INVESTIGATING THE RELATIONSHIP BETWEEN READING ACHIEVEMENT, AND STATE-LEVEL ECOLOGICAL VARIABLES AND EDUCATIONAL REFORM: A HIERARCHICAL ANALYSIS OF ITEM DIFFICULTY VARIATION

by

MICHELE YURECKO

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

written under the direction of

Gregory Camilli, Ph.D.

and approved by

Gregory Camilli, Ph.D.

Melanie R. Kuhn, Ph.D.

Douglas A. Penfield, Ph.D.

Jon S. Twing, Ph.D.

New Brunswick, New Jersey

January, 2009

ABSTRACT OF THE DISSERTATION

Investigating the Relationship between Reading Achievement, and State-Level Ecological Variables and Educational Reform:

A Hierarchical Analysis of Item Difficulty Variation

By MICHELE YURECKO

Dissertation Director:

Gregory Camilli, Ph.D.

This study identifies profiles in fourth grade reading achievement across states as measured by the 2002 National Assessment of Educational Progress (NAEP), and examines the link between these profiles and state policy and ecological variables. A series of multilevel models (MLM) that extend procedures traditionally employed in the analysis of differential item functioning (DIF) were applied in order to evaluate state-level performance at the *item level*. The variability of states' performances on individual items was estimated while controlling for overall state reading proficiency, and a residual variance statistic, item difficulty variation (IDV), was estimated for each item. A subset of items with relatively large IDVs was included in a second tier of analyses, and aggregated "parcel scores' were estimated. These parcel scores represent empirical clusters of items for which state membership influences student performance beyond what would be expected given state-level reading proficiency estimates. Two parcel scores were constructed in this analysis. Parcel 1 represented a cluster of items associated with long, fictional reading passages. Parcel 2 represented a cluster of items associated

with short, non-fictional reading passages. Parcel scores can be interpreted as value-added scores, suggesting that states with high Parcel 1 scores performed better on items associated with long, fictional passages than would have been predicted by overall state reading proficiency. Similarly, states with low Parcel 1 scores performed worse than expected. The interpretation of Parcel 2 follows in kind. As rates of non-native speakers and poverty increased across states, scores on Parcel 1 decreased and scores on Parcel 2 increased. These quantitative results, coupled with a qualitative case study of Maryland, New York and Texas, suggest the following major theme: larger, more populous states, with higher levels of poverty and non-native speakers of English exhibit a distinctive pattern in parcel score performance, scoring lower than predicted on items associated with long, fictional reading passages, and higher than predicted on items associated with short, non-fictional passages.

Acknowledgements

A frequently unconsidered benefit to completing a dissertation is the opportunity to formally and publicly thank those whose support and sacrifice made it possible. For as long as this dissertation survives, on a bookshelf or in cyberspace, it will be preceded by words of appreciation and gratitude for the people who enabled its completion. To my dissertation chair and committee, Gregory Camilli, Melanie R. Kuhn, Douglas A. Penfield and Jon S. Twing, I thank you for your support throughout this process. Your encouragement and praise buoyed me during times of frustration and doubt, and your constructive criticism impelled me to reach for the high standard to which you deemed me worthy of being held. In particular, I wish to thank my advisor, Gregory Camilli, for guiding me on this journey. Greg, you are like the Cheshire Cat, materializing with a vexing question or tantalizing comment, and sending me further down the rabbit hole. This process has been a true adventure, sometimes exhilarating, sometimes trying, but always challenging and worthwhile. Many members of the faculty and staff of the Graduate School of Education provided indispensable help at critical junctures during my studies. In particular, I wish to thank Jeffery K. Smith for first suggesting that I take a course in educational measurement; Angela M. O'Donnell for having an open door; Kris Spaventa for looking out for my best interests; and Carrie Armbrecht for assisting in far too many ways to list here. To my friends, particularly Jill Cermele, thank you for you support and encouragement. I also thank fellow GSE doctoral student, Serina Ting-Wei Chui, for her generosity in sharing her SAS expertise, and friendship. For my mentor, the late Joyce Zimmerman, I possess an abiding appreciation and affection, tinged with the regret that she did not live to celebrate the completion my Ph.D. Joyce never failed to

lead by example, and I aspire to grow into the scrupulously ethical and careful researcher she embodied. Thank you, Doug, for sending me to her. Finally, to my family, thank you for your love, patience and sacrifice. To my sister, Marybeth Yurecko, you humble and inspire me, and remind me everyday of what is truly important and special in my life. To my children, Eva and James Endahl, thank you for your unlimited supply of love and support. The time I have spent away from you comprises the greatest sacrifice I have had to make. You have exhibited patience, maturity and independence beyond your years, and I couldn't be more proud of you. To my husband, John Endahl, thank you for being my partner in this seemingly endless endeavor. You have shown your love for me through so many acts of patience, accommodation and support. Lastly, I thank my parents, Edward and Rose Marie Yurecko. I cannot craft sufficient words to express my gratitude for all you have done for me. Your contributions have ranged from the mundane to the sublime: from cooking dinner for my family, to inspiring me to believe that I am worthy and capable of my dreams. By loving and cherishing me, you inspired me to have enough confidence in myself to see this process through to the end. Thank you.

Dedication

For my parents.

Table of Contents

List of tables

List of figures

CHAPTER I. STATEMENT OF THE PROBLEM

In 1983, the National Commission on Excellence in Education issued the report, *A Nation at Risk*, which criticized public education in the United States as not only inadequate, but potentially dangerous by risking the security and prosperity of our society in its failure to produce citizens equipped to succeed in the coming new century. The commission called for fundamental changes in how and what we teach our children and how we run our public schools. Some critics of the report accused its authors of "cherry picking" facts in order to propagate a myth of achievement decline and mount a political assault on public education (Berliner & Biddle, 1995, 1996; Kosar, 2005). In spite of this controversy, many states responded to the report by launching efforts to influence both academic content and teaching in public schools.

Soon after the publication of *A Nation at Risk*, John Jacob Cannell, a West Virginia medical doctor, self-published the first of two reports which described what became know as the "Lake Woebegone effect" (Cannell, 1987, 1989). In these reports Cannell documented that all fifty U.S. states reported implausible standardized achievement test scores above the national average. He asserted that these anomalous score patterns were the result of a conspiracy of statistical manipulation, dubious test preparation practices, and outright cheating, rather than any authentic improvement in academic achievement. Cannell's findings were widely publicized in print and televised media, and further spurred the national debate on education reform and testing (Cannell, 2006).

In 1986, the National Governors Association issued the report, *Time for Results*, and demanded higher expectations and accountability measures for our nation's public

schools. In 1989, President George H. W. Bush presided at the first National Education Summit where he echoed these earlier calls for greater accountability and higher standards. Six National Education goals were subsequently adopted, and congress commissioned the National Education Goals Panel in 1990 to track state efforts to meet the National Education Goals by the year 2000 (Baron, 1999).

In 1994, President William Jefferson Clinton signed the Goals 2000: Educate America Act into law:

> Goals 2000 became the most pervasive national K-12 education policy in a generation. It provided federal incentives for states to create new systems of accountability be setting their own standards and creating new assessments, which the states did. At the start of the decade only a handful of states had academic standards. By the end of it, close to fifty states had developed standards. (Kean, 2003, p. 327).

In his 1997 State of the Union Address, President Clinton further championed "a national crusade for education standards" (para. 27) and voluntary national achievement tests in reading and mathematics. Specifically regarding literacy standards, he asserted that every child must be able to read by the end of third grade.

At that time, Goals 2000 was regarded as a broad and sweeping national education policy. It was frequently criticized by Republicans and conservatives as overreaching and intrusive in its influence (Kean, 2003; Rabb, 2004). The core of their argument lay in an ideological commitment to states' rights and local control of education. The subsequent Republican president, George W. Bush, signed an arguably more pervasive act into law with the passing of No Child Left Behind (NCLB),

reauthorization of the Elementary and Secondary School Education Act (Rabb, 2004). President Clinton's earlier remarks regarding elementary school readers were codified by the Reading First program (established by Title I, Part B, Subpart 1 of NCLB), which intended to ensure that by the end of the third grade, all American children could read at or above grade level (Schenck, Walker, Nagel, & Webb, 2005). NCLB also mandated that each state adopt challenging academic standards for all public elementary and secondary school students in mathematics, reading/language arts, and science as of the 2005-2006 school year. In addition, each state must establish a formal accountability system with measurable indicators of adequate yearly progress (AYP). Furthermore, these accountability systems must include rewards and sanctions for local agencies and schools with regard to AYP. (For more information, see Public Law 107-110, No Child Left Behind Act of 2001, Sec. 1111. State Plans.)

Given this climate of political will and public support, the standards-based reform movement has become characterized by state-level policy initiatives intended to "elicit, encourage, or demand changes in teaching and learning," (Valencia & Wixson, 1999, p. 1). Polices have taken many different forms, including: implementation of state content standards; implementation of aligned state testing programs; increased stringency in promotion and graduation requirements; accountability systems intended to hold students, teachers, administrators and districts responsible for inadequate achievement outcomes; professional development for teachers regarding pedagogy and content; higher standards for teacher certification; and changes in school organization and management (Baron, 1999; Cizek, Trent, Crandell, Hirsh, & Keene, 2000; Goertz, 2001; Valencia & Wixson, 1999). These initiatives have been implemented in order to encourage coherence and

uniformity in educational expectations and goals, which can then be directed to students, teachers and local districts. Nevertheless, various stake-holders (e.g., parents, teachers, local administrators, politicians) can hold divergent opinions. For example, Dutro (2002) noted that finding consensus among stake-holders as to what constitutes crucial and appropriate content for academic standards can be problematic.

As reported by a number of researchers, a wide variety of policy factors have been found to impact student academic performance as measured by large-scale achievement tests (see Amrein-Beardsley & Berliner, 2007; Baron, 1999; Dutro, 2002; Grissmer & Flanagan, 1998; Grissmer et al., 2000; Lee, 2006; Monfils, 2004; Nichols, Glass, & Berliner, 2006; Olson, 2006). These factors include instructional time, duration of school year, student/faculty ratio, tracking, placement procedures for students with limited English proficiency (LEP) or disabilities, behavioral climate, teacher education and certification, curricula alignment to state standards, staff professional development, preschool participation, full-day kindergarten, supplemental programs for at-risk students, teacher salaries, availability of instructional materials, and school size. Efforts to investigate the effects of these different policy initiatives have frequently been inconsistent and the results contradictory. For example, while many researchers have investigated the relationship between academic achievement and the use of high-stakes testing as an accountability tool (Amrein-Beardsley & Berliner, 2003; Grissmer & Flanagan, 1998; Lee, 2006; Nichols et al., 2006), comparatively fewer studies attempt to investigate the potential influence of other systematic reform initiatives, such as the adoption of state-wide academic content standards (Bracey, 2000; Dutro, 2002; Olson 2006).

The current study examines the potential link between reading performance and state-level correlates, including, but not limited to, systematic educational reform. The design of this study also anticipates the potential influence of state-level, non-policy ecological variables (such as social and demographic factors) on educational outcomes. These non-policy factors are not only useful in providing context for the interplay of policy initiatives, but may also prove to be as relevant and important as state-level policy initiatives in explaining patterns in reading performance.

Purpose

The purpose of this study is to identify different profiles in reading achievement across states as measured by the 2002 fourth grade reading test of the National Assessment of Educational Progress (NAEP), and subsequently to examine the link between particular achievement profiles and state-level policy and ecological variables. The end goal of this analysis is to identify noteworthy patterns in reading achievement and uncover possible state-level correlates. In addition, this study may also provide specific information to direct and improve state reading policies.

Using techniques elaborated by Camilli, Monfils and others (Camilli & Monfils, 2003; Camilli et al., 2006; Camilli, Prowker, Vargas, & Waszkielewicz, 2005; Monfils, 2004; Prowker & Camilli, 2006), a series of multilevel models that extend procedures traditionally employed in the analysis of differential item functioning (DIF) is applied to the 2002 NAEP fourth grade reading test data. Though it is common practice in policy analysis to examine state effects at the total score level, in this study, state performance is evaluated at the *item level*. The variability of state performance on individual items is estimated while controlling for overall state reading proficiency levels. A residual

variance statistic called item difficulty variation (IDV) is estimated for each item. Items with IDVs of relatively large magnitude may indicate the presence of differential item functioning across states. These items are then inspected in order to determine if their value-added (or value-subtracted) effects can be attributed to state membership beyond what would be expected given state reading proficiencies. Items with large IDV may indicate that in addition to reading proficiency, state-level factors (such as the presence of particular policies) influence performance on particular items.

A subset of items with relatively large IDVs is selected for further inspection and included in a second tier of analyses. A factor analysis is conducted using those items in order to obtain "parcel scores" that represent empirical clusters of items for which state membership influences student performance beyond what would be expected based on state-level ability estimates. Unlike total test and subscale scores, these parcel scores may not necessarily be aligned with any underlying cognitive constructs or processes. Parcel scores may also have diagnostic value in highlighting item-level characteristics (e.g., academic content, cognitive process, item format) that are affected by state-level factors, such as accountability pressure, teacher preparation, or poverty (Camilli et al., 2006). A correlational analysis is then conducted with a battery of moderator variables in order to backward-map parcel score performance patterns onto state-level social and demographic characteristics, and policy initiatives. These results may suggest specific item dimensions that are associated with particular external variables (e.g., specific state policies such as implementation of content standards and state assessments; demographic variables; social variables), and offer an enriched description of the interplay of state-level ecological and policy variables, and reading test performance. Based on these results, three states are

selected for an exploratory collective case study. This qualitative aspect is included to demonstrate how parcel scores may be used to explore the potential link between individual state characteristics and reading performance.

The following chapters provide more details of this analysis and the results. Chapter II contains a review of the literature, and examines two primary bodies of research. The first section presents an overview of studies that have examined the link between reading achievement and systematic reform in the form of academic content standards. In the second section, the methodological legacy that underlies the statistical procedures proposed for this study is explicated. Finally, given the context provided by this discussion of the literature, a brief outline of the current study is provided, including a discussion of its place among the previous research. The third chapter outlines the study design and methodology. It also includes a detailed description of the data, including a discussion of the history of NAEP and an overview of the 2002 NAEP reading assessment of fourth graders. The research results of this analysis are presented in Chapter IV, followed by a summary and discussion in Chapter V.

CHAPTER II. LITERATURE REVIEW

Efforts by researchers to link systematic educational reform and academic achievement have been inconsistent. In order to provide an illustration of these circumstances, this chapter begins with a review of the literature that examines the relationship between reading achievement and one prominent example of systematic educational reform, the implementation of state-level academic content standards. Following this discussion, the analytic methods that provide the statistical framework for this dissertation are described. This chapter ends with a brief overview of the proposed study and placement of this study in the context provided by the previous research.

Academic Content Standards and Reading Achievement

Both popular opinion and the assertions of some policy-makers endorse as a *fait accompli* a causal relationship between the adoption of "high-quality" content standards and increased academic achievement. Nevertheless, there is relatively little evidence-based research to support this assertion, let alone a consensus as to what constitute "high-quality" academic standards (Olson, 2006). A number of large studies that rate the quality of literacy or reading standards (e.g., Gottlieb, 2001; Otuya & Krupka, 1999; Schenck et al., 2005; Stotsky, 2000; Stotsky, 2005) ignore academic achievement outcomes altogether. Ironically, as some proponents of standards-based reform call for increased measures of educational accountability, they fail to apply stringent accountability tests to the policies they support (Bracey, 2000). Arguably, the systematic evaluation of state policies should include an examination of relevant educational outcomes, and could serve as a useful tool in designing and fine-tuning effective policy implements. Furthermore, given the intent of NCLB with regard to basing educational practice on scientifically

based reading research, policy-makers could be obliged to demonstrate the efficacy of systematic reform on student reading performance with soundly constructed research studies. (See Camilli, Wolfe, & Smith, 2006, for a discussion of meta-analysis and scientifically based literacy research.) A thorough review of the literature revealed merely a handful of studies that have examined the link between state-level reading performance and the implementation of academic content standards (Baron, 1999; Dutro, 2002; Lee, 2006; Olson, 2006).

In fulfillment of its 1998 Congressional mandate to report on promising national, state, and local educational initiatives, the National Education Goals Panel commissioned a study to explore the link between Connecticut's educational policies and its high levels of reading achievement (Baron, 1999). A qualitative case study was conducted to address six research questions related to reading achievement. Two of those questions directly addressed state-level policy variables:

- To what extent can Connecticut's high and improved reading scores [as measured by the Connecticut Mastery Test (CMT) and the NAEP] be explained by its educational policies rather than its wealth, race/ethnicity, and parental education?
- What state-level policies and practices are likely to have contributed to the improved reading scores? (Baron, 1999, p.3).

Responses to these questions were obtained through interviews with approximately two dozen stake-holders within the Connecticut education system. Participants included district superintendents, school board members, principals, language arts coordinators,

reading consultants, classroom teachers, and professional development providers. These interviews served as the primary data source for the analysis.

During interviews, subjects identified six policies and practices they believed contributed to Connecticut's outstanding reading achievement as measured by the Connecticut Mastery Test (CMT) and NAEP from 1992 to 1998. These six policies and practices fell into two categories: state-level accountability pressure, and educational resources and support. None of the favored policies pertained directly to academic content standards. With regard to accountability, stake-holders identified two state policies that they believed contributed to improved reading achievement. The mandatory requirement of district participation in the CMT for grades four, six and eight was identified as a major agent in focusing classroom instruction and aligning local curricula with the test. In addition, test results for local schools were publicly reported to boards of education and released to local newspapers. "Many administrators expressed that this highly public school-by-school reporting…had a strong impact on their instruction and student achievement" (Baron, 1999, p.27).

The four other factors identified by stake-holders pertained to state-level support and resources. First, varied and flexible reporting of CMT scores by the state allowed districts to use their own test data to inform local curricular decisions. The state also provided each district with their own data files and a tailored software package to disaggregate and reanalyze local test data. Second, supplemental CMT testing in grades three, five and seven was made available to districts. These tests were aligned with the required testing in grades four, six and eight. Subjects felt that this supplemental testing reinforced consistency in the curriculum across grade levels. Third, state-level support in

the form of financial and human resources were increased for the neediest and lowest achieving districts. Local stake-holders emphasized their belief that this support contributed to gains in reading scores among the lowest performing students in the state. Finally, as a result of the Education Enhancement Act (EEA) of 1986, Connecticut teachers were among the highest paid teachers in the world by 1999. In addition to raising salaries, the state also raised standards for new teachers and provided a system of professional supports. Stakeholders believed that these initiatives (higher salaries, higher standards and professional support) spurred the hiring and retention of highly qualified and competent classroom teachers resulting in increased reading achievement scores on both the CMT and NAEP reading and mathematics tests from 1992 to 1998.

Several methodological concerns temper Baron's (1999) conclusions. The interview data were not triangulated with external sources, such as independent research reports, observations or public documents. Findings were not validated with additional analyses that attempted to establish a measurable relationship between identified policy variables and reading performance on the CMT or NAEP. The sample selection procedures used to identify the stake-holders were not described, and the extent to which their opinions adequately and accurately reflected the educational climate in Connecticut is unknowable. In addition, Baron accepts a causal link between the factors cited by her subjects and Connecticut's high test scores without mention of the possibility of undesirable sources of score inflation, such as teaching to the test or cheating.

In a qualitative case study of a single elementary school in Morretown, California, Dutro (2002) examined the relationship between state content standards and student reading achievement in the first grade. Dutro found that the influence of state standards

on reading test scores could not be properly analyzed without consideration of a panoply of local variables, such as "district initiatives, curriculum adoption, shifting district demographics, and the individual expertise of teachers" (Dutro, 2002, p. 6). Characterization of the relationship between state standards and reading achievement was difficult due to the complex array of influences embedded in the context. "The impact of macro-level policies [was] dependent on numerous micro-level issues such as district decision-making, teacher beliefs, and social dynamics among school staff" (Dutro, 2002, p. 3). Furthermore, as documented by Spillane (1998) in a case study of two school districts, the unique composition of contextual and cultural variables across different school districts (or even across schools within the same district) may result in dramatically different approaches to implementation of the same educational policies.

Dutro suggested that increased reading scores on the California state achievement test may be due to the interactive effects of two sets of variables: state-level variables (e.g., highly detailed literacy content standards, and intense accountability pressure), and district-level curricular and ecological variables (e.g., local curriculum content, and student demographics). The influence of state-level systematic reform was manifested locally in a number of ways. The district adopted a state-approved reading curriculum (Open Court) that was aligned with the content standards. Teachers in the district reported that they actively used state literacy standards in both long-term and short-term planning for instruction. In addition, Dutro suggested that state-level accountability pressure instigated local initiatives to boost reading achievement, such as the organization of book clubs before and after school; tutoring of struggling readers; increased instructional time allocated to test preparation; and active recruitment of higher achieving

students from more affluent socio-economic strata via the implementation of a specialty arts program in the school. While these conclusions do not unequivocally establish or explain a relationship between standards-based reform and reading achievement, several potential variables of interest were highlighted, and some insightful descriptions were provided concerning how state-level policy variables may manifest at the local level where achievement testing takes place.

Using hierarchical linear growth models, Lee (2006) investigated the effects of two types of state policy emphases on academic achievement. Data from all fifty states were used in an analysis of the effects of test-driven accountability policies and state support for school resources on student achievement, as measured by the NAEP reading (1992-2003 for fourth grade, 1998-2003 for eighth grade) and mathematics (1992-2003 for fourth grade, 1990-2003 for eighth grade) assessments. Lee found that state pressure in the form of high-stakes testing was not necessarily accompanied by dedicated resources or fiscal support (e.g., per-pupil expenditures, class size reduction, and in-field teaching). Only three states, Indiana, New Jersey and New York, were classified as consistently strong on both accountability and school resource measures.

Regarding the potential link between test-driven accountability policies and student achievement, achievement gains were positively and significantly related to accountability policies for mathematics only. Although these gains were statistically significant, Lee characterized the size of the effect as “slight.” A similar relationship was detected between support for school resources and mathematics achievement, with modest, but significant, gains associated with state-level support in the form of dedicated resources. The interaction between state accountability and school support was also

significant for mathematics achievement. State accountability had a greater effect on mathematics achievement gains in those states with stronger support for school resources. Mathematics gains for states with both high accountability pressures and high resources were estimated to be about one third of a standard deviation higher than average expected gains. “When the state support for school resources was relatively low, state activism in school accountability policy hardly made a difference in the size of achievement gains” (Lee, 2006, p. 59). This outcome may suggest that educational policy, accompanied by adequate resources and support, offers states the best opportunity to increase mathematics achievement. No significant relationships were detected for reading achievement.

Although the national growth rate in reading achievement as measured by the NAEP fourth and eighth grade tests increased significantly over the 1992-2003 period, these gains were considerably smaller than those in mathematics. “The average amount of gain per year was about five times larger for math than for reading,” (Lee, 2006, p. 56). In addition, no significant effects were detected for accountability, school resources, or the accountability-by-resource interaction on reading achievement gains. Lee suggests that these findings may be attributed to concentrated efforts by states to focus school reform on mathematics, to the neglect of reading. However, Valencia and Wixson (1999) caution against the assumption that the application of policy reforms will have a uniform effect across different content domains. Fry (1998) suggests that mathematics may lend itself more easily to standards-based reform efforts, while reading is a less discretely defined domain that is inherently more challenging to address with content standards and standardized testing. In addition, Dutro (2002) cautions that “research on current and previous reform movements has shown, [that] state, district, and teacher-level issues

interact in unique ways that make it difficult to ascribe change [in reading achievement] to any one element" (p. 2). Although Lee did not find evidence supporting a policy effect on reading achievement, this may not be the result of the neglect of state policy-makers to attend to reading, but rather evidence of the inherent difficulty in effectively addressing the complexity and nuance of literacy with broad, state-level policy tools.

In a special edition of *Education Week, Quality Counts at 10: a Decade of Standards-Based Education*, Olson (2006) posed a key question: is there any evidence to suggest that the past decade of standards-based educational policy has improved academic achievement? To answer this question, the Editorial Projects in Education Research Center (EPE) was commissioned to conduct a series of regression analyses examining the relationship between standards-based educational policy and student achievement as measured by the NAEP fourth and eighth grade mathematics and reading scores from 1992 to 2005. Consistent with Lee's (2006) findings, the EPE analysis revealed a moderate positive relationship between overall implementation of standards-based policies (as measured by a policy composite variable) and gains in NAEP mathematics achievement. Consistent with Lee's (2006) findings and Fry's (1998) predictions, EPE discovered a slightly negative relationship between standards-based policies and NAEP reading achievement. In other words, the advent of standards-based reform as described in the EPE analysis occurred with small losses in NAEP reading achievement.

In a subsequent analysis to unpack the effects of mingled state policies on reading achievement, EPE deconstructed the policy composite variable and examined the relationships between individual policy variables (e.g., presence of teacher quality

initiatives, content standards, assessment programs, accountability systems) and reading achievement. Depressed NAEP reading scores were determined to be the result of "a negative relationship between state efforts to improve teacher quality and gains in student [reading] achievement" (Olson, 2006, p.10). Three other major policy components included in the composite variable (state implementation of content standards; implementation of state assessment programs; and state-level accountability systems) were found to be positively related to both mathematics and reading achievement. Olson did not offer an interpretation or explanation of the negative relationship between reading performance and efforts to improve teacher quality.

Inspection of the results of the previous four studies reveals a number of important aspects of the potential relationship between state educational policy and reading achievement. First, studies attempting to link educational policy to reading achievement have yielded mixed results. Of the four studies discussed in this literature review, only two attempted to quantify the relationship between policy and reading achievement. Both of these quantitative studies reported mixed results regarding the relationships between some standards-based policies and reading achievement, while the two case studies (Baron, 1999; Dutro, 2002) asserted that standards-based reform was associated with positive gains in reading achievement. Second, relationships between reading performance and state-level policy and ecological variables may differ from those documented for other content areas, such as mathematics. In fact, the efficacy and appropriateness of state content standards in reading may qualitatively differ from other academic disciplines (Valencia & Wixson, 1999). It is unclear what lies at the root of this difference: the unique nature of literacy development in children, the manner of reading

instructional practices in U.S. schools, or other unknown factors. Third, composite variables of education policies may obscure differential effects across individual state-level variables, as indicated by Olson (2006). Consequently, state-level variables may need to be considered individually, in conjunction with data reduction or variable aggregation.

Much of the research that examines academic achievement both within and across states utilizes standardized test data in the form of total test scores or, occasionally, nominal subscale scores. A number of major drawbacks in using total and subscale test scores exist. Differences in performance across states may be obscured because states with similar total scores may actually possess very different response patterns across individual items. As noted by Schmidt and colleagues (1997), schools may disproportionately focus on particular aspects of curriculum over others in response to state-level accountability pressures, and differences in curricular emphasis and classroom practice could be absorbed by the total score metric. Several scholars have suggested that item level analyses may provide more power in discerning differences in group performance than total or subscale scores (Camilli & Monfils, 2003; Monfils, 2004; Porter, 1988; Swaminathan & Rogers, 2000). Furthermore, item level analyses may be more effective in linking state-level policy initiatives to student achievement because these analyses are better able to detect subtle differences in state performance patterns (Camilli et al., 2005; Prowker & Camilli, 2006).

The current study addresses the limitations of the previous research through application of techniques elaborated on by Camilli, Monfils and others (Camilli & Monfils, 2003; Camilli et al., 2006; Camilli et al., 2005; Monfils, 2004; Prowker &

Camilli, 2006). A series of multilevel models that adapt procedures traditionally employed in the analysis of differential item functioning (DIF) are applied to a national sample of reading achievement data from the 2002 NAEP of fourth graders. This analysis includes state, school, teacher, and student level moderator variables in an interpretation of effects. Because state policies do not exist in a vacuum, but function in a context that also includes social and demographic variables, as well as local policies and characteristics (Dutro, 2002), this study may provide an advantage over previous research due to the incorporation of multiple contextual variables. In addition, data is evaluated at the item level in an effort to link state-level variables directly to student achievement patterns. In the following section of the literature review, the evolution of this analysis is explained, and relevant studies that have employed this methodology to examine school and state-level effects are discussed.

An Overview of Multilevel Modeling for Estimating Group Effects

The following section provides a review of the recently developed models and statistical techniques that are employed in this study, beginning with overviews of differential item functioning (DIF) and multilevel modeling (MLM). This section ends with a brief discussion of how these two methodologies can be combined to form multilevel item response models for estimating item difficulty. This class of multilevel models is used in the main analysis of this dissertation.

Differential Item Functioning (DIF)

Differential Item Functioning occurs when the probability of correctly answering a test item varies across two or more groups of examinees of comparable proficiency (Holland & Thayer, 1988). A battery of statistical procedures can be applied in a DIF

analysis in order to detect potentially biased test items. When an item exhibits DIF, members of a particular group (or groups) enjoy an advantage in answering an item correctly that is not associated with their proficiency, but rather with a construct-irrelevant group characteristic such as gender, socio-economic status, or ethnicity. It is important to note that statistical methods used for identifying DIF cannot determine item bias. Rather they can merely flag items based on statistical anomaly. Subsequent expert review must be undertaken to determine if differentially functioning items are truly biased, or merely reflect a legitimate group difference on the underlying construct(s).

Over the past twenty-five years, methods of DIF analysis have grown more sophisticated to accommodate various testing conditions and practices. The simplest DIF analysis involves the comparison of two groups (a reference group and a focal group) on a dichotomously scored test item. Methods have been expanded to include the detection of DIF across more than two groups; for polytomously scored items; for subsets of items; for items developed within an Item Response Theory (IRT) framework; and for situations in which items are nested within a hierarchy of potential effects (Penfield & Camilli, 2007). In the case of multiple groups, the variance of the item difficulty estimate can be calculated, in lieu of a comparison of individual item difficulties. Items with large difficulty variation may be flagged for further examination. The methodology of the current study adapts DIF techniques in the analysis of item performance across multiple groups (i.e., across states) within an IRT framework with nested contextual variables.

Multilevel Modeling

Within educational research, variables of interest are frequently embedded in a nested or hierarchical context. For example, students may be nested within classrooms, nested within schools, nested within districts, nested within states.

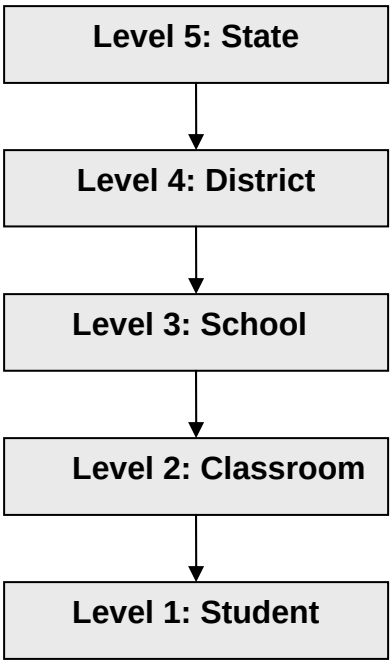


Figure 2.1. A Sample Multilevel Model

A multilevel modeling framework can accommodate variables in a hierarchical structure by explicitly modeling the manner in which variables are grouped within nested levels (Goldstein, 2003). Such models permit detection of patterns in the data and estimate effects that might be overlooked or obscured by traditional methods that cannot accommodate the intercorrelation of layers of variables.

Multilevel modeling (MLM) frameworks are comprised of nested families of regression equations. The Level 1 model specifies the regression equation of a random

outcome variable at the primary unit of analysis. Such variables can be described as random because they are permitted to vary randomly across units. In the example above, the Level 1 outcome variable would be a student outcome variable, such as reading performance, which varies randomly across individuals. It is important to note that multilevel models can accommodate random outcomes (i.e., dependent variables) of different forms. For example, in the case of total or subscale test scores, this variable could be viewed as continuous. In the case of dichotomously scored item data, this variable would be discrete, taking on a value of 0 (incorrect) or 1 (correct). For noncontinuous outcome variables, a link function must be specified to linearize the relationship between the random outcome variable and the set of predictors. For dichotomously scored data, the logit link function is commonly used (Goldstein, 2003; Fielding, 2003).

The Level 2 equations within a MLM framework model the regression coefficients obtained by the Level 1 model as random variables, and also include contextual variables associated with the Level 2 unit of analysis as fixed coefficients. Recall the previous example (see Figure 2.1) where students (Level 1) are nested within classrooms (Level 2). The Level 2 equations may include explanatory variables associated with classroom characteristics, such as class size. Similarly, Level 3 equations model the Level 2 effects (given the associations with the Level 1 variables), and can also include another set of contextual variables at the Level 3 unit of analysis. If the previous example is extended to include a third level (students nested within classrooms nested within schools), the Level 3 equations may include school level explanatory variables, such as school size or organizational structure.

When analyzing nested data, multilevel models have several advantages over traditional methods of analysis, such as multiple linear regression. “First, [a multilevel model] enables data analysts to obtain statistically efficient estimates of regression coefficients. Secondly, by using the clustering information, it provides correct standard errors, confidence intervals and significance tests, and these generally are more ‘conservative’ that the traditional ones that are obtained simply by ignoring the presence of clustering” (Goldstein, 2003, p. 3). Thirdly, the impact of higher level covariates on lower level variables can be teased apart and explored in a multilevel analysis, whereas in a traditional analysis the influence of layers of nested variables may be difficult, or even impossible, to detect. Finally, the relative performance of subjects on the primary outcome variable can be compared in light of the embedded contextual effects of higher level variables. An understanding of the relative performance of individual units will be enriched by the information that a multilevel model can provide about the effects of nested contextual variables.

Multilevel DIF analyses incorporate MLM techniques in order to estimate group membership effects on item performance within a hierarchy of contextual variables. Such analyses accommodate correlations among variables, and allow for the separation and estimation of effects at different levels within the hierarchy. In a new approach to diagnosing differences in academic achievement as measured by standardized tests, a number of authors (Camilli & Monfils, 2003; Monfils, 2004; Prowker & Camilli, 2006) have applied multilevel DIF techniques within an IRT framework to identify school and state effects on academic achievement. These techniques are modifications of previously developed three-level hierarchical generalized linear models (Kamata, 1999a, 1999b,

2001; Rogers, Swaminathan & Egan, 1999; Rogers & Swaminathan, 2000) and provide the methodological framework for this study.

Camilli and Monfils (Camilli & Monfils, 2003; Monfils, 2004) extended multilevel DIF procedures within an IRT framework in an analysis of school effects on fourth grade mathematics achievement as measured by the 2001 New Jersey Elementary School Performance Assessment (ESPA). In traditional DIF analyses of multiple groups, the goal is to identify items with large item difficulty variation across groups, while controlling for overall ability. In spirit, this study is similar, in that the authors identified items with comparatively large IDVs, and outlined an approach for examining potential relationships between individual items and school contextual factors, such as school size, class size and school wealth. In an analysis of school effects, items with highly variable difficulty were linked to school level factors, such as class size and faculty/student ratio. In a separate set of publications (Camilli et al., 2006; Camilli et al., 2005; Prowker & Camilli, 2006), the same MLM techniques were used in an analysis of the 2000 NAEP mathematics assessment of fourth graders.

The MLM frameworks discussed above utilized discrete item response data and incorporated a one-parameter (1PL) or Rasch IRT model that only included item difficulty parameters, i.e., IRT *b* parameters. Multilevel item response models can be generated for 2PL IRT formulations that include both *b* (item difficulty) and *a* (item discrimination) parameters. In a comparative analysis, Camilli and Monfils evaluated both 1PL and 2PL multilevel models using simulated data (Camilli & Monfils, 2003; Monfils, 2004). For the 1PL model, item difficulty estimates from the multilevel model appeared to be quite stable and varied little from the original IRT parameter estimates;

however parameter estimates from the 2PL multilevel model were not consistent with original IRT parameter estimates. The authors speculated that until scale indeterminacy issues associated with item discrimination are resolved, the 1PL model is most suitable for formulating multilevel analyses of item difficulty variation. Omission of the IRT *a* parameter may result in model misspecification since variability associated with item discrimination would not be partitioned within the model, and interpretation of the results of the 1PL model should acknowledge this omission. However, given the instability of *a* parameter estimates with the current methodology, the 1PL or Rasch model is employed in this study.

The Current Study

The current study addresses a significant gap in our national discussion of standards-based education reform, through a quantitative analysis of the relationship between state-level policy and ecological variables, and reading achievement. Using techniques elaborated by Camilli and Monfils (Camilli & Monfils, 2003; Monfils, 2004), this study examines how the effects of state-level policies and social and demographic variables bear upon the reading performance of fourth grade students as measured by the 2002 NAEP. By focusing the analysis at the item level, rather than the subscale or test level, the results of this study should yield a fine-grained analysis that allows for the identification of specific item dimensions (e.g., cognitive processes, content areas, item formats) that are associated with particular external variables (e.g., specific state policies such as implementation of content standards and state assessments; demographic variables; social variables).

The design and methodology of this current study differs from previous research that has examined differences among states in reading achievement in several significant ways. First, by using an adaptation of multilevel modeling, this study allows for the estimation of state-level effects in a hierarchical context. Much of the previous research neglects to examine or account for the layers of related variables frequently associated with academic performance. Second, unlike much previous research on the effects of educational policies and state ecological factors, the current study uses item-level achievement test data (rather than total score or subscale results) in order to isolate and specify items that reveal meaningful between-state differences in reading achievement. Item level analyses may provide more diagnostic clarity than those involving total score or subscale scores. Third, this study uses "parcel scores" to characterize state-level performance. Typically in achievement test data, total test and subscale scores are engineered to conform to an underlying latent trait or traits. Unlike total test and subscale scores, parcel scores are not necessarily aligned with any underlying cognitive construct or process. Rather, they represent organic clusters of test items that arise from differential state performance on individual items. Finally, this current study employs both quantitative and qualitative research methods in an attempt to identify and explain differences in reading achievement. This combination of statistical techniques and qualitative research methods enables a richer, more detailed description of the nature of between-state differences in reading achievement.

CHAPTER III. STUDY DESIGN AND METHODOLOGY

Chapter III begins with the posing of the central research questions, and a very brief overview of the study design. This section is followed by a detailed description of the data, including a discussion of the history of NAEP and current assessment characteristics. The study design and methods are then described in detail, including full explication of the four steps of the proposed analysis.

Research Questions

This study is essentially guided by two broad research questions:

(a) What kinds of profiles of reading achievement can be detected across states?

(b) What kinds of state-level contextual factors can be identified that are associated with those profiles?

An analysis using the NAEP 2002 reading assessment of fourth graders is conducted in order to identify and explain differential performance in reading achievement across all participating jurisdictions.

A Brief Overview of the Study Design

The proposed study design can be divided into four steps.

Step 1: Identification of items with large item difficulty variation (IDV) using multilevel item response models,

Step 2: Factor analysis of identified item residuals,

Step 3: Parcel score construction and estimation,

Step 4: Moderator variable analysis.

In the first step of the analysis, techniques elaborated by Camilli, Monfils and others (Camilli & Monfils, 2003; Camilli et al., 2006; Camilli et al., 2005; Monfils, 2004;

Prowker & Camilli, 2006) are applied in a series of multilevel models that adapt procedures traditionally employed in the analysis of DIF. A set of target items with significantly large IDV estimates are identified for the 2002 NAEP fourth grade reading test.

An exploratory factor analysis is then conducted on the residuals obtained from the target items. The goal of this portion of the analysis is to identify common variance among the residuals and examine the underlying factor structure among the target items. These factors indicate additional shared variance among items, above and beyond that accounted for by overall state reading proficiency. Clusters of related items such as those indicated by the factors may be used diagnostically with regard to differential state performance (Camilli et al., 2006).

In the third step of the analysis, parcel scores are constructed based on the obtained factor structure. Parcel scores are analogous to item bundles (Douglas, Roussos & Stout, 1996), or market baskets (Mislevy, 1996). They are aggregated scores for sets of items with correlated residuals, and represent clusters of items for which state membership influences student performance beyond what would be expected based on state-level proficiency estimates. Unlike total test and subscale scores, parcel scores are not necessarily aligned with any underlying cognitive construct or process as defined in a set of test specifications. Rather, they represent organic clusters of test items that arise from differential state performance on individual items. Interpretation of these parcel scores may inform our understanding of differential reading achievement across states.

Following the estimation of the parcel scores, a correlational analysis is conducted in order to relate parcel score performance patterns to a battery of moderator variables

that represent state-level policy and ecological factors. In particular, state-level moderator variables are backward-mapped onto parcel score results. This analysis provides a detailed description that allows for the identification of specific item dimensions (e.g., cognitive processes, content areas, item formats) that may be associated with external variables, such as educational policies, demographic variables, and social variables. Finally, these results are used in the selection and examination of three states in a collective case study.

Data: The National Assessment of Educational Progress (NAEP)

One of the main goals of this study is to compare differences in reading performance among states. In order to make meaningful state-by-state comparisons, a common metric for measuring reading achievement is necessary. Although all states currently administer some form of standardized achievement testing in their public schools, both these tests and the standards with which they may be aligned vary from state to state (Olson, 2006). While these different assessments may allow for the evaluation of academic performance with regard to particular educational standards *within* states, they do not necessarily permit valid comparisons in academic performance *between* states. In addition to a common metric, sufficiently large and representative samples within states are required in order make state-by-state comparisons in academic performance. The National Assessment of Educational Progress (NAEP) represents a common metric across sufficiently large, representative samples that permit such comparisons.

Prior to the 1960s, education was viewed by most Americans as a local issue; but with the changing social climate after World War II, the onset of the Cold War, and the

launch of Sputnik in 1957, the federal government increased its role in public education (Brain, 1969, 1971; Vinovskis, 1998). Previously, the primary duties of the federal Department (and later, Bureau) of Education concerned the collecting and reporting of statistics about the state of education in the United States, per the 1867 bill that created the department. Over the past 35 years, the role of the federal government has expanded dramatically beyond its role as compiler and reporter of statistics (Kean, 2003; Vinovskis, 1998).

During 1963 and early 1964, then U.S. Commissioner of Education, Francis Keppel, in conjunction with the Carnegie Foundation, sponsored two conferences to plan a large-scale, national assessment program. From those meetings, NAEP's precursor, the Exploratory Committee on Assessing the Progress of Education (ECAPE) was formed, with Ralph W. Tyler, "psychologist, and the nation's most prominent educational evaluator" as chair (Vinovskis, 1998, p. 6). Over the next several years, plans were developed for a testing program based on a representative national sample that would provide statistics about the educational achievement of Americans. In addition to national statistics, Keppel and Tyler advocated for the collection and reporting of state-level statistics. These plans were abandoned due to opposition from states and professional organizations that feared the data would be used to make unfair and inappropriate comparisons (Vinovskis, 1998). As a compromise, ECAPE's original plans included the presentation of assessment results at the national level and for four large geographic regions (Northeast, Southeast, West and Far West). Ten key subject areas were to be tested (mathematics, science, reading, writing, literature, social studies, art, music,

career/occupational development, and citizenship), for four groups of U.S. residents (ages 9, 13 and 17 years, and young adults between 26 and 35).

After undergoing several administrative changes, ECAPE eventually evolved into the National Assessment of Educational Progress (NAEP). The first assessments were administered in 1969-1970 in citizenship, writing and science. Since then "NAEP has regularly collected, analyzed, and reported valid and reliable information about what American students know and can do in a variety of subject areas" (Grigg, Daane, Jin & Campbell, 2003, p. 1). Over the past 30 years the program has grown considerably. To date all 50 states participate in some form of NAEP testing. Currently, NAEP encompasses periodic testing in grades four, eight and twelve, across a variety of subjects, including reading, writing, mathematics, science, U.S. history, world history, civics, economics, geography, and the arts. Basic academic subjects like math and reading are assessed with the greatest frequency, approximately every other year. Other subjects are tested less frequently; for example, U.S. history and civics are assessed every five years or so. The National Center for Education Statistics (NCES) is responsible for the design and administration of the NAEP testing program. This division of the federal Department of Education also disseminates NAEP results to the public via public reports and a comprehensive website (http://www.nces.gov). It also administers a licensing program for independent researchers to obtain access to NAEP data.

Although the NAEP may currently be the most appropriate data source available for the comparison of state-by-state differences in academic achievement (Prowker & Camilli, 2006), a number of threats exist to the validity of such comparisons. Caution is

warranted regarding inferences about state educational policies, ecological variables, and academic performance. Two noteworthy threats to validity are discussed below.

First, the NAEP is a "low stakes" test. These scores are generally not linked to rewards or penalties, unlike scores on "high stakes" tests which are tied to consequences such as grade promotion for students, publication of school performance in local newspapers, and tracking of adequate yearly progress (AYP) in accordance with NCLB. Since NAEP performance is not likely to be linked to such consequences, students may be less compelled to perform. Thus test scores may not be reasonable representations of academic proficiency. Given this, it is prudent to question the extent to which a "low stakes" test like the NAEP can allow for valid comparisons of academic achievement.

Second, reading content standards and attendant state testing programs vary across individual states, resulting in the likely event that standards, curriculum and testing for some states will be more closely aligned with the NAEP than for others. As a result, inferences about academic proficiency may be influenced by the extent of alignment between NAEP content strands and individual state programs. In particular, if state policy decisions are driven by the goal of aligning content standards, curriculum and testing, any inferences about individual policy initiatives could be sensitive to the degree of agreement between state standards and NAEP content. Subsequently, those inferences may not be generalizable in this regard.

Of the two threats discussed above, the second is of the greater concern. While it is likely that issues concerning motivation and NAEP performance will be similar across states, the same cannot be said for the alignment of state content standards with the NAEP. In fact, alignment is likely to vary across states and thus compromise state-by-

state comparisons. Consequently, the degree to which state curricula are aligned with the NEAP must be considered in the course of interpreting the results of this study.

NAEP Format and Content

The 2002 NAEP reading assessment of fourth graders was comprised of both selected- and constructed-response items. Items could take on one of three different formats: multiple choice, short open-ended response and extended open-ended response. All multiple choice items consisted of four options (one correct answer and three distracters), and were scored dichotomously. Short open-ended items were evaluated according to a scoring rubric and were either scored dichotomously or using partial credit. All extended open-ended items were evaluated with a rubric and scored using partial credit. There were a total of 82 items on this test.

Table 3.1
Distribution of Item Formats on the NAEP 2002 Reading Assessment of Fourth Graders

Item Format (Abbreviation)	No. of Items	Percent
Multiple Choice (MC)	37	45%
Short Open-Ended (OS)	37	45%
Extended Open-Ended (OE)	8	10%
Total	82	100%

All items were presented within testlets, or mini-tests, in which examinees were presented with a reading passage and a corresponding set of nine to twelve items. There were no “stand alone” items on the fourth grade reading NAEP; each item within a testlet dealt directly with material presented in the associated reading passage. Passages were

drawn from authentic books and publications that children might find at school, home or their local library (Grigg et al., 2003).

Academic content on each NAEP assessment is guided by a subject-area framework document (Vinovskis, 1998). The National Assessment Governing Board (NAGB), assembles the NAEP frameworks via "a comprehensive process involving a broad spectrum of interested parties, including teachers, curriculum specialists, subject matter specialists, school administrators, parents, and members of the general public" (Grigg et al., 2003, p. 2). These frameworks function as blueprints that specify the content of each NAEP assessment. In addition, frameworks may also describe relevant processes, skills or aspects of learning at a particular grade level. The NAEP reading assessment of 2002 was guided by the framework document developed in 1992, and updated in 2002 "to provide more explicit detail regarding assessment design" (Grigg et al., 2003, p. 3).

On the 2002 NAEP Reading Assessment of fourth graders, all items were cross-classified according to two dimensions: contexts for reading, and aspects of reading. Two types of reading contexts were included in the assessment of fourth graders: reading for literary purpose and reading for information. Test items were divided equally between the two types of reading contexts, with 41 items apiece. In addition to the NAEP total reading score, subscale scores were reported for each reading context.

For items classified as reading for literary purpose, examinees were presented with exercises in which they were required to "explore themes, events, characters, settings, plots, actions, and the language of literary works. Various types of texts are associated with reading for literary purpose, including novels, short stories, poems, plays,

legends, biographies, myths and folktales" (Grigg et al., 2003, p. 4). All items within a testlet containing a fiction passage were classified as reading for literary purpose. For items classified as reading for information, examinees were presented with texts that engaged them "with aspects of the real world. Reading for information is most commonly associated with textbooks, primary and secondary sources, newspapers and magazine articles, essays and speeches" (Grigg et al., 2003, p. 4). All items within a testlet containing a non-fiction passage were classified as reading for information.

Table 3.2
Distribution of Items across Testlets and Reading Contexts for the NAEP 2002 Reading Assessment of Fourth Graders

Testlet	Reading Context	No. of Items
Beetle	Literary	9
Box in Barn	Literary	12
Money Makes	Literary	11
Goodall	Information	9
Ellis Island	Information	10
Space Pioneer	Information	10
River	Literary	9
Wombats	Information	12
	Total	82

The aspects of reading dimension classifies NAEP items according to four different types of reading comprehension: forming a general understanding; developing an interpretation; making reader/text connections; and examining content and structure. "As readers attempt to develop understanding of text, they focus on general topics or themes, interpret and integrate ideas, make connections to background knowledge and

experiences, and examine the content and structure of the text. The [NAEP] framework accounts for these different approaches to understanding text by specifying four 'aspects of reading' that represent the types of questions asked of students" (Grigg et al., 2003, p. 4). Item classifications for aspects of reading varied from item to item, and were not uniform within each testlet.

Items classified as forming a general understanding present exercises in which examinees are required to demonstrate a broad understanding of the entire text. For example, examinees could be asked to identify or provide the main topic of a passage, describe the theme of a story, or explain the purpose of an article (Grigg et al., 2003). For items classified as developing an interpretation, examinees are challenged to further develop their broad impressions of a passage and construct a deeper more complete level of understanding. To complete these kinds of tasks successfully, examinees must be able to link specific, disparate aspects of their general understanding. For example, examinees may be prompted to make inferences about the relationship between two pieces of information, or be asked to explain the reasons behind a particular action by citing specific information in the passage (Grigg et al., 2003). To successfully respond to an item classified as making reader/text connections, the examinee must relate aspects of their own knowledge and experience to specific information in the passage (Grigg et al., 2003). Finally, items classified as examining content and structure challenge examinees to engage in critical evaluation of a piece of text; activities may include comparing, contrasting, evaluating the organization of a passage, and exploring the use of literary devices such as irony and humor (Grigg et al., 2003). These kinds of questions encourage the examinee to engage in a critical and objective evaluation of the text regarding its

overall quality; purpose; appropriateness for a particular use; quality of language and textual elements; and author's writing style.

NAEP Sampling Design

Prior to 1990, the scope of NAEP involved describing the academic achievement of all U.S. students by analyzing data from a single, large national sample. Over time the national will drifted as the social climate changed, and both citizens and politicians exhibited increasing interest in state-level achievement results (Vinovskis, 1998). Publication of state-level results has become more acceptable, consistent with Keppel's and Tyler's original vision though not without controversy or criticism (Jones, 1996; Vinovskis, 1998). As a result, in 1990 NAEP began collecting data from a second set of samples from individual states. These "state-by-state" samples allowed for the publication of robust estimates for individual states, thus providing valuable information about their individual academic profiles.

After maintaining separate national and state-by-state samples for a number of years, NAEP has begun merging the two programs into a single sample for analyses at both the national and state level. In 2000, the NAEP Reading Assessment included separate national and state-by-state samples, whereas the 2002 and subsequent assessments included a single sample designed to accommodate both national and state-by-state comparisons. In 2002, 140,487 students from 5,518 schools participated in the NAEP fourth grade reading assessment (Rogers & Skoekel, 2004). Forty-five states participated in this assessment. Although nine of these states failed to meet one or more NAEP participation guidelines, all forty-five are included in this analysis. In addition, this analysis includes data from the District of Columbia, Guam, the United States Virgin

Islands, and foreign and domestic United States Department of Defense schools, for a total of fifty participating jurisdictions.

NAEP is distinguished by its large scope, scale and mission. In addition, several innovative features that were implemented in the design of the NAEP bear particular relevance to the current study, in particular NAEP sampling procedures and design (Jones, 1996; Rogers & Stoekel, 2004). The original program specifications endorsed by ECAPE in the 1960s included a recommendation to use matrix sampling. In matrix sampling assessments include large numbers of items in order to sample a variety of relevant academic tasks; however, individual examinees only receive a subset of test items. (The original ECAPE recommendations suggested a subset of one tenth or fewer of the total test items per examinee. See Jones, 1996). Subsets of items are distributed systematically across examinees, ensuring a large and varied sample of respondents for any given question. Matrix sampling of test items enables the NAEP to administer large numbers of items to a sufficient sample of respondents without burdening individual students with extensive assessment time.

The NAEP employs a matrix sampling procedure called partially balanced incomplete block (PBIB) sampling (Rogers & Stoekel, 2004). In PBIB sampling, items are organized into blocks, and a variety of test booklets are constructed with different combinations of item blocks. Blocks of items are assigned to test booklets such that their positioning across booklets is balanced with regard to NAEP reading context (Rogers & Stoekel, 2004). Test booklets are distributed to examinees according to a cyclical pattern, in order to ensure that within any given assessment session very few students received the same booklet. As a result, position and contextual effects are minimized.

For the 2002 NAEP reading assessment of fourth graders, eight 25-minute blocks were constructed, each consisting of a single reading passage and 9 to 12 questions (i.e., each block was comprised of a single testlet). Four of the blocks were classified as Reading for Literary Purpose, and four were classified as Reading for Information. Each test booklet contained two blocks of items (i.e., two testlets). Booklets contained approximately 9 to 14 multiple choice questions, 8 to 10 short constructed-response questions, and two extended constructed-response question.

NAEP's use of PBIB sampling is particularly relevant with regard to this study. These procedures provide a large and representative sample of students within each state, while minimizing the impact of test fatigue. The size and representativeness of the sample allows for the estimation of state-level proficiencies as well as state-by-state comparisons. In addition, the large pool of test items that comprise the entire assessment provides comprehensive and varied coverage of the entire content domain, and also allows for the fine-grained analysis of individual item performance. (For more details regarding the technical aspects of the NAEP sampling design, see Rogers & Stoekel, 2004).

Method

Dependent Variable

The primary dependent variables in this study are the individual item scores on the 2002 NAEP reading assessment of fourth graders for participating jurisdictions. These scores are calculated as the proportion correct within a jurisdiction on a particular item. These scores are adjusted by the rescaled NAEP weighting variable (Origwt), which accounts for the distribution of various population characteristics within the sample. (See

Rogers & Stoekel, 2004, for a more detailed discussion). In order to accommodate SAS processing requirements, Origwt is rescaled downward for this analysis by multiplication by a constant equal to one divided by the entire national sample size, since use of the unadjusted Origwt variable may result in estimation problems.

Steps and Procedures

Using techniques elaborated by Camilli, Monfils and others (Camilli & Monfils, 2003; Camilli et al., 2006; Camilli et al., 2005; Monfils, 2004; Prowker & Camilli, 2006), this study uses individual items from the NAEP reading assessment of fourth graders as a "toolkit" for disentangling the effects of different state-level variables on reading achievement. The eventual goal is to link patterns in item performance to specific state policies or ecological variables. The heart of this analysis is the estimation of item difficulty variance (IDV) and the subsequent derivation of parcel scores. As noted previously, the analysis can be divided into four steps:

Step 1: Identification of items with significant item difficulty variation (IDV) using multilevel item response models,

Step 2: Factor analysis of identified item residuals,

Step 3: Parcel score construction and estimation,

Step 4: Moderator variable analysis.

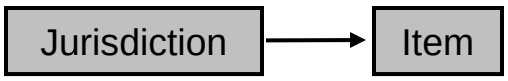
Step 1: Identification of Items with Significant Item Difficulty Variation (IDV) Using Multilevel Item Response Models

The data used in this analysis are hierarchical (or nested) in nature, and can be conceptualized in a variety of different ways. For example, using student as the first level subject of interest, the data can be conceptualized as students nested within schools,

nested within jurisdictions (or states). On the other hand, if we take test item as the first level unit of interest, the data can be conceptualized as items, nested within testlets, nested within states; or items nested within students, within schools, within districts. Depending upon the nature of the research question and how the structure of the data is conceptualized, different effects can be parsed. For this study, three, two-level, generalized linear mixed models (GLMM) are examined that account for slightly different underlying hierarchical structures.

The main analysis, Model 1, presents the simplest configuration of effects, with items (Level 1) nested within jurisdictions (Level 2). This model is carried through all four steps of the main analysis, and addresses the two core research questions posed in this study. In a secondary analysis, two slightly more complicated two-level models, Model 2 and Model 3, are examined. These two models include NAEP testlet as a Level 1 variable along with item. Analyses for Models 2 and 3 are not carried out for all four steps. They are included as exploratory analyses of the potential effects of NAEP testlets within the proposed model framework, given the possibility of local dependency among items within the same testlet.

Model 1. Model 1 is the primary model used in this dissertation. Residuals from this model are used in the subsequent factor analysis and parcel score analysis. In this first model, the first level unit is item and the second level unit is jurisdiction (typically a state).



(Level 2) (Level 1)

Figure 3.1. Model 1

In this two-level analysis of items nested within jurisdictions, success on a given item, *i*, is modeled as a function of proficiency and item difficulty. Suppose we have a test with *i*=1…*I* items administered across *j*=1….*J* jurisdictions. The Level 1 model follows as,

$$f(n_{+ji} / n_{ji}) = \mu_j - \delta_{ji} + \varepsilon_{ji} \tag{3.1}$$

where n_{+ji}/n_{ji} equals the ratio of correct to total responses for item *i* in jurisdiction *j*; μ_j represents the overall reading proficiency for jurisdiction *j*; δ_{ji} represents the difficulty of item *i* for jurisdiction *j*; ε_{ji} represents the item-level error term specified as $\varepsilon_{ji} \sim N(0,1)$; and f(•) represents the logit link function.

Within Level 1 (i.e., at the item level), both μ_j and δ_{ji} are constant. However, each of these effects can be decomposed into fixed and random components across jurisdictions (Level 2):

$$\mu_j = \mu + \theta_j, \tag{3.2}$$

$$\delta_{ji} = \delta_i + \upsilon_{ji}, \tag{3.3}$$

where θ_j represents the jurisdiction level ability component; δ_i represents the fixed item difficulty component for item *i*; and υ_{ji} represents a random item difficulty component for item *i*. The component υ_{ji} can be conceptualized as a value-added (or value-subtracted)

effect unique to a particular jurisdiction for item *i*. The Level 2 effects (θ_j and υ_{ji}) are then specified by

$$\theta_j \sim \mathrm{N}(0, \sigma_\theta^2), \tag{3.4}$$

$$\upsilon_{ji} \sim \mathrm{N}(0, \tau_i^2). \tag{3.5}$$

Following from these equations, τ_i^2 represents the variance of the jurisdiction effects on item difficulty for a given item *i*. If τ_i^2 is close to zero, υ_{ji} does not vary significantly across all jurisdictions for item *i*. In other words, after taking into account jurisdiction proficiency estimates, item difficulty is consistent across jurisdictions in this case. In a traditional DIF analysis, this result would indicate no DIF across the reference and focal groups.

Solutions for υ_{ji} are referred to as best linear unbiased predictors or BLUPs (Little, Milliken, Stroup, & Wolfinger, 1996) and τ_i^2 is called item difficulty variation (IDV). Given an IDV estimate greater than zero, a large positive individual BLUP estimate indicates that the item proved to be more difficult than expected given the overall proficiency level of that particular jurisdiction. Conversely, a large negative BLUP estimate would indicate that an item was easier than expected for that jurisdiction.

For this analysis, the dependent variable is the logit of the probability that item *i* is answered correctly by a student in jurisdiction *j*, (n_{+ji}/n_{ji}), adjusted by the rescaled NAEP weighting variable (Origwt). The predictors are expressed as I-1 dummy variables that represent the items on the test. For the 2002 data set, 81 dummy variables are used. The following table represents the coding scheme for the analysis of the 2002 data (Kamata, 1999a, 1999b, 2001).

Table 3.3
Coding Scheme for Model 1

Jurisdiction	D01	D02	D03	…	D79	D80	D81	*p*
01	1	0	0	…	0	0	0	$p_{1,01}$
01	0	1	0	…	0	0	0	$p_{1,02}$
01	0	0	1	…	0	0	0	$p_{1,03}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
50	0	0	0	…	0	1	0	$p_{50,80}$
50	0	0	0	…	0	0	1	$p_{50,81}$
50	0	0	0	…	0	0	0	$p_{50,82}$

The variable p_{ji} represents the weighted proportion correct for jurisdiction *j* on item *i*; there are a total of JxI outcome variables represented in the table. The number of dummy variables are I-1, with I=82. A response for the first item is indicated by a row consisting of the numeral one followed by 80 zeros, while a response for the final item is indicated by a row of 81 zeros. Following this coding format, the design matrix can only achieve full rank when one item is designated as the reference item and omitted from the matrix (Kamata, 1999a, 1999b, Kamata). In doing so, the reference item's difficulty is ostensibly set at zero, and all other item difficulties are interpreted relative to the reference item.

Models 2 and 3. The second and third models in this study are somewhat more complex than the previous one, and use both items and testlets in a two-level design.

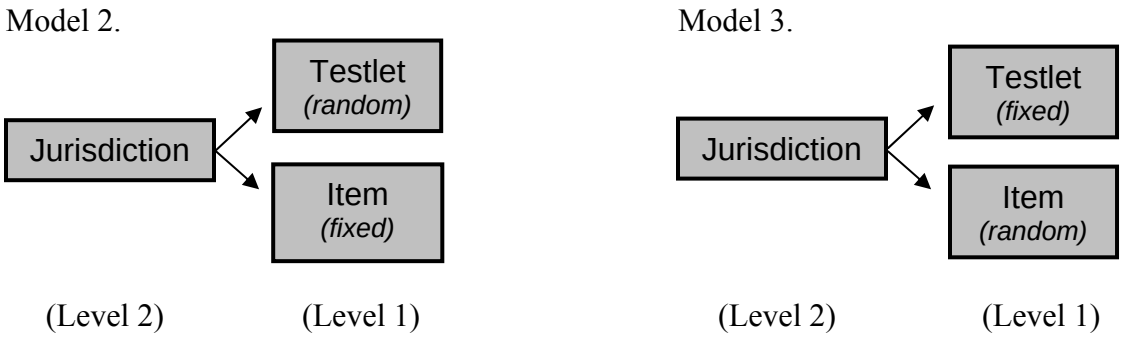


Figure 3.2. Models 2 and 3

Two separate models are run in these analyses. In Model 2, item difficulties are held fixed, while testlet difficulties vary randomly. In Model 3, testlet difficulty is fixed, and item difficulty is random. The dependent variable in both analyses is the state-level item responses, calculated as the adjusted proportion correct within each state for each item.

In Model 2, testlet difficulty is random, while item difficulty is fixed. Using this model, testlet difficulty variation is estimated across all states, given state-level proficiency and fixed item difficulty estimates. The differentiating feature of this model is that it estimates testlet difficulty variation, rather than item difficulty variation, while taking into account item difficulty and state-level proficiency estimates.

Success on a given item, i, is modeled as a function of ability, item difficulty (fixed) and testlet difficulty (random). Suppose we have a test with i=1…I items, and t=1…T testlets, administered across j=1…J jurisdictions. The Level 1 model follows as,

$$f(n_{+jti} / n_{jti}) = \mu_j - \delta_i - \lambda_{jt} + \varepsilon_{jti}, \quad (3.6)$$

where n_{+jti}/n_{jti} equals the ratio of correct to total responses for item i for testlet t in jurisdiction j; μ_j represents the overall reading proficiency for jurisdiction j; δ_i represents the fixed difficulty of item i across all jurisdictions; λ_{jt} represents the difficulty of testlet t for jurisdiction j; ε_{jit} represents the error term specified as $\varepsilon_{jit} \sim$ N(0,1); and f(•) represents the logit link function.

In the context of the Level 1 model, both μ_j, λ_{jt} are fixed effects. However, each of these effects can be decomposed into fixed and random components.

$$\mu_j = \mu + \theta_j, \quad (3.2)$$

$$\lambda_{jt} = \lambda_t + \upsilon_{jt}, \quad (3.7)$$

where θ_j represents the jurisdiction level ability component; λ_t represents the fixed difficulty component for testlet t; and υ_{jt} represents a random difficulty component for testlet t that varies across jurisdictions. The testlet BLUP, υ_{jt}, can be conceptualized as a value-added (or value-subtracted) effect unique to a particular jurisdiction for testlet t.

The Level 2 effects (θ_j and υ_{jt}) are specified as

$$\theta_j \sim \text{N}(0, \sigma_\theta^2), \quad (3.4)$$

$$\upsilon_{jt} \sim \text{N}(0, \tau_t^2), \quad (3.5)$$

Following from these equations, τ_t^2 represents the variance of the jurisdiction effects on testlet difficulty for a given testlet, t, after taking into account jurisdiction ability and fixed item difficulty estimates. Unlike the estimated BLUPs from Model 1, υ_{jt} represents a *testlet* BLUP, not an item BLUP, and τ_t^2 is referred to as *testlet* difficulty variation (TDV), with an interpretation similar to that of IDV.

In Model 3, testlet difficulty is viewed as fixed, while item difficulty is viewed as random. Using this model, item difficulty variation is estimated for all items, given state-level proficiency and fixed testlet difficulty. The differentiating feature of this model is that it takes testlet difficulty into account, in addition to state-level proficiency, before assessing differential item functioning via estimation of IDV.

For this model, success on a given item, *i*, is modeled as a function of proficiency, item difficulty (random) and testlet difficulty (fixed). Suppose we have a test with i=1…I items, and t=1…T testlets, administered across j=1…J jurisdictions. The Level 1 model follows as,

$$f(n_{+jti} / n_{jti}) = \mu_j - \delta_{ji} - \lambda_t + \varepsilon_{jti}, \tag{3.8}$$

where n_{+jti}/n_{jti} equals the ratio of correct to total responses for item i of testlet t in jurisdiction j; μ_j represents the overall reading proficiency for jurisdiction j; δ_{ji} represents the difficulty of item i for jurisdiction j; λ_t represents the fixed difficulty of testlet t; ε_{jti} represents the error term specified as $\varepsilon_{jti} \sim N(0,1)$; and f(•) represents the logit link function.

In the context of the Level 1 model, both μ_j, δ_{ji} are fixed effects. However, each of these effects can be decomposed into fixed and random components.

$$\mu_j = \mu + \theta_j, \tag{3.2}$$

$$\delta_{ji} = \delta_i + \upsilon_{ji}, \tag{3.3}$$

where θ_j represents the jurisdiction level ability component; δ_i represents the fixed item difficulty component for item i; υ_{ji} represents a random item difficulty component for item i that varies across jurisdictions.

The Level 2 effects (θ_j and υ_{ji}) can be estimated and defined as

$$\theta_j \sim \mathrm{N}(0, \sigma_\theta^2), \tag{3.4}$$

$$\upsilon_{ji} \sim \mathrm{N}(0, \tau_i^2). \tag{3.5}$$

It follows that τ_i^2 represents the variance of the jurisdiction effects on item difficulty as defined above, and υ_{ji} is the value-added effect. Model regressors include I-1 dummy variables representing item difficulty, and T-1 dummy variables representing testlet difficulty. The dependent variable *p* is defined as above.

Table 3.4
Sample Testlet Design Matrix for Models 2 and 3

Jurisdiction	T01	T02	…	T06	T07	D01	D02	…	D80	D81	p
01	-1	0	0	0	0	1	0	0	0	0	$p_{1,01}$
01	-1	0	0	0	0	0	1	0	0	0	$p_{1,02}$
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
50	0	0	0	0	0	0	0	0	0	1	$p_{50,81}$
50	0	0	0	0	0	0	0	0	0	0	$p_{50,82}$

A sample design matrix for Models 2 and 3 is illustrated in Table 3.4. The variables representing the testlets, T01-T07 take on values of zero or *negative* one, rather than zero or positive one. By coding the testlet dummy variables in this manner, estimated testlet BLUPs obtained by Model 2 can be interpreted as testlet *scores*. A negative testlet score indicates that a jurisdiction performed worse than expected on a give testlet (i.e., the testlet was more difficult than expected), while a positive value indicates that a jurisdiction performed better than expected (i.e., the testlet was easier than expected).

This interpretation is different from that of the estimated item BLUPs, which are interpreted similarly to IRT item difficulty coefficients, or *b* parameters.

Step 2: Factor Analysis of Identified Item Residuals

For Model 1, an exploratory factor analysis is conducted on the residual BLUP estimates of a subset of target items with large IDVs, following procedures elaborated by Camilli et al. (2006). The goal of this portion of the analysis is to consolidate data across test items by identifying common variance among item residuals. In addition, this analysis may provide conceptual clarity toward an interpretation of effects by suggesting an underlying factor structure across the target items that indicate additional shared variance. In order to construct a substantive explanation of the factor structure, related items are logically inspected across a variety of characteristics, including item format; item content; NAEP context for reading; NAEP aspect of reading; content of reading passage; and presence of graph, figures or pictures.

Step 3: Parcel Score Derivation

Based on the obtained factor structure from Model 1, parcel scores are derived that correspond to the retained factors. Parcel scores are estimated in the framework of a multidimensional Rasch model that conforms to the underlying factor structure identified in the factor analysis (Camilli et al., 2006). Given *K* factors, the model conforms to the following general format:

$$\eta_{ji} = \mu + \theta_j + \sum_{k=1}^{K} d_{ik}\varphi_{jk} - \delta_i \,, \tag{3.9}$$

where η_{ji} represents the propensity score for jurisdiction j on parcel i; μ represents the average state reading proficiency; θ_j represents the proficiency estimate for jurisdiction j; d_{ik} corresponds to the set of indicator coefficients for each factor; φ_{jk} represents the

proficiency of jurisdiction j on parcel k; and δ_i corresponds to the fixed difficulty of item i. Recall that a negative parcel score indicates that a jurisdiction performed worse than expected on the parcel given its overall proficiency estimate, while a positive value indicates that a jurisdiction performed better than expected. Parcel scores cannot be interpreted as typical proficiency estimates. Similar to the DIF estimates discussed earlier, parcel scores can be thought of as relative performance measures, above (or below) what would be expected given a state's proficiency.

Step 4: Moderator Variable Analysis

A correlational analysis is conducted with the parcel scores and a battery of moderator variables in order to backward map state-level correlates onto patterns in parcel score performance. In backward mapping, variables are included that exist outside the scope of specific educational polices, and can include organizational characteristics or aspects of local environments (Recesso, 1999). Backward mapping permits the examination of a wide variety of potential effects that may, or may not, have been considered during policy design or implementation.

This battery of moderator variables used in this analysis is based on a review of relevant studies and reports, including an examination of NAEP published materials and data sources. Moderator variables are drawn from several different sources including NAEP background questionnaires of students and teachers; the United States Census;

Table 3.5
Proposed Moderator Variables

Dimension	Variable Description
State	Poverty level
	Proportion of non-English Speakers
	Median household income
	Alignment between state testing and the NAEP
	State educational policy profile
	State content standards program (quality, duration)
	State testing program (quality, duration)
School	School location (urban, suburban, rural)
	Type of school (public, private)
Teacher	Instructional time in reading and/or language arts
	Homework frequency
	Testing frequency
	Perceived teacher instructional control
	Availability of classroom resources
	Reading certification
	Teacher education
	Teaching experience
	In-service and pre-service training
Student	Proportion with Individualized Education Plans (IEPs)
	Proportion receiving free or reduced lunch
	Parental education level

research reports; policy documents; and other public domain sources documenting the educational policies and standards associated with individual states. Any interpretation of the potential effects of moderator variables must include consideration of the extent to

which individual state curricula and testing may be aligned with NAEP content strands. Table 3.6 contains the original selection of moderator variables proposed for this study.

Following the correlational analysis, a brief collective case study of three states is conducted in order to illustrate how parcel scores may be used to explore the potential link between individual state characteristics and reading performance. This analysis is not intended to provide generalizable or summative commentary regarding all fifty participating jurisdictions. Selection of the three states included in the case study is based on inspection of the IDV estimates, parcel scores, and correlations of parcel scores with moderator variables. In addition, relevant characteristics, (such as state population composition, size, and history of educational reform) are also included in the selection process, as well as the extent to which state standards align with NAEP content strands. This analysis provides a more detailed description of specific item dimensions (e.g., cognitive processes, content areas, item formats) that may be associated with particular external variables (e.g., specific state policies such as implementation of content standards and state assessments; curricular alignment; demographic variables; social variables) for the selected states.

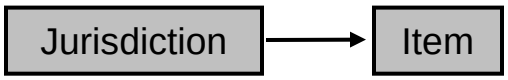
CHAPTER IV. RESULTS

In this chapter, a detailed discussion is given of the results obtained following application of the methodology outlined in Chapter III. Specifically, the application of MLM Model 1 for identifying items with relatively large IDV is discussed, followed by an interpretation of the performance of NAEP jurisdictions on the target items. Results of the factor analysis and parcel score derivation are then presented, as well as interpretation of the performance of NAEP jurisdictions on parcel scores. A discussion of Model 2 follows, focusing on NAEP jurisdiction performance on individual testlets. Next, results of an analysis of the association of potential state-level moderator variables with performance on parcel scores are given, followed by a brief case study of three jurisdictions of interest. Detailed statistical outputs from theses analyses, as well as the results of Model 3, are given in the appendices.

The following conventions have been adopted in the reporting of results. Estimated item BLUPs, testlet scores, and parcels scores are rounded to three decimal places. Their corresponding standard errors are rounded to four decimals. Item difficulty variation (IDV) and testlet difficulty variation (TDV) are also rounded to four decimals. Overall reading proficiency estimates (θ), and IRT a, b and c parameters are rounded to two decimals, and their corresponding standard errors are rounded to three decimals. Factor loadings are rounded to three decimals, while factor eigenvalues (λ) are rounded to two decimals. Correlation coefficients (r) are rounded to two decimals. For the sake of clarity, the names of all moderator variables are written in capital letters in the tables and text.

Model 1

Model 1 was the primary model in this study. In this two-level analysis of items nested within jurisdictions, success on a given item *i*, was modeled as a function of jurisdiction proficiency and item difficulty, with *i*=1…82 test items administered across *j*=1…50 jurisdictions.



(Level 2) (Level 1)

Figure 4.1. Model 1

Item BLUPs, τ_i^2, were estimated for each of the 50 jurisdictions across all 82 test items, as well as 82 separate IDVs. Given a significantly large IDV estimate, a positive BLUP estimate indicates that the item is more difficult than expected in light of the overall proficiency level of the jurisdiction. An estimated BLUP close to zero indicates that a jurisdiction performed as expected. Conversely, a large negative BLUP estimate indicates that an item was easier than expected for that jurisdiction.

Consider an example using item D71, which had the largest estimated IDV in this analysis. This item was a multiple choice item that accompanied a short, non-fiction reading passage entitled "Watch Out for Wombats," a descriptive essay about the biology, behavior and habitat of Australian wombats. Based on the specifications of the NAEP design, this item required examinees to read for the purpose of gathering information and forming an initial understanding.

This article mostly describes how

A) the wombat's special body parts help it to grow and live
B) highway signs help to save the wombats
C) the wombat is like the koala and the North American badger
D) wombats feed and raise their young

Figure 4.2. Item D71

This item proved to be surprisingly difficult for students, as indicated by the IRT *b*-parameter of 3.03. Examinees falling three standard deviations above their average peers in reading proficiency had just about a 50% likelihood of answering item D71 correctly. The residual variance associated with item D71 was the largest among all the 82 test items, as shown in Figure 4.3. This item's IDV of 0.0841 was more than twice that of the next largest IDV of 0.0385. Estimated jurisdiction BLUPs varied greatly for item D71 compared to the other 81 test items.

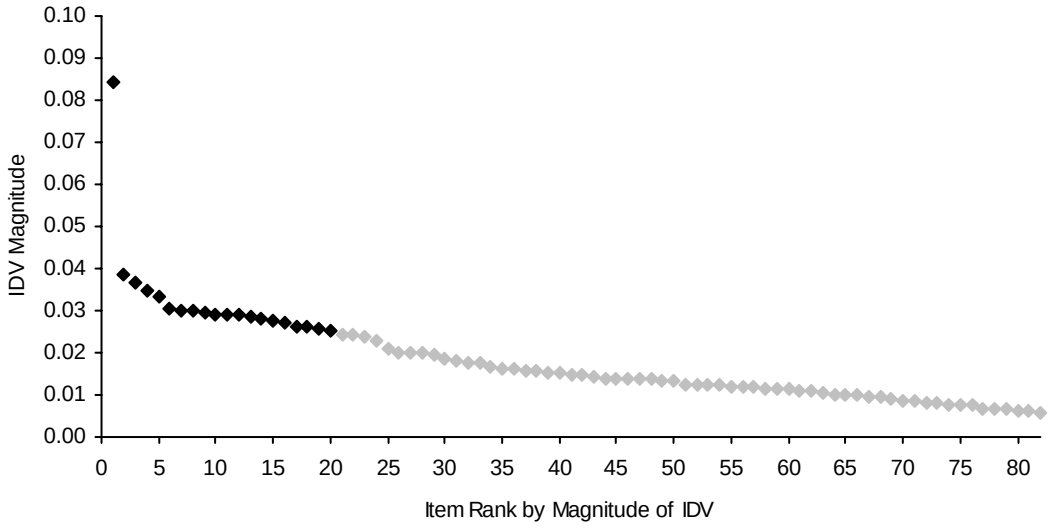


Figure 4.3. Plot of Item IDVs

Idaho had the largest estimated BLUP for item D71 with a value of 0.256 (SE=0.1561). This implies that Idaho's performance on item D71 exceeded what would have been expected given Idaho's overall reading level of θ_{ID}=0.11. Conversely, Texas had the smallest BLUP for D71, with a value of –0.506 (SE=0.0640), indicating that students in Texas scored lower on D71 than would have been expected given θ_{TX}=0.20. In contrast to both of these jurisdictions, Pennsylvania's BLUP of 0.004 (SE=0.0817) was almost to zero, which indicates that students in Pennsylvania performed about as well on item D71 as would have been expected given θ_{PA}=0.13

Identification of Target Items

In selecting items for further analysis, items were rank ordered by magnitude of IDV, which ranged from 0.0841 to 0.0056. An arbitrary goal of 20% was set as a threshold for retaining items. Use of this criterion, along with interpretation of the plot of item IDVs (see Figure 4.3), led to the retention of twenty items which represented 24% of the total. The twenty target items selected for further analyses were drawn from all eight testlets included on the 2002 NAEP reading assessment of fourth graders. The average IDV of the target items was 0.0325, compared to 0.0132 for the remaining sixty-two items and 0.0179 for all 82 test items.

The 2002 NAEP reading assessment of fourth graders was comprised of 37 (45%) multiple choice (MC) items, 37 (45%) short open-ended response (OS) items, and eight (10%) extended open-ended response (OE) items. The distribution of item formats among the twenty target items was similar, with eight (40%) multiple choice items, and eleven (55%) short open-ended response items. One (5%) extended open-ended response item

was also included. In addition, the target items also reflected a similar distribution of the cross-classification of items according to NAEP reading context and aspect.

Table 4.1
Distribution of NAEP Item Classifications: Target Items vs. Entire Assessment

Item Characteristic	All Items (*i*=82)		Target Items (*i*=20)	
Item format				
Multiple choice	37	(45%)	8	(40%)
Short open-ended	37	(45%)	11	(55%)
Long open-ended	8	(10%)	1	(5%)
NAEP reading context				
Literary purpose	41	(50%)	10	(50%)
Information	41	(50%)	10	(50%)
NAEP aspect of reading				
Forming a general understanding	8	(10%)	3	(15%)
Developing an interpretation	45	(55%)	8	(40%)
Forming reader-text connections	12	(15%)	2	(10%)
Examining content and structure	17	(21%)	7	(35%)

The target items were evenly split between the two possible reading contexts, Reading for Literary Purpose (i.e., fiction passage) and Reading for Information (i.e., non-fiction passage). This pattern was identical to that found on the entire 82-item assessment. On the second dimension, the target items were distributed across the four different reading aspects in a pattern that was generally similar to that of the entire 82-item test. Three target items (15%) required the examinee to form a general understanding of the passage. Eight items (40%) required the development of an interpretation. Two items (10%) required the examinee to form reader-text connections,

and seven items (35%) required examination of passage content and structure. Target items varied in difficulty, from very easy to very difficult, in terms of IRT *b*-parameters (*b* ranging from -2.18 to 3.03).

An outlier analysis was conducted on the twenty target items. Scatter plots of estimated item BLUPs and overall proficiency estimates were constructed across all fifty participating jurisdictions. In addition, tests of normality were run for each set of estimated item BLUPs and overall reading proficiency estimates. Item residuals and proficiency estimates were generally normal, however, three extreme outliers were identified: the U.S. Virgin Islands, Guam, and the District of Columbia. These three jurisdictions had the lowest overall reading proficiency estimates of θ_{VI} = -0.91, θ_{GU} = -0.73, and θ_{DC} = -0.69. The jurisdiction with the next lowest reading proficiency estimate was Mississippi, with θ_{MS} = -0.34. The potential impact of these three outliers was considered in the subsequent factor analysis, and is discussed in more detail in the next section.

Factor Analysis of Target Item BLUPs

An exploratory factor analysis was conducted on the estimated BLUPs obtained from the target items for Model 1. The goal of this portion of the analysis was to consolidate data across test items by identifying common variance among estimated BLUPs, and enabling parsimonious interpretations of effects. Initially, principal axis factoring was used in this analysis with varimax rotation. A second factor analysis was then run allowing for oblique rotation. Since the results using oblique rotation were similar to those obtained using orthogonal rotation, the initial principal axis factoring

with varimax rotation was retained in order to allow for more cleanly defined factors. These results are presented below.

The factor analysis yielded four factors with eigenvalues greater than one (λ_1=8.79, λ_2=3.32, λ_3=2.40, and λ_4=1.65), explaining 44%, 17%, 12% and 8% of the total variance respectively. Of the twenty target items, fourteen displayed factor loadings that clearly placed them within the first or second factor. Seven items loaded on Factor 1: D09, D21, D50, D67, D68, D69 and D70. Seven items also loaded on Factor 2: D26, D52, D53, D55, D63, D71 and D74. Of the remaining 6 items, only item D23 loaded on Factor 3, while items D51, D72 and D10 loaded on Factor 4. Communality estimates for the fourteen items that loaded on Factors 1 and 2 ranged from 0.60 to 0.93. Communality estimates for the remaining six items were generally smaller and ranged from 0.49 to 0.87.

The four factors obtained in this analysis were reviewed in order to determine which would be retained for the calculation of the parcel scores. Based on the skree plot of all eigenvalues, the item communality estimates, the proportion of explained variance of each factor, and the pattern and magnitude of the factor loadings, Factors 1 and 2 were retained. The paucity of information provided by the few items which loaded on Factors 3 and 4 made interpretation problematic, particularly for Factor 3 which represented only a single item. Due to the relatively few items loading on these factors, their comparatively smaller proportions of explain variance, and the smaller communalities of their associated items, Factors 3 and 4 were dropped.

Table 4.2
Model 1 Factor Loadings

Item	Factor 1 λ_1=8.79	Factor 2 λ_2=3.32	Factor 3 λ_3=2.40	Factor 4 λ_4=1.65	Communality
D71	-0.077	0.891	0.120	-0.003	0.81
D69	0.912	-0.184	0.042	0.022	0.87
D70	0.937	-0.153	-0.150	-0.005	0.92
D67	0.777	-0.494	-0.008	0.030	0.85
D09	0.917	-0.027	0.018	-0.059	0.84
D63	-0.317	0.556	0.245	-0.358	0.60
D52	-0.503	0.662	-0.006	0.080	0.70
D51	0.340	0.034	-0.352	0.616	0.62
D50	0.816	-0.044	0.074	-0.100	0.68
D21	0.898	-0.277	0.103	-0.015	0.89
D74	-0.372	0.605	-0.477	0.675	0.79
D72	-0.067	0.118	0.107	0.675	0.49
D53	-0.136	0.830	0.061	0.189	0.75
D23	-0.208	0.421	0.796	-0.119	0.87
D49	-0.448	0.020	-0.520	-0.405	0.64
D55	-0.555	0.566	-0.401	0.064	0.79
D26	0.257	0.666	0.448	-0.249	0.77
D33	-0.797	0.008	-0.189	-0.066	0.67
D10	-0.262	-0.070	-0.073	0.726	0.60
D68	0.870	-0.290	-0.281	-0.105	0.93

The fourteen items comprising Factors 1 and 2 were included in an interpretation of factors and the calculation of parcel scores. These items were compared across various dimensions, including: testlet; reading passage word count; presence of picture or graphic accompanying the passage; relatedness of picture or graphic to the test item; item format;

NAEP reading context; NAEP aspect of reading; NAEP difficulty label (easy, medium, difficult); IRT *a* parameter; and IRT *b* parameter. These explanatory variables were either provided in the NAEP dataset (e.g., IRT *a* and *b* parameters, NAEP aspect of reading), or derived from the testlets themselves (e.g., presence of picture or graphic, relatedness of picture or graphic). A written protocol was prepared for the derived variables in order to formalize procedures for determining their value. An exploratory analysis of these variables included use of scatter plots, correlation matrix (factor loadings with explanatory variables), partial correlations, histograms, frequency tables, and descriptive statistics in order to describe the two primary factors. None of the item-based characteristics, such as item format or difficulty, were associated with the resultant factors. Two passage-level characteristics, word count and NAEP reading context, did suggest explanations of the underlying factor structure.

The preceding analyses were rerun excluding the U.S. Virgin Islands, Guam, and the District of Columbia, which had been identified as outliers. The same four factors were recovered, with comparable factor loadings for each item. In addition, the same pattern of correlations between explanatory variables and factor loadings were obtained, although all correlations increased in magnitude when the outliers were dropped. Since the exclusion of the outliers from the analysis did not change any of the conclusions, they were retained in the sample for the final analysis.

Table 4.3
Model 1 Factor Patterns by Passage Word Count and NEAP Reading Context

	Passage Word Count		NAEP Reading Context	
Item	Factor 1	Factor 2	Factor 1	Factor 2
D71		**684**		**Non-fiction**
D69	**1184**		**Fiction**	
D70	**1184**		**Fiction**	
D67	**1184**		**Fiction**	
D09	*840*		**Fiction**	
D63		*1184*		*Fiction*
D52		**779**		**Non-fiction**
D50	**1011**		*Non-fiction*	
D21	**1029**		**Fiction**	
D74		**684**		**Non-fiction**
D53		**779**		**Non-fiction**
D55		**779**		**Non-fiction**
D26		*1366*		*Fiction*
D68	**1184**		**Fiction**	

Using passage length as the guiding explanatory variable suggests that Factor 1 represents a Long Passage Factor, while Factor 2 represents a Short Passage Factor. NAEP reading context provides an alternative interpretation, with Factor 1 as a Reading for Literary Purpose (i.e., Fiction) Factor and Factor 2 as a Reading for Information (i.e., Non-Fiction) Factor. Ironically, passage word count and NAEP reading context were confounded for this assessment, with non-fiction passages tending to be shorter in length than fiction passages. (For the entire sample of 82 test items, the correlation between word count and reading context was $r = 0.654$. For the pool of 20 target items, $r =$

0.753.) For the seven testlets that contributed items to Factors 1 and 2 (the Goodall testlet did not contribute items to either factor), the non-fiction passages had an average length of 825 words while the fiction passages had an average length of 1105 words, constituting a difference of about two paragraphs.

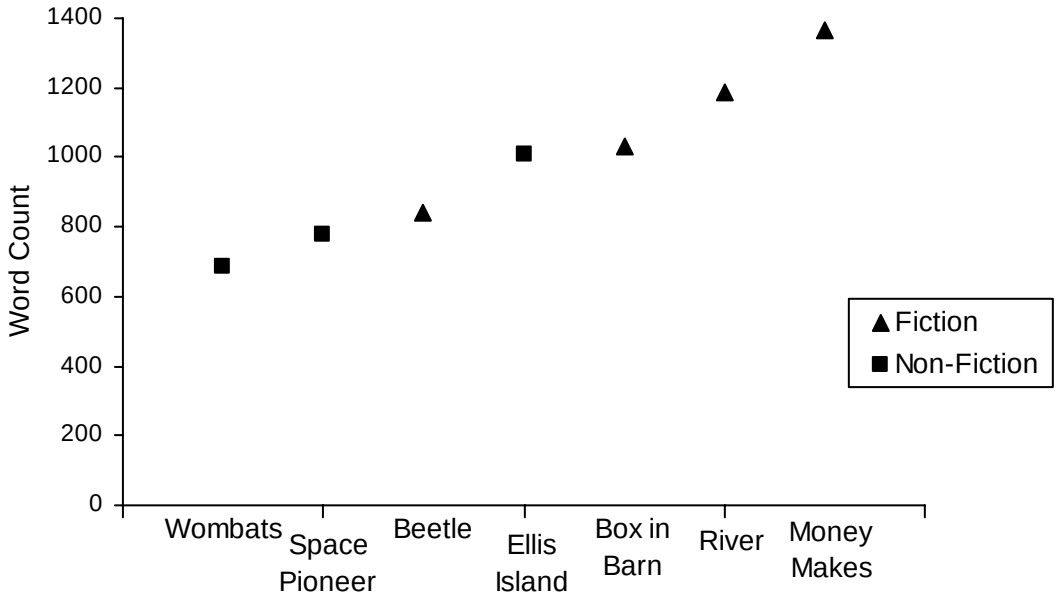


Figure 4.4. Plot of Passage Word Count

The confounding of passage length and reading context makes interpretation of the factors somewhat difficult, since both variables provide plausible explanations of the results. On one hand, the factors may represent differences in reading short texts versus long texts. On the other hand, the two factors may capture the use of different strategies in reading fiction versus non-fiction passages. It is also possible that additional item or passage characteristics which were not considered in this analysis might also provide plausible explanations of the two factors.

Parcel Scores

Based on the factor analysis, two parcel scores were derived by constructing a multidimensional Rasch model (Camilli et al., 2006) that conformed to the underlying structure identified in the factor analysis. Parcel 1 reflects the underlying structure of Factor 1, and can be interpreted as a Long Passage Parcel or Fiction Parcel. Parcel 2 is comprised of those variables associated with Factor 2, and can be interpreted as a Short Passage Parcel or Non-Fiction Parcel. The process and design matrix used in the estimation of the parcel scores were similar to those used in the estimation of testlet scores from Models 2 and 3. In the design matrix for the Model 1 parcel scores, parcels are indicated by dummy variables which take on a value of zero or negative one.

Table 4.4
Design Matrix for Model 1 Parcel Scores

Item	Parcel 1	Parcel 2	Item	Parcel 1	Parcel 2
9	-1	0	26	0	-1
21	-1	0	52	0	-1
50	-1	0	53	0	-1
67	-1	0	55	0	-1
68	-1	0	63	0	-1
69	-1	0	71	0	-1
70	-1	0	74	0	-1

Values for Parcel 1 ranged from -0.304 to 0.211, while values for Parcel 2 ranged from -0.156 to 0.212. By design, the mean of each parcel was zero. Texas had the lowest Parcel 1 score among all fifty participating jurisdictions ($\varphi_{1=}$-0.304). Students in Texas performed more poorly on those items associated with longer, fiction passages than

would have been expected given θ_{TX} = -.20. Tennessee had the highest Parcel 1 score ($\varphi_{1=}$0.211), indicating that students in Tennessee performed better than expected on those items associated with longer, fiction passages given θ_{TN} = 0.02. Interpretation of Parcel 2 scores follows similarly. For example, Utah, Montana, Idaho, Kansas and Minnesota all performed more poorly than expected on the items associated with short, non-fiction passages that were captured by Parcel 2. Conversely, Louisiana, Texas, Mississippi, New York and Maryland performed better than expected on the same parcel of items associated with short, non-fiction reading passages.

Table 4.5
Jurisdictions with Extreme Parcel Scores

	Jurisdiction (Parcel Score Magnitude)	
Parcel	Low	High
Parcel 1 - Long/Fiction	**Texas (-0.304)** **Maryland (-0.245)** Massachusetts (-0.231) **New York (-0.218)** Washington (-0.167)	Tennessee (0.211) West Virginia (0.185) Iowa (0.172) Indiana (0.148) Nebraska (0.148)
Parcel 2 - Short/Non-fiction	Utah (-0.156) Montana (-0.134) Idaho (-0.115) Kansas (-0.110) Minnesota (0.104)	Louisiana (0.212) **Texas (0.206)** Mississippi (0.159) **New York (0.136)** **Maryland (0.134)**

Parcel 1 and Parcel 2 scores were moderately negatively correlated (r = -0.43), indicating that jurisdictions that tended to score low on Parcel 1 scored comparatively high on Parcel 2, and vice versa. Three states, Texas, Maryland and New York, were among the extremes for both Parcels 1 and 2. Given their overall reading proficiency estimates, these states scored relatively low on items associated with Parcel 1 (i.e., long

passage or fiction items) and relatively high on items associated with Parcel 2 (i.e., short passage or non-fiction items).

The correlation between Parcel 1 scores and overall jurisdiction reading proficiency was trivial (r = 0.121). The distribution of Parcel 1 scores was slightly skewed with five outliers, (Tennessee, Texas, Maryland, Massachusetts and New York). Removal of these outliers only minimally affected the correlation between Parcel 1 and overall proficiency (r = 0.208). On the other hand, Parcel 2 was *negatively* correlated with overall reading proficiency (r = -0.60). Jurisdictions with low overall reading proficiencies tended to perform better than expected on items accompanying shorter, non-fictional reading passages. Conversely, states with higher reading proficiencies tended to perform worse than expected on those same items. Parcel 2 was more normally distributed than Parcel 1, with only 2 extreme (low) observations, (Louisiana and Texas). Removal of these two outliers only minimally affected the correlation between Parcel 2 and reading proficiency (r = -0.595).

Model 2

As noted above, Model 2 included both items and testlets in a two-level design. Rather than focusing on item BLUPs and item difficulty variation, the Model 2 analysis focused on testlet scores and testlet difficulty variation (TDV). In Model 2, testlet was viewed as a random variable, while item difficulty was viewed as fixed. Using this model, TDV was estimated for each testlet across all states, given state-level proficiency and fixed item difficulty estimates. The differentiating feature of this model is that testlet

Model 2.

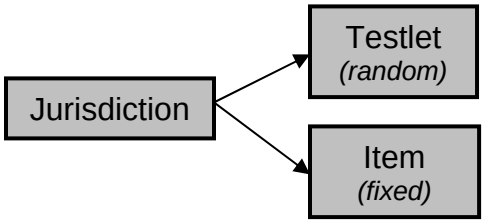


Figure 4.5. Model 2

---

difficulty variation was estimated, rather than item difficulty variation, in order to detect differential state performance on testlets, while taking into account item difficulty and state-level proficiency estimates.

Table 4.6
Model 2: Rank Order of TDV

Testlet	TDV	Context	Word Count
River	0.0027	Fiction	1184
Box in Barn	0.0012	Fiction	1029
Ellis Island	0.0011	Non-fiction	1011
Money Makes	0.0007	Fiction	1366
Goodall	0.0004	Non-fiction	993
Space Pioneer	0.0004	Non-fiction	779
Wombats	0.0003	Non-fiction	684
Beetle	0.0001	Fiction	840

Testlet difficulty variation was estimated for each of the testlets, and ranged from 0.0001 for the Beetle testlet to 0.0027 for the River testlet. River, Box in Barn, Ellis Island and Money Makes had the relatively largest estimates of TDV, indicating that

those testlets were more likely to capture differential testlet performance. Goodall, Space Pioneer, Wombats and Beetle had relatively smaller TDVs, indicating that there was less variability across corresponding testlet scores.

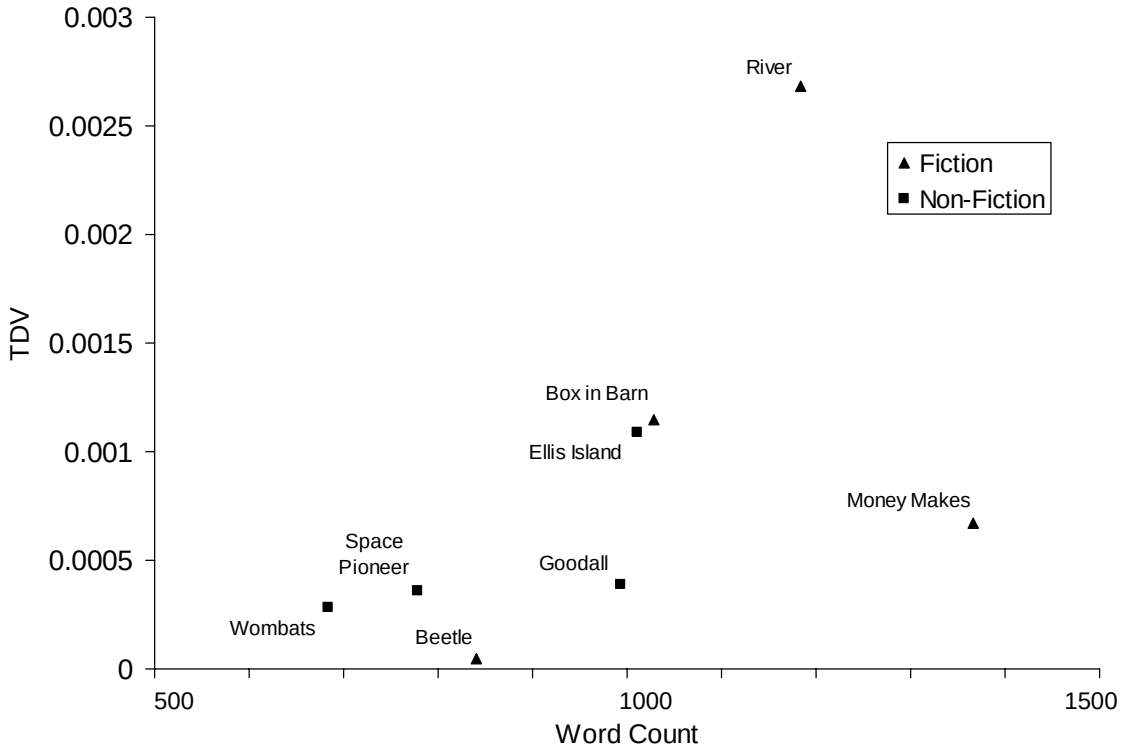


Figure 4.6. Testlet Difficulty Variation (TDV) by Word Count

Interestingly, passage length was correlated with TDV (r = 0.55). Longer passages were associated with larger TDV, while shorter passages were associated with smaller TDV. In other words, jurisdiction performance was more consistent with expected levels when reading passages were short. When reading passages were longer, jurisdiction performance varied more widely from expected values. States exhibited differential

performance in testlet scores when students were required to read longer passages, supporting previous speculation that one or more state-level factors may be influencing performance on items associated with longer reading passages. Reading context was also associated with TDV magnitude, although the relationship was weaker (r = 0.38) than that of passage length and TDV.

For the full complement of 82 test items, estimated item BLUPs were not highly correlated with their parent testlet scores. In fact, for the Space Pioneer and River testlets, item BLUP by testlet score correlations were generally large and negative. Most of the testlet scores were not correlated with overall jurisdiction reading proficiency. The largest positive correlation was with the estimated BLUPs from the Goodall testlet, (r = 0.32). As overall proficiency increased, performance on the Goodall testlet exceeded what would have been expected, indicating that jurisdiction with higher levels of reading proficiency scored even higher on the Goodall testlet than expected, while districts with lower reading proficiency scored even worse than expected on the Goodall testlet. Conversely, the Money Makes testlet was moderately negatively correlated with jurisdiction proficiency (r = -0.43), indicating that as overall jurisdiction reading proficiency rose, performance on this testlet fell short of what would have been predicted.

Factor Analysis of Model 2 Testlet Scores

A factor analysis was run on the testlet scores of all fifty participating jurisdictions using principle axis factoring and varimax rotation. This factor analysis partially confirmed the underlying structure of the two factors retained from the Model 1 analysis. A similar pattern was obtained with Factor 1 representing longer, fiction passages and Factor 2 representing shorter, non-fiction passages. The passage Money Makes did not

load on Factor 1; however, this passage did have a negative loading of -0.777 on Factor 2. The negative loading is not surprising given that Money Makes is a long, fiction passage, and the passages loading on Factor 2 are shorter, non-fiction passages. In addition to the first two factors, a third factor, with (λ_3=1.24, 15.4% of explained variance) emerged during this analysis. Only the Space Pioneer testlet loaded on Factor 3.

Table 4.7
Model 2 Testlet Factor Loadings, Eigenvalues and Percent of Explained Variance

	Factor 1 (λ_1=2.25, 28.1%)			Factor 2 (λ_2=1.97 24.6%)		
Testlet	Loading	Context	Words	Loading	Context	Words
Beetle	-0.628	Fiction	840	-0.238		
Box in Barn	0.707	Fiction	1029	0.093		
Money Makes	-0.289			-0.777		
Goodall	-0.313			0.674	Non-fiction	993
Ellis Island	0.737	Non-fiction	1011	-0.101		
Space Pioneer	-0.176			-0.322		
River	0.749	Fiction	1184	-0.145		
Wombats	-0.180			0.849	Non-fiction	684

Analysis of Moderator Variables

Profiles of jurisdiction characteristics were created for each of the fifty participating jurisdictions. Jurisdiction profiles were comprised of forty-two moderator variables drawn from the United States Census 2000; the NAEP background questionnaires of students, teachers and principals; the NAEP website; *Quality Counts 2002: The State of the States* (Meyer, Orlofsky, Skinner & Spicer, 2002); and two published research reports (National Center for Education Statistics [NCES], 2007;

Peterson & Hess, 2006). These moderator variables were consolidated into eight classes of jurisdiction characteristics.

Table 4.8
Moderator Variable Classes

Class	Description
I	English proficiency
II	Socio-economic status
III	Demographics
IV	Learning resources
V	Student characteristics
VI	Teacher preparation and development
VII	Content standards and NAEP alignment
VIII	Educational policy

Following creation of the jurisdiction profiles, a correlational analysis was conducted in order to backward map jurisdiction characteristics onto parcel score outcomes. A brief case study of three jurisdictions is presented last, in order to explore the potential utility of parcel scores in the interpretation state-level reading performance patterns, and provide a unique description of the specific conditions within the selected states.

Major Findings

The correlational analysis yielded two major findings. The most consistent finding pertained to variables associated with English proficiency and usage. As rates of non-native speakers increased, scores on Parcel 1 decreased and scores on Parcel 2 increased. In other words, jurisdictions with higher levels of non-native speakers scored

lower on items accompanying longer, fiction passages and higher on those accompanying shorter, non-fiction passages than would have been expected given overall jurisdiction reading proficiency. Deviation from expected performance was clearly associated with rates of non-native speakers, and similar patterns were seen across all variables associated with English language use. Consequently, the proportion of non-English speakers (NENG) in a jurisdiction was used as a control variable in a secondary analysis of partial correlations. These results are presented and discussed alongside the main findings of this analysis.

A less consistent, but similar finding was observed among some measures of jurisdiction wealth. As average income decreased, scores on Parcel 1 tended to increase. As poverty levels increased scores on Parcel 2 tended to increase. Taken together, these observations suggest that poorer jurisdictions scored lower on longer, fiction passages and higher on shorter, non-fiction passages that would have been expected given overall reading proficiency. This finding was not as consistent as those observed among variables associated with English proficiency. Among the measures of jurisdiction wealth only poverty rate correlated with the proportion of non-English speakers (r = 0.36).

Class I: English Proficiency

Two moderator variables directly measured use of English language. NENG was culled from the U.S. Census 2000 and reflects the proportion of non-native speakers of English within a jurisdiction. ENG is estimated from the 2003 NAEP student questionnaire, and reflects the proportion of fourth grade students who reported living in households where only English is spoken. Not surprisingly, NENG and ENG were highly negatively correlated (r = -0.94).

Table 4.9
Correlations: English Proficiency and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
NENG	-0.38**	0.32*	–	–
ENG	0.49**	-0.30**	0.43**	0.01
LEP	-0.10	0.01	0.14	-0.21
LEP LOG	-0.31*	0.01	-0.08	-0.28

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

NENG was negatively correlated with Parcel 1 and positively correlated with Parcel 2. Not surprisingly, the inverse of these correlations was observed for ENG. A similar pattern was observed with the demographics variables, FOREIGN and HISPANIC, which reflect the proportions of residents identifying them selves as foreign born, or Hispanic or Latino in origin. (For a more detailed discussion of FOREIGN and HISPANIC, see the subsequent section addressing Class III: Demographics.) These variables reflect the same underlying relationship: jurisdictions in which languages other than English are more likely to be spoken tended to perform better on short passage, non-fiction items than expected, while performing worse than expected on long passage, fiction items.

The variable LEP, which refers to the proportion of students classified as having limited English proficiency, did not reflect the same pattern as the other variables. This inconsistency can be explained by two aspects of the variable LEP. First, LEP was not normally distributed. Values were heavily skewed toward the low end of the distribution.

In an attempt to counter this, a new variable was created by taking the log of LEP. The correlation between LEP LOG and Parcel 1 was consistent with other measures of English language proficiency and usage; however, the correlation with Parcel 2 failed to conform.

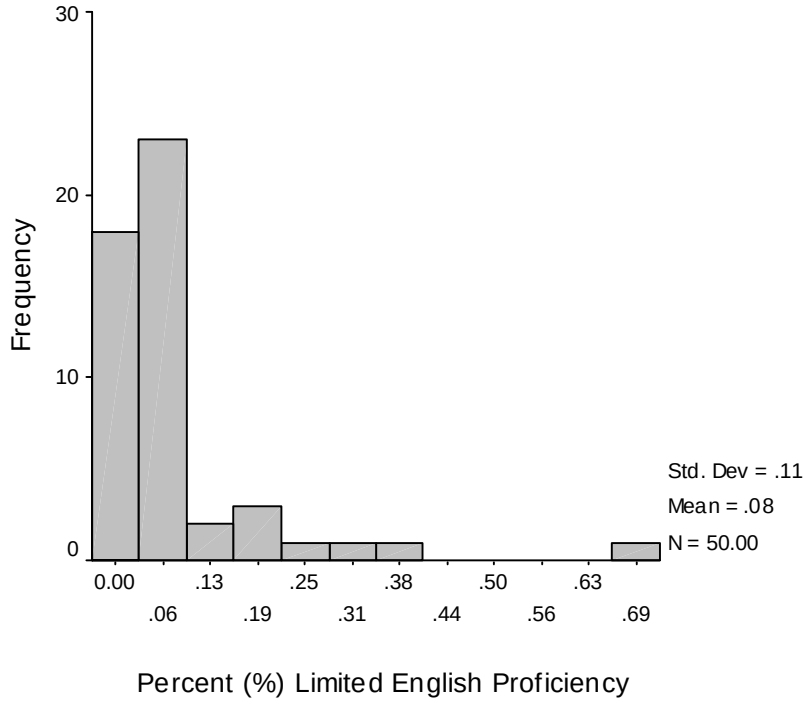


Figure 4.7. Histogram of Limited English Proficiency (LEP)

Second, concerns have been raised by other sources regarding how jurisdictions classify students as having limited English proficiency. Notably, Abedi (2004), in a paper commissioned by the National Assessment Governing Board (NAGB) for the NAGB Conference on Increasing Participation of SD and LEP Students in NAEP, reported inconsistencies in how limited English proficiency is defined across the nation. Furthermore, performance on reading and language arts tests could not reliably predict

state classifications, indicating that factors other than language proficiency may influence how children are classified as English language learners across jurisdictions.

Given these concerns the variable LEP was eschewed, and the variable NENG was selected as a control variable for the analysis of partial correlations. When NENG was used as a control variable, the correlation between ENG and Parcel 1 remained roughly the same. The relationship between ENG and Parcel 2 changed when NENG was introduced as a control variable, dropping from $r = -0.30$ ($p < 0.01$) to $r = 0.01$ (NS). This observation suggests that for Parcel 2 in particular, rates of English proficiency may be an important mediator in the pattern of observed effects.

Class II: Socio-Economic Status

Five moderator variables captured different aspects of socio-economic status (SES) and jurisdiction wealth. Four variables (INCOME, PER CAPITA, POVERTY, and HOMEOWN) were obtained from Census 2000. The variable FREE LUNCH, which reflects the proportion of fourth graders eligible for free/reduced lunch, was derived from the NAEP database. The results of this analysis suggest that poorer districts perform worse on Parcel 1 (fiction/long passage) and better on Parcel 2 (non-fiction/short passage) than predicted by overall reading proficiency, and vice versa for wealthier districts. While a clear trend is evident, these findings are not consistent across all SES measures.

Parcel 1 (fiction/long passage) was positively correlated with jurisdiction wealth as measured by median and per capita income: INCOME ($r = -0.42$) and PER CAPITA ($r = -0.42$). Jurisdictions with lower measures of income performed worse than expected

on items associated with longer, fiction passages. These correlations were fairly robust and shrunk only slightly when NENG was introduced as a control variable.

Table 4.10
Correlations: SES and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
INCOME	-0.42**	-0.08	-0.36*	-0.24
PER CAPITA	-0.42**	0.30	-0.38*	-0.06
POVERTY	0.04	0.43**	0.17	0.58**
HOMEOWN	0.41**	-0.36*	0.13	-0.22
FREE LUNCH	-0.02	0.51**	0.11	0.63**

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

Conversely, the proportion of home ownership was negatively correlated with Parcel 1 ($r = 0.41$), although this correlation evaporated when NENG was introduced as a control. The correlation between HOMEOWN and Parcel 1 is not entirely inconsistent with the previous results. Inspection of the census data indicates that high rates of home ownership are not necessarily associated with greater jurisdiction wealth, and the relationships among these variables are far more complex. For example, both poor, rural communities and affluent suburbs can have high rates of home ownership.

Correlations associated with the remaining two moderator variables suggest that as rates of poverty and eligibility for free/reduced lunch increase, Parcel 2 performance exceeds what would be predicted by overall jurisdiction reading proficiency. As poverty rates increased, Parcel 2 scores increased as well ($r = 0.43$). Likewise, as the proportion of children eligible for free/reduced lunch rose across jurisdictions, Parcel 2 scores rose

as well (r = 0.51). This pattern of correlations was consistent with partial correlations obtained when NENG was used as a control variable.

Class III: Demographics

Correlations between with parcel scores and measures of population size and density were generally trivial, excepting those associated with total population. POPULATION was negatively correlated with Parcel 1 (r = -0.38), and positively correlated with Parcel 2 (r = 0.43). Jurisdictions with larger total populations tended to score lower on long passage, fiction items and higher on short passage, non-fiction items than would be predicted based on their overall reading proficiency. Conversely, states with smaller populations tended to do better on long passage, fiction items and worse on short passage, non-fiction items, than would have been predicted. When the proportion of non-native speakers of English was introduced as a control variable, these two correlations dropped significantly. The proportion of minors, the proportion of senior citizens, and population density were not associated with either parcel score.

Table 4.11
Correlations: Population Size and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
POPULATION	-0.38**	0.43**	-0.07	0.37
POPULATION<18	0.02	0.03	0.10	0.24
POPULATION≥65	0.21	0.09	0.20	-0.18
DENSITY	-0.12	0.22	-0.07	0.19

[a] Controlling for NENG. * p<0.05. ** p<0.01.

A number of variables were drawn from the US Census 2000 that related to gender and ethnicity. Parcel 1 was negatively correlated with the proportion of residents identifying themselves as Hispanic or Latino (r = -0.42), and the proportion of foreign born residents (r = -0.63). Jurisdictions with higher levels of Hispanic and foreign-born residents scored lower on long passage, fiction items than predicted by overall reading proficiency. Interestingly, when the proportion of non-native speakers is included as a control variable, the correlation between Hispanics and Parcel 1 changes dramatically from r = -0.42 to r = 0.31. When NENG is used as a control variable, the correlation between HISPANIC and Parcel 1 becomes positive, and more similar to the simple correlation between the proportion of Whites and Parcel 1 (r = 0.17). The correlation between Parcel 1 and the proportion of foreign-born residents shrinks as well (r = -.03). Thus the significant correlations of HISPANIC and FOREIGN with Parcel 1 are largely driven by rates of English proficiency within the jurisdictions.

Table 4.12
Correlations: Gender and Ethnicity, and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
FEMALE	-0.12	0.39**	-0.25	0.50**
WHITE	0.17	-0.52**	0.11	-0.58**
BLACK	-0.10	0.61**	-0.18	0.73**
HISPANIC	-0.42**	0.32*	0.31*	-0.10
FOREIGN	-0.63**	0.41*	-0.28	0.11

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

Parcel 2 was correlated with all five measures of gender and ethnicity. The proportion of White residents was negatively correlated with Parcel 2, (r = -0.52), indicating that states with smaller proportions of White residents tended to perform better on the short passage, non-fiction items than would have been predicted based on overall reading proficiency. This general finding is supported by correlations observed between Parcel 2 and the proportions of Blacks/African-Americans (r = 0.61), and Hispanics (r = 0.32). Controlling for NENG resulted in stronger correlations for WHITE and BLACK, while correlations for HISPANIC and FOREIGN shrunk dramatically, (r = -0.10 and r = 0.11 respectively), suggesting that the rate of English proficiency is mediating the relationship between Parcel 2 (short passage/non-fiction) and these moderator variables.

Class IV: Learning Resources

Five variables representing learning resources were drawn form the NAEP questionnaires of students and teachers. The variable RESOURCES represents the proportion of students with "A lot" of classroom reading resources. LANG ARTS represents the proportion of students receiving ten or more hours of Language Arts instruction per week, while SKILLS represents the proportion spending a majority of Language Arts time on reading skills instruction (at least 60%). The proportions of

Table 4.13
Correlations: Learning Resources and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
RESOURCES	0.02	-0.43**	-0.20	-0.33*
LANG ARTS	-0.03	0.23	-0.06	0.29
SKILLS	0.16	0.42**	0.38*	0.34*
BOOKS	-0.01	-0.65**	-0.29	-0.64**
COMPUTER	-0.15	-0.49**	-0.43**	-0.45**

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

students reporting more than 25 books in their homes and possessing a home computer are captured by BOOKS and COMPUTER respectively. The simple correlations between Parcel 1 and this class of variables were small. Four variables were significantly correlated with Parcel 2.

The proportions of fourth graders with "A lot" of classroom reading resources, more than 25 books in the home, and a home computer were negatively correlated with Parcel 2 (r = -0.43, r = -0.65, and r = -0.49 respectively). Conversely, the proportion of students receiving more reading skills instruction correlated positively with Parcel 2 (r = 0.42). This pattern of correlations may be explained be examining the correlations of these four resource variables with measures of SES.

Table 4.14
Correlations: Learning Resources and SES

	INCOME	PER CAPITA	POVERTY
RESOURCES	0.38**	0.42**	-0.55**
SKILLS	-0.41**	-0.32*	0.67**
BOOKS	0.48**	0.42**	-0.81**
COMPUTER	0.59**	0.54**	-0.80**

* $p<0.05$. ** $p<0.01$.

Students in wealthier jurisdictions had more access to learning resources in the form of classroom reading resources, books in the home, and computers in the home, indicating that the variables RESOURCES, BOOKS and COMPUTER may represent aspects of Opportunity to Learn (OTL) associated with jurisdiction wealth. The observed correlation coefficients between Parcel 2 and these three resource variables suggest that as access to learning resources decreases, jurisdiction performance on the short passage, non-fiction items associated with Parcel 2 exceeds that which would be predicted by overall jurisdiction reading proficiency. This corroborates the observation that poorer jurisdictions score better than expected on short passage, non-fiction items.

Jurisdictions with higher rates of reading skills instruction tended to have lower levels of income and higher rates of poverty. This suggests that children in poorer jurisdictions may receive more skills-based instruction than their peers in wealthier jurisdictions. Parcel 2 was negatively correlated with SKILLS ($r = 0.42$), indicating that jurisdictions reporting higher rates of reading skills instruction scored higher than expected on short passage, non-fiction items.

Class V: Student Characteristics

Where applicable, student characteristics were included in variable classes that were conceptually related to the characteristic being measured. For example, the variable LEP is included in Class I: English Proficiency. The three remaining student characteristic variables are discussed in this section were drawn from the NAEP database. IEP represents the proportion of students within a jurisdiction that have been classified as having a disability. ACCOMODATION represents the proportion of students receiving accommodations during NAEP administration. The most common accommodation was extended testing time (Grigg et al., 2003). ABSENTEEISM represents the proportion of students missing one or more days of school during the month before NAEP administration. Correlations between these three variables and the parcels were trivial, and did not suggest a relationship with parcel performance.

Table 4.15
Correlations: Student Characteristics and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
IEP	0.00	-0.14	-0.20	-0.02
ACCOMODATION	-0.18	-0.13	-0.13	-0.21
ABSENTEEISM	0.08	-0.02	0.21	-0.12

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

Class VI: Teacher Preparation and Development

The nine variables included in this section are divided into two sets in order to facilitate explanation. The first set of variables pertains to teacher quality and training. The second set addresses compensation and professional development.

Teacher Quality and Training

Teacher quality and training was captured by four variables. QUALITY is an overall composite score of improving teacher quality as reported in *Quality Counts 2002: The State of the States* (Meyer et al., 2002). EXPERIENCE represents the average number of years of experience for teachers within a state. CERTIFICATION represents the proportion of students with a state certified teacher, and DEGREE represents the proportion of students with a teacher who has a major or minor in reading or literacy.

Both EXPERIENCE and CERTIFICATION were positively correlated with Parcel 1 (r = 0.33 and r = 0.38 respectively), and negatively correlated with Parcel 2, (r = -0.50 and r = -0.40 respectively). Jurisdictions with less experienced and fewer credentialed teachers scored lower than expected on long passage, fiction items, and higher than expected on short passage, non-fiction items. The converse was found for jurisdictions with more experienced and certified teachers. These correlations generally shrank when NENG was used as a control variable, although the basic pattern was maintained.

Table 4.16
Correlations: Teacher Quality and Training, and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
QUALITY	0.00	0.17	-0.02	-0.20
EXPERIENCE	0.33*	-0.50**	0.14	-0.41**
CERTIFICATION	.38**	-0.40**	0.20	-0.29
DEGREE	-0.13	-0.01	-0.05	-0.08

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

In unpacking the relationship between teaching experience and certification, and Parcels 1 and 2, it is useful to examine measures of SES and rates of non-native speakers. EXPERIENCE and CERTIFICATION were both negatively correlated with POVERTY and NENG, indicating that students in jurisdictions with higher rates of poverty and non-native speakers were more likely to be in classrooms with less experienced and non-certified teachers. It is not surprising then that the pattern of correlations between Parcels 1 and 2 and EXPERIENCE and CERTIFICATION replicates that between Parcels 1 and 2 and SES, and between Parcels 1 and 2 and English proficiency variables. In summary, jurisdictions with higher levels of poverty, non-native speakers, less experienced teachers, and non-certified teachers scored lower on longer, fiction passages and higher on shorter, non-fiction passages that would have been expected given overall reading proficiencies.

Table 4.17
Correlations: Teacher Quality, Poverty and English Proficiency

	POVERTY	NENG
EXPERIENCE	-0.35*	-0.22
CERTIFICATION	-0.32*	-0.49**

* $p<0.05$. ** $p<0.01$.

Compensation and Professional Development

Four variables measured compensation and professional development of teachers. The variable SALARY is the average salary for all teachers within a jurisdiction adjusted for the cost of living (Meyer et al., 2002). With regard to professional development, the variables RELEASE, STIPEND and TUITION provide the proportion of students with

teachers eligible for release time, stipends, and tuition remission. The correlations between these variables and Parcels 1 and 2 were generally small and not significant, and do not suggest an association between the parcel scores and compensation and support for professional development.

Table 4.18
Correlations: Teacher Compensation and Development, and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
SALARY	-0.24	0.04	-0.32*	0.06
RELEASE	-0.11	-0.09	-0.06	-0.14
STIPEND	-0.03	0.05	-0.11	0.13
TUITION	-0.12	-0.45**	0.28	-0.47**

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

Class VIII: Content Standards and NAEP Alignment

The eighth class of moderator variables examined in this analysis consisted of seven variables related to state content standards and alignment with the NAEP. These variables are presented in two sets. The first set addresses quality and usage of content standards. The second set addresses alignment and rigor of content standards with respect to the NAEP.

Content Standards Quality and Usage

The variables STANDARDS was drawn from *Quality Counts 2002*, a large-scale evaluation of state standards and accountability published annually in *Education Week*, (Meyer et al., 2002). STANDARDS is a composite variable representing an overall score with regard to the quality of state content standards and accountability. In constructing this score, the authors evaluated the clarity and specificity of core content standards; the

quality of assessments used to evaluate student or school performance; and the extent to which schools were held accountable for performance.

Table 4.19
Correlations: Content Standards and Usage, and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
STANDARDS	-0.40**	0.40**	-0.32*	0.33*
USAGE	-0.00	0.14	-0.04	0.20

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

STANDARDS was negatively correlated with Parcel 1 (r = -0.40) and positively correlated with Parcel 2 (r = 0.40). States with higher STANDARDS scores tended to perform worse than expected on long passage, fiction items, and better than expected on short passage, non-fiction items. Partial correlations were similar but smaller when controlling for the proportion of non-native speakers.

STANDARDS was not associated with measures of SES or English proficiency. The correlations between STANDARDS and POVERTY (r = 0.16), and STANDARDS and NENG (r = 0.26) were small and non-significant. These small correlations indicate that it is unlikely that jurisdiction wealth or English proficiency is driving the relationship between the variable STANDARDS and the parcel scores. This suggests that in their evaluations of state standards and accountability programs, Meyer and colleagues (2002), may be assigning higher scores to programs that favor instruction and testing of non-fiction or short texts.

Regarding the usage of state Language Arts standards, there was no association between parcel score performance and the proportion of teachers who reported using the standards to guide classroom practice. This is likely due to the fact that the variable USAGE had very low variability, since most jurisdictions were similar in having very high percentages of teachers who reported using the state standards.

NAEP Alignment

Two sources (NCES, 2007; Peterson & Hess, 2006) provided moderator variables that estimated the alignment of state content standards or assessments with the NAEP. NCES (2007) compared discrepancies between proficiency estimates based on individual state assessments compared to performance on the NAEP. State assessment cut-scores were mapped onto the NAEP scale to create NAEP equivalency estimates. These NAEP equivalency scores can be compared to actual NAEP scores to estimate how closely state assessments (and their corresponding standards) are aligned with the NAEP. The authors assert that the "relative ranking of the NAEP score equivalents to the states' proficiency standards offers (a) a credible indicator of the relative stringency of the [states'] standards, and (b) a more useful basis for policy discussion in the differences in percentages [of students deemed proficient by individual state tests compared to the NAEP]" (NCES, 2007, p. 1). Two moderator variables were culled from this study. EQUIVALENT contains NAEP equivalent scores for each jurisdiction. EQUIVALENT SE contains the estimated standard errors of the NAEP equivalency scores.

Peterson and Hess (2006) compared NAEP jurisdiction proficiency estimates to proficiency estimates measured by individual state assessments. They awarded letter grades ranging from A to F, according to how well state proficiency estimates coincided

with NAEP estimates of proficiency. The authors assert that states with more rigorous content standards are more likely to be aligned with the NAEP, and proficiency estimates should be similar. States with proficiency estimates that coincided with those obtained by the NAEP (or whose standards were deemed more stringent than NAEP requirements) were awarded high grades by the authors. States whose proficiency estimates were much larger than those estimated by NAEP were awarded very low marks. For the current study, the variable NAEP Alignment was created by converting the grades awarded by Peterson and Hess (2006) to numeric grade equivalents, (e.g., A=4, B=3, C=2, D=1, F=0). This variable can be loosely interpreted as an indicator of the stringency of state content standards and alignment with the NAEP.

Table 4.20
Correlations: NAEP Alignment and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
EQUIVALENT	-0.29	-0.01	-0.17	-0.15
EQUIVALENT SE	0.26	-0.11	0.01	0.06
ALIGNMENT	-0.31	-0.01	-0.18	-0.07

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

Correlations between the parcel scores and the NAEP alignment variables were small and non-significant. NAEP alignment variables suggested by the NCES (2007) and Peterson and Hess (2006) do not appear to predict parcel score performance.

Class IX: Educational Funding

The final class of moderator variables pertained to different aspects of educational funding. All three were drawn from *Quality Counts 2002* (Meyer et al., 2002). The

variable EQUITY was a composite score of resource equity that included evaluations of states' efforts to equalize funding across districts, and the extent to which funding actually varied across districts. ADEQUACY was a composite variable that captured the extent to which states provided adequate resources for public education. EXPEDITURES represented the annual per-pupil expenditures adjusted for regional cost differences. The correlations between these three variables and the parcels scores were all small and non-significant. Education funding did not appear to be associated with parcel score performance across jurisdictions.

Table 4.21
Correlations: Educational Funding and Parcel Scores

	Correlation		Partial Correlation[a]	
Moderator Variable	Parcel 1	Parcel 2	Parcel 1	Parcel 2
EQUITY	-0.04	0.09	0.12	-0.01
ADEQUACY	-0.02	-0.16	-0.28	-0.04
EXPENDITURES	-0.01	-0.10	-0.17	-0.08

[a] Controlling for NENG. * $p<0.05$. ** $p<0.01$.

Case Study: Maryland, New York and Texas

A brief case study was conducted on a subset of three states, Maryland, New York and Texas. This analysis is provided in order to illustrate how parcel scores may be used to explore the potential link between jurisdiction characteristics and reading performance.

Sample Selection

Three jurisdictions were selected for this analysis, in order to provide a sample that was small and manageable given the amount of data, yet large enough to provide

some variation across observations. These three states were selected based on inspection of the IDV estimates, parcel scores, and moderator variable profiles.

A quadrant grid of parcel scores was employed in the selection of this sample. (See Table 4.5.) Simple inspection of this grid revealed that three states, Maryland, New York and Texas, exhibited a very similar score pattern. All three states were among the lowest extreme scores for Parcel 1 and the highest extreme scores for Parcel 2. Students in Maryland, New York and Texas score lower than expected on items from longer, fiction passages, and higher than expected on items from shorter, non-fiction passages. This pattern illustrates an extreme of the relationship between Parcel 1 and Parcel 2. Due to this pattern of parcel scores, Maryland, New York and Texas were selected for further investigation.

Data and Analysis

The primary data in this analysis were the individual state profiles that were originally constructed for the moderator variable analysis. These variables were drawn from the U.S. Census 2000; the 2002 NAEP reading assessment of fourth graders; 2002 NAEP questionnaires of students and teachers; *Education Week*: Quality Counts 2002; Peterson & Hess (2006); and NCES (2007). These profiles were supplemented with additional information about reading achievement from The Nation's Report Card: Reading 2002 (Grigg et al., 2003). Official websites of the departments of education of Maryland (http://mdk12.org, http://marylandpublicschools.org), New York (http://emsc.nysed.gov/) and Texas (http://www.tea.state.tx.us/) also provided more information about state academic standards and assessments.

The two main goals of this analysis were to provide:

(1) a brief description of each state, with regard to reading achievement as measured by the 2002 NAEP, and key ecological and policy variables; and

(2) a discussion of shared ground among the three states, in order to provide context for the observed pattern of parcel scores.

A descriptive summary of reading achievement was generated for each state based on quantitative achievement outcomes (e.g., parcel scores, testlet scores, and mean NAEP reading score). These descriptions were supplemented with additional information provided in *The Nation's Report Card: Reading 2002* (Grigg et al., 2003). In addition to achievement variables, moderator variable profiles were examined and descriptive summaries were prepared for classes of variables. These descriptive summaries were then supplemented with information from the *The Nation's Report Card: Reading 2002* (Grigg et al., 2003), *Education Week: Quality Counts 2002*, and websites of the departments of education of Maryland, New York and Texas. The information provided from these additional sources was both quantitative and descriptive.

A second review of moderator variable profiles was conducted in order to identify similar patterns in variables, or classes of variables, across the three states. In this analysis both raw scores and *z*-scores of all quantitative variables were compared in an attempt to identify clusters of variables that suggested overlap among the three states. In addition, several new variables were examined, (e.g., year of implementation of content standards in Reading/Language Arts; the extent of state academic achievement testing). It is important to note that the moderator variables included in this analysis were not all obtained during the same year. For example, although much of the Census data was

drawn from 2000, some summary information was not available for that year. In that event, data from year closest to 2002 was used. This general rule was followed for all sources of data.

Results

Maryland

Among the three states included in this analysis, Maryland had the smallest population, with 5,296,486 residents, and ranked 19th among the 50 states. Compared to a national poverty rate of 12.7%, Maryland's rate of 8.8% was relatively small and ranked 46th. With regard to reported ethnicity, Maryland was one of the most diverse states in the country. Only 64.2% of Maryland residents reported their ethnic group as White, yielding a small White majority (47th out of 50). In spite of this diversity, most residents of Maryland spoke primarily English. Compared to the national average of 17.9%, only 13.2% of Maryland residents spoke another language at home. In addition, Maryland had a comparably low proportion of foreign born residents, ranking 14th in the nation with only 10.8% of the total population born outside the U.S.

With regard to reading achievement, Maryland's performance on the 2002 NAEP was about average. The national average scale score for fourth graders was 219, while Maryland's average was 217. The percentage of Maryland students scoring at the *Proficient* level or higher was not found to be significantly different from the nation (Grigg et al., 2003).

Compared to other states, Maryland's implementation of systematic, standards-based reform appears to have occurred later and on a smaller scale. Maryland did not adopt state-wide academic content standards until 2001, compared to some other states

that had already adopted core content standards in the mid 1990s. In addition, Maryland has content standards in only four academic subject areas: English/Language Arts, Mathematics, Science, and Social Studies. Some other states have a broader range of standards, and address subjects such as the arts, physical education, and foreign language.

The Maryland School Assessment (MSA) is the state assessment in reading, mathematics, and science. The reading and mathematics tests are administered annually in grades three through eight. The science test is administered in fifth and eighth grade. The MSA is aligned with the Maryland Voluntary State Curriculum, in which the core content standards are embedded.

Mixed results surface regarding the quality of Maryland's reading content standards and assessment. With regard to the implementation of general academic standards and accountability, Maryland received the highest score of A in Quality Counts 2002: The State of the States (Meyer et al., 2002). With regard to rigor, Peterson and Hess (2006) awarded Maryland's fourth grade reading content standards a grade of C. In addition, the MSA reading test was found to be misaligned with the NAEP for fourth graders (NCES, 2007). Using data from 2005, the authors calculated a NAEP equivalent score of 187 based on MSA proficiency rates. This score of 187 is well below the NAEP cut-off scores for *Basic* (208) and *Proficient* (238). If the MSA were perfectly aligned with the NAEP, one would expect the NAEP equivalent score to match the NAEP *Proficient* score. The extent to which the NAEP equivalent score differs from the NAEP *Basic* and *Proficient* cut scores indicates misalignment between the NAEP and the MSA. "Most of the heterogeneity in score equivalents can be attributed to differences in the stringency of the proficiency standards set by the states" (NCES, 2007, p.2); however,

one cannot rule other reasonable differences, such as divergent emphases with regard to content or use of different item formats on the assessments. While we cannot conclusively state that Maryland core content standards and states assessment are less stringent than the NAEP, both Peterson and Hess (2006) and NCES (2007) arrived at that conclusion.

New York

New York is a highly populous and diverse state. With 18,976,457 residents, New York ranked third in the nation in population size. Just over 14% of New York's population was living below the poverty level, ranking it 16th among the states. While not among the poorest in the nation, this observation does place New York within the poorest third. With regard to reported ethnicity, New York was also one of the most diverse states. Only 68.8% of New York residents reported their ethnic group as White (43rd out of 50). This diversity is reflected in the percentage of foreign born New Yorkers (20.9%, second largest percentage in the nation), and the percentage of residents that speak a language other than English at home (27.4%, fourth largest percentage in the nation).

Fourth graders from New York scored slightly above the national average on the 2002 NAEP reading assessment, 222 versus 219. The percentage of New York students at or above *Proficient* of the NAEP was found to be significantly higher than that of the nation (Grigg et al., 2003).

With regard to modern, systematic education reform, New York was among the earliest states to implement academic content standards. In 1996, content standards were adopted in English/language arts; mathematics, science and technology; social studies; the arts; and languages other than English. In 2001, New York added standards for

physical education. From 2002 to 2007, New York students were tested in English and language arts in grades four and eight. In 2008, the state assessment was changed to include annual testing in grades three through eight.

With regard to the implementation of general academic standards and accountability, New York received a grade of A and ranked second (just behind Maryland) in *Quality Counts 2002: The State of the States* (Meyer et al., 2002). Peterson and Hess (2006) awarded New York state proficiency standards a strength grade of C. Although the New York assessment of English language arts (ELA) was found to be somewhat misaligned with the NAEP (NCES, 2007), its NAEP equivalent score was among the highest of participating states. This score was very close to the NAEP cut score of 208 for *Basic* however, New York's NAEP equivalent was well below the NAEP *Proficient* cut score of 238. Taken together, these observations suggest that New York English language arts standards and assessment to be somewhat aligned with the de facto rigor of the NAEP.

Texas

According to Census 2000, Texas is the second most populous state in the nation after California, with 20,851,820 residents. Texas was the 8th poorest state in the nation, with 16.6% living below the poverty level, compared to the national average of 12.7%. Texas ranked 36 among states with regard to the percentage of White residents (73%); however, Texas holds high percentages of foreign born residents (15.2%, 7th nationally), and residents who speak a language other than English at home (31.5%, 3rd nationally). With regard to reading achievement, Texas's performance on the 2002 NAEP was about average (217 compared to the national average of 219). The percentage of Texas students

scoring at the *Proficient* level or higher was not found to be significantly different from the nation (Grigg et al., 2003).

Texas implemented systematic education reform with the introduction of content standards in English language arts, mathematics, science, heath education, physical education, and fine arts in 1997. Standards in languages other than English followed in 1998, and social studies followed in 2000. In 2002 the Texas Assessment of Knowledge and Skills (TAKS) was introduced, with annual testing in reading taking place in grades 3 through 9. In addition, writing tests are administered to fourth and seventh graders, and an English language arts assessment is given to students in grades 10 and 11.

Mixed results surface regarding the quality of Texas's reading content standards and assessment. In *Quality Counts 2002: The State of the States* (Meyer et al., 2002), Texas was placed in the middle of the pack and received a grade of B- for the implementation of general academic standards. Peterson and Hess (2006) awarded Texas state proficiency standards a strength grade of F in 2003. With respect to NAEP alignment, Texas received a NAEP equivalent score of 190 (NCES, 2007), well below the NAEP cut-off scores for *Basic* (208) and *Proficient* (238). This equivalent score indicates misalignment between the NAEP and the TAKS, and suggested that Texas' content standards may be less rigorous than those associated with the NAEP.

Discussion

Careful inspection of the moderator variable profiles and additional data sources revealed few clear similarities across all three states, although New York and Texas were generally more similar than Maryland. While some patterns can be teased from the data, the emergent and overriding theme is that the multiplicity and complexity of factors make

it difficult to link parcel score performance patterns to state ecological or policy variables. In addition, the likelihood that some salient variables have not been included in this analysis is great, and the influence of these unknown factors cannot be accounted for.

With regard to demographics and SES, New York and Texas were more similar to each other than Maryland. Both New York and Texas were highly populous, diverse states, and ranked among highest third in the nation with regard to poverty. Maryland's profile was different. Across all three states, only rates of ethnic diversity were similar; however, patterns of English language usage and the proportion of foreign born residents in Maryland indicate the possibility that minority populations in Maryland may be different from New York and Texas.

Table 4.22
Demographics, SES and English Proficiency: Maryland, New York and Texas

	Maryland		New York		Texas	
Moderator Variable	Estimate	Rank	Estimate	Rank	Estimate	Rank
POPULATION	5,296,486	19	18,976,457	3	20,851,820	2
POVERTY	8.8%	46	14.2%	16	16.6%	8
WHITE	64.2%	47	68.8%	43	73.0%	36
FOREIGN	10.8%	14	20.9%	2	15.2%	7
NENG	13.2%	17	27.4%	4	31.5%	3

Note. Shaded areas indicate similarity.

Texas and New York were also similar in the general structure of their state reading proficiency standards. Both New York and Texas implemented content standards in the 1990s (1996 and 1997 respectively), while Maryland did not implement content standards until 2001. In addition, the content standards adopted by New York and Texas

address a much broader body of academic subjects. Regarding state assessments in reading, all three states currently test reading annually in at least grades three through eight, although New York just recently started in 2008. Between the years 2002 and 2007, the New York state assessment of reading was only administered to fourth and eighth graders. Evaluations of the quality, stringency and alignment of state content standards varied.

Table 4.23
Content Standards and Assessment: Maryland, New York and Texas

	Maryland	New York	Texas
Content standards			
Year reading standards adopted	2001	1996	1997
Number of subject areas	4	8	8
State assessment in reading			
Fourth grade test 2002	Yes	Yes	Yes
No. elementary grades tested 2002	6	2	6
Content standards quality & alignment			
QUALITY	A	A	B-
EQUIVALENT	187	207	190
ALIGNEMENT	C+	C	F
Average 2002 NAEP score	217	222	217

Note. Shaded areas indicate similarity.

Regarding proficiency as measured by the NAEP, Texas and Maryland were identical. The national average scale score on the 2002 NAEP reading assessment of fourth graders was 217.5. Average scores for Maryland and Texas were 217. New York

differed from Maryland and Texas in both average NAEP scores (222), and the proportion of students above *Proficient*, which was significantly higher than the nation.

CHAPTER V. SUMMARY AND CONCLUSIONS

This study was guided by two broad research questions:

(a) What kinds of profiles of reading achievement can be detected across states?

(b) What kinds of state-level contextual factors can be identified that are associated with those profiles?

A series of multilevel models that extended procedures traditionally employed in the analysis of differential item functioning (DIF) were applied to the 2002 NAEP fourth grade reading data. Variability in jurisdiction performance on individual items was estimated while controlling for overall state reading proficiency levels. Twenty target items with the largest IDV estimates were selected for a second tier of analyses, and an exploratory factor analysis was conducted on the twenty sets of estimated item BLUPs. Two factors were recovered that indicated shared variance among items beyond that accounted for by jurisdiction reading proficiency. Interpretation of the underlying factor structure was problematic. None of the item-based characteristics, such as item format or difficulty, were associated with the resultant factors; however, two passage-level characteristics, word count and NAEP reading context, suggest divergent explanations. Using passage length as the guiding explanatory variable, Factor 1 represented a Long Passage Factor, while Factor 2 represented a Short Passage Factor. Using NAEP reading context provides an alternative interpretation, with Factor 1 as a fiction factor and Factor 2 as a non-fiction factor. Ironically, passage word count and NAEP reading context were confounded for this assessment, and the factors could not be unequivocally defined. A subsequent analysis of testlet difficulty variation (TDV) confirmed the two-factor

structure of the original analysis, as well as the competing explanations tied to NAEP reading context and passage length.

Based on the obtained factor structure, two parcel scores were constructed, and values were estimated for each state across all jurisdictions. Parcel 1 represented a long passage or fiction parcel, while Parcel 2 represented a short passage or non-fiction parcel. The parcel scores were moderately negatively correlated (r = -0.43, $p<0.01$), indicating that jurisdictions that tended to score low on Parcel 1 scored comparatively high on Parcel 2, and vice versa.

Supposing Factor 1 is a fiction factor and Factor 2 is a non-fiction factor, the parcels may reflect differences in the focus of instruction across jurisdictions regarding how students are taught to interact with fiction and non-fiction texts. Those differences in focus may be the result of any number of effects, including state education policies. Alternatively, Parcel 1 can be defined as a long passage parcel and Parcel 2 as a short passage parcel. In this case, differential parcel performance may not be linked to systematic education reform, but to student characteristics such as English proficiency. For example, states with higher levels of non-English speakers tended to score lower on long passage items and higher on short passage items. It is possible that items associated with longer passages may result in underestimation of reading proficiency for those jurisdictions because non-native speakers reach a cognitive processing threshold beyond which their ability to apply their reading skills and strategies decays. Both NAEP reading context and passage length provide plausible explanations of the parcels; however, due to the confounding of passage length and NAEP reading context, it is difficult to assign labels to Parcel 1 and Parcel 2. The parcels may represent different reading contexts,

different passage lengths, a combination of both, or an alternative explanatory variable that was not considered in this analysis.

Following estimation of the parcels, profiles were created for each of the fifty participating jurisdictions and included data from over forty variables across nine classes of characteristics. A correlational analysis was conducted in order to backward map jurisdiction characteristics onto parcel score outcomes. Performance on both parcels was clearly associated with rates of non-native speakers, and similar patterns were seen across all variables associated with English language use. As rates of non-native speakers increased, scores on Parcel 1 decreased and scores on Parcel 2 increased. In other words, jurisdictions with higher levels of non-native speakers scored lower on longer, fiction passages and higher on shorter, non-fiction passages that would have been expected given overall jurisdiction reading proficiency. A similar, but less consistent finding was observed among some measures of jurisdiction wealth with poorer jurisdictions scoring lower on Parcel 1 and higher on Parcel 2 than would have been expected given overall reading proficiency.

A collective case study was then conducted on Maryland, New York and Texas. These three states were selected because of the pattern of their parcel scores: they were among the lowest extreme scores for Parcel 1 and the highest extreme scores for Parcel 2. Students in Maryland, New York and Texas scored lower on items from longer, fiction passages, and higher on items from shorter, non-fiction passages than expected.

Taken together, the mapping of moderator variables onto parcel score performance and the collective case study of Maryland, New York and Texas suggest the following major theme: larger, more populous states, with higher levels of poverty,

diversity and non-native speakers of English may exhibit a distinctive pattern in parcel score performance. The evidence suggests that states fitting this description score lower on Parcel 1 and higher on Parcel 2 than would be predicted by overall jurisdiction reading proficiency. In other words, these states score lower on parcel items associated with longer fiction passages, and higher on parcel items associated with shorter, non-fiction passages than expected.

Due to the multiplicity of potential effects, any number of different explanations of this phenomenon can be constructed. Assuming that the parcel scores represent NAEP reading context, one could speculate that larger, poorer, more diverse states are enacting similar policies that result in unexpectedly high levels of performance with regard to reading non-fiction texts. For example, states with high-stakes testing programs that are aligned with relatively weak content standards may over represent skills that are more closely associated with non-fiction passages. In response to this situation, it would not be unreasonable to expect that teachers might emphasize reading strategies best adapted for non-fiction texts. The parcels could reflect a subsequent focus of instruction regarding how students interact with the task of reading for information compared to reading for literary purpose. Passage length can provide an equally interesting, and speculative, explanation that rests with the proportion of non-native speakers of English across states. States with higher levels of non-English speakers tended to score lower on long passage items and higher on short passage items. An explanation for this pattern can be drawn from the field of Information Processing. It is possible that non-native speakers perform closer to their potential on short passage items, because the cognitive processing requirements associated with shorter passages are less taxing than those associated with

longer passages. In this case, longer passage items may result in underestimation of reading proficiency because non-native speakers reach a processing threshold beyond which their ability to apply their reading skills and strategies decays.

It is also possible that a better explanatory variable than either reading context or passage length may exist. For example, passage characteristics such as text concreteness or complexity may provide alternative explanations. In addition, explanatory variables that relate cultural characteristics to passages characteristics may be especially appealing given the association of the parcels with English proficiency. Future research could include an investigation of possible cultural bias regarding the content of fiction reading passages, or the possibility that non-fiction texts are more likely to contain Latin-based root words that could be more accessible to non-native speakers of English than words likely to be found in a non-fiction text.

Limitations and Implications for Future Research

There are several limitations to this study pertaining to the utility of the methodology for future research and the validity of the results. First, the statistical models are complex and require use of massive databases with thousands of observations. Due to the limited availability of such databases, use of these models is somewhat restricted. There may be interesting and important research questions that can be addressed by this methodology, but those questions can only be answered if a sufficiently large database containing the relevant information is available. Second, results based on these models may be less accessible to the typical consumer of education policy research due to the complexity of the statistical methods. In this case, effects may require longer explanation and the conclusions may not be straightforward. The duty of deftly

interpreting these results for lay people resides with the researcher and requires an uncommon combination of analytic and communication skills.

Regarding the validity of the results, several aspects of this study must be considered. Two sets of variables were constructed in order to facilitate interpretation of the parcels. The first set was a database of item characteristics that was used to interpret the two factors obtained in the factor analysis, and label the subsequent parcels. The second set was a collection of potential moderator variables that comprised the district profiles and was used to construct an interpretation of effects with regard to parcel score outcomes. These variables were drawn from multiple sources, and many were included based on previous research that suggested their potential utility. The process of identifying these variables was extensive, but not exhaustive. It is not only possible, but likely, that there exist other effects that were never included in this study. Omission of these variables may be the result of oversight; however, in many cases, potential variables were logistically difficult or even impossible to collect. For example, this study does not adequately address the potential influence of education policy effects. In order to thoroughly address policy effects, a complex set of variables would have to have been collected for each of the fifty jurisdictions. Compiling such a large and detailed collection of data would have been too onerous a task in addition to the other requirements of this dissertation, so a truncated set of policy variables was used instead.

The data used in this study were drawn from the 2002 NAEP reading test of fourth graders, and the results present a snapshot of reading performance. Whether the findings of this study generalize over time, or to older children, or to other subject areas, is unknowable without further research. An expansion of the current study to include

reading data from the 2003, 2005 and 2007 assessments of fourth graders would be a first step in examining whether the effects of this study are robust over time. In particular, it would be interesting to see if the same two parcels are recovered. Subsequent studies including older children or other subject area tests can similarly be constructed.

Conclusion

This study addressed a significant gap in our national discussion of standards-based education reform through an analysis of the relationship between state-level policy and ecological variables, and reading achievement. It also allowed for the estimation of state-level effects in a hierarchical context typical of educational settings. Unlike much previous research, the current study used *item-level* achievement test data (rather than total score or subscale results) in order to isolate and specify items that revealed meaningful between-state differences in reading achievement.

By focusing the analysis at the item level, this study produced fine-grained results that presented a more unique perspective than that allowed by typical analyses of total or subscale scores. This study also used parcel scores in an empirically driven characterization of jurisdiction-level performance that drew on the rich item-level data provided by the first set of analyses. Finally, while the methodology used in this study was primarily quantitative, it also employed qualitative research methods in an attempt to identify and explain differences in reading achievement. This combination of novel, sophisticated statistical techniques with traditional qualitative research methods led to a rich, detailed and compelling description.

APPENDIX A. Model 1 Target Items: IDVs

Item	IDV (SE)	Parcel
D09	0.03337 (0.007233)	1
D10	0.02576 (0.006055)	.
D21	0.02919 (0.006316)	1
D23	0.02820 (0.006343)	.
D26	0.02611 (0.006415)	2
D33	0.02598 (0.005740)	.
D49	0.02781 (0.006386)	.
D50	0.02975 (0.006441)	1
D51	0.02984 (0.006994)	.
D52	0.03006 (0.006852)	2
D53	0.02855 (0.006474)	2
D55	0.02707 (0.006043)	2
D63	0.03050 (0.007180)	2
D67	0.03470 (0.007610)	1
D68	0.02535 (0.005796)	1
D69	0.03851 (0.008522)	1
D70	0.03670 (0.008078)	1
D71	0.08408 (0.018210)	2
D72	0.02894 (0.006997)	.
D74	0.02916 (0.006398)	2

APPENDIX B. Model 1 Target Items: Descriptive Statistics

Item	Testlet	Format	Context	Word Count	IRT Parameters		
					A	B	C
D09	Beetle	OS	Literary	840	0.6502	-2.1783	0.0000
D10	Box in Barn	MC	Literary	1029	1.6837	-1.0137	0.2644
D21	Box in Barn	OS	Literary	1029	0.7056	-0.7139	0.0000
D23	Box in Barn	MC	Literary	1029	1.4837	1.0802	0.1744
D26	Money Makes	MC	Literary	1366	0.9270	0.8985	0.2889
D33	Goodall	OS	Information	993	0.7498	-1.0246	0.0000
D49	Ellis Island	OE	Information	1011	0.5881	1.5700	
D50	Ellis Island	MC	Information	1011	0.6109	-0.0162	0.3597
D51	Ellis Island	OS	Information	1011	0.9023	0.7299	0.0000
D52	Space Pioneer	OS	Information	779	0.4313	-1.4416	
D53	Space Pioneer	MC	Information	779	0.7182	0.1769	0.3397
D55	Space pioneer	OS	Information	779	0.3666	1.0947	
D63	River	OS	Literary	1184	0.4588	-1.0762	0.0000
D67	River	OS	Literary	1184	0.6813	0.2611	
D68	River	OS	Literary	1184	0.5854	-0.0114	
D69	River	MC	Literary	1184	0.6097	-1.4604	0.2892
D70	River	OS	Literary	1184	0.7879	1.3004	0.0000
D71	Wombats	MC	Information	684	0.5973	3.0252	0.3122
D72	Wombats	MC	Information	684	0.9302	-2.1458	0.2138
D74	Wombats	OS	Information	684	0.6815	-0.4531	0.0000

Note. MC = multiple choice; OS = short open-ended response; OE = extended open-ended response.

APPENDIX C. Jurisdiction Parcel Scores

	Proficiency (SE)		Parcel 1 Score (SE)		Parcel 2 Score (SE)	
ALABAMA	-0.1535	(0.04155)	0.1112	(0.04818)	0.0326	(0.04072)
ARIZONA	-0.2928	(0.04104)	0.0090	(0.04442)	0.0314	(0.03813)
ARKANSAS	-0.0643	(0.04341)	0.0965	(0.05820)	0.0344	(0.04905)
CALIFORNIA	-0.3149	(0.03895)	-0.0561	(0.02640)	0.1179	(0.02187)
CONNECTICUT	0.3694	(0.04255)	-0.0001	(0.05487)	-0.0690	(0.04502)
DELAWARE	0.1068	(0.05450)	-0.1121	(0.08903)	0.0151	(0.07317)
DIST. COLUMBIA	-0.6924	(0.06060)	-0.0385	(0.09727)	0.1077	(0.07952)
DODD/DOMESTIC	0.1948	(0.07389)	-0.0174	(0.10960)	-0.0065	(0.08685)
DODD/OVERSEAS	0.1810	(0.05859)	-0.0305	(0.09566)	0.0056	(0.07744)
FLORIDA	-0.0587	(0.03955)	-0.1455	(0.03285)	0.0923	(0.02771)
GEORGIA	-0.0097	(0.04013)	0.0164	(0.03820)	0.0073	(0.03225)
GUAM	-0.7299	(0.07817)	0.0146	(0.11070)	0.0371	(0.08878)
HAWAII	-0.2726	(0.04903)	-0.0821	(0.07667)	0.0596	(0.06441)
IDAHO	0.1186	(0.04690)	0.0144	(0.07223)	-0.1146	(0.05954)
ILLINOIS	0.0107	(0.03969)	0.0534	(0.03437)	-0.0209	(0.02878)
INDIANA	0.1892	(0.04088)	0.1482	(0.04475)	-0.0802	(0.03705)
IOWA	0.2449	(0.04344)	0.1720	(0.06006)	-0.0687	(0.04881)
KANSAS	0.1499	(0.04329)	0.0379	(0.05787)	-0.1097	(0.04834)
KENTUCKY	0.0662	(0.04211)	0.0323	(0.05179)	-0.0061	(0.04352)
LOUISIANA	-0.3109	(0.04182)	0.0131	(0.04926)	0.2118	(0.04185)
MAINE	0.2065	(0.04829)	0.0157	(0.07611)	-0.0858	(0.06267)
MARYLAND	-0.0857	(0.04133)	-0.2454	(0.04646)	0.1343	(0.03974)
MASSACHUSETTS	0.3924	(0.04098)	-0.2309	(0.04443)	0.0124	(0.03767)
MICHIGAN	-0.0066	(0.04001)	-0.0802	(0.03687)	0.0107	(0.03128)

APPENDIX C. Jurisdiction Parcel Scores

	Proficiency (SE)		Parcel 1 Score (SE)		Parcel 2 Score (SE)	
MINNESOTA	0.2405	(0.04128)	0.0534	(0.04739)	-0.1038	(0.03901)
MISSISSIPPI	-0.3575	(0.04280)	0.1260	(0.05448)	0.1586	(0.04639)
MISSOURI	0.1353	(0.04117)	0.0491	(0.04626)	-0.0381	(0.03877)
MONTANA	0.1979	(0.05035)	0.0190	(0.08172)	-0.1335	(0.06668)
NEBRASKA	0.1658	(0.04642)	0.1478	(0.07129)	-0.0711	(0.05828)
NEVADA	-0.2341	(0.04467)	0.0064	(0.06269)	0.0401	(0.05368)
NEW MEXICO	-0.2361	(0.04535)	-0.0356	(0.06518)	0.0003	(0.05561)
NEW YORK	0.0704	(0.03944)	-0.2184	(0.03172)	0.1363	(0.02679)
NO. CAROLINA	0.1135	(0.04037)	-0.0029	(0.04033)	0.0327	(0.03383)
NO. DAKOTA	0.2382	(0.05585)	0.1461	(0.09336)	-0.0580	(0.07490)
OHIO	0.1349	(0.03999)	-0.0107	(0.03724)	-0.0087	(0.03120)
OKLAHOMA	-0.0309	(0.04214)	0.1281	(0.05186)	-0.0537	(0.04378)
OREGON	0.0325	(0.04258)	-0.0911	(0.05395)	-0.0630	(0.04571)
PENNSYLVANIA	0.1258	(0.03991)	0.0003	(0.03654)	0.0035	(0.03057)
RHODE ISLAND	0.0589	(0.05042)	-0.0644	(0.08099)	-0.0115	(0.06708)
SO. CAROLINA	-0.0868	(0.04217)	0.0175	(0.05128)	0.0858	(0.04381)
TENNESSEE	0.0118	(0.04101)	0.2105	(0.04516)	-0.0281	(0.03786)
TEXAS	-0.1901	(0.03921)	-0.3035	(0.02936)	0.2055	(0.02468)
UTAH	0.1411	(0.04337)	0.0139	(0.05818)	-0.1559	(0.04890)
VERMONT	0.2551	(0.05662)	-0.0197	(0.09313)	-0.0853	(0.07552)
VIRGIN ISLANDS	-0.8954	(0.09396)	-0.0487	(0.11650)	0.0435	(0.09243)
VIRGINIA	0.2572	(0.04078)	0.0210	(0.04383)	-0.0729	(0.03634)
WASHINGTON	0.1480	(0.04089)	-0.1667	(0.04363)	-0.0839	(0.03715)
WEST VIRGINIA	0.1361	(0.04676)	0.1854	(0.07231)	-0.0417	(0.05928)
WISCONSIN	0.2199	(0.04153)	0.1087	(0.04926)	-0.0215	(0.04037)
WYOMING	0.1097	(0.05805)	0.0326	(0.09476)	-0.0243	(0.07732)

APPENDIX D. Testlet Difficulty Variation (TDV) and Characteristics

	Testlet	TDV	Reading Context	Word Count	Item	
					Numbers	Count
1	Beetle[a]	0.00005	Literary	840	D01-D09	9
2	Box Barn[b]	0.00115	Literary	1029	D10-D21	12
3	Money Makes[c]	0.00067	Literary	1366	D22-D32	11
4	Goodall[c]	0.00039	Information	993	D33-D41	9
5	Ellis Island[d]	0.00109	Information	1011	D42-D51	10
6	Space Pioneer[e]	0.00036	Information	779	D52-D61	10
7	River[f]	0.00268	Literary	1184	D62-D70	9
8	Wombats[g]	0.00028	Information	684	D71-D82	12

[a] How the Brazilian Beetles Got their Coats, retold by Elsie Eells. [b] The Box in the Barn, by Barbara Eckfeld Conner. [c] Details of Money Makes and Goodall have not yet been released to the public. [d] Ellis Island: Doorway to America, by Bill Walter. [e] Dr. Shannon Lucid: Space Pioneer, by Vicki Oransky Wittenstein. [f] The River, by Yetti Frenkel. [g] Watch Out for Wombats!, by Caroline Arnold.

APPENDIX E. Jurisdiction Testlet BLUPs: Testlets 1-3

	T1: Beetle		T2: Box in Barn		T3: Money Makes	
	BLUP	SE	BLUP	SE	BLUP	SE
ALABAMA	0.0001	0.01407	0.0104	0.03391	0.0572	0.03142
ARIZONA	0.0001	0.01396	-0.0011	0.03223	0.0599	0.03046
ARKANSAS	-0.0034	0.01425	0.0254	0.03798	0.0030	0.03441
CALIFORNIA	0.0062	0.01263	-0.0584	0.02052	0.0795	0.01992
CONNECTICUT	0.0003	0.01418	0.0176	0.03680	-0.0398	0.03296
DELAWARE	0.0008	0.01450	0.0072	0.04592	-0.0065	0.03957
DIST. COLUMBIA	0.0005	0.01454	-0.0132	0.04709	0.0094	0.04050
DODD/DOMESTIC	-0.0002	0.01457	0.0058	0.04864	0.0008	0.04114
DODD/OVERSEAS	-0.0004	0.01453	-0.0036	0.04690	0.0016	0.04013
FLORIDA	0.0099	0.01337	-0.0908	0.02548	0.0155	0.02455
GEORGIA	0.0037	0.01369	-0.0509	0.02875	-0.0134	0.02722
GUAM	-0.0003	0.01457	-0.0090	0.04878	0.0036	0.04143
HAWAII	0.0021	0.01445	-0.0095	0.04351	-0.0035	0.03824
IDAHO	0.0013	0.01440	0.0317	0.04233	-0.0099	0.03710
ILLINOIS	-0.0054	0.01346	0.0264	0.02632	0.0197	0.02523
INDIANA	-0.0027	0.01391	0.0596	0.03214	-0.0051	0.02962
IOWA	-0.0036	0.01426	0.0366	0.03849	-0.0192	0.03415
KANSAS	0.0011	0.01425	0.0188	0.03798	-0.0041	0.03421
KENTUCKY	0.0037	0.01413	-0.0119	0.03558	-0.0063	0.03244
LOUISIANA	-0.0001	0.01410	-0.0157	0.03453	0.0043	0.03219
MAINE	-0.0003	0.01443	0.0127	0.04332	-0.0152	0.03775
MARYLAND	-0.0027	0.01402	-0.0421	0.03327	0.0158	0.03105
MASSACHUSETTS	-0.0037	0.01394	-0.0190	0.03260	-0.0092	0.02975

APPENDIX E. Jurisdiction Testlet BLUPs: Testlets 1-3

	T1: Beetle		T2: Box in Barn		T3: Money Makes	
	BLUP	SE	BLUP	SE	BLUP	SE
MICHIGAN	-0.0063	0.01364	0.0010	0.02814	0.0305	0.02688
MINNESOTA	0.0019	0.01401	-0.0464	0.03338	-0.0010	0.03065
MISSISSIPPI	0.0019	0.01421	0.0083	0.03664	0.0100	0.03386
MISSOURI	-0.0002	0.01400	0.0013	0.03277	-0.0242	0.03057
MONTANA	-0.0004	0.01446	0.0109	0.04444	-0.0036	0.03855
NEBRASKA	-0.0052	0.01439	-0.0198	0.04167	-0.0051	0.03681
NEVADA	0.0006	0.01433	-0.0172	0.03967	0.0229	0.03567
NEW MEXICO	-0.0014	0.01436	0.0260	0.04061	0.0187	0.03636
NEW YORK	0.0109	0.01324	-0.0743	0.02471	0.0330	0.02373
NO. CAROLINA	0.0096	0.01378	0.0493	0.03011	-0.0622	0.02803
NO. DAKOTA	-0.0005	0.01452	0.0115	0.04628	-0.0035	0.03970
OHIO	-0.0011	0.01363	0.0043	0.02833	-0.0515	0.02649
OKLAHOMA	0.0009	0.01414	0.0312	0.03547	-0.0110	0.03262
OREGON	0.0023	0.01419	0.0314	0.03644	0.0032	0.03321
PENNSYLVANIA	0.0030	0.01357	-0.0707	0.02785	-0.0089	0.02630
RHODE ISLAND	0.0011	0.01447	0.0123	0.04436	-0.0008	0.03858
SO. CAROLINA	-0.0017	0.01413	-0.0375	0.03557	0.0283	0.03274
TENNESSEE	-0.0039	0.01396	0.0893	0.03249	-0.0016	0.03023
TEXAS	-0.0037	0.01301	0.0238	0.02298	-0.0236	0.02237
UTAH	0.0054	0.01425	0.0379	0.03814	-0.0027	0.03422
VERMONT	-0.0003	0.01452	-0.0020	0.04655	0.0056	0.03988
VIRGIN ISLANDS	0.0000	0.01458	-0.0113	0.04932	0.0022	0.04173
VIRGINIA	-0.0060	0.01390	-0.0064	0.03190	-0.0358	0.02930
WASHINGTON	-0.0031	0.01394	-0.0086	0.03179	-0.0074	0.02969
WEST VIRGINIA	-0.0025	0.01440	0.0004	0.04199	0.0046	0.03699
WISCONSIN	-0.0072	0.01406	0.0213	0.03412	-0.0535	0.03155
WYOMING	-0.0005	0.01453	0.0071	0.04675	-0.0008	0.04011

APPENDIX E. Jurisdiction Testlet BLUPs: Testlets 4-6

	T4: Goodall		T5: Ellis Island		T6: Space Pioneer	
	BLUP	SE	BLUP	SE	BLUP	SE
ALABAMA	-0.0290	0.02986	-0.0315	0.03667	0.0076	0.02811
ARIZONA	0.0110	0.02888	-0.0269	0.03534	-0.0025	0.02730
ARKANSAS	-0.0058	0.03193	0.0347	0.04048	-0.0103	0.03015
CALIFORNIA	-0.0320	0.02014	0.0008	0.02240	0.0228	0.01910
CONNECTICUT	0.0234	0.03137	0.0497	0.03833	-0.0267	0.02928
DELAWARE	0.0042	0.03527	0.0015	0.04747	0.0010	0.03331
DIST. COLUMBIA	-0.0080	0.03564	0.0021	0.04892	0.0020	0.03379
DODD/DOMESTIC	0.0033	0.03616	0.0012	0.04984	0.0004	0.03422
DODD/OVERSEAS	-0.0002	0.03558	0.0086	0.04834	0.0009	0.03363
FLORIDA	0.0189	0.02438	-0.0630	0.02776	0.0119	0.02288
GEORGIA	-0.0053	0.02668	-0.0464	0.03133	-0.0068	0.02510
GUAM	-0.0021	0.03619	-0.0067	0.05019	-0.0022	0.03431
HAWAII	-0.0063	0.03435	-0.0292	0.04580	-0.0024	0.03248
IDAHO	0.0002	0.03385	-0.0211	0.04412	0.0073	0.03186
ILLINOIS	-0.0480	0.02499	0.0919	0.02884	0.0256	0.02345
INDIANA	0.0004	0.02880	0.0061	0.03413	-0.0081	0.02698
IOWA	-0.0180	0.03209	0.0450	0.04005	-0.0060	0.03012
KANSAS	0.0078	0.03200	0.0517	0.03993	-0.0325	0.03001
KENTUCKY	-0.0018	0.03073	-0.0079	0.03785	0.0050	0.02890
LOUISIANA	-0.0259	0.03020	0.0259	0.03734	0.0284	0.02846
MAINE	0.0083	0.03432	-0.0098	0.04494	-0.0047	0.03228
MARYLAND	0.0310	0.02954	0.0302	0.03584	-0.0071	0.02784
MASSACHUSETTS	0.0242	0.02911	0.0014	0.03411	0.0500	0.02724

APPENDIX E. Jurisdiction Testlet BLUPs: Testlets 4-6

	T4: Goodall		T5: Ellis Island		T6: Space Pioneer	
	BLUP	SE	BLUP	SE	BLUP	SE
MICHIGAN	0.0122	0.02624	0.0320	0.03047	0.0052	0.02466
MINNESOTA	-0.0064	0.02962	0.0680	0.03517	-0.0397	0.02758
MISSISSIPPI	-0.0252	0.03134	-0.0068	0.03975	-0.0046	0.02965
MISSOURI	0.0239	0.02927	-0.0069	0.03540	0.0078	0.02758
MONTANA	-0.0021	0.03476	-0.0012	0.04603	-0.0005	0.03273
NEBRASKA	0.0209	0.03370	0.0168	0.04378	-0.0038	0.03171
NEVADA	-0.0084	0.03278	-0.0137	0.04243	-0.0019	0.03104
NEW MEXICO	-0.0122	0.03309	0.0296	0.04311	-0.0081	0.03133
NEW YORK	0.0130	0.02385	-0.0525	0.02686	-0.0312	0.02233
NO. CAROLINA	-0.0105	0.02754	-0.0534	0.03225	0.0061	0.02577
NO. DAKOTA	-0.0008	0.03536	-0.0057	0.04775	-0.0032	0.03346
OHIO	0.0109	0.02642	0.0281	0.02995	-0.0428	0.02463
OKLAHOMA	-0.0313	0.03074	-0.0131	0.03804	0.0075	0.02893
OREGON	0.0156	0.03124	-0.0423	0.03903	-0.0126	0.02940
PENNSYLVANIA	-0.0345	0.02583	0.0316	0.02968	0.0237	0.02435
RHODE ISLAND	-0.0068	0.03467	0.0073	0.04609	0.0028	0.03274
SO. CAROLINA	0.0173	0.03068	-0.0365	0.03837	0.0063	0.02898
TENNESSEE	-0.0082	0.02892	0.0463	0.03489	-0.0237	0.02724
TEXAS	-0.0167	0.02220	-0.0261	0.02507	0.0607	0.02111
UTAH	-0.0025	0.03199	-0.0368	0.04014	-0.0208	0.03009
VERMONT	0.0056	0.03546	-0.0088	0.04784	-0.0046	0.03351
VIRGIN ISLANDS	-0.0047	0.03638	-0.0024	0.05063	0.0008	0.03448
VIRGINIA	0.0509	0.02860	0.0024	0.03375	-0.0031	0.02679
WASHINGTON	0.0518	0.02886	-0.0573	0.03431	0.0016	0.02703
WEST VIRGINIA	-0.0015	0.03382	0.0093	0.04392	0.0021	0.03185
WISCONSIN	-0.0014	0.02991	-0.0034	0.03626	0.0224	0.02803
WYOMING	0.0008	0.03554	-0.0128	0.04821	0.0003	0.03364

APPENDIX E. Jurisdiction Testlet BLUPs: Testlets 7-8

	T7: River		T8: Wombats	
	BLUP	SE	BLUP	SE
ALABAMA	0.0118	0.04226	-0.0141	0.02645
ARIZONA	-0.0194	0.03969	-0.0307	0.02563
ARKANSAS	0.0279	0.04855	-0.0028	0.02818
CALIFORNIA	-0.0452	0.02399	-0.0426	0.01835
CONNECTICUT	-0.0097	0.04658	0.0072	0.02770
DELAWARE	-0.0298	0.06197	-0.0001	0.03087
DIST.COLUMBIA	-0.0227	0.06423	-0.0040	0.03119
DODD/DOMESTIC	-0.0122	0.06727	0.0003	0.03162
DODD/OVERSEAS	-0.0022	0.06382	0.0017	0.03117
FLORIDA	-0.0967	0.03033	-0.0002	0.02189
GEORGIA	0.0053	0.03488	0.0378	0.02389
GUAM	0.0029	0.06747	-0.0020	0.03164
HAWAII	-0.0086	0.05735	0.0119	0.03012
IDAHO	0.0101	0.05570	-0.0122	0.02968
ILLINOIS	0.0104	0.03142	-0.0192	0.02238
INDIANA	0.0575	0.03963	-0.0133	0.02556
IOWA	0.0957	0.04937	-0.0011	0.02833
KANSAS	0.0553	0.04829	-0.0184	0.02810
KENTUCKY	-0.0230	0.04469	-0.0035	0.02713
LOUISIANA	0.0034	0.04317	-0.0151	0.02675
MAINE	0.0423	0.05722	0.0009	0.03007
MARYLAND	-0.0975	0.04142	0.0104	0.02614
MASSACHUSETTS	-0.0378	0.04023	-0.0173	0.02575

APPENDIX E. Jurisdiction Testlet BLUPs: Testlets 7-8

	T7: River		T8: Wombats	
	BLUP	SE	BLUP	SE
MICHIGAN	-0.0549	0.03374	-0.0030	0.02345
MINNESOTA	0.0470	0.04188	0.0161	0.02626
MISSISSIPPI	0.0872	0.04638	-0.0152	0.02762
MISSOURI	-0.0099	0.04095	-0.0036	0.02592
MONTANA	0.0394	0.05937	-0.0033	0.03035
NEBRASKA	0.0504	0.05494	0.0088	0.02960
NEVADA	0.0189	0.05091	-0.0030	0.02882
NEW MEXICO	-0.0073	0.05207	-0.0119	0.02905
NEW YORK	-0.0955	0.02939	0.0166	0.02135
NO. CAROLINA	-0.0375	0.03656	0.0031	0.02455
NO. DAKOTA	0.0540	0.06314	-0.0022	0.03100
OHIO	-0.0028	0.03416	0.0528	0.02364
OKLAHOMA	0.0873	0.04470	-0.0064	0.02716
OREGON	-0.0520	0.04626	0.0009	0.02759
PENNSYLVANIA	0.0426	0.03367	0.0065	0.02320
RHODE ISLAND	-0.0273	0.05925	-0.0034	0.03045
SO. CAROLINA	-0.0092	0.04422	0.0038	0.02715
TENNESSEE	0.0648	0.03969	-0.0213	0.02567
TEXAS	-0.1588	0.02699	0.0233	0.02008
UTAH	-0.0024	0.04850	-0.0037	0.02824
VERMONT	0.0142	0.06312	0.0029	0.03104
VIRGIN ISLANDS	-0.0110	0.06861	-0.0033	0.03180
VIRGINIA	-0.0243	0.03909	0.0239	0.02531
WASHINGTON	-0.0698	0.03926	0.0205	0.02556
WEST VIRGINIA	0.0846	0.05534	-0.0105	0.02970
WISCONSIN	0.0268	0.04296	0.0381	0.02645
WYOMING	0.0278	0.06353	0.0001	0.03113

APPENDIX F. Model 3 Results and Discussion

The third model in this study used both items and testlets in a two-level design, with testlet difficulty viewed as fixed and item difficulty viewed as random. Using this model, item difficulty variation was estimated for all items, given state-level proficiency and fixed testlet difficulty. The differentiating feature of this model is that it takes testlet difficulty into account, in addition to state-level proficiency, before assessing differential item functioning via estimation of item difficulty variation (IDV).

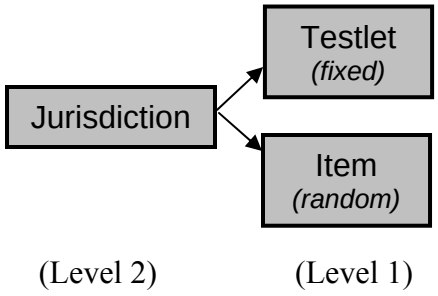


Figure F.1. Model 3.

For this model, success on a given item *i* was modeled as a function of proficiency, item difficulty (random) and testlet difficulty (fixed), with *i*=1…82 items, and *t*=1…8 testlets, administered across *j*=1…50, jurisdiction. The Level 1 model follows as,

$$f(n_{+ji} / n_{ji}) = \mu_j - \delta_{ji} - \lambda_t + \varepsilon_{ji}, \quad \text{(F.1)}$$

where n_{+ji} / n_{ji} equals the ratio of correct to total responses for item *i* in jurisdiction *j*; μ_j represents the overall reading proficiency for jurisdiction *j*; δ_{ji} represents the difficulty of item *i* for jurisdiction *j*; λ_t represents the difficulty of testlet *t*; ε_{ij} represents the item-level

error term specified as $\varepsilon_{ij} \sim N(0,1)$; and f(•) represents the logit link function. This model is a simple elaboration of Model 1 with the addition of fixed testlet difficulties.

For each of the 50 jurisdictions included in the Model 3 analysis, BLUPs were estimated for each of the 82 test items. In addition, IDVs were calculated for each test item, and compared to those obtained by Model 1.

Table F.1
Rank Order of Items by Magnitude of IDV from Models 1 and 3.

	Target Items 1-10					Target Items 11-20			
	Model 1		Model 3			Model 1		Model 3	
Rank	Item	IDV	Item	IDV	Rank	Item	IDV	Item	IDV
1	D71	0.0841	D71	0.0732	11	D74	0.0292	D65	0.0252
2	D69	0.0385	D74	0.0302	12	D72	0.0290	D67	0.0251
3	D70	0.0367	D52	0.0285	13	D53	0.0286	D63	0.0247
4	D67	0.0347	D33	0.0279	14	D23	0.0282	D49	0.0246
5	D09	0.0334	D23	0.0271	15	D49	0.0278	D10	0.0241
6	D63	0.0305	D55	0.0267	16	D55	0.0271	D25	0.0239
7	D52	0.0301	D69	0.0265	17	D26	0.0261	D50	0.0227
8	D51	0.0298	D72	0.0254	18	D33	0.0260	D21	0.0226
9	D50	0.0297	D09	0.0253	19	D10	0.0258	D03	0.0217
10	D21	0.0292	D53	0.0252	20	D68	0.0253	D39	0.0214

Note. Shading indicates items that rank among the highest 20 IDV estimates for both Models 1 and 3.

Model 3 IDVs were smaller than Model 1 IDVs for all but one test item (D01). With the introduction of fixed Testlet difficulty as a control variable, item IDVs shrank an average of 19%. The pattern of IDVs, however, was similar for both models, as indicated by the large correlation between Model 1 and Model 3 IDVs ($r = 0.96$). Comparison of the original 20 target items from Model 1 and the 20 items from Model 3

with the largest IDV, reveals an overlap of sixteen items. Although the introduction of fixed testlet difficulty reduced estimates of item difficulty variation, a similar set of target items was recovered by the third model.

REFERENCES

Abedi, J. (2004). *Inclusion of Students with Limited English Proficiency in NAEP: Classification and Measurement Issues.* Paper presented at the NAGB Conference on Increasing the Participation of SD and LEP Students in NAEP. Washington, DC. Retrieved 8/10/08 from http://www.nagb.org/pubs/conferences/abedi.pdf.

Amrein-Beardsley, A. & Berliner, D. C. (2007). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Educational Policy Analysis Archives, 11(25)*. Retrieved 6/19/07 from http://epaa.asu.edu/epaa/v11n25.

Baron, J. B. (1999). *Exploring high and improving reading achievement in Connecticut: Lessons from the states*. Washington, DC: National Education Goals Panel.

Berliner, D. C. & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud and the attack on America's public schools*. Reading, MA: Addison-Wesley.

Berliner, D. C. & Biddle, B. J. (1996). Making molehills out of molehills: Reply to Lawrence Stedman's review of "The manufactured crisis." *Education Policy Analysis Archives, 4(3)*. Retrieved 6/19/2007 from http://epaa.asu.edu/epaa/v4n3.html.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives, 12(1)*. Retrieved 6/19/07 from http://epaa.asu.edu/epaa/v12n1.

Bracey, G. W. (2000). *A review of "The state of state standards."* Center for Education Research, Analysis, and Innovation (CERAI), School of Education, University of Wisconsin-Milwaukee.

Brain, G. B. (1971). National assessment moves ahead. *Today's Education, 60(2)*, 45.

Brain, G. B. (1969). What's the score on NAEP? *Today's Education, 58(1)*, 18-21.

Camilli, G. & Monfils L. (2003). *Studying school effects with item difficulty variation.* Paper presented at the Annual Meeting of the American Educational Research Association: Chicago, IL.

Camilli G., Prowker, A., Dossey, J. A., Lindquist, M. M., Chiu, T., & Vargas, S. (2006). Summarizing item difficulty variation with parcel scores. In review.

Camilli, G., Prowker, A., Vargas, S. & Waszkielewicz, I., (2005). Studying state educational policies with NAEP mathematics items: Results from the 2000 fourth grade assessment. In review.

Camilli, G., Wolfe, P. M. & Smith, M. L. (2006). Meta-analysis and reading policy: Perspectives on teaching children to read. *The Elementary School Journal, 107(1)*, 27-36.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 state are above the national average.* Albuquerque, NM: Friends for Education.

Cannell. J. J. (1989). *How public educators cheat on standardized achievement tests: the "Lake Woebegone" report.* Albuquerque, NM: Friends for Education.

Cannell, J. J. (2006). "Lake Woebegone," twenty years later. *Third Education Group Review, 2(1)*, 1-14.

Cizek, G. J., Trent, E. R., Crandell, J., Hirsh, T. & Keene, J. (2000, April). *Research to inform policy: An investigation of pupil proficiency testing requirements and state*

education reform initiatives. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Clinton, W. J. (1997). *State of the Union address.* [Transcript]. Retrieved November 14, 2008 from www.usa-presidents.info/union/clinton-5.html.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33(4)*, 465-484.

Dutro, E. (2002). Do state content standards make a difference? An illustration of the difficulties of addressing that pressing question. *Mid-Western Educational Researcher, 15(4)*, 2-6.

Fielding, A., (2003). Ordered category responses and random effects in multilevel and other complex structures. In N. Duan & S. P. Reise (Eds.), *Multilevel modeling: methodological advances, issues and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fry, E., (1998). An open letter to United States President Clinton. *The Reading Teacher, 51(5)*, 366-370.

Goertz, M. E. (2001). Standards-based accountability: Horse trade or horse whip? In S. H. Furhman (Ed.), *From the capitol to the classroom: Standards-based reform in the states, Part II*. Chicago, IL: National Society for the Study of Education.

Goldstein, H. (2003). *Multilevel statistical models*. 3rd Ed. London & New York: Oxford University Press, Inc.

Gottlieb, S. S. (2001). *A review of state reading and language arts standards.* Bloomington, IN: ERIC Clearing House on Reading, English, and Communication. (ERIC Document Reproduction Service No. ED456425)

Grigg, W. S., Daane, M. C., Jin Y., & Campbell, R. J. (2003). *The nation's report card: Reading 2002* (NCES 2003-521). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Grissmer, D. & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas: Lessons from the states.* Retrieved August 18, 2008 from www.negp.gov/reports/grissmer.pdf.

Grissmer, D., Flanagan, A., Kawata, J. & Williamson, S. (2000). *Improving student achievement: What state NAEP scores tell us.* Santa Monica, CA & Arlington, VA: Rand.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H. & Braun, H. I. (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Jones, L. V. (1996). A history of the national assessment of educational progress and some questions about its future. *Educational Researcher, 25(4)*, 15-22.

Kamata, A. (1999a). Some generalizations of the Rasch model: An application of the hierarchical generalized linear model. *Dissertation Abstracts International. A (Humanities and Social Sciences), 60(30-A)*, 0715.

Kamata, A. (1999b, November). *Multilevel DIF analysis via hierarchical generalized liner modeling*. Paper presented at the Annual Meeting of the Florida Educational Research Association, Deerfield Beach, FL.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of*

Educational Measurement, 38(1), 79-93.
Kean, M. H. (2003). Educational assessment in a reform context. In Wall, J. E. & Waltz, G. R. (Eds.) *Measuring up: Assessment issues for teachers, counselors, and administrators*. Greensboro, NC: ERIC Counseling and Student Services Clearinghouse.
Kosar, K. R. (2005). *Failing grades: The federal politics of education standards*. Boulder, CO: Lynne Rienner Publishers.
Lee, J. (2006). Input-guarantee versus performance-guarantee approaches to school accountability: Cross-state comparisons of policies, resources, and outcomes. *Peabody Journal of Education, 81(4)*, 43-64.
Little, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.
Meyer, L., Orlofsky, G. F., Skinner, R. A. & Spicer, S. (2002). The state of the states [Special issue]. *Education Week: Quality Counts 2002, 21(17)*, 68-169.
Mislevy, R. J. (1996, January). *Revitalizing the NAEP design*. Paper presented at the meeting of the NAEP Design and Analysis Committee, San Diego, CA.
Monfils, L. F. (2004). Multilevel item analysis of a standards-based assessment: Using item difficulty variation to study school effects. *Dissertations Abstracts International, 65(06),* 3217B. (UMI No. AAT 3134864).
National Center for Education Statistics (2007). *Mapping 2005 state proficiency standards onto the NEAP scales*. Retrieved 5/15/2008 from http://nces.ed.gov/nationsreportcard/pubs/studies/2007482.asp.
Nichols, S. L., Glass, G. V. & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Educational Policy Analysis Archives, 14(1)*. Retrieved 6/19/2007 from http://epaa.asu.edu/epaa/v14n1/.
Olson, L. (2006). A decade of effort. *Education Week, 35(17)*, 8-16.
Otuya, E. & Krupka, S. (1999). *Federal and state strategies to support early reading achievement*. Washington, DC: Educational Testing Service, State and Federal Relations.
Penfield, R. D. & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of Statistics, Volume 26*: Psychometrics. New York: North Holland.
Peterson, P. E. & Hess F. M. (2006). Keeping an eye on state standards. *Education Next, 6(3)*. Retrieved 5/1/2007 from http://www.hoover.org/publications/ednext/3211601.html.
Porter, A. (1988). Indicators: Objective data or political tool. *Phi Delta Kappan, 69(7)*, 503-508.
Prowker, A. & Camilli G. (2006). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journal of Educational Measurement, 44(1)*, 69-87.
Rabb, T. K. (2004). "No child" left behind historical literacy. *Education Digest, 70(2)*, 18-21.
Recesso, A. M. (1999). First year implementation of the school to work opportunities act

policy: an effort at backward mapping. *Education Policy Analysis Archives, 7(11)*. Retrieved 9/29/2008 from http://epaa.asu.edu/epaa/v7n11.html.

Rogers, A. M. & Stoekel, J. J. (2004). *NAEP 2002 reading and writing assessments secondary-use data files data companion*.(NCES 2004-553/554). Washington, DC: U.S. Department of Education, Institute of Educations Sciences, National Center for Education Statistics.

Rogers, H. J., & Swaminathan, H. (2000, April). *Identification of factors that contribute to DIF: A hierarchical modeling approach*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Rogers, H. J., Swaminathan, H., & Egan, K. (1999, April). *A multi-level approach for investigating differential item functioning*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Schenck, E.A., Walker, D. R., Nagel, C. R. & Webb, L. C. (2005). *Analysis of state K-3 reading standards and assessments*. Washington, D.C U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, Education Publications Center.

Schmidt, W. H., McKnight, C., & Raizen, S. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht: Kluwer.

Spillane, J. P. (1998). State policy and the non-monolithic nature of the local school district: Organizational and professional considerations. *American Educational Research Journal, 35(1)*, 33-63.

Stotsky, S. (2005). *The state of state English standards*. Washington, DC: Thomas B. Fordham Foundation.

Stotsky, S. (2000). The state of literary study in national and state English language arts standards: Why it matters and what can be done about it. In, S. Stotsky (ed.) *What's at stake in the K-12 standards wars: A primer for educational policy makers*. New York: Peter Lang.

Swaminathan, H., & Rogers, H. J. (2000). *Identification of factors that contribute to differential item functioning*. Amherst, MA: Center for Educational Assessment, Research and Evaluation Methods Program, University of Massachusetts.

Valencia, S. W. & Wixson, K. K. (1999). *Policy-oriented research on literacy standards and assessment.* Ann Arbor, MI: Center for the Improvement of Early Reading Achievement (CIERA).

Vinovskis, M. A. (1998). *Overseeing the nation's report card: The creation and evolution of the national assessment governing board (NAGB)*. Washington, DC: National Assessment Governing Board.

Curriculum Vita

MICHELE YURECKO

Education

The Graduate School of Education – **Rutgers, the State University of New Jersey,**
Ph.D. – Educational Psychology, 2009
Ed.M. – Educational Statistics and Measurement, 1993

Georgetown University, College of Arts and Sciences
B.S. – Major: Mathematics, Minor: English Literature, 1990

Professional Experience

2008, 2006 **Drew University** – Department of Psychology
Adjunct Instructor of Educational Psychology

2002-2003 **National Institute of Early Education Research (NIEER), The Graduate School of Education – Rutgers, the State University of New Jersey**
Research Associate

2000-2001 **Center for Education Policy Research (CEPA), The Graduate School of Education – Rutgers, the State University of New Jersey**
Graduate Research Assistant

1999-2004 **The Graduate School of Education – Rutgers, the State University of New Jersey –** Department of Educational Psychology
Adjunct Professor of Educational Psychology, Classroom Assessment for Teachers, and Psychometric Theory I.

1999 **Kean University** – Department of Psychology
Adjunct Professor of Introduction to Statistics.

Hoechst Marion Roussel Pharmaceuticals Inc.
1998-1999 *Senior Statistician* – Global Pharmacoepidemiology
1996-1998 *Senior Statistician* – Biostatistics/Drug Development
1994-1996 *Associate Statistician* – Biostatistics/Drug Development
1993-1994 *Assistant Statistician* – Biostatistics/Drug Development

Merrill Lynch Financial Data Services
1991-1992 *CDSL Correction Specialist*
1990-1991 *Trade Corrections Representativ*

Research and Publications

Yurecko, M. (2008). Investigating the Relationship between Reading Achievement, and State-Level Ecological Variables and Educational Reform: A Hierarchical Analysis of Item Difficulty Variation. *Doctoral dissertation.*.

Yurecko, M. (2008). Basic Statistical Concepts: Enriching, Validating and Evaluating Previous Discovery. *Manuscript in progress.*

Yurecko, M. (2008). Understanding Basic Statistical Concepts: a Collective Case Study. *Manuscript in progress.*

Camilli, G., Vargas, S. & **Yurecko, M**. (2003). "Teaching Children to Read": The Fragile Link between Science and Federal Education Policy. *Educational Policy Analysis Archives,* 11(15).

Firestone, W. A., Camilli, G., **Yurecko, M.**, Monfils, L. & Mayrowetz, D. (2000). State Standards, Socio-Fiscal Context and Opportunity to Learn in New Jersey. *Educational Policy Analysis Archives*, 8(35).

Monfils, L., Camilli, G., Firestone, W. A., **Yurecko, M.** & Mayrowetz, D. (2000). Multidimensional Analysis of Scales Developed to Measure Standards-Based Instruction in Response to Systematic Reform. *Paper presented at the annual meeting of the American Educational Research Association.*

Yurecko, M. (2000). The Effects of Between-Class Ability Grouping on Self-Concept: a Meta-Analysis. *Poster presented at the annual meeting of the American Educational Research Association.*

Yurecko, M. (1999). The Effects of Between-Class Ability Grouping on Self-Concept: Preliminary Results of a Meta-Analysis. *Paper presented at the annual meeting of the North Eastern Educational Research Association.*