the important role that fantasy plays in helping us understand the complexities of human nature, or the presence of good and evil within us. The controlling idea is Surface Multistructural Solid. <u>Organization</u> The effective opening sentence sets the stage for the writer's ideas that follow while the direct address to the selection committee in the final paragraph brings the piece to a close. Transitions that move the reader forward are present between the movie and school supporting points but not elsewhere. Organization is Deep Relational Low. <u>Elaboration of Ideas</u> The writer has two sets of points that appear to be support for Tolkein's nomination. The first set is that Tolkein's books have been made into films and are taught as literature in school. In this section of the text the writer provides more support for the film director's successes than for the great author. In other words, the elaboration of Tolkein's achievements detour into Jackson's accomplishments. The second set of potential support for the nomination, that Tolkein contributed to literature and history, reads as an aside. There is no elaboration of the author's contributions to history. When read carefully, there is little support in the text beyond the repetition that Tolken is "the best." Elaboration is Surface Multistructural Low. <u>Voice</u> In this feature, the writer shines, addressing the selection committee with authority ("The

	choice is quite simple. . . Tolkein is arguably the best author of all time. . . As you can see, the late J.R.R. Tolkein is easily the best choice. . .”). Word choice is sophisticated and varied, adding to the impression that the speaker is well-informed and should therefore be listened to (“inducted,” “enticed,” “arguably.”) While strong, the text falls short in terms of pace and authenticity (perhaps due to the lack of specifics about Tolkein and his works). Voice is Deep Relational Solid. <u>Sentence Formation</u> Sentences are not only clear, they demonstrate an integration of ideas (“Not only have readers been enticed by the tales of fantasy for pleasure, many schools across the nation now use. . .”). Relationships between the ideas with complex sentences are clear. Sentence Formation is Deep Relational Solid.

Relational Writing Exemplar
Case: Middle Childhood/Generalist 33488

Student Writing Sample: Rosa Parks	**Commentary Related to SOLO Evaluation**
The person I have selected for the History Hall of Fame is Rosa Parks. One of the main reasons I chose her is for her courage. She refused to give up her seat to another because she believed it wasn’t skin color that mattered. I agree. Judging somebody by their skin color is like saying, “She’s a cheerleader, so she must be preppy,” or “She’s dumb because she’s a blond.” “It’s wrong*. The second reason I selected her was because her refusal started the bus boycott. The bus boycott played a large roll in the Civil Right Movement. She started that & partly because of her, today all U.S. Citizens have equal human rights The third reason I nominated her is because, as our world develops it is becoming more important to learn about our history. She played a large roll in history & is a good example for kids to look up to. Some	The writer’s response to the assigned task, to explain to the selection committee why a certain individual should become a member of the History Hall of Fame, demonstrates that the writer can make a connection between an historical figure and contemporary life. The writing is characterized by connections: then and now, linking judgments about skin color to other physical characteristics such as blond hair and the truism “never judge a book by it’s cover,” and how history affects today. More than merely listing reasons for accepting Rosa Parks into the Hall of Fame (a Multistructural response) this writer integrates the reasons into the broader context of how heroes make a difference not only in their own lifetime, but in ours. The response is Deep Relational Solid.

important characteristics we look for in romodels are courage, kindness & the ability to show what we believe. The fourth and final reason I picked her is because she understood that diversity isn't necessarily a bad thing. There are some advantages. For example, you can bring in knowledge of other cultures & ideas. Today we look for diversity in a person & she had that trait. She is a hero in my eyes. I hope she is in your eyes, too. * to think that way. Never judge a book by its cover.	
Student Writing Sample: Rosa Parks	**Commentary Related to Writing Evaluation**
The person I have selected for the History Hall of Fame is Rosa Parks. One of the main reasons I chose her is for her courage. She refused to give up her seat to another because she believed it wasn't skin color that mattered. I agree. Judging somebody by their skin color is like saying, "She's a cheerleader, so she must be preppy," or "She's dumb because she's a blond." "It's wrong*. The second reason I selected her was because her refusal started the bus boycott. The bus boycott played a large roll in the Civil Right Movement. She started that & partly because of her, today all U.S. Citizens have equal human rights The third reason I nominated her is because, as our world develops it is becoming more important to learn about our history. She played a large roll in history & is a good example for kids to look up to. Some important characteristics we look for in romodels are courage, kindness & the ability to show what we believe. The fourth and final reason I picked her is because she understood that diversity isn't necessarily a bad thing. There are some advantages. For example, you can bring in knowledge of other cultures & ideas. Today we look for diversity in a person & she had that trait. She is a hero in my eyes. I hope she is in your eyes, too.	<u>Controlling Idea</u> The controlling idea of this piece is a general theme, that heroes make a difference forever, with lessons to teach us long after they and the specific events are gone. This implicit controlling idea is Deep Relational Low, as the final point of diversity is not clearly articulated and hence not clearly related to the controlling idea. <u>Organization</u> The organization of the piece, a variation of the five paragraph theme, has a clear beginning and middle and an ending that returns to the task and the beginning of the piece. The ideas, however, have a list-like quality. There does not appear to be a thoughtful plan for the placement of details or transitions that move the reader through the text. Organization is Surface Multistructural Solid. <u>Elaboration of Ideas</u> Each of the writer's four reasons for nominating Rosa Parks (courage, role in the boycott, need for role models, diversity) addresses the topic and task. These supporting points are, however, discrete. For example, the writer does not link the first point of courage in how Rosa Park's judged character by more than skin color to the second point, the bus boycott. It must have taken courage to survive losing her job, being threatened, and all the

* to think that way. Never judge a book by it's cover.	events that occurred for the two years after she refused to give up her seat on the bus. The general nature of the writer's word choice throughout the text adds little detail. Elaboration is Surface Multistructural Solid. <u>Voice</u> The writer maintains enthusiasm for the topic, Rosa Parks' contributions to our lives. The tone demonstrates an awareness that the selection committee is to be addressed respectfully ("She is a hero in my eyes. I hope she is in your eyes, too."). Varied sentence structure for effect is demonstrated though limited ("She refused to give up her seat to another because she believed it wasn't skin color that mattered. I agree." Sentences are varied within the body paragraphs, adding a sense of flow. The writer's decision to conform to the "three reasons" organizational plan results in repetition, but the tedium is mitigated by varied verbs within the topic sentences. Voice is Deep Relational Low. <u>Sentence Formation</u> Sentences are clear and demonstrate an understanding of more than simple sentence structure. Many of the sentences demonstrate an integration of ideas. Sentence Formation is Deep Relational Low.

The student work sample and writing assessment exemplars provide descriptive evidence and commentary about the depth of student understanding of classroom curricula and writing. Student responses at a variety of SOLO levels are provided to illustrate the continuum represented within the SOLO Taxonomy.

Section III

Comparative Teaching Practices: Teachers' Intention to Foster Deeper Student Outcomes

Quantitative Analyses

Based on teachers' work samples, our scoring team members used a holistic SOLO Taxonomy rubric to assign values to the implicit and explicit aims of teachers' instruction. In other words, scorers determined if participants' instruction was designed to elicit surface or deeper outcomes. Scorers had no knowledge about the participants' Certification status. After all scoring was completed, the research team began to compare these ratings to participants' Certification status. Table 5.14 shows the number and percentage of Certified and non-Certified teachers whose instructional aims and design were evaluated as "surface" or "deeper."

Table 5.14 *Participants' Depth of Instruction and Certification Status*

	Surface	**Deeper**	**Total**
Certified	18 (51% of Certified)	17 (49% of Certified)	35
Non-Certified	23 (79% of non-Certified)	6 (21% of non-Certified)	29
Total	41 (64% of participants)	23 (36% of participants)	64

The findings indicated that a majority of the teachers (64%) aimed instruction and assignments toward surface learning outcomes. However, the NBCTs were more than two times as likely to aim instruction at deeper learning outcomes (Certified: 49%, or 17 of 35; Non-Certified: 21%, 6 of 29). There was a statistically significant difference between the aims related to the depth of student learning of NBCTs and those who had attempted, but did not receive certification. NBCTs more often intended to foster deeper student understanding.

Some might speculate that the findings in this study that suggested that the teachers in this sample taught in similar teaching contexts could be attributed to a small sample of 64 teachers. However, for a database of the same size, teacher instruction has been categorized into a deep or surface category. The National Board Certification scores are significantly different for teachers between the deep and surface categories [$t(61) = 2.85$, $p < 0.01$]. In other words, teachers whose instruction is categorized as "deep" are likely to score higher in the National Board assessment than their peers in the "surface" category. This finding provides empirical evidence for validation of the National Board assessment.

A frequency table is provided below to check the linkage between the Certification status and the deep/surface levels of instruction.

Table 5.15 *Frequencies of Teachers Across the Certification Status and the SOLO Categorizations*

	Deep	**Surface**
Certified	17	18
Non-Certified	6	23

The contingency table analysis reveals a significant association between these two variables: $\chi^2(1) = 5.36$, $p = 0.02 < 0.05$. In other words, the relationship is significant between the quality of instruction (as categorized by the "deep" or "surface" levels on the SOLO dimension) and National Board Certification status (Certified vs. non-Certified). Because $0.01 < p = 0.02 < 0.05$, the claim of a significant relationship is at $\alpha = 0.05$, instead of a highly significant level of $\alpha = 0.01$.

Qualitative Analyses

In addition to the previous quantitative analysis related to comparative teaching practices, qualitative analyses were also conducted. A type of content analysis was used to evaluate each

set of teacher work samples as "surface" or "deep." Merriam (1998) proposed that category construction is a type of analysis that moves beyond basic description. Here, the challenge is to "construct categories or themes that capture some recurring pattern that cuts across 'the preponderance' (Taylor and Bogdan, 1984, as cited in Merriam) of the data" (p. 179).

Comparative Teaching Practices Exemplar
Case 110007: The 1920s and Today

The following information and descriptions are provided to clarify the SOLO Taxonomy evaluations related to comparative teaching practices. The primary data sources for the evaluation focusing on the teachers' instructional aims and design were the Unit Context responses, Profile of Instruction information, and Profile of Work Samples information provided by the teacher. The student work samples provided confirmatory evidence related to the teachers' instructional aims and design.

Data Source: Unit Context

Description of teacher practice	Commentary related to SOLO Evaluation
Ms. Douglas (a pseudonym) titled her unit "The 1920's and Today." She depicted the context of her unit by describing the students in her Advanced Placement American History Class. Students in the class were in eleventh grade and in the top 5% of their class, with a mean SAT score of 1280 (1140-1420 range). Ms. Douglas indicated that the majority of the students were white, with three African-American and one Hispanic student (out of 10 students). She described the students as "extremely motivated, competitive, and vocal" and indicated further that these students process information in complex ways, learn rapidly, and love to hypothesize about issues. As part of the research design and data collection request, Ms. Douglas described the major principles underlying her teaching. She stated that the students in the class were very creative and intelligent, but that they needed to develop more complex and highly developed thinking skills. She indicated that her challenge has been to get them to use these skills in studying history, especially through an examination of primary documents. Her belief	In statements related to her beliefs about teaching, Ms. Douglas focused on the importance of fostering high-level critical thinking skills, including "generalization, analysis, and synthesis, rather than just a simple recitation of facts." She explained that her instruction allows students to "relate the specific facts of American history to the broad-

is that the most important skill she can teach her students is how to think critically. Therefore, she uses open-ended questioning techniques that "emphasize generalization, analysis, and synthesis, rather than just a simple recitation of facts." Ms. Douglas believes that, through the use of assignments that incorporate high-level critical thinking skills, her students are able to explore American history and relate the specific facts of American history to the broad-based themes and issues of American history.	based themes and issues of American history." Ms. Douglas's description of her teaching was consistent with the deeper aims of instruction in the Teacher SOLO rubric. At the relational level, the teacher requires students to "think about many details and ideas at once and use information in a broader conceptual context" (Teacher SOLO Rubric). Also, at the relational level, the teacher requires students to "organize details and ideas into moderately complex combinations within a specifically defined context" (Teacher SOLO Rubric). In evaluating this case, raters would note Ms. Douglas's descriptions of her practice as evidence; then, they would determine if, in fact, her instructional design fostered the deeper learning aims she identifies. Evaluation was not based on one piece of information or teachers' self-report. Rather, the preponderance of the evidence from all data sources determined the final rating.
Ms. Douglas stated that she had two objectives that dictated the instructional time frame for this class: the High School Graduation Test and the AP American History exam. She described the requirements and importance of each assessment and indicated that she has to "maintain a very rigorous pace for the entire year" so that students are prepared.	Ms. Douglas indicated that two summative assessments for students "dictated the instructional time frame for this class." This statement provides evidence about the influence of testing on teachers' instructional goals. If, in fact, these assessments are designed to foster deeper learning outcomes, perhaps teachers will design instruction to foster deeper outcomes.
Regarding the unit of interest, Ms. Douglas explained that "The 1920s and Today" examines the economic, social, and cultural aspects of the 1920s. Her goal for the students was to gain an understanding of life during the 1920s. Ms. Douglas hoped that students would see the similarities between the 1920s and today. She provided examples of these similarities. She explained that like in the 1920s, today America is a consumer culture. As the people in the 1920s were recovering from the devastation of World War I, Americans today are recovering from the devastation of September 11. She explained that conservatism and liberalism affect American society today in similar ways that	In Ms. Douglas's description of the unit, she included goals that were aimed at deeper learning outcomes. In fact, if her instruction was found to be consistent with these goals, this instructional design may be categorized as extended abstract, the highest level on the SOLO Taxonomy continuum. Ms. Douglas indicated that she wanted students to see the relationship between the conservatism and liberalism of the 1920's and American society today. She also wanted students to examine similarities between pop cultures of the 1920s and today. These goals are consistent with the deeper instructional intentions described in the

these ideologies influenced the 1920s. Finally, Ms. Douglas wanted students to examine the similarities between pop cultures of the 1920s and today.	Teacher SOLO rubric. At the extended abstract level, the teacher "allows students to make generalizations to situations not experienced" (Teacher SOLO Rubric, here, the 1920s) and at this level, the teacher demonstrates a "genuine interest in students understanding the principles of the unit" (Teacher SOLO Rubric). Again, scorers for this case would examine all the evidence to determine if Ms. Douglas instruction is consistent with her stated goals.

Data Source: Profile of Instruction

Description of teacher practice	**Commentary related to SOLO Evaluation**
Ms. Douglas submitted "Profile of Instruction" forms for five lessons in this unit. Across the unit, Ms. Douglas used a variety of teaching activities, including film/video, a graphic organizer, small-group document examination, lecture, seminar/discussion, computer-assisted instruction, reading, and essay assessment. <u>Day 1</u> On the first day of the unit, Ms. Douglas showed the students a film about The Roaring Twenties. As students viewed the film, they were to complete a graphic organizer prepared by Ms. Douglas. The graphic organizer included categories related to the 1920s: fads, culture, people, conflict, culture/race, politics, and entertainment. The students were to write details from the video with the appropriate category on their copies of the graphic organizer. <u>Day 2</u> On the second day of the unit, Ms. Douglas had students examine six pages of advertisements based on some of the major themes of advertising of the 1920s. On this day, students were learning about the consumer culture of the 1920s. Students were directed to examine the advertisements for themes such as easy credit, self-improvement, snob appeal, sex appeal, and	While this activity was surface in nature, it did provide students important facts and information that could help them develop conceptual understandings later in the unit. On this day, Ms. Douglas was beginning to have students make connections between primary documents and major themes of advertising in the 1920s. Her final goal in the lesson required students to "evaluate why the consumer culture became so prominent during

popular culture. After students examined the advertisements, Ms. Douglas led a discussion in which students reported their observations about the advertisements. As a result of this lesson, Ms. Douglas wanted students to learn the following: • To analyze advertisements as primary documents • To analyze documents for obvious as well as hidden meaning • To identify how companies use human motivation to sell products, and • To evaluate why the consumer culture became so predominant during the 1920s. <u>Day 3</u> On the third day of the unit, Ms. Douglas's focus was "Culture in the 1920s." Through reading of a chapter in *Only Yesterday: An Informal History of the Nineteen-Twenties* and analysis of a website, students examined various fads of the 1920s, including mahjongg, Freudianism, men's fashions, flapper fashions, radio, movies, such as *Gold Rush*, *The Kid*, *The Circus*, *Dr. Jekyll and Mr. Hyde*, and *The Jazz Singer*, and musicians such as Bessie Smith, George Gershwin, and Duke Ellington. Ms. Douglas expected students to use information from the website to critique the reading assignment. She also expected students "to learn about the American culture during the 1920s by making inferences about social and cultural changes based on an analysis of the website information and the reading." Additionally, she expected students to learn how to interpret information from other sources besides textbooks. Finally, Ms. Douglas hoped that photographs of the era would give students a more personal understanding of life in the 1920s. <u>Day 4</u> The topic for the fourth day of the unit was "Rural and Urban Conflicts of the 1920s." For this lesson, students read three pages related to	the 1920s." With this goal, Ms. Douglas was having students consider the relationship between the details of the 1920s that they learned about in the previous lesson and the consumer culture that emerged during the decade. Her goal was consistent with the relational level of the Teacher SOLO rubric because she was asking students to consider and apply the details "to a single context" (the 1920s). If Ms. Douglas had required the students to think about the impact of the consumer culture in other contexts, including other decades in American history or in other countries, her goals would move toward extended abstract. The aims of this lesson were consistent with the relational level of the SOLO Taxonomy because Ms. Douglas was requiring students "to integrate multiple independent details" (Teacher SOLO Rubric) related to trends of the 1920s to develop an understanding of the culture of the decade. This task also required students to "think about many details and ideas at once and use information in a broader conceptual context" (Teacher SOLO Rubric). Students were considering details about the 1920s from a variety of sources as they made "inferences about social and cultural changes." In this lesson, Ms. Douglas had students read

polarizing conflicts of the 1920s in their US History textbook. The focuses of the text included fundamentalism, Prohibition, Nativism, and the Ku Klux Klan. After students read, Ms. Douglas lectured about these conflicts and focused on the Scopes Monkey trial, efforts to undermine prohibition, quotas and immigration, the Sacco and Vanzetti case, and the Ku Klux Klan. As a result of the reading and lecture, Ms. Douglas expected students to develop an understanding of the different conflicts that polarized America during the 1920s and to evaluate human qualities that led to the conflict (generational conflict compared to geographical conflict). In addition to this, Ms. Douglas listed several procedural skills related to note-taking that she wanted students to learn or practice: • To be active listeners during the lesson. In order to understand the lecture, one must listen. • To raise appropriate lecture-related questions about the 1920s and to respond to teacher questions about the 1920s conflicts. • To develop and use a consistent method of note-taking. • To listen for and recognize cues as to important points and to recognize transition from one point to the next.	about some major conflicts of the 1920s. She used specific examples of these conflicts in her lecture and required students to integrate these examples into an understanding of the "different conflicts that polarized America during the 1920s." She also asked students to "evaluate human qualities that led to the conflict" and wanted students to understand the difference between generational and geographical conflict. In this lesson, Ms. Douglas's stated goals were more procedural than informational. She emphasized skills such as active listening, question formulating, and effective note-taking. These goals demonstrated Ms. Douglas's attention to the Domain of Mental Procedures that was incorporated into the Teacher SOLO Rubric. In this lesson and the next, Ms. Douglas required students to integrate multiple mental skills simultaneously into a macroprocedure (Marzano, 2000).
<u>Day 5</u> On the final day of this unit, groups of students were assigned a Document-Based Question (DBQ) related to the 1920s. In small groups, students spent 20 minutes discussing the documents. After the small group discussions, individual students were instructed to construct essays integrating their knowledge of the 1920s with the information in the documents. Essays were due at the end of the class period. Ms. Douglas indicated that during this lesson she expected students to work together in small groups to analyze the series of documents from the 1920s. Individually, she expected students to integrate the group discussion of the	In this lesson, students were required to compose a response to a Document-Based Question. Here, Ms. Douglas instructed students to construct essays "integrating their knowledge of the 1920s" with information they encountered for the first time during this lesson. Again, students were required to "integrate multiple details" into a "coherent whole or generalization" (Teacher SOLO rubric). Ms. Douglas again emphasized mental procedures that required students to strategize their collaborative efforts, manage their time,

document with their own knowledge of the 1920s to answer the DBQ effectively. As a result of this lesson, Ms. Douglas expected students to learn the following: • To work successfully with a group • To interpret and use primary documents • To distinguish relevant information from the documents that is pertinent to the document • To recognize relationships between a group of documents, and • To organize, analyze, synthesize, and write a coherent DBQ Essay in 35 minutes.	organize information, and write a response in a specified amount of time. The successful combination and execution of these skills would represent a macroprocedure. The profile of this lesson provided additional evidence of the relationship between the teacher's instruction and the high-stakes AP Exam. The DBQ Essay was, in fact, a practice for the AP Exam. However, this task, no matter the source, did require higher level thinking and a deeper understanding of unit concepts for students to be successful.

Data Source: Profile of Student Work Samples

<table>
<tr><th>Description of teacher practice</th><th>Commentary related to SOLO Evaluation</th></tr>
<tr><td>Ms. Douglas collected work from her students on three of the unit days. The assignments she collected included the introductory graphic organizer from Day 1, an essay critique of a chapter from Only Yesterday (text) collected on day Day 4, and the DBQ essays collected on Day 5. The work samples were profiled as follows:
<table>
<tr><th>Assign-ment</th><th>Where completed</th><th>How long?</th><th>Organi-zation</th><th>Whose work?</th><th>Guidance, feedback, support</th></tr>
<tr><td>Graphic Organizer</td><td>In class</td><td>20-40 mins.</td><td>Individual</td><td>Individual</td><td>Some</td></tr>
<tr><td>Only Yesterday essay critique</td><td>At home</td><td>1-2 hrs.</td><td>Individual</td><td>Individual</td><td>None</td></tr>
<tr><td>DBQ essay</td><td>In class</td><td>20-40 mins.</td><td>Group, then individual</td><td>Individual</td><td>None</td></tr>
</table>
</td><td>The profile of work samples typically provided evidence about the consistency of teacher stated goals and intentions. Sometimes this evidence was confirmatory (consistent with stated intentions); sometimes this evidence was contradictory (inconsistent with stated intentions). The profile of work samples was typically not the primary source of evidence for making evaluations about the Teacher's SOLO level. Here, the raters can see an overview of the</td></tr>
</table>

	assignments students were required to complete in response to the teacher's instruction.

Data Source: Student Work Samples

Description of teacher practice	Commentary related to SOLO Evaluation
As is described in the Profile of the Student Work Samples section above, the student work samples for this unit included six graphic organizers, six essay critiques, and six DBQ essays. The graphic organizers were nearly identical; however, the essays were unique and distinct from each other. On these work samples, the teacher provided feedback and questions to help students make connections to the broader themes and issues in the unit.	The student work samples were the primary source for evaluating students' SOLO level. For the teacher evaluation, the work samples provided important evidence to help raters assess the consistency between teachers' stated intentions and instructional practices. With this data source, the raters might have seen students performing at the level the teacher expected. Raters might have also seen evidence that the students were not making the connections the teacher had hoped. Student work samples in this unit suggested that students' understanding of the unit concepts became increasingly sophisticated. In the first assignment, students merely organized facts and information from the film onto a graphic organizer. Most of the students' had exactly the same information on this assignment, which suggested that they reproduced the information from a common source, rather than creating their own observations. In the second assignment, students composed individual essays related to the 1920s as a period of "Revolution in Manners and Morals" (chapter title). In these essays, the patterns of response indicated that students were drawing similar conclusions about the 1920s related to issues such as women's rights, the effect of World War I, technology and transportation, fashion, prohibition, and media. However, the students used a wide variety of specific examples to support their generalizations, thereby making the conceptual understandings more individual. In the final assignment, students integrated ideas and information from the unit into document-based essays. For these essays, many students underlined their topic sentences.

	Though the topic sentences represented similar themes from the unit, each student's topic sentences and subsequent supporting details were unique, but accurate. Again, this provided evidence that students were individually integrating the details into a conceptual understanding of the decade under study.

Evaluation Summary of Comparative Teaching Practices Exemplar

The SOLO Taxonomy rating for this teacher was Deep Relational Solid. This case was evaluated by three separate expert scorers before it was labeled as a benchmark case for the relational level of performance. Each of the raters scored this teacher's materials as relational. The rationale for the final score assignment, developed by the three raters collectively, was as follows:

> Students had to think about many facts and ideas at once in order to compose essays synthesizing their knowledge about the 1920s. The instruction was hierarchic; the teacher began with a video introduction and a simple graphic organizer, then introduced other resources and primary documents. The culminating activity required students to organize details within a defined context (the 1920s).

Had the teacher led students to apply conceptual understandings to a different context (i.e., the current decade, a different country or culture), this unit might have been evaluated as extended abstract. However, while the teacher stated in her unit goals that she was going to have students compare the 1920s to today, there is no evidence that this happened, except for a couple statements by students in the final essay. Still, students' understanding of the constructs for the 1920s was sophisticated, and students were required to support their understandings with specific events and examples from the time.

Summary

Fortunately, the parametric and non-parametric methods resulted in similar findings from this validation study. On one hand, Certified and non-Certified teachers are not significantly different in their work settings as described by various personal and/or contextual factors. On the other hand, the analyses indicate significant relationships between teaching quality (deep/surface) and Certification assessment outcomes, regardless of the scaling of the outcomes as Certification scores or into a Certified vs. non-Certified dichotomy. When comparing teaching outcomes using student work samples, the results indicated that students of NBCTs were much

more likely to achieve deep student learning outcomes. While this relationship was not statistically significant, it is an important finding. The results from the writing assessment confirmed a positive relationship between NBC and student outcomes. All six indicators assessed from the writing samples were highly statistically significant, with the students of NBCTs performing at deeper levels. When comparing teaching practices, the results indicated that a statistically significant difference existed between NBCTs and their non-Certified counterparts in regards to their intentions and goals to teach at deeper levels.

Although sample size might have played a role in the statistical testing, the concern is mainly focused on an issue of relatively small teacher samples, which might have undercut the statistical power to detect significant differences on these contextual factors. The already significant relationship from this study is likely to sustain future verifications from large databases. The triangulation of the research findings between parametric and non-parametric approaches also reinforces the conclusion of significant relationships between teaching quality and the National Board Certification**.**

The qualitative data included in the exemplars of comparative teaching practices and comparative teaching outcomes provides rich descriptions of "surface" and "deep" instructional aims and outcomes. The implications of these data are far-reaching and will be explained in the Discussion chapter that follows.

DISCUSSION AND CONCLUSIONS

Summary

This chapter provides an overview of the study, including its limitations. In addition, this chapter provides a discussion about the implications of this study for the assessment of student learning, research, teacher development, and public policy.

This was a construct validation study of the National Board's vision of accomplished teaching, the instantiation of that vision in its assessments, and ultimately the certification decisions resulting from that assessment. The intent of the research design was to determine whether students of National Board Certified teachers (NBCTs) exhibited to a measurably greater degree deeper learning outcomes than students of teachers who had attempted but had not achieved Certification. A second, related intent of the research design was to determine the extent to which NBCTs exhibited to a measurably greater degree instructional aims, plans, and assignments designed to deepen student understanding.

This was a mixed-method study comparing the teaching practices and outcomes of National Board Certified teachers (NBCTs) and their non-Certified counterparts. The study addressed two questions:

- *Comparative Teaching Outcomes.* Do students taught by National Board Certified teachers produce deeper responses (to class assignments and standardized writing assessments) than students of teachers who attempted National Board Certification but were not Certified?

- *Comparative Teaching Practices.* Do National Board Certified teachers develop instruction and class assignments designed to produce deeper student responses than teachers who attempted National Board Certification but were not Certified?

Evidence of accomplished teaching was sought by examining teachers' stated instructional goals, instructional resources and materials, descriptions of lessons, descriptions of

students' assignments, and student responses to those assignments. To examine the comparative teaching outcomes, we collected student work samples connected to regular class curricula as well as standardized writing assessment responses. The writing assessment was administered in the classrooms of participants in the EA/ELA and MC/Gen certificate areas. Six dimensions of writing were scored and analyzed on the writing assessment (see the Results chapter).

Participants included 64 teachers from 17 different states. Participants had sought National Board Certification in one of four certificate areas: Middle Childhood/Generalist, Early Adolescence/ English Language Arts (EA/ELA), Adolescence/Young Adulthood Science (AYA/Science), and AYA Social Studies-History (AYA/SS-H). Participants who completed all requirements for the study submitted a description of the unit context, including characteristics of their teaching situation, an overview of a self-selected unit of study, as well as their stated expectations for student learning. They also submitted a profile of each lesson taught during the unit and of each work sample they collected from students. Finally, they submitted all work associated with the unit from six randomly-selected students in one class. Participants who had sought certification in the MC/Gen and EA/ELA certificate areas also submitted responses of all students in one class to a standardized writing assessment.

In every single comparison between NBCTs and non-NBCTs, NBCTs obtained higher means scores. While the potential for statistical significance was somewhat limited by the sample size, the differences were nonetheless statistically significant in 7 of the 8 measures. The conclusion seems clear: the Certified teachers in this sample developed and implemented, to a considerably greater degree than non-Certified teachers, instructional plans and assignments aimed at fostering deeper student understanding. In addition, the students of NBCTs

accomplished deeper learning outcomes more frequently than did students of non-Certified teachers.

On the comparative teaching practices dimension, results of the quantitative analysis indicated statistically significant differences between the Certified and non-Certified teachers' stated and implemented instructional aims, goals, and assignments. These results indicate that teachers who have achieved National Board Certification design instruction with the intent to foster deeper student understanding than do teachers who have sought and not achieved Certification.

On the comparative teaching outcomes dimension, a comparison of the depth of student understanding on the work samples collected from students did not yield statistically significant differences, although students of NBCTs were more likely to achieve deeper learning than the students of the non-Certified teachers. Results of the analysis of the writing performance indicators, however, did show statistically significant differences on all indicators separately and as a single indicator of "depth of knowledge of writing." In other words, the students of NBCTs exhibited a deeper understanding of writing than the students of the non-Certified teachers.

Limitations

A discussion of the limitations of this investigation is appropriate as results and conclusions are discussed. As has been mentioned, the original design of this study called for 200 participants, 50 for each certificate. Besides concerted efforts to accommodate participant mortality, the final sampling outcome was influenced by several extraneous factors. In particular, many teachers did not meet eligibility criteria because they were no longer classroom teachers. Some had retired; others had moved to different positions or professions. Another prohibiting factor was that teachers did not respond to the original letter and survey that were sent to them.

Finally, many teachers who did agree to participate later withdrew from the study. Of course, pursuant to the Institutional Review Board guidelines, participants could withdraw participation at any time without penalty or explanation. Some teachers explained to us that they could not participate because they did not have time to complete the requirements. Others told us that they had difficulty obtaining administrator permission to participate or that parents were unwilling to consent to having their children participate in the study (via work samples or writing assessment). The truth is that participation was fairly burdensome. It is doubtful that professionals in any field would be enthusiastic about doing something like this for so little compensation.

Despite the inadvertent sample attrition, significant difference was found on the comparative teaching practices dimension and on all six dimensions of the writing assessment. One might, therefore, expect that the significant findings would persist when the analyses were conducted on a larger sample with an increase of statistical power to detect significant differences.

Another sampling issue was related to participation by non-Whites. While more than 13% of the original sample was non-White, no participants for whom race is known were non-White. Again, we cannot assume to know why non-White teachers did not participate in this study. However, we do know that the resultant sample is not representative of the population based on race and ethnicity. This finding may contribute to the increasing literature related to the adverse impact of National Board Certification on minorities (see Goldhaber, Perry, & Anthony, 2003; Ladson-Billings & Darling-Hammond, 2000). Additional research in the future may enrich understanding of the general participants of the National Board assessment, and thus, shed more light on the Certification outcomes across the ethnicity dimension.

In summary, the sample achieved is not a representation of all teachers, National Board Certified Teachers, or all teachers who have been candidates for certification. Rather, we sampled from four certificate areas and experienced high participant mortality. To help reduce the sample attrition rate in future investigations, we wish to make a final note related to the possible connection between participation and compensation. While Certification is often rewarded with substantial pay increases, bonuses, and professional kudos, participants in this study received only a small honorarium (no more than $200). Though we used a random sampling mechanism, participants could choose not to participate without any substantial loss of benefit or current status.

This being said, the authors believe that if the present investigation were replicated with perhaps improvements that will be discussed below, results obtained would be similar to those described herein.

In the course of this investigation, we have learned what we did well, what we should have done but did not do, and perhaps most painfully, what we should not have done at all. We drew from experience with previous similar studies (Bond et al, 2000, Smith 1999, 2004), and we conducted two pilot investigations before beginning the operational phase of the study. We believe the research design, including the assessment model, scoring rubrics, training materials and protocols, data collection materials, and the scorers were exceptional. It is easy to defend the quality of the data obtained. That said, there is also considerable room for improving and refining many of the measures in the study. Also, given the number of factors studied, a much larger sample of teachers would have been desirable. In addition, the evaluation of student work samples would have been more accurate if we could have collected demographic data about each individual student, including reliable measures of entering achievement. One difficult issue here

is protecting students' confidentiality. Already, with a guarantee of confidentiality for students and teachers, some parents were hesitant to consent to participation. If we had asked for information related to family education levels, standardized test outcomes, or classification of students by disability, we might have had even fewer participants who would agree to participate in this study.

Another factor that might have influenced participation was the participants' relationship to the researchers. In almost all cases, the participants did not know the researchers. For some teachers, this lack of familiarity might have been reassuring. On the other hand, because of this lack of familiarity, some participants seemed to demonstrate little concern when they withdrew from the study. In fact, some withdrew participation without informing the researchers. If the potential participants had known the researchers, they might have felt more compelled to complete all requirements. However, such familiarity would have certainly introduced additional methodological issues.

While we have acknowledged multiple limitations, particularly related to sample size, that could have affected the outcomes, it should be noted that there were also forces in this study which tended to militate ***against*** the likelihood of finding positive results and, therefore, make this study a very rigorous examination and the results actually obtained even more persuasive. First, the non-Certified teachers in this study took the very demanding assessment of the National Board. Almost all teachers who have undergone this year-long, rigorous assessment, including those who were ***not*** Certified, testify to the power of the experience as a professional development activity. Most report that it changed fundamentally the way they think about and approach teaching, and that they are better teachers for the experience (Anderson, Hancock, & Jaus, 2001; National Board for Professional Teaching Standards, 2001). Second, many of the

non-Certified teachers in the study enthusiastically agreed to participate because they felt they had "something to prove." This is, of course, the well-known "John Henry" effect, the exertion of unusual effort by subjects in a study who are aware of their status as members of a control or comparison group.

Comparative Teaching Outcomes

Writing Assessment Results

The results regarding comparative teaching outcomes in the writing assessment represent a particularly important finding in this investigation. Prior to this study, most student learning measures that have been used to assess the impact of National Board Certified teachers on student learning have been large-scale, multiple-choice tests (Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Vandevoort et al., 2004). The writing assessment, designed by an external test development organization, provided an "external" indicant of achievement, not directly tied to the teacher's specific instructional objectives during the teaching unit. The decision to use writing as one of the measures of student outcomes was motivated by a desire to gauge the effects teachers have on a universally-valued student outcome that is common to virtually all school curricula, but does not encourage the counterproductive practices of externally imposed paper and pencil tests. The production of written texts allows for students to "go deeper" in ways that multiple-choice tests cannot. In this study, we only collected a single writing measure (a typical strategy in standardized writing assessment administration). The research team felt justified in collecting the single sample because a traditional pre- and post-test model would not mitigate for intervening variables unrelated to the teacher. For example, students may be getting additional help through a school or community tutoring program. Some students may have more help at home. In this study, with teachers from a variety of schools, school districts, and states,

we could not control for these factors. Therefore, we used the teachers' instructional context to determine that the Certified and non-Certified teachers taught in similar situations. Our finding suggests that the variability of the writing assessment outcomes was more related to the teacher than to any contextual variable. This assessment also illustrates the difficulty of creating assignments that elicit depth of learning.

In this study, a measure of student writing in which students had to *compose* a unique response and persuade or inform an audience was used to assess student achievement in writing. This assessment was not linked to a particular curriculum or writing assessment model in any particular state.

Student Work Samples

The data collected on the depth of students' responses of the objectives of particular instructional units were very informative, particularly in relationship to other findings in the study. For example, it may seem contradictory that students of NBCTs would produce deeper outcomes on a standardized writing measure and yet produce mostly surface responses to classroom assignments and tasks. One issue that scorers mentioned repeatedly was that the students seemed limited by teachers' expectations or by the instructional materials themselves. It was often difficult to determine students' depth of understanding because the tasks and questions they were given were aimed only at surface outcomes (i.e., reproduction or categorizing of information, recall of facts, replication of a simple procedure). Only rarely did students demonstrate a deeper understanding when the tasks were not aimed at fostering deeper learning outcomes.

This finding has implications for curriculum planners and writers, for professional development providers, and for classroom teachers who are interested in fostering deeper student

understanding of content. Many of the resources used by teachers in the sample were commercially made. We worked with our scorers to defuse the bias that often accompanies the observation of "worksheet-driven" instruction. Scorers were trained to assess the value and intent of those materials for eliciting deeper student understanding. Even when teachers did not create the materials, we assumed that they *selected* them for the particular lessons. If the teaching resources were designed to elicit surface responses, usually students responded in a like manner. If, however, the instructional materials were designed to foster the understanding of concepts, relationships, and other deeper outcomes, students were more apt to make connections among the facts and details presented to arrive at novel or more sophisticated understandings. A promising contribution of this study is that it provides a way to think about the design and selection of instructional goals, approaches, and resources that foster deeper student understanding. The Teacher SOLO Rubric provides language that could facilitate a more careful selection of instructional resources.

Also deserving mention is the deliberate research design decision to use measures of student achievement other than commercially or state-developed multiple-choice tests of generic academic subjects such as reading and mathematics. As efficient monitors of general academic skills and abilities (e.g., NAEP), such measures have a place. It is, however, in their uses as measures of individual teacher effectiveness and quality that such measures are questionable. Many researchers and assessment experts have discussed the problems of using standardized achievement scores as a means of assessing teacher effectiveness (Darling-Hammond, 1997, 1998; Millman & Schalock, 1997, Popham, 1997; Webster, 1995).

Comparative Teaching Practices

The results of the Comparative Teaching Practices portion of this study strongly suggest that accomplished teachers, as exemplified by the National Board Certified Teachers in this sample, are demonstrably more intent on fostering in their students a level of understanding that is richer, more elaborated, and more meaningfully interconnected with related concepts. This finding is not restricted to a particular grade level or to a particular subject matter. It appears to be a skill of accomplished teachers at grade levels from middle childhood to high school. The variety of content and ability objectives in the 64 instructional units in the study span the spectrum: from understanding literary genres, to historical and social movements; from the complexities of interdependent ecosystems, to basic concepts in Mendelian genetics.

To the extent that the NBCTs in this sample are representative of the larger population of NBCTs, the evidence from this investigation seems clear: Certified teachers possess, to a considerably greater degree than non-Certified teachers, attributes of teacher expertise that are consistent with the emerging body of research on teaching and learning. Two of these attributes were particularly relevant to this study: Expert teachers focus on student learning, and expert teachers have a deep understanding of content. Demonstrating their focus on student learning, expert teachers use a student-centered instructional approach, employ flexible and diverse strategies, monitor student performance consistently, and understand that pedagogical expertise is situated in an understanding of students as individuals and learners (Berliner, 2004, Hattie et al., 1996, Smith, 2004; Stronge, 2002). Further, while expert teachers have a thorough understanding of domain-specific knowledge, they also understand that knowledge is contextually-bound (Berliner, 2004; Hattie et al., 1996). NBCTs in this study more often demonstrated a focus on student learning and a deep understanding of content as they planned

and implemented instruction based on contextual factors to promote deep learning than their non-Certified counterparts.

Implications

The methods and findings from this investigation have important implications for policy, research, and practice related to teacher quality and student learning. This section provides a discussion of implications for future research topics, methods of analysis, and the professional development of practicing teachers.

The Significance of National Board Certification

Without question, the findings from this study contribute to the evidence that NBCTs have a positive impact on student learning (see Berliner, 2004; Cochran-Smith, 2004; Darling-Hammond & Loewenberg-Ball, 1997). While many recent studies have demonstrated the positive effect of NBCTs on student achievement (see Goldhaber & Anthony, 2004; Vandevoort et al., 2004; Cavalluzzo, 2004), most of these studies have limited their data collection to the outcomes of NBCTs' students on standardized tests. In contrast, this study provides descriptions of teacher practices and student outcomes that are associated with deeper learning. Both types of studies are critical to our understanding of the relationship between National Board Certification and student learning. Collectively, these studies provide evidence that the National Board for Professional Teaching Standards is realizing its goal of identifying accomplished teachers who influence student achievement.

Assessment of Student Learning

Perhaps one of the greatest values of this study is that it provides a promising model for accomplishing a critical aim of assessment: improving student understanding and performance.

Wiggins (1998) suggested that "the aim of assessment is primarily to *educate and improve* student performance, not merely to *audit* it" (italics in original, p. 7). He continued:

> People do not run their businesses only to satisfy an auditor's requirement for records that appear accurate. But schools too often worry about the equivalent: we focus on teaching students to pass simplistic, often multiple-choice tests composed of "items" that neither assess what we value nor provide useful feedback about how to teach and how to learn.
>
> We sacrifice our aims and our children's intellectual needs when we test what is easy to test rather than the complex and rich tasks that we value in our classrooms and that are at the heart of our curriculum. That is, we sacrifice information about what we truly want to assess and settle for score accuracy and efficiency. That sacrifice is possible only when all of us misunderstand the role assessment plays in learning. In other words, the greatest impediment to achieving the vision described is not standardized testing. Rather, the problem is the reverse: we use the tests we do because we persist in thinking of assessment as not germane to learning, and therefore best done expediently (p. 7).

Perhaps such a close examination of student learning and the structure of student understanding is not feasible to replace large-scale assessments; however, implementing such models is critical to improving teaching and learning. If we do not study *how* students learn and demonstrate their learning, we can never understand how to help them learn better. The SOLO Taxonomy represents a learning cycle and continuum. It acknowledges the importance of facts and information in the first two levels (referred to in the model as the quantitative phase); however, the model also provides a way to think about the *quality* of student learning at the relational and extended abstract levels (referred to as the qualitative phase).

Emerging literature related to assessment reform challenges educators and policy makers to think about the *purpose* of assessment (Berry, 2004; Darling-Hammond & Snyder, 2000). Most advocates of assessment reform do not recommend that we abandon all large-scale assessment of student achievement and learning, particularly if we have no plan for replacing them. Here, again, Wiggins explained:

> Assessment reform is thus neither as easy nor as simple as throwing out conventional tests. Before we can change our system into one that serves all our needs, we require something more educative and exemplary to which to aspire – something vivid and provocative that makes us see the deficiencies in our time-honored practices, something designed to promote excellence, not just to measure efficiently. (p. 7)

Wiggins's comments suggest that we need to re-think our assessment and instructional practices. He challenges us to think about assessment as a means, not an end. Hattie and Jaeger (1998) also argued for an approach to assessment that acknowledges its importance in the learning process. They contended that "assessment needs to be an integral part of a model of teaching and learning if it is to change from its present status as an adjunct to 'see' if learning has occurred, to a new status of being part of the teaching and learning process" (p. 111). The data collected in this study as well as data that can be subsequently collected and analyzed have the potential to inform efforts to improve the teaching-learning process.

Professional Development

This study also has implications for teacher preparation and professional development for practicing teachers. The SOLO Taxonomy and Marzano's New Taxonomy of Educational Objections offer insights about how teachers can examine and evaluate the quality of student learning. Teachers can use these tools to design instruction that fosters deeper student learning.

Teachers and teacher candidates can examine their learning goals for students to determine if they have set expectations for students and then provided appropriate instruction to help students learn the facts and information they need to form relationships and generalizations associated with the content.

Not only can teachers use the research tools from this study to develop appropriate learning goals and instruction, but also they can use the SOLO Rubrics to evaluate teaching materials for their disciplines. In the Teacher SOLO Rubric, for example, the relational level states, "Task or sequence of tasks requires students to organize details and ideas into moderately complex combinations within a specifically defined context." Using this language, teachers could evaluate texts and ancillary materials to determine if those materials will support students' acquisition of deeper learning outcomes. The authors anticipate that ideas related to the professional development implications of this study will continue to emerge over time.

Policy

It seems appropriate to speak to policymakers directly because they are often given the responsibility (and opportunity) to make decisions related to how teachers are prepared, evaluated, and compensated. While it may seem attractive to make decisions using criteria related to efficiency and/or economy, such decisions do not always benefit the members of the constituency. In the short run, perhaps the expeditious decision will be valued. In the long run, however, student learning and potential may be compromised. When discussing 'promising models' for assessing teacher quality, Cochran-Smith (2004) questions the 'Teaching at Risk' report (The Teaching Commission, 2004) which calls for better assessments such as the new American Board for the Certification of Teacher Excellence and collaborations between ETS and NCATE which focus on Praxis II scores. Cochran-Smith ponders why the large-scale report (a

follow up to Nation at Risk) hardly mentions standards-based approaches such as the NBPTS and NCATE. This issue is also discussed in the work of Berry (2004). Perhaps the reason standards-based approaches are disregarded is because they are complex. Embracing a complex model rather than a simplistic one requires commitment on behalf of educators and policymakers to understand the models and the issues. In this case, the relationship between teaching and learning is indeed complex, with many mitigating factors that can influence the success of teachers and students. We need assessment systems that acknowledge and account for these complexities – and we need political and educational leaders who are committed to and enthusiastic about the potential of such systems to improve teaching and learning.

Research

While the researchers involved in this study believe that the research tools and findings offer a significant contribution to our understanding about the relationship between student learning and accomplished teaching, we also know that this single study is limited in scope. We are hopeful that future studies will continue to examine alternatives for evaluating quality teaching and student learning. Specifically, we recognize the need for studies that involve more comprehensive data collection, including classroom observation data and appropriate and reliable measures of students' entering ability, with larger samples of teacher and student participants.

Regarding the importance of research related to deeper learning outcomes and teaching practices that foster such outcomes, the authors understand the need for longitudinal studies that assess how deeper learning approaches and outcomes benefit students. If more nationwide studies are conducted, and if adequate compensation can be provided to participants, educators,

policymakers, and the American public will be more likely to realize the potential of authentic assessment to improve teaching and learning.

References

Airasian, P. W. (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, Inc.

Anderson, K. M., Hancock, D., & Jaus, V. (2001). *Program evaluation report for The Charlotte Collaborative Project*. Retrieved March 29, 2005, from http://www.nbpts.org/pdf/charlotte_collaborative_rpt.pdf

Arter, J. & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.

Atwell, N. (2002). *Lessons that change writers*. Portsmouth, NH: Heinemann.

Baldwin, D. (2004). A guide to standardized writing assessment. *Educational Leadership*, *62*(2), 72-75.

Bartel, R. (1983). *Metaphors and symbols: Forays into language*. Urbana, IL: National Council of Teachers of English.

Benton, S. L., Corkill, A. J., Sharp, J. M., Downey, R. G., & Khramtsove, I. (1995). Knowledge, interest, and narrative writing. *Journal of Educational Psychology, 87*, 66-79.

Berliner, D. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society 24*(3), 200-212.

Berry, B. (2004, February). *Taking action to improve teaching a quality: Addressing shortcomings in the Teaching Commission Report*. Retrieved February 24, 2005 from The Southeast Center for Teaching Quality Web site: http://www.teachingquality.org/resources/SECTQpublications.htm

Biggs, J. (1987). *Student approaches to learning and studying*. Melbourne: Australian Council for Educational Research.

Biggs, J., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO Taxonomy*. New York: Academic Press.

Biggs, J., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. In H. Rowe (Ed.), *Intelligence, reconceptualization, and measurement* (pp. 57-76). New Jersey: Laurence Erlbaum Assoc.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Drathwohl, K. (1956). *Taxonomy of educational objectives: The cognitive domain*. New York, New York: David McKay Co.

Bond, L., Smith, T. W., Baker, W., & Hattie, J. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Retrieved February 7, 2004, from the National Board for Professional Teaching Standards Web site: http://www.nbpts.org/research/research_archive.cfm

Boulton-Lewis, G. M., (1994). Tertiary students' knowledge of their own learning and a SOLO Taxonomy. *Higher Education, 28*, 387-402.

Boulton-Lewis, G. M., Dart, B., & Brownlee, J. (1995). Student teachers' integration of formal and informal knowledge of learning and teaching. *Research and Development in Higher Education, 18*, 136-142.

Boulton-Lewis, G. M., Smith, D. J. H., McCrindle, A. R., Burnett, P. C., & Campbell, K. J. (2001). Secondary teachers' conceptions of teaching and learning. *Learning and instruction, 11*, (1) 35-51.

Boulton-Lewis, G. M., Wilss, L., & Mutch, S. (1996). Teachers as adult learners: Their knowledge of their own learning and implications for teaching. *Higher Education, 32,* 89-106.

Burnett, P. C. (1999, April). *Assessing the outcomes of counseling within a learning framework.* Paper presented at the Annual Conference of the American Educational Research Association, Montreal, Canada.

Calkins, L. (1994). *The art of teaching writing*. Portsmouth, NH: Heinemann.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. W. Williamson & B. A. Huot (Eds.) *Validating holistic scoring for writing assessment*. Cresskill, NJ: Hampton Press.

Campbell, J., Smith, D., Boulton-Lewis, G., Brownlee, J., Burnett, P. C., Carrington, S., & Purdie, N. (2001). Students' perception of teaching and learning: The influence of students' approaches to learning and teachers' approaches to teaching. *Teachers and Teaching, 7*, 173-187.

Carew, A. L., Mitchell, C. A. (2002). Characterizing undergraduate engineering students' understanding of sustainability. *European Journal of Engineering Education, 27,* 349-61.

Carnegie Forum on Education and the Economy. (1986). *A nation prepared: Teachers for the 21st Century*. Washington D.C.: The Task Force on Teaching as a Profession. (ERIC Document Reproduction Service No. ED268120).

Cavalluzzo, L. (2004, November). *Is National Board Certification an effective signal of teacher quality?* Retrieved February 24, 2005, from the CAN Corporation Web site: http://www.cna.org/expertise/education/

Cawelti, G. (1999). Improving achievement: Finding research-based practices and programs that boost student achievement. *The American School Board Journal, 186*(7), 34-37.

Chan, C. C., Tsui, M. S., Chan, M. Y. C., & Hong, J. H. (2002). Applying the Structure of the Observed Learning Outcomes (SOLO) Taxonomy on student's learning outcomes: An empirical study. *Assessment and Evaluation in Higher Education, 27*(6), 511-517.

Chick, H. (1998). Cognition in the formal modes: Research mathematics and the SOLO Taxonomy. *Mathematics Education Research Journal, 10*(2), 4-26.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education,* 175-218.

Cochran-Smith, M. (2004). The report of the Teaching Commission: What's really at risk? *Journal of Teacher Education, 55*(3), 195-200.

Cohen, F. (2003). Mining data to improve teaching. *Educational Leadership, 60*(8), 55-56.

Dale, E. & O'Rourke, J. (1976). *The living word vocabulary*. Elgin, IL: Dome Press.

Darling-Hammond, L. (1996). The quiet revolution: Rethinking teacher development. *Educational Leadership, 53*(6), 4-10.

Darling-Hammond, L. (1997). Toward what end? The evaluation of student learning for the improvement of teaching. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 248-263). Thousand Oaks, CA: Corwin Press, Inc.

Darling-Hammond, L. (1998). Standards for assessing teacher effectiveness are key: A response to Schlalock and Myton. *Phi Delta Kappan, 79*(6), 471-472.

Darling-Hammond, L. (1999). Teacher quality and student achievement: A review of state policy evidence. Retrieved April 19, 2005 from the Center for the Study of Teaching and Policy Web site: http://depts.washington.edu/ctpmail/Reports.html#TeacherQuality

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives, 8*(1). Retrieved November 2001, from http//epaa.asu.edu/epaa/v9n1

Darling-Hammond, L. (2003). Keeping good teachers: Why it matters, what leaders can do. *Educational Leadership, 60*(8), 6-13.

Darling-Hammond, L., & Loewenberg-Ball, D. (1997). *Teaching for high standards: What policymakers need to know and be able to do*. Retrieved March 1999 from http://www.negp.gov/Reports/highstds.htm

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

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, *16(4-5)*, 523-545.

Dart, B., & Boulton-Lewis, G. (1998). *Teaching and Learning in Higher Education*. Melbourne, Victoria: The ACER Press.

Duziban, C. D., Cornett, J. W., Moskal, P. D., & Gyori, D. (2000). A three year evaluation of citizen in a democracy. *The Hungarian CIVITAS Program Evaluation Report.*

Engelhard, G., Gordon, B., & Gabrielson, S. (1992). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English, 26* (3), 315-336.

Entwistle, N. (1988). Motivational factors in students' approaches to learning. In R.R. Schmeck (Ed.), *Learning strategies and learning styles*. New York: Plenum.

Entwistle, N. (2001). Conceptions, styles and approaches within higher education: Analytic abstractions and everyday experience. In R. Sternberg & L. F. Zhang (Eds.), *Perspectives on Cognitive, Learning, and Thinking Styles* (pp. 103-136). Mahwah, NJ: Erlbaum.

Fletcher, R. (1993). *What a writer needs*. Portsmouth, NH: Heinemann.

Fletcher, R., & Portalupi, J. (2001). *Writing workshop: The essential guide*. Portsmouth, NH: Heinemann.

Floden , R. E. (2002). Research on effects of teaching: A continuing model for research on teaching. In Richardson, V. (Ed.), *Handbook of Research on Education* (4th ed.). Macmillian: New York.

Freeman, M. (2003). *Building a writing community*. Gainesville, FL.: Maupin House Publishing.

Glasswell, K., Parr, J., & Aikman, M. (2001). Development of the asTTle writing assessment rubrics for scoring extended writing tasks. *Technical Report 6, Project asTTle*, University of Auckland.

Goldberg, G. L., Roswell, B. S., & Michaels, H. (1998). A question of choice: The implications of assessing expressive writing in multiple genres. *Assessing Writing, 5*(1), 39-70.

Goldhaber, D., & Anthony, E. (2004). Can teacher quality be effectively assessed? Retrieved on June 24, 2004, from www.crpe.org/workingpapers/pdf/NBPTSquality_report.pdf

Goldhaber, D., & Brewer, D. (2001). Evaluating the evidence on teacher certification: A rejoinder. *Educational Evaluation and Policy Analysis, 23*(1), 79-86.

Goldhaber, D., Perry, D., & Anthony, E. (2003). *NBPTS Certification: Who applies and what factors are associated with success?* The Urban Institute, Education Policy Center, Working Paper.

Gradwohl, J. M., & Schumacher, G. M. (1989). The relationship between content knowledge and topic choice in writing. *Written Communication, 6*(2), 181-195.

Graves, D. (2000). *Writing: teachers and children at work*. Portsmouth, NH: Heinemann.

Harper, G., & Kember, D. (1989). Interpretation of factor analyses from the approaches to studying inventory. *British Journal of Educational Psychology, 59*, 66-74.

Hattie, J. A. (1998). *Evaluating the Paideia Program in Guilford County Schools: First Year Report: 1997-98*. Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina, Greensboro.

Hattie, J. A. C. (2002). What are the attributes of excellent teachers? In *Teachers make a difference: What is the research evidence?* (3-26). Wellington: New Zealand Council for Educational Research.

Hattie, J. A., Clinton, J. C., Thompson, M., & Schmitt-Davis, H. (1996). *Identifying expert teachers*. Technical Report presented to the National Board for Professional Standards, Detroit. www.nbpts.org/research/archive_3cfm?id=6

Hattie, J., & Clinton, J. (2001). The assessment of teachers. *Teaching Education, 12*(3), 279-297.

Hattie, J., & Jaeger, R. (1998). Assessment and classroom learning: A deductive approach. *Assessment in Education*, 5, 111-121.

Hattie, J., & Purdie, N. (1998). The SOLO model: Addressing fundamental measurement issues. In B. Dart & G. Boulton-Lewin (Eds.), *Teaching and learning in higher education* (pp. 145-176). Melbourne, Austalia: ACER Press.

Hilgers, T. L. (1982). Experimental control and the writing stimulus: The problem of unequal familiarity with content. *Research in the Teaching of English, 16*(4), 381-390.

Hillocks, G. (1986). *Research on written communication*. Urbana, Illinois; ERIC Clearing House on Reading and Communication Skills.

Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. New York: Teachers College Press.

Holland, R. (2002). *National teaching certification: Advanced quality or perpetuating mediocrity*? Lexington Institute Archives (Online). Available: http://www.lexingtoninstitute.org/education/default.asp

Huot, B. A. (1993) The influence of holistic scoring procedures on reading and rating student essays. In M. W. Williamson & B. A. Huot (Eds.) *Validating holistic scoring for writing assessment.* (pp. 206-232). Cresskill, NJ: Hampton Press.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Utah: Utah University Press.

Jaeger, R. (1998). Evaluating the psychometric qualities of the National Board for professional teachers standards' assessment: A methodological accounting. *Journal of Personal Evaluation in Education, 12*(2), 189-210.

Johnson, R. L., Penny, J., & Gordon, B. (in press). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores?

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and interrater reliability of holistic scores in rating essays. *Written Communication, 18*(2), 229-249.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relationship between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*(2), 121-138.

Johnson, T. S., Smagorinsky, P., Thompson, L., & Fry, P. G. (2003). Learning to teach the five-paragraph theme. *Research in the Teaching of English, 38*(2), 136-176.

Kember, D., Jones, A., Loke, A., McKay, J., Sinclair, K., Tse, H., Webb, C., Wong, F., Wong, M., & Yeung, E. (1999). Determining the level of reflective thinking from students' written

journals using a coding scheme based on the work of Mezirow. *International Journal of Lifelong Education, 18*, 18-30.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, (25)*3, 287-298.

Ladson-Billings, G., & Darling-Hammond, L. (2000). *The validity of National Board for Professional Teaching Standards (NBPTS)/ Interstate New Teacher Assessment and Support Consortium (INTAC) assessments for effective urban teachers: Findings and implications for assessment.* Washington, D.C.: Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED448152).

Langer, J. A. (1984). Where problems start: The effects of available information on responses to school writing tasks. *Research in the Teaching of English, 18*(1), 27-44.

Larson, R. (1971). "Rhetorical writing" in elementary school. *Elementary English, 48*, 926-931.

Larson, R. (1992). Classes of discourse, acts of discourse, writers, and readers. *English Journal, 81*(8), 32-36.

Lawless, C. (1994). Investigating the cognitive structure of students studying quantum theory in an open university history of science course: A pilot study. *British Journal of Educational Technology, 25*, 198-216.

Levins, L. (1995*). A comparison of the developmental patterns in students' responses to the question in two science topics*. Australian Association for Research in Education, Hobert, Australia. Retrieved February 26, 2005 from http://www.aare.edu.au/95pap/levil95152.txt

Linn, R. (2000). Assessment and Accountability. *Educational Researcher, 29*, 4-16.

MacNealy, M. (1999). *Strategies for Empirical Research in Writing*. Memphis, TN: Allyn and Bacon.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: Outcome and process. *British Journal of Educational Psychology, 46*, 4-11.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: Outcome as a function of learners' conception of task. *British Journal of Educational Psychology, 46*, 115-27.

Marton, F., & Säljö, R., (1984). Approaches to learning. In F. Marton, D. Hounsell, D. N. Entwistle (Eds.), *The experience of learning*. Edinburgh: Scottish Academic Press.

Marton, F., Dall'Alba, G., & Beaty, E. (1993). Conceptions of learning. *International Journal of* Educational Research, *19*(3), 277-300.

Marzano, R. J. (1992). *A different kind of classroom: Teaching with dimensions of learning.* Alexandria, VA: Association for supervision and curriculum development.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives*. Thousand Oaks, CA: Corwin Press.

Marzano, R. J., Brandt, R. S., Hughes, C. S., Jones, B. F., Presseisen, B. Z., & Rankin, S. C., (1998). *Dimensions of thinking: A framework for curriculum and instruction*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., Pickering, D. J., Arredondo, D. E., Blackburn, G. J., Brandt, R. S., & Moffett, C. A. (1997). *Dimensions of learning: Teacher's manual* (2nd ed). Alexandria, VA: Association for Supervision and Curriculum Development.

McAlpine, I. (1996). A qualitative study of learning from CAL programs in two tertiary education courses. *Learning Technologies: Prospects and Pathways*. Retreieved April 15, 2005 from the

Australian Society for Educational Technology Web site: http://www.aset.org.au/confs/edtech96/edtech96_contents.html

Mendro, R. L. (1998). Student achievement and school and teacher accountability. *Journal of Personnel Evaluation in Education, 12*(3), 257-267.

Millman, J., & Schalock, H. D. (1997). Beginning and introduction. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp.3-10). Thousand Oaks, CA: Corwin Press, Inc.

Mogilner, A. (1992). *Children's writer's word book*. Cincinnati, OH: Writer's Digest Books.

Murray, D. (2002). *Write to learn*. (7th ed.). Boston, MA: Thomson Heinle.

Myford, C. M., & Englehard, G. (2001). Examining the psychometric quality of the National Board for Professional Standards Early Childhood/Generalist Assessment System. *Journal for Personnel Evaluation in Education, 15*(4), (253-85).

National Assessment of Educational Progress. (1998). *Writing framework and specifications for the 1998 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

National Board for Professional Teaching Standards. (2004). *About NBPTS: State and local support & incentives*. Retrieved March 28, 2005 from http://www.nbpts.org/about/state.cfm

National Board for Professional Teaching Standards. (2001). *I am a better teacher: What candidates for National Board Certification say about the assessment process*. Retrieved March 29, 2005, from http://www.nbpts.org/pdf/better_teacher.pdf

National Commission on Teaching and America's Future. (2003, January). *No dream denied: A pledge to America's children*. Washington, DC: Author. Retrieved January 14, 2005 from http://www.nctaf.org/article/?c=4&sc=16

National Commission on Writing in America's Schools and Colleges. (2003). *The Neglected "R": The Need for a Writing Revolution.* The College Board.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Nunally, T.E. (1991). Breaking the five-paragraph-theme barrier. *English Journal*, *80*(1), 67-71.

Paris, S. G. & McEvoy, A. P. (2000). Harmful and enduring effects of high-stakes testing. *Issues in Education, 6*, 145-159.

Pask, G. (1988). Learning strategies, teaching strategies, and conceptual or learning style. In R.R. Schmeck (Ed.), *Learning strategies and learning styles* (pp. 83-100). New York: Plenum Press.

Pegg, J., & Davey, G. (1989). Clarifying level descriptors for childrens' understanding of some basic 2-D geometric shapes. *Mathematics Education Research Journal*, *1*, 16-27.

Penny, J., Johnson, R. L. & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessment Writing*, *7*, 143-164.

Penny, J., Johnson, L., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education, 68*(3), 269-287.

Podgursky, M. (2001a). Defrocking the National Board. *Education Next, 1*(2), 79-82.

Podgursky, M. (2001b, April 11). Should states subsidize national certification? *Education Week on the WEB*, Retrieved December 7, 2004 from http://www.edweek.org/ew/articles/2001/04/11/30podgursky.h20

Popham, J. (1997). The moth and the flame: Student learning as a criterion of instructional competence. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 264-273). Thousand Oaks, CA: Corwin Press, Inc.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

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

Scholten, I., Keeves, J. P., Lawson, M. J. (2002). Validation of a free response test of deep learning about the normal swallowing process. *Higher Education*, 44, 233-55.

Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley & Sons.

Singer, J., Marx, R. W., Krajcik, J., & Chambers, J. C. (2000). Constructing extended inquiry projects: Curriculum materials for science education reform. *Educational Psychologist*, 35(3), 165-178.

Smith, T. (2004). Toward a prototype of expertise in teaching. *Journal of Teacher Education 55*(4), 357-371.

Stone, J. E., (2002). *The value-added achievement gains of NBPTS-Certified Teachers in Tennessee*: Retrieved January 23, 2005 from the Information for Decision Making Web site: http://www.financeprojectinfo.org/ProfDevelop/valueadded.asp

Stronge, J. H. (2002). *Qualities of effective teachers.* Alexandria, VA: Association for Supervision and Curriculum Development.

Stronge, J. H., & Hindman, J. L. (2003). Hiring the best teachers. *Educational Leadership, 60*(8), 48-52.

Stronge, J. H., & Tucker, P. D. (2000). *Teacher evaluation and student achievement*. Washington, DC: National Education Association.

Svensson, L. (1984). Skill in learning. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning*. Edinburgh: Scottish Academic Press.

The Teaching Commission. (2004). *Teaching at risk: A call to action*. New York: Author.

Thirunarayanan, M. O. (2004). National Board certification for teachers: A billion dollar hoax. *Teachers College Record*. Retrieved March, 2004, from http://www.tcrecord.org/Content.asp?ContentID=11266

Thompson, S. (2001). The authentic standards movement and its evil twin. *Phi Delta Kappan, 82*(5), 358-363.

Vandervoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board Certified Teachers and their students' achievement. *Education Policy Analysis Archives 12*(46), 1-117.

Vermunt, J. (1998). The regulation of constructive learning processes. *British Journal of Educational Psychology, 68*, 149-171.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for student learning. *Review of Educational Research*, *63*(3), 249-294.

Webster, W. J. (1995). The connection between personnel evaluation and school evaluation. *Studies in Educational Evaluation 21,* 227-254.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass Publishers.

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Columbus, Ohio: Merrill Prentice Hall.

Wilcox, D. W., & Finn, C. E. (1999, August 9). Board games: Failure of the National Board for Professional Teaching Standards to accomplish objectives of improving quality of teaching in the US: Business backs a losing education strategy. *National Review*. Retrieved December 7, 2004 from http://www.findarticles.com/p/articles/mi_m1282/is_15_51/ai_55234304

Wilson, S., Floden, R., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations*. Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and the classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.