Examining NAEP: The Effect of Item Format on Struggling 4th Graders' Reading Comprehension

by

Vicki Benson Griffo

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Joint Doctor of Philosophy
with San Francisco State University
in
Special Education
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor P. David Pearson, Chair
Professor Karen Draney
Professor Art Shimamura
Professor Nicholas Certo

Spring 2011

Abstract

Examining NAEP: The Effect of Item Format on Struggling 4th Graders' Reading Comprehension

by

Vicki Benson Griffo

Joint Doctor of Philosophy in Special Education

with San Francisco State University

University of California, Berkeley

Professor P. David Pearson, Chair

Mixed item formats are in extensive use in large-scale assessment today and widely accepted as a means to improve assessment validity. Many studies have investigated the differential effect of *Item Format* on gender and ethnic subgroups in mathematics, yet few of these studies have attended to their impact on students with limited language proficiency and linguistic abilities in the area of reading. As educational policy increasingly mandates the inclusion of minorities such as English Language Learners (ELL) and students diagnosed with Specific Learning Disability (SLD) in federal and state assessment, many question the validity of achievement test scores because the degree to which the test score is a function of language proficiency is not clearly understood (Mahoney, 2008). To fill that research void, this study investigated *Item Format* effect and its interaction with (a) group membership as an ELL and/or SLD student and (b) assessment format differences related to genre and reading content. The software program ConQuest (Wu et al., 2007) was used to conduct Item Response Modeling on the 2007 NAEP 4th grade reading achievement data. Analysis showed an overall hierarchy of *Item Format* difficulty: MC<SCR<ECR items, and also demonstrated a significant effect of group membership on item proficiency. Yet, multidimensional regression analysis demonstrated no interaction between the two variables since the focal groups underperformed equally across all three *Item Formats*. Contrary to expectation, DIF analysis demonstrated that flagged CR items favored both ELL and SLD students, while flagged MC items generally favored only the Non-SLD group (but not the NE). To further probe into the nature of these differences, it would be necessary to access the full items. In looking to the future of assessment design, more research is needed to fully understand how *Item Format* differences contribute to assessment difficulty, the limited application of various *Item Formats*, specifically how they are suited to particular content, and how to fuse *Item Formats* in a manner that utilizes their unique benefits while also producing fair results (Hastedt & Sibberns, 2005).

Dedication

This dissertation is lovingly dedicated to my husband, Frankie, whose steadfast support, belief, patience, and love ushered me to this milestone. This one's for you, baby!

Table of Contents

List of Figures

List of Tables

Acknowledgements

This dissertation is the product of many hands and minds who have selflessly lent their time, support, and expertise, coupled with an abundance of kindness and encouragement. The mentorship that I have received from UC Berkeley has been unparalleled in my life. It has been a privilege to have pursued this doctorate in Education, and an honor to have done so at UC Berkeley. I will forever fondly cherish the learning, the experiences, and above all, the friends that I have come to know and love –professors and students alike. This has been one incredible journey!

First and foremost, I owe an unbelievable amount of gratitude to Professor David Pearson. Reflecting back on the ten years we have worked together, I am astounded at all that he has taught me, and the love, respect, and patience he has shown in the process, influences which have all funneled directly into this scholarly endeavor. I am thankful for the myriad opportunities he has given me to conduct a diversity of educational research. Even as a novice, he treated me as a colleague, listening to my opinions and putting trust in my judgments. As a mentor, he put my learning first and foremost. He guided me through project planning, conducting research, and academic writing, but always with a willingness to turn over the reins, giving me opportunity to take the lead in project management, and in the publication and presentation of our findings. He was always incredibly selfless, readily carving out time in a compacted schedule of administration, travel, teaching, and advising. He has a keen understanding of the financial burden of graduate school and was always willing to find a well-suited role on a project and to dig into his coffers in time of need. It is from one of our research projects that the idea for this dissertation was born. And, it is through his guidance, knowledge and insight that I was able to complete it. It has been an incredible privilege to have studied under Professor Pearson and never an individual whom I have respected more! His mentorship will no doubt remain unsurpassed.

To Professor Karen Draney, whose mentorship helped me untangle many of the challenges in working with a large data set and in running the specialized software necessary to analyze it. She was always willing to talk though complicated ideas; it was not unusual for me to leave her office hours with some form of a scribbled didactic in hand. She challenged my data analysis and interpretation, and demonstrated a knack for making complex content accessible. In submitting this dissertation, I am proud of the final product, and have her in large part to thank for that. Her unwavering patience, support, and mentorship have made this dissertation not only possible, but both interesting and enjoyable as well. But most importantly, my exchanges with Professor Draney have made a large contribution to my education at Berkeley and to my knowledge of measurement and quantitative methodology -and for that I am forever grateful!

I would like to express deep appreciation for my two supporting dissertation committee members, Art Shimamura (UCB) and Nick Certo (SFSU), thank you for all the feedback, encouragement, and flexibility. It was been a pleasure working together. Professor Certo, special thanks to you for stepping in as a replacement on such short notice.

To my mother, Carmen, who single-handedly laid the foundation for me to launch into such a challenging endeavor. Her innumerable sacrifices and emphasis on education unknowingly

started me on this path many years ago. Always placing my needs ahead of her own, she clearly had one goal in mind: to ensure my health, happiness, and success.

And finally, to my most prized team member, the outstanding man I married, Frank, whose behind-the-scene support and sacrifice has made this journey possible. While the list is endless, above all, he has lent a ready and constant supply of love and support that has given me the courage and confidence to pursue even the most challenging ideas. He has shown an incredible willingness to frequently pick up the slack, to take the lead, and to make many sacrifices. His steady supply of patience has grounded me in the hardest of times. Fortunately, he has also demonstrated a unique gift in lifting my spirits along the way. Through these years he has been an amazing husband, and in the frequency of my absence, an even more incredible father! On behalf of our two children, Gianna (2.75 yrs.) and Torino (9 months), I thank him. Frankie, we love you and are so lucky to have you! Here's to the beginning of many more happy years to come!

Examining NAEP: The Effect of Item Format on Struggling 4th Graders' Reading Comprehension

Chapter 1: Introduction and Literature Review

The National Assessment of Educational Progress (NAEP) has been a national monitor of what American students in 4th, 8th, and 12th grade know and can do across a variety of subjects since 1969. Reported subject matter scores of NAEP are nationally disseminated to help the public, policy makers, and education professionals understand the strengths and weaknesses in student performance and to inform educational policy decisions (National Assessment Governing Board, 2007). Recently the stakes have been raised since NAEP has become a congressionally legislated piece of No Child Left Behind (P.L. 107-110, 2002) and the receipt of Title 1 funds contingent upon state participation (Stedman, 2009).

Reading Comprehension data collected every two years suggests that there is a subtle closing of the achievement gap since 1992. Most recent 2009 data indicates that while fourth-graders' reading comprehension scores remained unchanged from 2007, overall scores have slightly risen: the average reading score of 221 is up 2 points since 2005 and 4 points since 1992. While the growth in gain scores is encouraging, the national portrait of overall student reading comprehension ability is a grim one. Higher percentages of students are performing at or above the *Basic* and *Proficient* achievement levels than in previous years, yet significant numbers of students continue to perform at very low proficiency levels: one-third of fourth-graders performed below the *Basic* level and two-thirds performed below the *Proficient* level.

For several years, NCES has been concerned with the fact that so many fourth-graders are performing below *Basic* levels (NAEP Validity Panel, 2008). In response, NCES charged a NAEP Validity Panel to devise a series of "easy" booklets that might increase NAEP accessibility for low-achievers. Through the creation of a set of "easy books" the NAEP Panel has hoped to refine measurement at the lower ends of the scale by increasing the amount of information generated about low-achieving student ability and thereby increase validity as a whole.

The issue of low performance on the NAEP is an increasingly complex issue in light of the evolving profile of American students. For example, English language learners who now comprise a large proportion of the school population are growing at unprecedented rates (U.S. Census, 2010). In turn, federal regulators are requiring national, state, and district assessments to be reflective of this change in school populations by increasing representation of students with disabilities (SPED) and English language learners (ELL). Special attention to the academic progress of these populations is greatly warranted since a large number of SPED and ELL students demonstrate low academic performance and high drop-out rates (Pelligrino, Jones, & Mitchell, 1999).

Assessments intended to measure subject matter knowledge can often be inadvertently confounded by a participant's English language proficiency (National Research Council, 2002; Mahoney, 2008). Linguistic minorities have underperformed in comparison to their English-proficient counterparts (The National Council of Educational Statistics, 2007). Abedi (2002) found that the performance gap between native and non-native speakers was even greater in content areas that have a high language demand. As growing numbers of non-native English speakers participate in the NAEP assessment, it is essential to understand just how subject matter proficiency is influenced by student linguistic ability.

Assessments of Reading Comprehension are one such example of a content area that inevitably carries a heavy language load; participants must read first the selected text, then a set of subsequent questions and in the case of MC items, 4-5 answer options per question in order to select the best answer choice. But reading isn't the only language demand present in this content area. It has become increasingly commonplace in large-scale assessment design to couple the use of MC items with constructed-response (CR) in order to strengthen test validity. Constructed-response items carry the added language demand of producing a written response, which inadvertently places a high premium on written language ability, a skill that is clearly related but arguably outside of the perimeter of reading comprehension.

Given the ubiquitous nature of mixed response formats in assessment today, it is troublesome that little is understood about the nature of their differences (Rodriguez, 2003), the added informational value of constructing written responses (Campbell, 2005; Pearson & Garavaglia, 1997; Lukehele, Thissen & Wainer, 1994), and how differing response modes impact student performance (Barnett-Foster & Nagy, 1996). For example, little empirical evidence is available to examine the impact item format has on linguistic subpopulations of test takers, specifically for constructed-response items. Given the limited linguistic abilities of many ELL and SPED students (specifically SPED students who are diagnosed with a Specific Learning Disability), it is important to question whether linguistically demanding test items are biasing assessment performance for certain populations. And, it raises an important concern in the valid reporting of scores for students with limited linguistic proficiency (Mahoney, 2008), an important concern given that NAEP is a widely trusted vehicle for benchmarking and interpreting change in student performance and state assessments over time.

The purpose of this study is to fill that void by exploring the relative effect of item response format (i.e. MC versus CR) specifically with an eye focused on the degree to which the effect interacts with a) status as an ELL and/or a person with SLD and b) other assessment format differences (e.g. *Context* and *Aspect*). My research questions are as follows:

1. What are the overall mean student performance differences (i.e. significant main effects) between *Item Formats*?
2. Do items that appear in different formats (MC vs. CR) measure the same construct of reading comprehension? In other words, is it possible to be relatively better at one item format than the other?
3. Are there interaction effects between *Item Format* and sources of variation within:
    i. The participant population (e.g., ELL and SLD)? For example, do different language groups (ELL vs. Native English speakers) perform significantly differently on item formats (MC vs. CR) in comparison to their peers at similar ability levels?
    ii. The assessment format (*Context* and *Aspect*)? For example, is there a significant difference between *Aspect of reading* across the different item formats (MC vs. CR)? In other words, does the difficulty level of *Aspect of reading* vary depending on whether an item is MC or CR?

My hypothesis is that student performance on the 2007 4th grade NAEP Reading Comprehension assessment is affected by item design variations within the item pool. More specifically, I expect that CR items will have a detrimental effect on student performance for ELLs and students with SLD as compared to their equal ability peers because CR items rely on English language skill by requiring the production of a written response relative to MC items where respondents read and select a response option.

The focus of this dissertation also extends the work of the NAEP Validity Panel who in the interest of gathering more information about what low proficiency students are capable of doing, has launched an effort to increase measurement precision at the lower end of the performance continuum. Because there is such a great floor effect on information gathered from low ability students, NAEP results have yielded information about what students are not able to do, yet describe very little about what they can do in reading and mathematics. As a resolution, the Validity Panel has been engaged in the construction of "easy reading blocks" to include more items that are accessible to low-achieving students in order to increase the amount of information gathered. This dissertation complements these efforts of the Validity Panel to increase accessibility of the 4th grade Reading Comprehension assessment. Information regarding variables, such as item format, that may contribute to poor performance has important educational policy implications for diverse populations as it can aid in future development of more accessible, and hence more reliable and informative, assessments for students at all levels.

Inconclusive evidence for item format effect coupled with the high prevalence of constructed-response items on the NAEP reading assessment highlights the necessity -for validity's sake- of probing for the presence and nature of any performance discrepancies. Knowledge of performance differences between examinee subgroups can potentially lend invaluable aid to future assessment design, adaptation, and interpretation by identifying the factors that influence item statistics as well as by examining the value added by mixed item formats (Hambleton & Jirka, 2006; Pearson & Garavaglia, 1997; Bridgeman, 1992). By designing assessments that all examinees can participate in, we, as a research community, learn more about what it is that readers at all levels can and cannot do.

The NAEP dataset is an ideal object of inquiry since the assessment features a 50/50 mixed format of MC and CR items. The NAEP results to be analyzed are based on a nationally representative sample of fourth-grade students assessed with the 2007 NAEP Reading Comprehension Assessment (n=191,040 reported sample). My analysis will compare item format effect of the general population versus two populations of students who typically struggle with the linguistic demands of reading comprehension: ELLs (n=15,784) and SLDs (n=8,244). For my analyses, a Rasch simple logistic model (1980) will be used to analyze the dichotomous items and a partial credit model (Masters, 1982) used for the polytomous items. Using the computer program ConQuest (Wu, Adams & Wilson, 2007) I run a Rasch-type model to estimate individual item difficulties, run latent regression, as well as to examine individual items for statistically significant differential item functioning (DIF) across language groups, and to explore the presence of dimensionality across the three item formats. Item analysis will factor 100 items comprised of multiple-choice, short- and extended-constructed response formats (MC=57 items, SCR=11 items, ECR=32 items).

Item Response Models, such as the one-parameter Rasch model, are characterized by the ability to separate both item and person parameters. The benefit of such a model, as compared to more traditional factor analysis and mean comparisons, is that estimation of individual item difficulties can probe for differences in content tapped by item formats (i.e. do MC items typically tap lower cognitive abilities). This is essential as a potentially confounding variable in analyzing item formats is that constructed-response items are thought to be typically reserved for items demanding higher order cognition, thus linguistically demanding item types are commonly paired with higher level content (Garner & Engelhard, 1999). Item response methodology also has the ability to explore the presence of differential item functioning by comparing the probability of examinee subgroups correctly answering an item when matched on ability. This

level of analysis gets at the heart of the question of whether constructed-response items are consequently factoring in language ability when assessing reading comprehension.

Literature Review

Today, it has become increasingly commonplace in large-scale assessment design to use mixed response formats of MC and CR items (Hastedt, 2004; Sykes & Yen, 2000). Studies have shown that MC items are the optimal choice when efficiency is the goal because these items are highly efficient, economical to produce and to score, and in terms of test validity, produce highly reliable scores (U.S. General Accounting Office, 2003; Bennett & Ward, 1993). In a MC format, examinees are required to read a question stem and then select an answer from among three or more predetermined options. These items take much less time than CR items for the examinee to complete because they require only the selection of an answer rather than the construction of a written one. As a result, MC items potentially increase content-related validity because more items can be completed in a reasonable timeframe thus allowing the adequate sampling of a content domain. Additionally, MC are administratively more efficient since they can be scored via computer and result in little dispute about the correctness of the keyed answer (Downing, 2006). These characteristics make MC items very appealing in assessment design. The downside to these items is that students may get credit for answers they do not know by simply guessing (Hastedt, 2004) or by the process of elimination through working backwards from options to exclude incorrect answer choices (Bridgeman, 1992; Donoghue, 1994; Campbell, 1999).

Limitations of MC items and issues in validity brought about a coupling with CR items. In a CR format examinees are required to read a question and then generate an answer using several words, phrases, or sentences to explain or support their ideas with evidence from the text (NAGB, 2007).The production of a unique answer is perceived to be more effective tool for assessing deep understanding of content knowledge and higher-order thinking skills (Haladyna, 1997; Hollingworth, Beard, & Proctor, 2007; Manhart, 1996). Yet, these item types do not escape criticism either. These items take longer for examinees to complete. Secondly, assessments can only utilize a small quantity of CR items in the allotted testing time, thus reducing the amount of content covered. Additionally, these items are typically scored via a rubric, so they are vulnerable to the subjectivity and bias inherent in human judgment (Downing, 2006; Wainer & Thissen, 1993). All three of the elements make CR items impractical for use as the sole item format since they can lead to lower reliability. The greatest criticism of CR items, however, is that they introduce an element of construct-irrelevant variance (CIV). The added task demand of producing a written response places a premium on written verbal abilities within a construct purporting to assess a different (albeit related) ability such as reading comprehension (Haladyna, Downing, & Rodriguez, 2002). Despite the individual pros and cons of using MC versus CR formats, a widely held assumption in test design is that each makes a unique contribution to the assessment as a whole, while the combination of the two is a means to improve validity (Ercikan et al., 1998).

The issue of validity and reliability of item formats is very relevant to test construction and interpretation. The use of testing in education presupposes that an individual's reported score is an accurate reflection of content mastery (Martinello, 2008). ELLs provide a special challenge in this regard. Low test scores for ELLs may be more related to limited English proficiency than mastery of subject matter and could have a large impact on test validity (National Research Council, 2000, 2002).

Admittedly, issues such as construct-irrelevance variance in reading comprehension are complex. On one hand, writing is distinctly intertwined with the reading process and in the real

world people often write something as a way of sharing their understanding of a text. Yet the task demand of requiring a written answer in a reading comprehension framework certainly taps differing skills from merely reading and selecting a pre-manufactured response on a MC. As a result, CR items inherently take into account how well students write about texts thus confounding the measurement of a reading comprehension construct solely intended to gauge how well students read various texts and answer the questions about those texts (Downing, 2006; Zwick, Donoghue, & Grima, 1993). Inadvertently, the tapping of a skill outside of the intended construct poses an element of CIV (Messick, 1989) where factors not central to the construct are being assessed and accounted for in student proficiency.

Given the reliability and validity issues surrounding CR items, researchers have weighed the added information value of these item types. For one, CR items are perceived to be more effective tools for assessing deep understanding of content knowledge and higher-order thinking skills (Haladyna, 1997; Hollingworth et al., 2007; Manhart, 1996) and require the construction of new knowledge (Mazzeo & Yamamoto, 1993) whereas MC items are criticized for focusing solely on discrete skills or facts (Campbell, 1999), foster a correct-way mentality, and narrow the curriculum (Hambleton & Murphy, 1992). Additionally, CR items elicit the production of a unique answer that requires examinees to display and sometimes explain their thinking (Hollingworth et al., 2007; Nitko, 2004). As a result, these items are thought to more authentic assessments as they closely mirror classroom tasks (Manhart, 1996) thus increasing face validity of the assessment as a whole (i.e. do you appear to be measuring what you claim to measure). In a review of the literature, Hollingworth and colleagues (2007) note that general consensus in the educational psychology literature suggests that item format should be selected to reflect instructional intent. Downing (2006) agrees that the application of CR items should be reserved for when MC items cannot adequately measure the content skill area.

Item efficiency. Almost 100 years of research indicate that MC items are the best choice when efficiency is the goal. The ability of MC to be answered more quickly than CR items allows examinees to complete a higher quantity of items in an allotted time. Evidence of item efficiency is supported by studies such as Wainer and Thissen (1993) who compared MC and CR items on the College Board's Advanced Placement (CBAP) Chemistry Exam. Examining reliability using the Spearman-Brown formula, they determined that many CR items would be required to yield the same reliability as the MC section. More specifically, to equal the reliability of a 75-minute MC section, a CR section would require just over 3 hours of testing time, thus obviating the issue that consideration of time and expense would make such a test impractical. In support of Wainer & Thissen's findings, Donoghue (1994) found that MC items yielded approximately 1.33 times more information per minute than ECR items using an Item Response Theory (IRT) framework to analyze items from the 1991 field-tested NAEP reading assessment. These results are supported in a similar IRT study by Lukehele, Thissen, and Wainer (1994), which found sixteen MC items to be equivalent to 1 CR item and at a lower cost of time and resources when they analyzed DIF on the CBAP test of History and Chemistry. These studies lend support to the efficiency argument for MC items; in other words, the collective MC responses produce a more thorough and representative sampling of the cognitive domain being assessed and as result strengthen validity evidence by reducing the threat of construct underrepresentation (Messick, 1989).

Item format dimensionality. Cost-effectiveness and ease of scoring aside, researchers have examined the interdependency of item formats and whether the two are measuring the same construct. The most common approach to assessing dimensionality has been to apply exploratory

and confirmatory factor analyses (Ercikan et al., 1998). These studies raise the practical question of whether assessment scores should be calculated from a combination of the MC and CR item proficiencies, or whether -when these formats are found to belong to separate constructs- scores should be reported separately (Rodriguez, 2003). Five well known studies using factor analysis examined high-school, college and adult test performance on non-stem equivalent items in the areas of mathematics (Traub & Fisher, 1977), computer science (Bennett, Rock, and Wang, 1991), chemistry (Thissen, Wainer, & Wang, 1994), analytical reasoning (Bridgman & Rock, 1993), and reading comprehension (Ward, Dupree, & Carlson, 1987). They all concluded that the two item formats best fit a one-factor model suggesting that that the two items types belong to the same construct (Ercikan et al., 1998).

An early study by Traub and Fisher (1977), for example, found little format effect and weak evidence that CR verbal items measured a different construct when they used confirmatory factor analysis to examine different item formats in Mathematics. A later study by Bennett, Rock, and Wang (1991) found that both item formats measured the same characteristics, suggesting that the addition of CR items did not provide different information in the CBAP Computer Science examination. Thissen, Wainer, and Wang (1994) replicated the Traub and Fisher study examining sections of the CBAP Computer Science and Chemistry tests. They additionally observed a small amount of local dependence among the CR items that produced a small degree of multidimensionality. This pattern of results suggest that the CR items are measuring the same content as the MC items in addition to something unique, which perhaps could be attributed to format effects (Pearson & Garavaglia, 1997). Bridgman and Rock (1993) found converging evidence of a one-factor model when using data from the Analytical Reasoning scale of the Graduate Record Examination (GRE) General Test.

Thus, these studies show the data better fit a one-factor model and that the two item types were generally highly correlated. Practically speaking, several researchers note the advantages of combining MC and CR scores into a single content area, one of which is that when these item types are combined, they produce a total score that has higher reliability than separate scores because often there are too few CR items to produce consistent scores (Ercikan et al., 1998; Sykes & Yen, 2000).

While factor analysis is certainly one lens for examining item format differences, there have been many criticisms leveled about its limitations. Pearson and Garavaglia (1997) note three major issues with factor analysis: 1) interpretation may depend on the model specified for analysis, 2) item difficulty varies between MC and CR items, and 3) design flaws exist such a small number of CR items are analyzed. Another layer of complexity is that item difficulty may not only vary as a result of *Item Format,* but item difficulties are not necessarily equal across examinees either because there may be factors that make an item easier for one group of examinees and harder for another (Wei, 2008).

Information value. Given the greater efficiency of MC items and evidence suggesting unidimensionality of MC and CR formats, some researchers have questioned the added information value of CR items; in other words, what are the benefits of these items that ultimately are more time and resource consuming? These results are complex as some studies have focused on different aspects of item format differences such as: cognitive abilities tapped (e.g. Katz, Bennett, & Berger, 2000), added information value of CR items (e.g. Donoghue, 1994), and item format bias (e.g. Garner & Engelhard, 1999).

Few differences have emerged from studies examining whether there are different cognitive demands tapped by two formats (Martinez, 1999).Van den Bergh (1990), for example,

explored the possibility that different intellectual abilities were involved in answering MC and CR reading questions. Using LISREL to conduct a Structural Equation Model (SEM), he examined the relationship between Dutch third graders' reading comprehension scores on parallel MC & CR items and their score on test of semantic abilities based on Guilford's structure of the intellect (SI) model. He found that intellectual abilities explained 62% of the variance, but that it was not possible to demonstrate a substantial difference in intellectual abilities measured related to item type, thus suggesting that individuals construct answers similarly (although not identically) across item types. In examining math word problems on the SAT, Katz, Bennett, and Berger (2000) found that examinees used both traditional and nontraditional strategies (checking answer against item clues) to approach different item formats, although the respondents were just as likely to use nontraditional strategies equally across stem equivalent item formats.

Yet other studies examining added informational value suggest that CR items make a unique contribution to the assessment. Harking back to an early 1976 study, Samejima found when using a graded-response model that polytomously scored items (more than two score categories) yielded considerably more information than dichotomized items on an experimental mathematics exam. More evidence of the information value of polytomous items stems from Donoghue's more recent 1994 IRT study; he found that polytomously scored items yielded substantially more information than an equal number of dichotomous items. Examining fourth grade NAEP reading items, he reported that polytomous items yielded 2.33 to 3.66 times more information than the MC items. Short-constructed response items also yielded 1.66 to 2.33 times more information (although to a lesser degree than extended-constructed) compared to MC items. Even when the constructed response items were artificially dichotomized (i.e. scoring them as either right or wrong) in order to focus on item quality for purposes of analysis, these items still yielded more information than MC items. Another study using IRT by Ercikan et al. (1998) found that simultaneous calibration of MC and CR items did not lead to model fit problems, but did lead to loss of information on CR items on reading, language, mathematics, and science tests for 3rd, 5th, and 8th grades. These results suggest that CR items are assessing somewhat different skills, information that is lost when combining MC and CR scores together.

Item format bias toward subpopulations. Given the item format differences reviewed thus far, scholars have researched the question of whether particular item types exhibit bias for subgroups of test takers. The majority of these studies have examined gender bias. In a review of the literature, Traub and MacRury (1990) report that despite differences in test content of various studies, the performance of females relative to that of males was better on CR tests than on MC tests. Similar results emerged in Garner and Engelhard (1999) who found gender effects related to content and item format when using differential item functioning (DIF) to examine the mathematics portion of a high school exit exam. Overall, MC items favored men and CR favored women, who tended to offer most extensive explanations for their work. These studies unequivocally report the presence of performance differences across item types between gender subgroups. However, it is likely that other factors such as content, experience, and reading and writing ability also contribute to performance differences.

In a Beller and Gafni (2000) study exploring gender effect on the International Assessment of Educational Progress mathematics assessment for 4th and 8th graders, they too found that boys and girls respond differently to writing tasks, although they too assert that item format alone cannot explain the effect. Mazzeo et al. (1993) in a DIF study of four advanced placement (AP) examinations: American history, biology, chemistry, and English language and

composition found that the relatively better performance of females on CR tests might be related to the different construct measured by this item type. He suggests that CR tests probably require different sets of competencies than their MC counterparts, and gender-related differences in performance profiles across the two assessment formats most likely reflect disparities in the male and female proficiencies with regard to these different competencies.

This question of item bias is important in light of earlier discussion that CR items place a high demand on verbal abilities (when writing is not the construct focus), thus disadvantaging low ability examinees such as students with disabilities and English-language learners (ELLs; Haladyna & Downing, 2004). Hollingworth et al. (2007) cite an early study that showed MC items to favor ninth grade students who were highly test-wise, thus introducing another element of CIV because the variable of "test ability" was inadvertently tapped but not part of the construct purported to be measured.

Summary. In summary, both MC and CR item formats make a unique contribution to assessment construction. MC items demonstrate higher reliability and add to construct validity as their efficiency produces greater representation of the target construct (Wainer & Thissen, 1993). CR items, on the other hand, are thought to increase face validity (and perhaps constituent credibility) as they more closely match classroom tasks and demonstrate some added information value over CR items (Ercikan et al., 1998; Manhart, 1996). Factor analysis suggests that MC and CR item formats statistically tap the same construct (Pearson & Garavaglia, 1997). Yet, value added studies report the CR items contribute assessment information over and above MC items alone (Samejima, 1977; Donoghue, 1994; Ercikan et al., 1998). And, while studies of cognitive processing of item formats showed little difference in how individuals process information (e.g. Katz et al., 2000), studies of item bias suggest that item format differences influence subgroup test performance (e.g. Donoghue, 1994).

Chapter 2: Methodology

The purpose of this study is to explore the relative effect of item response format (i.e. multiple-choice versus constructed-response) specifically with an eye focused on differential item functioning (DIF) for ELL and SLD subgroups that may contribute to item bias, and, to examine to what degree the effect interacts with subject population (e.g. language status and reading disability) and/or assessment format differences (e.g. context and reading aspect)? This is especially important and relevant given the 2008 NAEP Validity Panel's goal to make the 4th grade Reading Comprehension assessment more accessible to the lower end of the performance continuum. Students performing at Below Basic level on the NAEP assessment are diagnostically problematic because there is such a great floor effect that the information gathered about students at this level describes much of what they cannot do, yet little about they are able to do. Put differently, there are so few NAEP items that measure the performance of low achievers reliably that NAEP yields little information about the performance of these students. Information regarding variables that contribute to poor performance can aid in the future development of more accessible, and hence more reliable and informative, assessments for students at all levels.

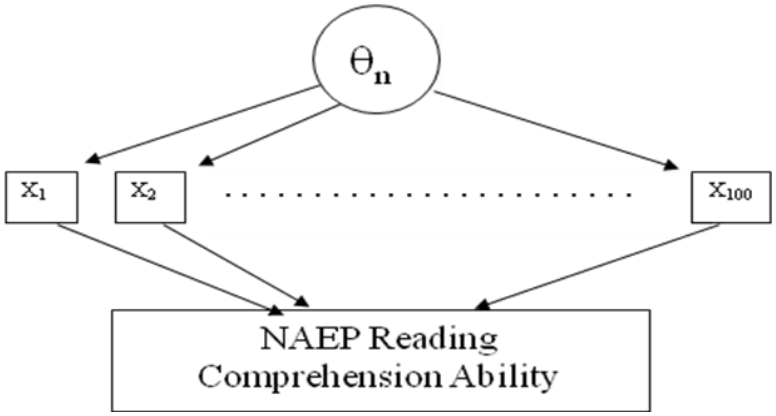
This study will use an exploratory approach to relate the NAEP reading item responses to both person predictors and item predictors. My hypothesis is that subgroup performance on the NAEP assessment is significantly affected by item format. More specifically, I predict that CR items will demonstrate greater difficulty for ELL and SLD students given the added task demand of producing a written English response relative to MC items which require the respondent to read and choose from preselected options. My research questions are as follows:

In general, does *Item Format* influence student performance on NAEP reading items?

1. What are the overall mean student performance differences (i.e. significant main effects) between *Item Formats*?
2. Do items that appear in different formats (MC vs. CR) measure the same construct of reading comprehension? In other words, is it possible to be relatively better at one item format than the other?
3. Are there interaction effects between *Item Format* and sources of variation within:
    i. The participant population (e.g., ELL and SLD)? For example, do different language groups (ELL vs. Native English speakers) perform significantly differently on item formats (MC vs. CR) in comparison to their peers at similar ability levels?
    ii. The assessment format (*Context* and *Aspect*)? For example, is there a significant difference between *Aspect of reading* across the different item formats (MC vs. CR)? In other words, does the difficulty level of *Aspect of reading* vary depending on whether an item is MC or CR?

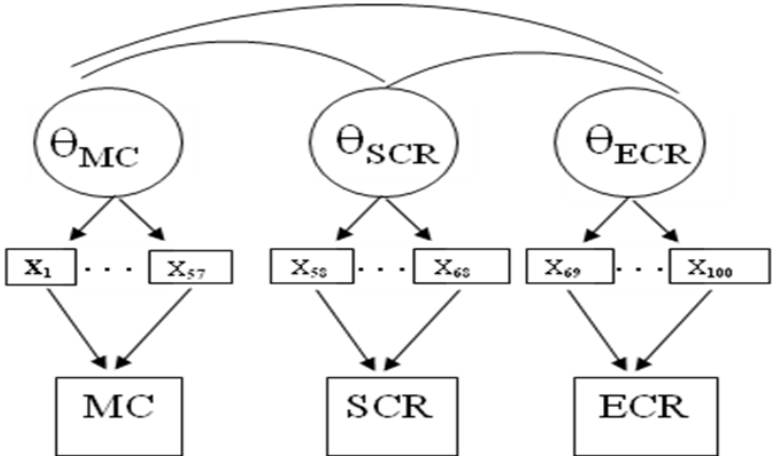
Figure 1. Conceptual Model. Illustrates the variables of the Unidimensional and Multidimensional model.

Unidimensional Approach



θ_n = *Single estimate of latent reading comprehension ability*

Multidimensional Approach



Student Characteristics

English Proficiency	Reading Proficiency
Native English Speakers	Specific Learning Disability (SLD)
English Lang. Learners	Non-SLD

Interaction Effects

Assessment Characteristics

Context	Aspect
Literary	General Understanding
Expository	Interpretation
	Making Connections
	Content & Structure

$\theta_{\text{item format}}$ = *Three independent estimates of latent reading comprehension ability*

Structure of the NAEP Reading Assessment

Participants. The NAEP results analyzed here are based on a nationally representative sample of fourth-grade students; the 2007 database reports reading comprehension proficiencies for 191,040 4th grade students. Of the original 204,394 students sampled for participation, 13,354 were "excluded" according to four strict NAEP criteria that deemed a participant incapable of completing the assessment. In brief, students with disabilities (SPED) and English language learners (ELLs) were given accommodations that matched, as closely as possible, their typical school testing situation, and they were evaluated according to the same criteria as students without accommodations.

The focus of this dissertation is on item format effects that may unfairly affect students who struggle with the print demands of the English language. Thus analysis will examine reported reading scores for NAEP subpopulations of students classified as ELLs (n=15,784) and students diagnosed with Specific Learning Disability (SLD, n=8,244). There is a small overlap in these populations of n=858 students who are classified as both SLD and ELL students.

This dissertation will consider only the SLD subgroup rather than the entire SPED population because these are the individuals classified with reading-related disabilities under the Individuals with Disabilities Education Improvement Act of 2004 (IDEA; P.L.108-466). IDEA defines SLD as: "a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which disorder may manifest itself in imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations".

Overall, there is a high incidence of SLD students as compared to the 12 other IDEA classifications; SLD constitutes ~50% of the population of students receiving special services (OSEP, 2003). On the NAEP assessment, students diagnosed with SLD comprise 44% of the entire SPED population, therefore, the 12 other SPED classifications are excluded from this analysis because these student profiles are not relevant to the analysis (e.g. hearing impairment, emotional disturbance, etc.).

NAEP samples students from a combination of public (n= 196,457) and non-public (n=3,481) schools in each U.S. state, the District of Columbia, Bureau of Indian Affairs (BIA; n=1117) and international Department of Defense (DOD; n=3339) schools. Approximately 30 students are selected from each school to complete the NAEP reading examination; they are randomly sampled with equal probability within schools from lists of enrolled students. Schools are sampled with the probability proportional to the size of enrollment (NAEP training seminar, 2009). NAEP uses a stratified multistage cluster sampling scheme in which students have differential probabilities of selection (the probability of selection is obtained by multiplying the probability of selection at each stage). As in most surveys, each respondent is assigned a sampling weight. Results are weighted to take into account the fact that states, and schools within states, represent different proportions of the overall national population. For example, results for students in less populous states are assigned smaller weights than the results for students in more populous states. (Campbell, 2001).

Instrument. The instrument analyzed was the 4th grade National Assessment of Educational Progress (NAEP) Reading Comprehension Assessment. Conceived as a project of the National Center for Educational Statistics (NCES), NAEP was developed and reviewed by a committee of reading and measurement experts based on the NAEP 2007 Reading Framework which describes the goals of the assessment and what kinds of exercises it should feature.

The Framework was developed through the twin tools of American Institutes for Research (AIR) and the National Assessment Governing Board (NAGB)- the policy-making body for NAEP- through a collaborative, comprehensive national effort involving a multitude of individuals such as testing and measurement experts, reading teachers, eminent reading scholars, curriculum specialists, local and state policymakers, and business and public representatives. Conceived in 1969 as a voluntary program through a privately funded initiative, it has more recently become a congressionally legislated program as part of No Child Left Behind (NCLB, 2001), requiring state participation in order to receive Title 1 funds (Stedman, 2009). The reading subject data was added in 1983 and is collected every two years (NAEP training seminar, 2009).

In terms of reading content, the NAEP assessment requires examinees to respond to a variety of texts, such as stories, poetry, articles, and advertisements. Reading passages range in length from 300 to 800 words and are drawn from typical grade-appropriate sources. Fourth Grade reading passages are classified into two *Contexts* (or genres): *Reading for Literary Experience* and *Reading for Information* (see Table 1).

Table 1
Description of Passage and Item Characteristics

CONTEXTS	ASPECTS
Reading for literary experience Readers explore events, characters, themes, settings, plots, actions, and the language of literary works by reading novels, short stories, poems, plays, legends, biographies, myths, and folktales.	**Forming a general understanding** The reader must consider the text as a whole and provide a global understanding of it.
Reading for information Readers gain information to understand the world by reading materials such as magazines, newspapers, textbooks, essays, and speeches.	**Developing interpretation** The reader must extend initial impressions to develop a more complete understanding of what was read.
	Making reader/text connections The reader must connect information in the text with knowledge and experience.
	Examining content and structure The reader must critically evaluate, compare and contrast, and understand the effect of such features as irony, humor, and organization.

Note. Adapted from "Reading Framework for the 2007 National Assessment of Educational Progress", by National Assessment Governing Board (2007).

In student booklets, each reading passage is typically followed by approximately 10 items (a fraction of the item pool) to be answered. The NAEP item pool consists of multiple-choice, short-, and extended-constructed response items. In terms of items content, items conform to one of the four *Aspects of reading* comprehension: forming a general understanding, developing interpretation, making connections, and examining content and structure (see Table 1). Because the NAEP is a secured test with copyrighted limitations, the reading passages and items are not released, and therefore cannot be specifically discussed here. The one exception are Block R11 items which are classified as publicly released and will be discussed for illustrative purposes.

NAEP uses matrix sampling design to ensure that each participating student takes only a portion of the complete set of cognitive items developed; one quarter of the student sample is exposed to each item. The NAEP Reading Comprehension assessment draws from a pool of 100 test items. Items are embedded within 9 total blocks (block = a passage and corresponding set of

questions). Each student test booklet contains two blocks (i.e. approximately 20/100 items). To ensure that few students get the same test booklet, NAEP organizes its assessments by a Balanced Incomplete Block (BIB) design. Test booklets pair each block with every other block, which results in approximately 50 different reading booklets where each block appears once in every position within each of the booklets (NAEP training seminar, 2009).

This dissertation will consider three variables that contribute to item difficulty*: Item Format, Context of reading*, and *Aspect of reading*. To reiterate, my hypothesis asserts that while *Context* and *Aspect* will account for some degree of mean student proficiency, I anticipate that *Item Format* will have the greatest impact. Table 2 outlines a breakdown of how items fit into each of the three characterizations of assessment items: *Item Format, Context,* and *Aspect*. *Item Format* is comprised of an approximate 50/50 mixture of multiple-choice and constructed-response items (MC=57, SCR=11, ECR=32; see Table 2). The same is true for *Context* with 51% literary and 49% informational. *Aspect* has a disproportionate number of items that fall into the Interpretation category, therefore, it is likely that analysis will also have disproportionate numbers appearing from this category.

This table also demonstrates the frequency of each of the Item Formats across the components of the Reading Framework. Item Format appears to be paired with almost of all the elements; there are 24 categories in total and only 5 categories in which a specified item format was not created for a particular Aspect within a Context (e.g. no ECR items appear under General Understanding content within Literary text). This indicates that there is no obvious pattern shift of *Item Format* pairing with *Context* and/or *Aspect* (i.e. we do not see a preponderance of one *Item Format* clustered on one end of the spectrum in contrast to another *Item Format* clustered on the opposite end).

Table 2
Item Format Count by Context (horizontal) and Aspect (vertical)

	MC		SCR		ECR		
	Literary	**Information**	**Literary**	**Information**	**Literary**	**Information**	**Total Aspect**
General Understanding	3	2	1	2		1	9
Interpretation	21	19	4		9	15	68
Connections	1		2	1	1	3	8
Content & Structure	5	6	1		3		15
Total Context	30	27	8	3	13	19	
Total Item Format	57 MC		11 SCR		32 ECR		100

NAEP reported scores. Reported scores allow student results to be placed on a common scale given the administration of different testing booklets. NAEP uses Item Response Theory (IRT) to convert percent of items answered correctly into scale scores. Results are reported as

average scale scores and as percentages of students performing at or above four NAEP achievement levels: Below Basic, Basic, Proficient, and Advanced. NAEP results include student background characteristics such as English language proficiency, special education categorization, gender, eligibility for free/reduced school lunch, and race/ethnicity.

The NAEP assessment is not intended to be a test of individual ability. Participants complete only a small fraction of test items from the pool. To avoid measurement error that would result from estimating individual ability, the technical innovations of marginal estimation and plausible values are used for reporting purposes by summarizing how well groups of students answered the NAEP questions and how well other students like them answered the rest of the test questions. For analysis purposes here, raw student scores from the database will be used for each item. Student responses in the database are coded as incorrect (0), correct (1) for dichotomous items, and range (1), (2), (3), and (4) for polytomous items in accordance with the item rubric. For purpose of analysis, student responses coded in the NAEP database as either "omitted", "not reached", "multiple", "illegible", "off task", and "non-rateable" were recoded in SPSS as incorrect (0).

Data Analysis/Procedures

Simple logistic and partial credit Rasch model. Item response theory (IRT) provides a framework for examining the responses of individuals to a set of items. For all my analyses, I will use a Rasch family model (1980). Several programs are available to calculate item parameters based on the Rasch model; for this dissertation, the computer program ConQuest (Wu, Adams, & Wilson, 2007) will be used to estimate individual item difficulties, run latent regression, as well as to examine individual items for statistically significant differential item functioning (DIF) across language groups, and to explore the presence of dimensionality across the three item formats. The basic Rasch measurement model is a one parameter model which assumes that items differ only in difficulty. The assumptions of the unidimensional model include local independence and that there is only one underlying latent trait. A Rasch simple logistic model will be fitted to the multiple-choice items and a partial credit model (Masters, 1982) used for the constructed-response items.

I will conduct secondary data analysis of the 2007 NAEP Reading Assessment to consider unplanned variation of the items as a product of item response mode. Below is a list of variables that I will use to analyze the effect of item response mode on participant latent ability (see Table 3). The ConQuest software combines an item response model and a multivariate regression model (latent regression model); therefore it is capable of estimating model parameters that control for ability level and population characteristics. This is required to adequately control for differences in mean group abilities. My analysis will focus on students who struggle with reading comprehension, thus I will consider item format effect on the performance of both English language learners (as opposed to native English speakers) and students classified with Specific Learning Disability (SLD). My primary item predictor will be item response format (multiple-choice, short-constructed response, and extended-constructed response). After my initial analysis, in order to isolate the effect of item format, I will factor in other item variables that may be contributing to item difficulty such as Reading *Context* and *Aspect* of Reading.

The Rasch model has the capacity to analyze individual item responses by estimating the probability of getting each item correct or of attaining a particular response level (in the case of polytomous items) depending on two factors: person ability (i.e., the underlying latent trait, which in this case is latent reading comprehension ability (θ_n) and item difficulty (δ_i)

(Embretson & Reise, 2000). As individual ability rises, the probability of answering an item correctly, or at the specified level, rises as well. Under IRT, the examinee ability is not a direct transformation of the number-correct, but is estimated by parameters that take into account test items (Rogers, 1999). In a dichotomous item, the probability (*P)* that a person (*n)* would get item (*i)* correct for a dichotomous item (1=correct) is expressed as:

$$\mathrm{P}(\mathrm{X}_{ni}=1 \mid \ \theta_n)=\frac{\exp(\theta_n-\delta_i)}{1+\exp(\theta_n-\delta_i)}$$

And for a polytomous item (k = number of steps) is expressed as:

$$\mathrm{P}(\mathrm{X}_{nix}=1 \mid \ \theta_n)=\frac{\exp\sum_{k=0}^{x}(\theta_n-\delta_{ik})}{\sum_{h=0}^{m}\exp\sum_{k=0}^{h}(\theta_n-\delta_{ik})}$$

where

k = particular item step

i = particular item in the test

$\mathrm{P}(\mathrm{X}_{ni}=1|\theta_n)=$ the probability that a randomly chosen examinee with ability θ_n answers item i correctly (or at the specified level)

θ_n = latent ability parameter for person n

δ_{ik} = difficulty parameter for response k to item i

exp = base of the natural logarithm whose value is approximately 2.718.

Model fit: mean, variance, and item fit. A preliminary step is to conduct traditional item analysis to assess whether all the items have satisfactory psychometric properties. This analysis produces classical difficulty, discrimination, and point biserial statistics for each item, as well as Cronbach's alpha for the test as a whole. As running classical difficulty and discrimination is standard procedure, I do not anticipate finding anything unusual since NCES has likely identified items for removal from the test pool using a three-parameter logistic model when items were not meeting conventional values for classical difficulty and discrimination. The point biserial statistics (also called discrimination indices) are produced for each item to indicate the relationship between individuals with a particular response (e.g. correct or incorrect) and their score on the rest of the assessment. Summary statistics will include calculations such as the estimated Person-Separation reliability coefficient for the whole test.

Individual item analysis. Individual item analysis will produce a mean estimate of latent reading comprehension ability and the variance of that ability distribution; the model produced here will serve as the baseline model for future comparisons. It will also produce item difficulty parameters, standard errors, and fit statistics for each of the 100 items and –in the case of polytomous items- individual item steps. ConQuest uses marginal maximum likelihood (MML) estimates for the item parameters of the model(s) using an EM (Expectation/Maximization) algorithm. In brief, ConQuest will alternately estimate the deltas (δ; item difficulty) and thetas (θ; latent ability) until estimations show little change. For purposes of model identification, ConQuest constrains the difficulty parameter estimate for the last item to ensure an average difficulty of zero. Item that are constrained are indicated by the asterisk (*) placed next to the parameter estimate (Wu et al., 2007).

For multiple-choice items, the simple logistic model will produce one difficulty parameter for each multiple-choice item. Constructed-response items, on the other hand, are

scored on a 1 to 4 rating scale depending on the levels of their corresponding rubric, and therefore the model for polytomous items must include an item*step parameter for each level of the items in addition to an overall difficulty parameter estimate. For example, the step parameter labeled as step 1 describes the transition from the score 0 to 1, where the probability of being in category 1 is greater than the probability of being in category 0, while the second step describes the transition from 1 to 2, and so on.

In addition to item parameters, ConQuest provides mean square fit statistics that operate like effect sizes (in terms of model misfit), and corresponding t statistics. If a mean square fit statistic lies near the value of 1.0 (the null hypothesis) then the item is acting as expected by the model. The further away a mean square statistic gets from 1.0, the less the item conforms to the model (more or less variation than expected); thus we would reject the null hypothesis that the data conforms to the model (Wu et al., 2007). Specifically, an infit mean square < 0.75 is an indicator of less randomness than expected while an infit mean square > 1.33 is an indicator of more randomness than expected. Because of the large sample size, all *unweighted* and *weighted t statistics* are expected to demonstrate significance at the > 2.0 threshold.

In addition to numerical values, ConQuest will produce a Wright map that graphically illustrates the relationship between latent ability and item difficulty. ConQuest produces item and step difficulties on a logit scale; the higher an item is on the Wright map, the more difficult the item is. Item difficulties are plotted as "Thurstonian" thresholds, in other words, the point where a student has a 50% chance of achieving at least the associated level of performance on an item.

The map also produces a histogram of participant latent ability (represented by x's; see Figure 2 for example). Participants (on the left hand side) whose proficiencies appear above all items (on the right hand side) of the Wright map have more than a 50% chance of answering all the respective items correctly, and those whose proficiencies lie below all the items have a less than 50% chance of answering all the respective items correctly.

Figure 2. Wright map example. This figure illustrates the elements of a Wright map.

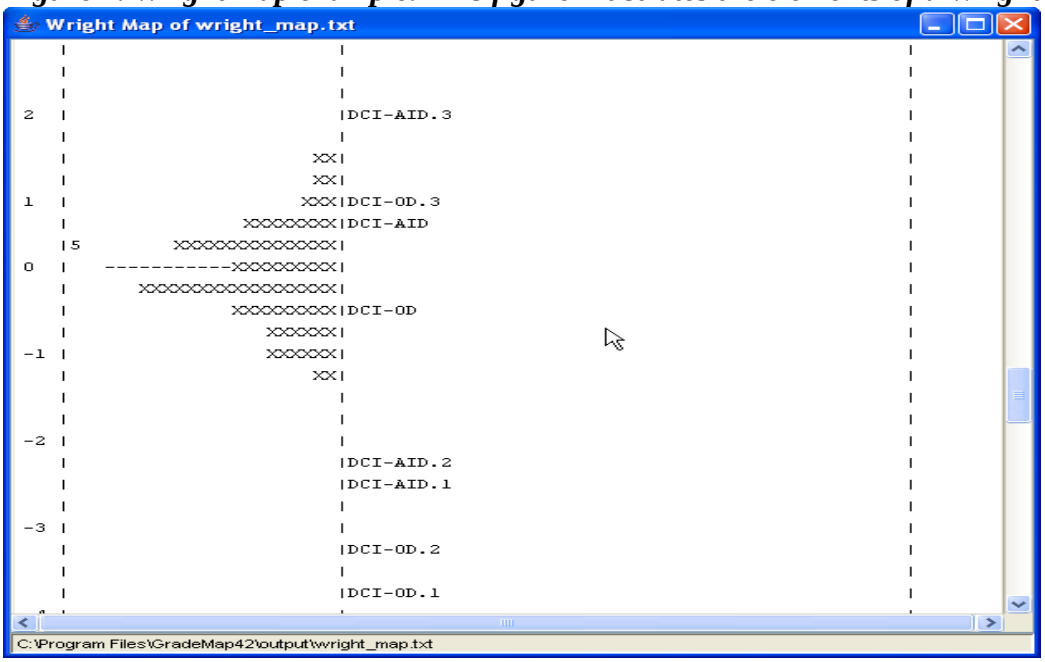


Image taken from: http://bearcenter.berkeley.edu/kennedy/GMOnline/Wright_Maps.html

Differential Item Functioning. Using the Rasch model, I will examine all 100 NAEP items for differential item functioning (DIF) to detect unexpected differences in performance across subgroups of test takers. DIF studies are generally concerned with the question of whether an item is 'fair' for a focal subgroup of individuals as compared to the reference group. DIF can reveal whether an item is particularly easy or difficult given subject proficiency and the general difficulty of the item for specific test taking groups such as limited English proficiency or Specific Learning Disability (Wilson, 2005). An item is considered to be unbiased if it is equally difficult for persons of the focal group in comparison to the reference group across individuals of equal ability level (Meulders & Xie, 2004). "When a test is functioning as intended, all examinees at the same level of ability will have scores that center on the same value" (Moore, 1996). If some test items are harder for ELLs than native English speakers of equal ability, for example, then ELLs will have lower test scores on average for DIF items than their counterparts. Thus, in detecting DIF, the model must accommodate information about group membership. DIF was first analyzed among NE and ELL speakers and then among students without SLD and students with SLD.

In fitting a DIF model, ConQuest will produce a mean parameter estimate for each of the items across the language groups. For dichotomous items, the DIF model will describe the probability of correct responses using two main effects, *item + language proficiency*, plus an interaction of *item *language proficiency*. The first term (*item)* will yield a set of item difficulty estimates.

$$\mathrm{P}(\mathrm{X}_{ni}=1\,|\,\theta_n,\vartheta=g)=\frac{\exp(\theta_n-\delta_i-\vartheta_g-\delta_{ig})}{1+\exp(\theta_n-\delta_i-\vartheta_g-\delta_{ig})}$$

The probability (P) of a correct response $(\mathrm{X}=1)$ depends on the ability of an individual (θ_n), the difficulty of an item (δ_i), the main effect of group membership (G_n), and the interaction between item difficulty and group membership (δ_{ig}). The group membership term (*language proficiency)* will give the difference in mean ability between the language proficiency groups. The interaction term (*item *language proficiency*) is what indicates DIF and this term will give an estimate of the difference in difficulty of the items between the language groups.

The amount of DIF is represented by the effect size, which is the weighted sum of differences between the proportion-correct on the items in the two groups across all score levels. The null hypothesis (Ho: $\delta_{ig}\neq 0$) is rejected if there are significant differences in the weighted means between the reference and focal groups.

In addition to the steps above, polytomous items also include a step term, *item *language proficiency*step* that models for each of the language proficiency groups, a probability within each of the items of reaching each step. The probability (P) of obtaining a particular score $(X=k)$ given ability of an individual (θ_n), group membership (ϑ_g), difficulty of an item step (δ_{ik}), the interaction of item and group membership δ_{ikg}, is expressed as:

$$\mathrm{P}(\mathrm{X}_{nix}=1\,|\,\theta_n,G=g)=\frac{\exp\sum_{k=0}^{x}(\theta_n-\delta_{ik}-G_n-\delta_{ikg})}{\sum_{h=0}^{m}\exp\sum_{k=0}^{h}(\theta_n-\delta_{ik}-G_n-\delta_{ikg})}$$

In this model, the abilities of ELLs and NEs are controlled statistically, thus DIF is indicated by any significant difference in item difficulty between the language groups

represented by the interaction term *item*language proficiency*. The parameter estimate is expressed on a scale in which positive values mean that the item is easier for members of the reference group (NE speakers and non-SLD individuals) than for matched members of the focal group (ELLs and individuals diagnosed with SLD). On the other hand, negative values indicate that the question is easier for members of the focal group than for matched members of the reference group.

In the following example, if ϑ_g = 0 for NEs and 1 for ELLs, a positive parameter estimate for δ_{ikg} indicates that the item was comparatively more difficult for ELLs, while a negative score would indicate that it as easier. The calculated difference between parameter estimates of 0.291 for ELLs and -0.291 for NEs (i.e. harder for ELLs vs. NEs) is 0.582 logits, meaning ELLs are lower than NEs by more than 50% of a standard deviation on that particular item (Wu et al., 2007). In addition, parameter estimates of an item for the reference group can be added to the parameter estimates previously produced by the term *item format* to indicate a "truer" item difficulty for the reference group; the same can be done for the other language groups as a comparison. For example, if parameter estimates indicate the aforementioned item has a difficulty parameter of 0.562, the DIF estimate for ELLs is added to that estimate (0.562 + 0.291), the difficulty changes from 0.562 to .853 logits for ELLs and the difficulty parameter reduces for NEs from 0.562 to 0.271.

While this analysis shows the existence of DIF between the selected terms, expressed as $\delta_{ig} = \delta_i(group1) - \delta_i(group2)$, it is the magnitude of the logit difference (i.e. the effect size) that determines substantive importance. To measure magnitude in the context of the Rasch model, this analysis will use Paek's (2002) translation of a standard effect size recommended by Longford, Holland, and Thayer (1993):

Level A: null hypothesis is retained if δ_{ig} <.426; indicates no DIF

Level B: null hypothesis is rejected if .426 $\leq \delta_{ig}$ < .638; indicates slight to moderate DIF.

Level C: null hypothesis is rejected if $\delta_{ig} \geq$ .638; indicates moderate to large DIF.

Level A contains the questions with little or no difference between the two matched groups. Level B contains questions with small to moderate differences. Level C contains the questions with the greatest differences.

In order to calculate the presence of DIF items within the three item formats, it is necessary to manually group items to see what patterns emerge. I will compare DIF results across items when grouped according to each of the three properties of assessment format: *Item Format*, *Context*, and *reading Aspect*.

Dimensionality across assessment and participant variables. To investigate the question of whether *Item Format* affects the construct of reading comprehension, my final analysis will test the basic unidimensional assumption of the data by fitting it to a multidimensional model that accounts for the subdomains of various assessment characteristics. It differs in that a unidimensional model assumes that the single latent trait of Reading Comprehension ability underlies all items and that a common set of item parameters estimates examinee ability; this implies that all examinees are similar in how they perceive and respond to items (see Figure 1) despite person properties or assessment characteristics.

While the NAEP Reading Comprehension assessment is modeled as a unidimensional test, there are several underlying domains that comprise its framework: *Item Formats* are designed to tap different structures of reading *Aspects* from two different *Contexts.* In contrast,

the multidimensional analysis challenges the single continuum narrative by evaluating whether some items are more difficult than others for all students and whether some students are more likely to give correct responses to all items than other students. While there is one primary dimension of interest in Reading Comprehension, there may also be other dimensions within that test, such as within *Item Format*, that produce construct-irrelevant variance.

This analysis will employ a *multidimensional between-item model* (Adams, Wilson, and Wang, 1997) by creating a statistic for each of the parallel unidimensional subscales derived from clusters of items that relate to only one of the latent dimensions (see Multidimensional in Figure 1). A covariation/correlation matrix of the multidimensional model illustrates how performance across the dimensions might be interrelated . "Treating student performance that is multidimensional in nature as unidimensional can have the effect of misrepresenting student ability" (p. 338, Briggs & Wilson, 2004).

This analysis will indicate if *Item Format* is playing a significant role in the estimation of overall reading comprehension ability. If the multidimensional model (that is, treating each of the item formats as a separate construct) fits better than the unidimensional model, it suggests that separate dimensions of *Item Format* are measuring different reading comprehension skills, although the test is modeled unidimensionally. Statistically this matters because if a multidimensional model fits better than a unidimensional model then the fundamental unidimensional assumption of local independence is violated. When violated, the implication is that item parameter estimates will be biased, and the associated standard errors of the ability estimates will be too small. But, a practical reason to care is that it is important that tests measure what they intend to measure. Some tests modeled as unidimensional can report subscores. And, multiple dimensions may prove useful diagnostically.

To test this theory, I will compare three multidimensional models based on *Item Format* to the baseline unidimensional model. The first model will investigate a two-dimensional model separating item format into MC and CR items (SCR and ECR items will be combined in this analysis). Secondly, to tease apart differences in CR items, I will analyze format by a three dimensional model of MC, SCR, and ECR items.

Comparison of these various models will help indicate whether some participants have observably different ability estimates on the construct of item format. Then latent regression can be used to determine whether individual variables of language proficiency or reading ability explain some of the variation in scores. For example, perhaps native English speakers are half a logit higher than ELLs on MC items, yet a full logit higher on CR items, implying an interaction between language status and item-type difficulty.

If earlier analysis demonstrates *Item Format* to play a significant role in reading comprehension proficiency, then it is important to compare its statistical influence to other possible multidimensional models. Two other potential sources of confounding variance are named within the NAEP Reading Comprehension Framework which characterizes items by two *Contexts of reading* (a. literary and b. informational) and four *Aspects of reading* (a. forming a general understanding, b. developing interpretation, c. making connections, and d. examining content and structure) as subdimensions of reading comprehension. In order to isolate the effect of *Item Format*, I will investigate the contribution of *Context* and *Aspect* to the variance. To this end, I will compare a two-dimensional model of *Context* and a four-dimensional model of *Aspect* to the aforementioned three-dimensional model of *Item Format*. I hypothesize that there may be some multidimensionality present within *Aspect and Context*, but that *Item Format* will have a greater impact on student performance.

In answering questions of best fit, I will consider whether the components of these various dimensions function as the same construct. In other words, does it make sense to combine student scores across these dimensions? If a low correlation across components of a dimension is derived from these analyses, then it suggests that these components do not belong together, and by combining scores we are ultimately combining two different constructs. If multidimensionality is established, to take this analysis one step further, I will also run latent regression to consider subgroup performance (e.g. ELL vs. NE) within each of the multidimensional models (i.e. *Item Format, Aspect, and Context*).

For dichotomous items, the probability (*P)* that a person (*n)* would get item (*i)* correct (1=correct) in dimension *(r)* is expressed as:

$$P(X_{ni}=1|\theta_{rn}) = \frac{\exp(\theta_{rn}-\delta_i)}{1+\exp(\theta_{rn}-\delta_i)}$$

For polytomous items, the probability (*P)* that a person (*n)* would answer item (*i)* at the expected level (1=correct or level 1) in dimension *(r)* is expressed as:

$$P(X_{nix}=1|\theta_{rn}) = \frac{\exp\sum_{k=0}^{x}(\theta_{rn}-\delta_{ik})}{\sum_{h=0}^{m}\exp\sum_{k=0}^{h}(\theta_{rn}-\delta_{ik})}$$

Model significance. The fit of all nested models will be compared to the Baseline model using two criterion in order to evaluate the coherence of data: Akaike's Information Criterion (AIC; Akaike, 1973) and Schwarz's Bayesian information criterion (BIC;Schwarz, 1978). The preferred model is the one with the lowest AIC and BIC values. The AIC formula is expressed as follows:

AIC = 2k-2ln(L)

The first component *k* is the number of estimated parameters, and the second component -*2ln(L)* is the deviance for the model. This formula is intended to discourage overfitting the data by rewarding goodness of fit (i.e. the deviance), while penalizing for overparameterization. The BIC formula has a greater penalty for the number of parameters used in a model. It is expressed as follows:

BIC = -2ln(L) + k*ln(n)

where -*2ln(L)* is the deviance, *k* is the number of estimated parameters and *n* is the sample size.

Summary. I will run four different types of ConQuest models. A model with traditional and individual item analysis will establish the reliability and fit of baseline model for which to compare all subsequent nested models. This model will also establish mean student proficiency for all 4th grade examinees and will illustrate the range of item difficulty for all 100 items. The Wright map in particular will help establish whether there is a hierarchy of difficulty across MC, SCR, and ECR *Item Formats.* A latent regression model will establish mean performance differences between the focal student subgroups (i.e. ELL and SLD examinees) and the reference group (NE and/or Non-SLD examinees). Differential item functioning will evaluate each and every item to determine if any items have a differential effect for group membership and –in the case that DIF is found- will also determine which membership group the flagged items tend to favor. Items that are flagged for DIF will also be evaluated for patterns of a differential effect related to *Item Format.* And, finally, multidimensional analysis will determine how closely

correlated the components of *Item Format* are, as well as to compute estimated performance differences between the reference and focal groups across each of the Item Format dimensions. This multidimensional analysis will be contrasted to subsequent multidimensional models of *Context* and *Aspect* by comparing AIC and BIC values.

Chapter 3: Results

The purpose of this dissertation is to explore the relative effect of *Item Format* on 4th grade reading comprehension proficiency. For this dissertation, the ConQuest software (Wu et al., 2007) was used to fit all models. A Rasch model was fitted to the multiple-choice items and a partial credit model (Masters, 1982) fitted to the constructed-response items for all 100 items from the 4th grade reporting sample in the NCES 2007 NAEP Reading Comprehension database. To investigate the research questions, four different types of ConQuest measurement models are used to analyze the item responses: a baseline model with traditional and individual item analysis, a latent regression model, a differential item functioning model, and three multidimensional models. Each model represents a particular perspective for thinking about the relationship between assessment format and student characteristics and delivers a significant story about the effect of *Item Format* on our perception of student reading comprehension proficiency.

Model 1: Baseline Model

In order to check for satisfactory psychometric properties and to establish a baseline model for future comparisons, a Rasch partial credit model with traditional and individual item analysis was run on all 100 NAEP items in the 4th grade reporting sample. The first step in addressing the research question begins with this analysis. In this preliminary step, we establish the overall mean performance on all items for all 4th grade students. Additionally, by way of the Wright map, we are able to compare the range of item difficulties across the subcategories of the assessment characteristics with primary focus on *Item Format.*

Traditional item analysis. Traditional item analysis produced IRT difficulty, classical discrimination, and point biserial statistics for each item, as well as Cronbach's alpha (Person-Separation reliability) as a measure of internal consistency for the test as a whole. Discrimination for MC items ranges from 0.29-0.60, SCR ranges 0.36-0.58, and ECR ranges 0.44-0.70. Point biserial correlations demonstrate an upward trend between thresholds for most polytomous items. Four ECR items demonstrate slight inversions between threshold (in bold):

Item 77: -0.54, **0.36, and 0.32;**
Item 79: -0.60, 0.30, **0.38, and 0.25;**
Item 84: -0.57, 0.24, **0.34, and 0.23;** and
Item 100: -0.55, 0.18, **0.37, and 0.28**.

Expected a posterior (EAP) estimate of reliability is 0.857. This estimate indicates an overall measure of internal consistency. The reliability is fair given that NAEP is not designed to be a test of individual ability; therefore we don't expect this estimate to be extremely high. When item formats are run separately, the EAP reliability drops (see Table 5), but most significantly for SCR comprised of only 10 items, therefore the drop is likely attributed to fewer items in this *Item Format* category.

***Fit statistics*.** Item fit indicates how well the IRT model represents the data on an item-by-item basis (Embretson & Reise, 2000). Fit is determined by an infit mean square, which is the ratio of the mean of the squares of the observed residuals to the expected squared residuals (the residual being the difference between the observed score and the expected score for an item). If a mean square fit statistic lies near the value of 1.0 (the null hypothesis) then the item is acting as expected by the model. The further away a mean square statistic gets from 1.0, the less the item conforms to the model (more or less variation than expected); thus we would reject the null hypothesis that the data is conforming (Wu et al., 2007). Specifically, an infit mean square < 0.75 is an indicator of less randomness than expected while an infit mean square > 1.33 is an

indicator of more randomness than expected. Items outside the acceptable range likely produce a corresponding *t* statistic (MNSQ fit statistic) with an absolute value exceeding 2.0; however, with large sample sizes –such as the 2007 NAEP Reading data- one can expect this statistic to typically show significant values for most items, hence a safer strategy is to consider only the items showing both a misfit by infit mean squares and a *t* statistic above 2.0 (Wilson, 2005). ConQuest will report both *unweighted* and *weighted t statistics*. The *unweighted t statistic* represents all respondents as treated the same, whereas the *weighted t statistic* gives more weight to respondents near the mean, and less weight to outliers (Wright & Stone, 1980; Wright & Masters, 1982).

The fit provided by the ConQuest output of the partial credit model ranged 0.63 to 1.43, thus indicating a good fit of the model at the item level. Analysis provides evidence that there are five misfitting or problematic items: three underfitting (Table 3) and two overfitting (Table 4; for full Infit details, see Appendix F). Overfitting items indicate less randomness than expected, but the underfitting items are typically more problematic because they demonstrate more randomness than expected and/or desired. NCES has utilized the many advantages of item response modeling to overcome sparse individual data points and to produce reliable results for subgroups. In the vetting process, it is likely that most misfitting items have been excluded from their item pool. Misfitting items produced in this analysis lay very close to the "acceptable" fit cut-off parameters and were probably a result of the application of a 1 PL model – as opposed to a 3 PL used by NCES to develop NAEP items. Both underfit and overfit items vary in terms of *Item Format*, *Context*, and *Aspect*.

Table 3
Baseline Model Underfitting Items; <.75 MNSQ

			Unweighted Fit			Weighted Fit		
Item	**Estimate**	**Error**	**MNSQ**	**CI**	***t***	**MNSQ**	**CI**	***t***
12	-1.533	0.016	0.64	(0.99, 1.01)	-57.9	0.82	(0.98, 1.02)	-19.9
28	-1.902	0.017	0.63	(0.99, 1.01)	-58.8	0.85	(0.98, 1.02)	-13.8
91	1.954	0.013	0.74	(0.99, 1.01)	-38.9	0.88	(0.98, 1.02)	-19.5

Table 4
Baseline Model Overfitting Items; >1.33 MNSQ

			Unweighted Fit			Weighted Fit		
Item	**Estimate**	**Error**	**MNSQ**	**CI**	***t***	**MNSQ**	**CI**	***t***
86	0.476	0.008	1.34	(0.99, 1.01)	42.5	1.25	(0.99, 1.01)	36.4
88	0.470	0.008	1.43	(0.99, 1.01)	52.4	1.27	(0.99, 1.01)	39.5

This finding of few misfitting items is not surprising considering that NAEP items have been subject to a vetting process by which they are designed and evaluated by the field's leading content experts, pilot tested, and the results analyzed using a 3PL Item Response Model (Pelligrino, Jones, & Mitchell, 1999).

Item analysis. The specified partial credit model produced two terms: a) item and b) item by steps. The partial credit model yielded statistical information related to both person and item

fit such as item difficulty parameters, standard errors, and fit statistics for each of the 100 items and –in the case of polytomous items- individual item steps. Final deviance of the unidimensional model is 4,875,366 with 136 total estimated parameters— 99 item parameters, 35 step parameters, one mean and one variance. Ninety-nine parameters are used to describe 100 items because identification constraints are applied to the last item. This is achieved by choosing the difficulty of the last item to be equal to the negative sum of the difficulties of the remaining items (Wu et al., 1997).

ConQuest reports the mean and standard deviation of latent reading comprehension ability for all students accumulatively across all *Item Formats* as 0.549 (1.140) logits. Individual models were run for each of the *Item Formats* (see Table 5). When averaging person proficiencies for all multiple-choice items the mean is 1.059 (1.281). The mean of the item difficulties for short-constructed response items is lower at 0.278 (1.217) and for extended-constructed items is even lower at -0.162 (1.038). Because each model is constrained to 0, the person proficiencies are not directly comparable across the MC, SCR, and ECR models because there is no way to tell if the decrease in means from MC>SCR>ECR is attributed to a drop in person proficiencies or a rise in item difficulty; although we can guess that it is probably both. Thus, to investigate the relationship between student proficiencies and item difficulties within the *Item Formats* the Wright map subsequently described is the ideal means for comparison. It provides a venue for which to examine the range of *Item Format* difficulties within the same model.

Table 5
Parameter Estimates of Baseline and Consecutive Models

Fixed Effects of:	Population mean (SE)	SD	EAP Reliability
Baseline Model: all formats	0.549 (0.003)	1.140	0.857
MC Model	1.059 (0.003)	1.281	0.721
SCR Model	0.278 (0.003)	1.217	0.311
ECR Model	-0.162 (0.002)	1.038	0.682

***Wright map*.** In addition to numerical values, ConQuest produced a Wright map that graphically illustrates the relationship between student latent ability (represented by x's on the left-hand side) and item difficulty (represented by item # on the right-hand side). ConQuest produces item and step difficulties on a logit scale; the higher an item is on the Wright map, the more difficult the item is, whereas the lower the item is, the easier it is. Ideally, a Wright map should illustrate a normal distribution of student proficiencies within the range of item difficulties. There should be items that are more difficult (represented at the top right of the map above student proficiencies) and items that are more easy (represented at the bottom right of the map below student proficiencies) in order to avoid ceiling and floor effects. Items and item steps are plotted on the Wright map at the point where the student with a particular logit ability has a 50% chance of scoring correctly (on dichotomous items) or getting a designated score or above

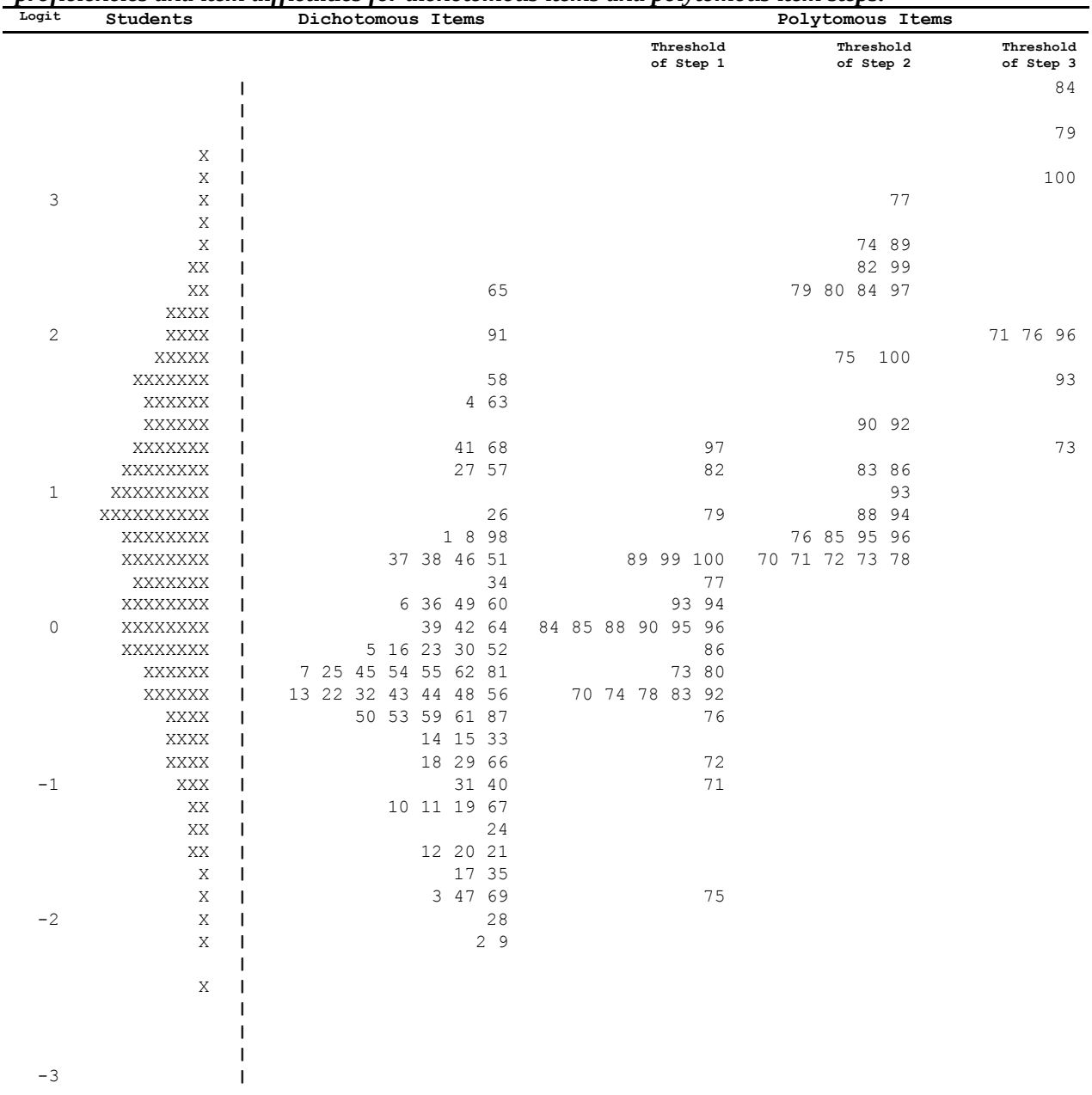
(in the case of polytomous items). So, while we know specifically how this 4th grade population performed on each of these NAEP Reading Comprehension items, their plotted difficulty on the Wright map indicates their relationship to one another and to overall student proficiencies. This illustration allows us to compare item difficulties across the three *Item Formats* as well as to later compare the relationship of items when organized by the other assessment formats of *Context* and *Aspect*.

Figure 3 is a Wright map of all items and student proficiencies from the 4th grade NAEP reading comprehension assessment. The estimated item difficulties have been manually separated into columns illustrating dichotomous items (multiple-choice and short-constructed response) versus the levels of polytomous items. Item 2 and 9 are estimated to have the lowest difficulty, so they are plotted at the bottom of the Wright map at approximately -2.0 logits. At the top of the map, threshold 3 of Items 77, 100, 79, and 84, respectively, have the highest difficulty estimates, with item 84 close to 4.0 logits.

One noticeable feature of this map is that estimated item difficulties (noted or the right hand side of the map by their item number) generally span the range of respondent abilities (noted on the left hand side of the map by X's) fairly well without any noticeable gaps that would indicate less coverage than expected. We also see that item difficulties exceed estimated latent student ability meaning that the assessment is tapping a broader range of abilities than present in the student abilities sampled. In particular, student proficiencies cap at ~3.5 logits while item difficulty edges towards ~4.0 logits. This range of item difficulty is important since it avoids possible ceiling effects which would result in greater standard error for proficiency estimates.

We do see, however, a group of students whose proficiencies at -2.5 logits fall below the difficulty of the two easiest items: Item 2 and 9 located closer to 2.0 logits. This 0.5 logit difference means that these students have a less than 38.5% chance of answering Item 2 and Item 9 correctly (Wilson, 2005). The Wright map also illustrates the floor effect that the NAEP Validity Panel is currently trying to resolve where even the easiest reading comprehension items are too difficult for low ability students; consequently because they are not able to answer them, information about what these students can do is lost.

Figure 3. Wright map for Baseline Model. This figure illustrates the distribution of person proficiencies and item difficulties for dichotomous items and polytomous item steps.



Note: X =3,294 students

In looking at the dichotomous column as compared to the three polytomous columns, we see a much wider range of item difficulty. The map also indicates that there is a very clear pattern of item format difficulty and that the item difficulties are acting as expected per the literature: dichotomous items are easier on average than polytomous items (Shohamy, 1984). Specifically, dichotomous items (i.e. MC and SCR items) span about -2.3 to 2.5 logits. MC items span -2.3 to 1.5 logits (Items 1-57) and SCR (Items 58-68) span about 1 logit higher at approximately -1.5 to 2.5 logits, indicating that MC items are easier on average than SCR.

Polytomous items (Items 69-100) span -2.0 to almost 4.0 logits, indicating that MC and SCR items are much easier on average than ECR (polytomous) items. Thus, MC < SCR < ECR.

Item difficulties are plotted on the Wright map as "Thurstonian" thresholds that indicate a probability of students at the corresponding proficiency answering a multiple-choice item correctly or answering a constructed-response item at a particular score level (x.1, x.2, x.3, or x.4) or above. Probability is exactly 50-50 when the student proficiency is aligned with item difficulty; this indicates that the individual has an equal chance of either getting a multiple-choice item correct or incorrect or, in the case of a constructed-response item, scoring above or below the indicated level. The number of students (indicated by an X) whose proficiencies appear above the plotted item difficulties of the Wright map have more than a 50% chance of correctly answering the respective items at the given levels, and those whose proficiencies lie below the plotted item difficulties have a less than 50% chance of correctly answering the respective items at the given levels (Wilson, 2005).

For example, at logit 0 which is the mean student proficiency, 8 X's represent 26,352 students who have a 50% chance of answering MC items 39, 42 and 64 correctly and ECR items 84, 85, 88, 90, 95, and 96 at step 1 or above on the rubric (note: 84.1 indicates Item 84, step 1). These same students have a greater than 50% chance of scoring correctly on Item 5 and 16, as well as scoring at level 1 or above on Item 86 because these item difficulties are plotted below their student proficiency. On the other hand, these same students have a less than 50% chance of scoring correctly on MC item 6 and 36, as well as scoring at level 1 or above on item 93 and 94 because these item difficulties are plotted above their proficiency.

The Wright map indicates that all students have a less than 50% chance of scoring at level 3 on Item 84 and 79 because these item difficulties are plotted above all student proficiencies (the X's in the left-hand column). We also see that for the easiest two items, Item 2 and 9, there is a set of 3,294 students who have less than a 50% chance of answering them correctly and a set of 3,294 students who have a 50-50 chance, and the rest of the students have a greater than 50% chance of answering the items correctly.

As seen in the fit statistics, there is an upward trend between item steps, meaning that difficulty increases between getting a score of 1 over a 0, or a 2 over 1, etc. This is also illustrated by item difficulty of steps on the Wright map. We see that the relative distances between item thresholds (i.e. steps) on the Wright map indicate that some items discriminate among respondent proficiencies better than others. Some items have larger logit differences between thresholds than others, meaning that there is an increase in difficulty between getting a score of 1 versus a 2 or a score of 2 versus a 3. For example Item 96.1 plotted at logit 0, has a half logit difference between threshold 1 and 2, yet these thresholds are not clustered so closely together that ability between the steps seems indistinguishable. On the other hand, item 75.1 that is plotted at almost -2.0 logits has an extremely large logit difference between threshold 1 and 2 of almost 4 logits indicating that it is a big leap in ability for a student to move from getting a score of 1 on the item to getting a higher score. Put differently, this indicates that for Item 96, it is only 0.5 logits more difficult to get a score of 2 than a 1 and this difference will have a minimal impact on a student's probability of giving the desired response, while, on Item 75, student probability drops significantly from the scoring a 1 to scoring a 2.

In order to graphically examine other assessment characteristics that may be contributing to item difficulty, I organized the Wright map by *Context* (i.e. genre) and then by *Aspect* (i.e. reading content) to detect whether any patterns of difficulty appear to arise in these alternative viewpoints as compared to the previous Wright map organized by *Item Format*. Logit differences

between subcategories of *Context* (see Figure 4) and subcategories of *Aspect* (see Figure 5) are marginal, and there does not appear to be a hierarchy of difficulty between subcategories on either map as opposed to the hierarchy that appeared across subcategories of *Item Format*.

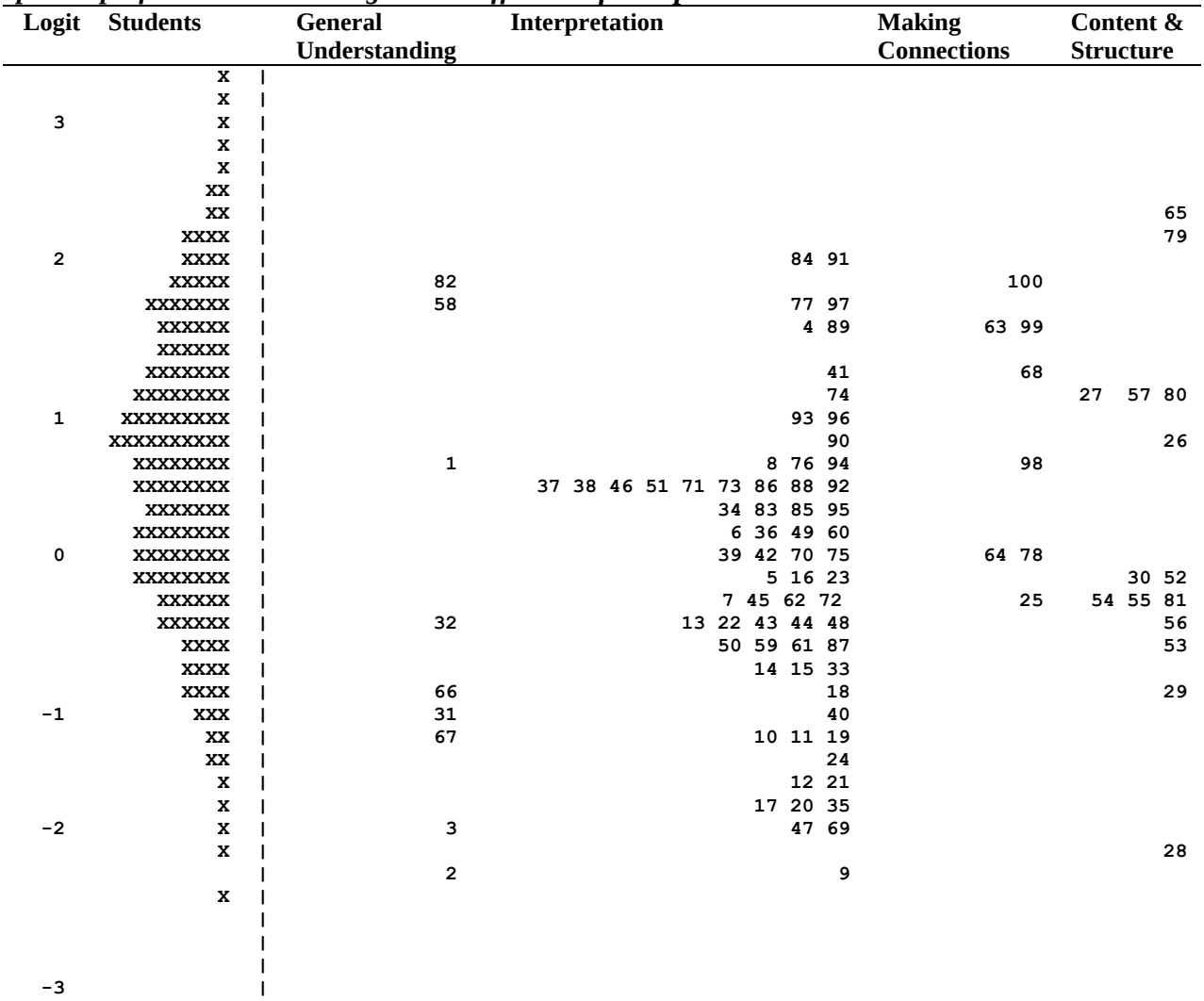
Across *Context*, estimated item difficulty spans ~-2.0 to 2.5 logits for Literary texts and spans ~-1.5 to 2 logits for Informational, meaning that items on the Literary texts have a greater span of difficulty than their counterpart: Literary spans ~0.5 logits easier to 0.5 logits more difficult than Informational (see Figure 4).

Figure 4. Wright map of item difficulty by Context. This figure illustrates the distribution of person proficiencies and averaged item difficulties for Literary and Informational items.

Logit	Students		Literary	Informational
	X	|		
	X	|		
3	X	|		
	X	|		
	X	|		
	XX	|		
	XX	|	65	
	XXXX	|	79	
2	XXXX	|		84 91
	XXXXX	|		82 100
	XXXXXXX	|	58 77	97
	XXXXXX	|	4 63	89 99
	XXXXXX	|		
	XXXXXXX	|		41 68
	XXXXXXXX	|	27 74 80	57
1	XXXXXXXXX	|		93 96
	XXXXXXXXX	|	26	90
	XXXXXXXX	|	1 8 76	94 98
	XXXXXXXX	|	71 73	37 38 46 51 86 88 92
	XXXXXXX	|		34 83 85 95
	XXXXXXXX	|	6 60	36 49
0	XXXXXXXX	|	64 70 75 78	39 42
	XXXXXXXX	|	5 16 23 30	52
	XXXXXX	|	7 25 62 72 81	45 54 55
	XXXXXX	|	13 22	32 43 44 48 56
	XXXX	|	59 61	50 53 87
	XXXX	|	14 15	33
	XXXX	|	18 29	66
-1	XXX	|		31 40
	XX	|	10 11 19	67
	XX	|	24	
	XX	|	12 21	
	X	|	17 20	35
	X	|	3 69	47
-2	X	|	28	
	X	|	2 9	
		|		
	X	|		
		|		
		|		
		|		
-3		|		

Across *Aspect*, there is even greater uniformity between categories: estimated item difficulty is generally -2.0 to 2 logits with the exception of Making Connections which has a couple more difficult items in the logit 2.5 range (see Figure 5). In later analysis, *Context* and *Aspect* will be further examined for multidimensionality and compared with a multidimensional *Item Format* model.

Figure 5. Wright map of item difficulty by Aspect. This figure illustrates the distribution of person proficiencies and averaged item difficulties for Aspect items.



***Summary*.** Model 1 creates a baseline model to which compare all subsequent alternative models. Traditional item analysis demonstrates that the NAEP instrument has good reliability and that the items are operating as expected with minimal problems. Of the 100 items, few items are misfitting: 3 items are underfit, 2 items overfit. Misfit analysis will be revisited when individual items are later analyzed for DIF. In the baseline analysis, we establish an overall mean performance for all students across all items of 0.549 (1.140) logits. The Wright map illustrates a hierarchy of item format difficulty with MC < SCR < ECR which evidences how item format may affect student reading proficiencies for all 4th graders. This suggests that we would expect to see students with lower proficiencies getting more MC than CR items correct while conversely see these same students having more difficulty with ECR items. As opposed to *Item* Format, no patterns of item difficulty emerge within the subcategories of reading *Aspect* or *Context* which suggests that students should perform equally well or equally poorly across all subcategories. The following section describes results from the latent regression partial credit models.

Model 2: Unidimensional Latent Regression

In this analysis we establish if there is a significant mean performance difference on all items between the reference and focal group. Both variables of English language proficiency and status as a student diagnosed with Specific Learning Disability were regressed on latent ability. The values reported here are the parameter estimates for each of the populations. The variables of ELL and SLD were coded as '0' for no and '1' for yes.

Table 6
Relative Model Fit: Regression versus Baseline Model

Model	Final Deviance	Estimated Parameters	AIC	BIC
Baseline	4,875,366	136	4,875,638	4,877,020
Regression ELL	4,923,533	137	4,923,807	4,925,199
Regression SLD	4,925,867	137	4,926,141	4,927,533
Regression ELL & SLD	4,915,082	138	4,915,358	4,916,760
Regression ELL*SLD	4,915,022	139	4,915,300	4,916,712

Note. Regression ELL compared with Baseline model
(AIC): 4,923,807 > 4,875,638 and (BIC): 4,925,199 > 4,877,020.
Regression SLD compared with Baseline model
(AIC): 4,926,141 > 4,875,638 and (BIC): 4,927,533 > 4,877,020.
Regression ELL&SLD compared with Baseline model
(AIC): 4,915,358 > 4,875,638 and (BIC): 4,916,760 > 4,877,020.
Regression ELL*SLD compared with Baseline model
(AIC): 4,915,300 > 4,875,638 and (BIC): 4,916,712 > 4,877,020.

Regression. Comparison of the AIC and BIC values indicate that none of the regression models are significantly better fitting than the Baseline model (see Table 6). What we are most interested in, however, is if there is any change in population mean as a result of group membership; these calculations are valid even though the regression models are not significantly better fitting. When status as an ELL student was regressed, the latent ability of NE is estimated at 0.634 (1.102) logits. Latent ability for ELLs is estimated -1.044 (1.102) logits below that of NE speakers placing ELLs at a mean ability of -0.410 logits (see Table 7). Similar results appear when status as a student diagnosed with SLD is regressed suggesting that the reference group outperforms the focal group, the lower proficiency group. Non-SLD students are estimated at 0.602 (1.110) logits. Latent proficiency falls -1.259 (1.110) logits for SLD from the reference group status placing their estimated ability at -0.657 logits.

Table 7

Parameter Estimates: Baseline versus Unidimensional Latent Regression Models

Model	Population Mean (SE)	Logit Difference(s)	SD
Baseline	0.549 (0.003)		1.140
Regression			
NE	0.634 (0.003)		1.102
ELL	-0.410	-1.044 (0.009)	
Regression			
Non-SLD	0.602 (0.003)		1.110
SLD	-0.657	-1.269 (0.013)	
Regression			
NE +Non-SLD	0.685 (0.003)		1.074
ELL	-0.343	-1.028 (0.009)	
SLD	-0.550	-1.235 (0.012)	
Regression			
NE +Non-SLD	0.686 (0.003)		1.073
ELL	-0.359	-1.045 (0.009)	
SLD	-0.583	-1.269 (0.013)	
ELL*SLD	-1.283	0.345 (0.040)	

When the variables ELL and SLD are entered into model simultaneously, the result is that latent ability on all items for Native English speakers and Non-SLD students is estimated at 0.685 (1.074) logits while the mean latent ability for ELLs drops a little over one logit and almost one and a quarter logits for SLDs. This indicates that group membership negatively affects latent ability on the reading comprehension assessment. Status as a student diagnosed with SLD has a slightly greater negative impact on estimated proficiency than status as an ELL student.

Interaction Effect: ELL*SLD. Eight hundred and fifty eight students were both ELL and SLD, so I calculated the interaction effect. This model estimated mean ability for NE and non-SLD students at 0.686 (1.073) logits. Latent regression indicates that the mean of ELLs is -0.359 logits and SLDs -0.583 logits, again demonstrating that group membership negatively affects latent ability on the reading comprehension assessment. The mean ability of an individual whose status is both ELL and SLD drops to -1.283 logits from the mean ability, which is a larger drop than ELL or SLD individually, though not quite as much as the sum of the ELL and SLD mean indicating that there is an interaction effect for ELL and SLD.

Summary. Regression analysis indicates that there are significant and detrimental effects of group membership for both ELL and SLD status on estimated student reading comprehension proficiencies, with SLD status having a slightly greater impact: SLDs < ELLs < Non-ELLs & SLDs. Up to this point we know that there is a difference in item format difficulty and that the focal groups are not performing overall as well as the reference group. From this point forward, what we want to know is how the focal groups perform across each of the *Item Format*s. To investigate the question, we follow two paths. One method is to see if item formats are differentially favoring one group over the other. The second method is to analyze the unidimensional model for dimensionality across the item formats; this process will highlight estimated differences in proficiency across the groups within each of the three item formats.

Model 3: Differential Item Functioning

DIF analysis reveals whether items are equally difficult for persons of the focal group compared to the reference group when pairing individuals of equal ability (Wilson, 2005). In this set of analyses, we are able to estimate –just as with regression- the overall mean performance on all items between the reference and the focal groups. We are also able to detect items that have a significantly differential function for subpopulations, and to look for patterns of this effect across the Assessment Characteristics of *Item Format*, *Context*, and *Aspect*. Both DIF models were better fitting than the Baseline model per the AIC and BIC values (see Table 8).

Table 8
Relative Model Fit: DIF versus Baseline Model

Model	Final Deviance	Estimated Parameters	AIC	BIC
Baseline	4,875,366	136	4,875,638	4,877,020
DIF ELL	4,861,245	236	4,861,717	4,864,115
DIF SLD	4,864,604	236	4,865,076	4,867,474

Note. DIF ELL compared to Baseline model
(AIC): 4,861,717 < 4,875,638 and (BIC): 4,864,115 < 4,877,020.
DIF SLD compared to Baseline model
(AIC): 4,865,076 < 4,875,638 and (BIC): 4,867,474 < 4,877,020.

Population mean. Items were analyzed for DIF in two models: the first model compared ELL and NE populations and the second model compared the SLD and non-SLD populations. The population means illustrated below note the mean performance difference between subgroups across all items. The DIF model for ELLs estimates that NEs scored 1.048 (1.104) logits higher than ELLs for a mean proficiency of NE = 0.636 and ELL = -0.412 logits. That difference is large at almost 1 student standard deviation. The DIF model estimates slightly lower estimated abilities for students diagnosed with SLD: non-SLDs scored 1.256 (1.112) logits higher than SLDs for a mean proficiency of Non-SLD =0.604 and SLD =-0.652 logits (See Table 9). This difference is also large as it is greater than 1 student standard deviation. Both analyses were consistent with the regression findings for the ELL and SLD models.

Table 9

Parameter Estimates: Baseline versus DIF Model for ELL and SLD Students

Model	Population Mean (SE)	Logit Difference (SE)	SD
Baseline	0.549 (0.003)		1.140
DIF ELL			1.104
0	0.636	0.524 (0.001)	
1	-0.412	-0.524 (0.001)	
		1.048	
DIF SLD			1.112
0	0.604	0.628 (0.001)	
1	-0.652	-0.628 (0.001)	
		1.256	

Note. 0 = Reference group (Non-ELL & Non-SLD);
1 = Focal group (ELL and SLD)

DIF significance. One goal of the DIF analysis is to investigate the presence of differential functioning between students of equal proficiency in the reference and focal group for each and every item. ConQuest calculates the logit difference for each item, and it is the magnitude of the logit difference that determines the level of DIF significance. To calculate the magnitude, I applied a standard effect size translated by Paek (2002) specifically for the Rasch model:

Level A: Little or no DIF when $\delta_{ig} < .426$,

Level B: Slight to moderate DIF when $.426 \leq \delta_{ig} < .638$, and

Level C: Moderate to large DIF when $\delta_{ig} \geq .638$.

The distinction between Slight to Moderate DIF at Level B and Moderate to Large DIF at Level C is most useful for assessment design. For purposes of exploring the presence of existing DIF the results discussed here collapse Level B and C DIF items into a single DIF category (see full DIF results in Appendix G and H). The ELL model flagged a total of 12 out of the 100 NAEP items as having DIF, and the SLD model flagged a total of 16 of the 100 items. Because the ELL and SLD models are subject to similar analyses it is useful to lay out results side-by-side but with the understanding that these models are mutually exclusive and comparisons cannot be made across the two models.

Table 10
Total Number of DIF Items Flagged in the ELL and SLD Model

Assessment Characteristics	Subcategory (# of items/100)	Number & Percentage of DIF items	
		ELL/NE Model	SLD/Non- Model
FORMAT	Multiple-choice 57	11 92%	13 81%
	Short-constructed 11	0	1 6%
	Extended-constructed 32	1 8%	2 13%
CONTEXT	Literary 51	7 58%	11 69%
	Informational 49	5 42%	5 31%
ASPECT	General Understanding 9	0	1 6%
	Interpretation 68	8 67%	12 75%
	Connections 8	0	0
	Content/structure 15	4 33%	3 19%
Total	100	12	16

Note. Reported percent is out of the total number of DIF items in each model: ELL (=12) and SLD (=16).

Patterns across Item Format subcategories. What is of great interest here is how DIF is distributed across the three *Item Formats.* Table 10 illustrates the DIF findings by reporting the number of flagged items across the different assessment characteristics in the ELL and the SLD models. In the ELL model, the majority of DIF items are in the MC format at 92% (11/12 items). There were no flagged SCR items. And, only 1 ECR item flagged. When comparing flagged DIF across for *Item Format* for SLD and non-SLD populations, just as with the previous analysis, a high proportion of the DIF items are the MC format; 81% (13/16 items). One SCR item and 2 ECR items are flagged for DIF. In subsequent analysis we will examine the directionality of the DIF items. Specifically, because so many MC items were flagged, it will be interesting to see if they tend to favor the reference or the focal group. The CR items are the *Item Format* of particular interest, but because so few items were flagged, any conclusions drawn can only be tentative.

Patterns across Context and Aspect subcategories. In order to account for effects that may be attributed to other assessment characteristics, I examine the distribution of DIF across subcategories of *Context* and *Aspect* in both the ELL and SLD models. As shown in Table 10,

DIF items in the ELL model are relatively split between the two Contexts. In the same model, a large number of DIF items fall into the Aspect subcategory of Interpretation, but this number, however, is in proportion to the large quantity of items in this category. On the other hand, more DIF items than the expected fall into the Aspect subcategory of Content and Structure: this category constitutes only 15% of the total items, whereas 33% of the DIF items fall into the same category. In the SLD model, we see a 70/30 split with 70% of the DIF items falling into Literary Context and only 30% into Informational; this is out of proportion to the relatively even split of items in each of these categories. What will be interesting in subsequent analysis is to examine which of the two membership groups Literary items favor. In the same SLD model, we see the presence of DIF items across the *Aspect* subcategories in relative proportion to their quantity, with the exception of Connections which constitute 8% of the total items, but here there are 0 DIF items.

DIF directionality. Once items have been flagged for DIF, it is of particular interest to evaluate which membership group the DIF items favor. The parameter estimate produced by ConQuest is expressed on a scale in which positive values mean that the item is easier for members of the reference group (NE speakers and non-SLD individuals) than for matched ability members of the focal group (ELLs and individuals diagnosed with SLD), whereas, negative values indicate the reciprocal. For points of comparison, Tables 11 and 12, respectively, illustrate the number of DIF items favoring each membership group in the ELL model and the SLD model across each of the assessment characteristics of *Format*, *Context* and *Aspect*.

In the ELL model, 7 DIF items favor the NE group while 5 DIF items favor ELLs (see Table 13). Items 14, 15, 20, 40, 44, 56, 91 favored ELL students and Items 4, 7, 26, 27, 57 favored NEs (refer to Table 1 for Passage and Item Characteristics). We saw earlier that the large proportion of MC items demonstrated DIF. Table 11 illustrates that favorability of the MC DIF items is fairly evenly split between group membership (55% to 45%). The one ECR item favors ELLs. In Context, a greater percentage of Informational DIF items favor ELLs at 80%. In Aspect, a greater percentage of Interpretation DIF items favor ELLs (75%), while 75% of Content & Structure DIF favor NEs.

Table 11
Directionality of the 12 DIF Items Favoring ELL versus NE Students

Assessment Characteristic	Subcategory (# of DIF)	ELL			NE		
		DIF Level B	DIF Level C	Total	DIF Level B	DIF Level C	Total
FORMAT	Multiple-choice 11	6 55%	0	6 55%	3 27%	2 18%	5 45%
	Short-constructed 0	0	0	0	0	0	0
	Extended-constructed 1	1 100%	0	1 100%	0	0	0
CONTEXT	Literary 7	3 43%	0	3 43%	2 29%	2 29%	4 57%
	Informational 5	4 80%	0	4 80%	1 20%	0	1 20%
ASPECT	General Understanding 0	0	0	0	0	0	0
	Interpretation 8	6 75%	0	6 75%	1 13%	1 13%	2 25%
	Connections 0	0	0	0	0	0	0
	Content/structure 4	1 25%	0	1 25%	2 50%	1 25%	3 75%
Totals	**12 Total DIF**	**7**	**0**	**7 Favor ELL**	**3**	**2**	**5 Favor NE**

Note. Reported percent is out of the total number of DIF items in the ELL model (=12).

In the SLD model, the 16 DIF items are evenly split between the two subgroups. Table 14 illustrates that Items 9, 12, 14, 17, 33, 59, 69, and 91 favored SLDs and Items 1, 4, 7, 26, 27, 38, 42, 57 favored Non-SLDs (refer back to Table 1 for Passage and Item Characteristics). In terms of *Item Format*, the SLD model has a higher percentage of MC DIF items favoring the Non-SLD group at 62% (see Table 12). The one SCR and two ECR items favor SLDs. *Context* is relatively split between the two groups. In *Aspect*, General Understanding and Content & Structure DIF favor Non-SLDs while 67% of Interpretation items favor SLDs.

Table 12
Directionality of the 16 DIF Items Favoring SLD versus Non-SLD Students

Assessment Characteristic	Subcategory (# of DIF)	SLD			Non-SLD		
		DIF Level B	DIF Level C	Total	DIF Level B	DIF Level C	Total
FORMAT	Multiple-choice 13	5 38%	0	5 38%	2 15%	6 46%	8 62%
	Short-constructed 1	1 100%	0	1 100%	0	0	0
	Extended-constructed 2	2 100%	0	2 100%	0	0	0
CONTEXT	Literary 11	6 55%	0	6 55%	1 9%	4 36%	5 45%
	Informational 5	2 40%	0	2 40%	1 20%	2 40%	3 60%
ASPECT	General Understanding 1	0	0	0	1 100%	0	1 100%
	Interpretation 12	8 67%	0	8 67%	1 8%	3 25%	4 33%
	Connections 0	0	0	0	0	0	0
	Content/structure 3	0	0	0	0	3 100%	3 100%
Totals	**16 Total DIF**	**8**	**0**	**8 Favor SLD**	**2**	**6**	**8 Favor Non-**

Note. Reported percent is out of the total number of DIF items in the SLD model (=16).

DIF crosstabulation across Item Format, Aspect, and Context. The following table crosstablulates the pattern of flagged DIF items for NE versus ELLs across *Item Format*, *Aspect*, and *Context* to more clearly illustrate the findings of the last two sections. This table emphasizes how most DIF items were of the MC format, yet are dispersed across the other subcategories of *Context* and *Aspect*. All flagged items have a corresponding (+) or (-) sign indicating directionality. A (+) sign next to the item number indicates the item favors NE and a (-) sign indicates it favors ELLs. Within the MC format, we see a mixture of DIF items favoring both subgroups indicating there is no real pattern of favorability. The only flagged ECR item favors the focal group, ELLs.

Table 13
DIF Items by Format, Context, and Aspect Characteristics for NE (+) and ELL (-)

	MC		SCR		ECR		Total Aspect
	Literary	Informational	Literary	Informational	Literary	Informational	
General Understanding	1 2 3	31 32	58	66 67		82	**0**
Interpretation	4+ 5 6 7+ 8 9 10 11 12 13 14- 15- 16 17 18 19 20- 21 22 23 24	33 34 35 36 37 38 39 40- 41 42 43 44- 45 46 47 48 49 50 51	59 60 61 62		69 70 71 72 73 74 75 76 77	83 84 85 86 87 88 89 90 91- 92 93 94 95 96 97	**8**
Connections	25		63 64	68	78	98 99 100	**0**
Content and Structure	26+ 27+ 28 29 30	52 53 54 55 56- 57+	65		79 80 81		**4**
Total Context	**7**	**4**	**0**	**0**	**0**	**1**	
Total Format	**11 MC**		**0 SCR**		**1 ECR**		**12**

The next table crosstablulates the pattern of flagged DIF items in the second model and also denotes directionality with a (–) or a (+) to indicate which subgroup the item favors. Again, this table emphasizes how most DIF items were of the MC format but spread across the subcategories of *Context* and *Aspect*. Overall, there is a spread in favorability of MC DIF items

between the two membership groups. But, all three flagged constructed-response items favor the focal group, SLDs.

Table 14
DIF Items by Format, Context, and Aspect Characteristics for SLD (-) and Non (+)

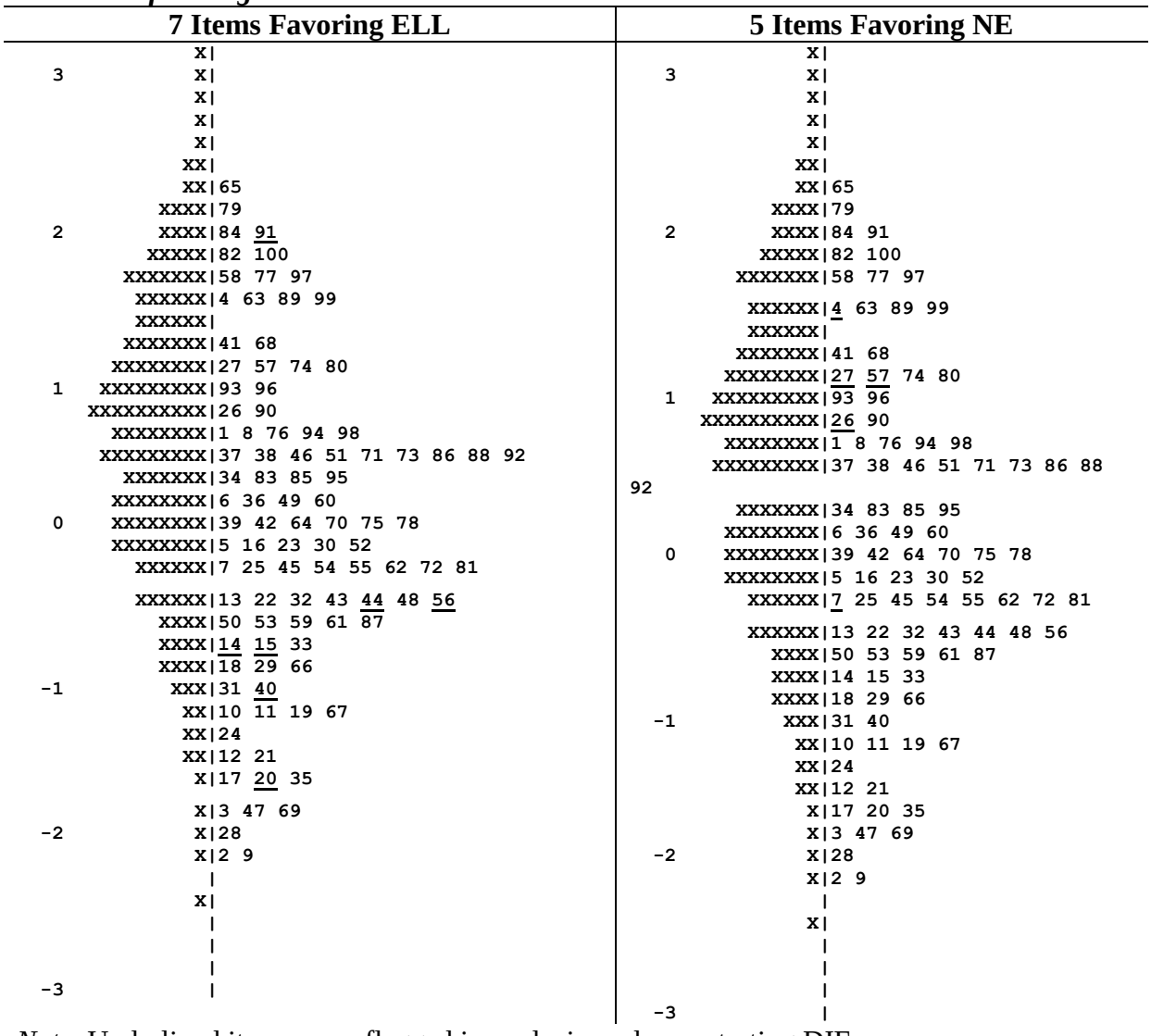
	MC		SCR		ECR		Total Aspect
	Literary	Informational	Literary	Informational	Literary	Informational	
General Understanding	1+ 2 3	31 32	58	66 67		82	**1**
Interpretation	4+ 5 6 7+ 8 9- 10 11 12- 13 14- 15 16 17- 18 19 20 21 22 23 24	33- 34 35 36 37 38+ 39 40 41 42+ 43 44 45 46 47 48 49 50 51	59- 60 61 62		69- 70 71 72 73 74 75 76 77	83 84 85 86 87 88 89 90 91- 92 93 94 95 96 97	**12**
Connections	25		63 64	68	78	98 99 100	**0**
Content and Structure	26+ 27+ 28 29 30	52 53 54 55 56 57+	65		79 80 81		**3**
Total Context	**9**	**4**	**1**	**0**	**1**	**1**	
Total Format	**13 MC**		**1 SCR**		**2 ECR**		**16**

DIF by item difficulty. The range of difficulty for flagged items was explored to see if there was any correlation between difficulty and DIF (Items underlined were flagged for Level 2 or 3 DIF). The Wright map used in this analysis (see Figure 6) illustrates the mean item difficulty across steps, meaning that ECR item steps are plotted according to average difficulties. Overall,

the range of item difficulty for all items is approximately -2 to 2.25 logits. In the ELL analysis, DIF items demonstrate a more narrow range of difficulty at -1.5 to 2.0 logits, meaning this group of items contained neither the most difficult nor contain the easiest items. DIF items favoring NE range from -0.5 to 1.5 logits and those favoring ELL range -1.5 to 2.0 logits.

While it may seem that on average that the DIF items favoring ELLs are harder, a closer look at where the majority of DIF items for each subgroup cluster paints a different picture. Figure 6 clearly illustrates a large logit difference in difficulty of DIF items between the two membership groups. If you exclude Item 91 (as it appears to be an outlier), the range of items favoring ELLs drops significantly down to -2.0 and 0.5 logits. Item 7 at 0.5 logits is the least difficult DIF item favoring NEs, but it marks the top range of difficulty for the ELL group.

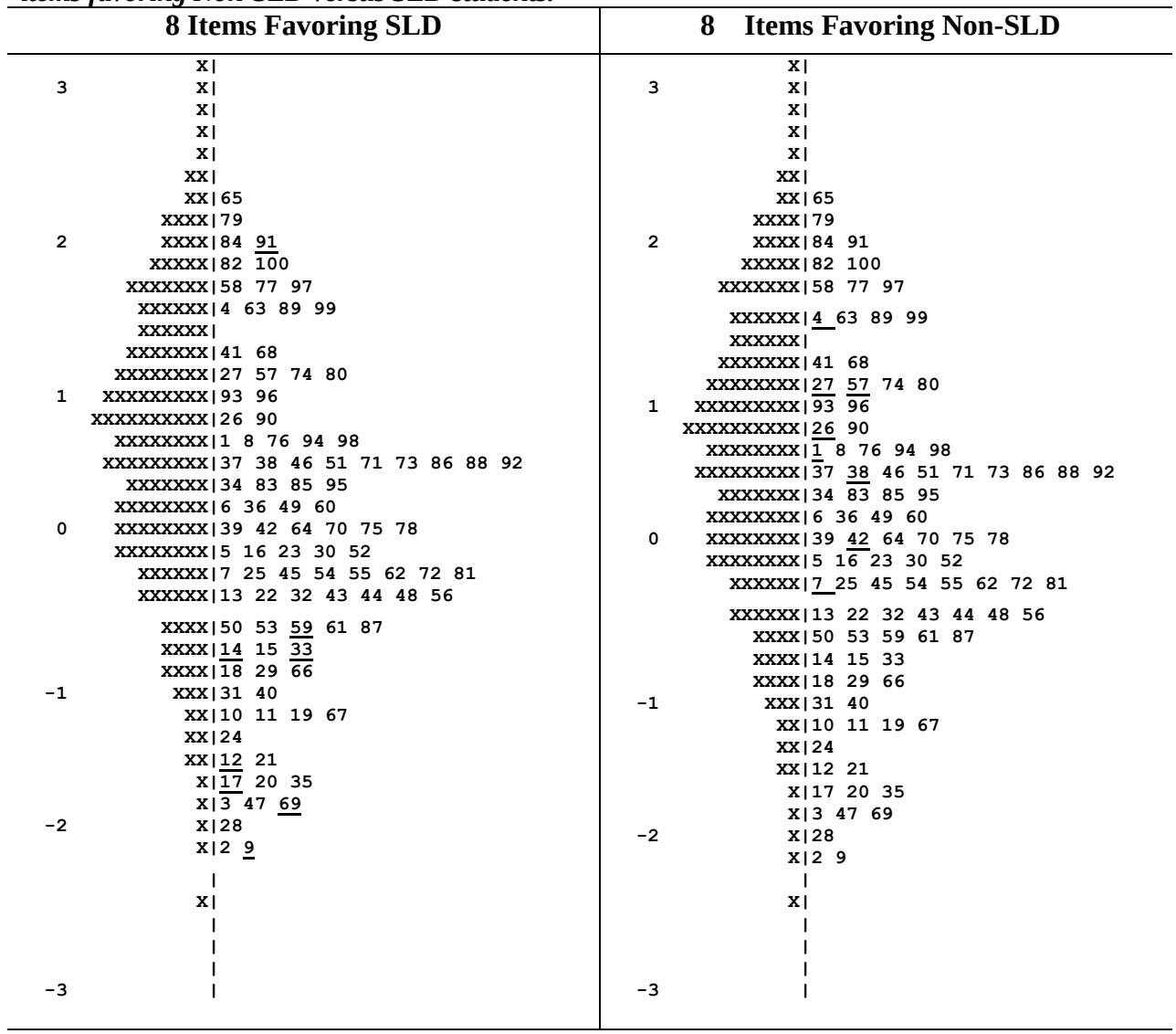
Figure 6. Wright map of ELL model DIF by difficulty. Compares the range of difficulty for DIF items favoring NE versus ELL students.



Note. Underlined items were flagged in analysis as demonstrating DIF.

In the SLD analysis, DIF items range from -2.0 to 2.0 logits. Just as in the ELL model, DIF items favoring the SLDs comprise the lower half of the range, and DIF items favoring the non-SLDs comprise the top half of the range (see Figure 7). We see that one of the flagged DIF items favoring SLDs is also one of the easiest items, Item 9. However, Item 91 is one of the most difficult items and also favors SLDs. If you exclude Item 91 (again based on the appearance that it is an outlier), the range of items favoring SLDs drop significantly down to -2.0 and 0.5 logits. Figure 7 clearly illustrates a large logit difference in difficulty of DIF items between the two groups. Similar to the previous analysis, Item 7 at 0.5 logits is the least difficult DIF item favoring Non-SLDs, but it marks the top range of difficulty for the SLD group.

Figure 7. Wright map of SLD model DIF by difficulty. Compares the range of difficulty for DIF items favoring Non-SLD versus SLD students.



Note. Underlined items were flagged in analysis as demonstrating DIF.

***DIF comparison to misfit analysis*.** Few misfitting items appear in the DIF analysis. Review of these items reveals no interesting patterns. One underfitting item is flagged for DIF in

both analyses: Item 91; it is also the item that appears to be an outlier. Item 91 is an ECR item from the Informational Context and Interpretation Aspect. The underfitting Item 12 also appears in the SLD analysis. It is an MC item from the Literary Context and is also from the Interpretation Aspect. Item 12 is a released item. The NAEP Codebook only gives the gist of the item. It reads, "When worker speaks to her why Rosa feels proud?"

***Summary*.** Consistent with the regression analysis, DIF analysis demonstrates that group membership affects population mean with both focal groups significantly underperforming by more than 1 logit. Twelve percent of the items in the ELL model demonstrate DIF, and 16% in the SLD model. DIF items were generally split in favor of the reference group and focal group (5 versus 7; 8 versus 8). Distinctive patterns of DIF difficulty emerge from the Wright map in both analyses demonstrating that DIF items favoring the reference group were generally more difficult items that clustered near to and well above the mean, while those favoring the focal group clustered at 0.5 logits below the mean. On one hand, this means that the focal groups are performing better than expected on easy items while the reference group is performing worse than expected on them. Conversely, the reference group is scoring higher than expected on the harder items, and the focal groups are scoring lower than expected on them. An investigation of *Item Format* differences demonstrates that in both models a disproportionately large number of DIF items were of the MC format, while few DIF items were of the SCR or ECR format. While only few in number, both models show all DIF SCR and ECR items favoring the focal groups. In contrast, two-thirds of the MC DIF items favor the Non-SLD group, although no patterns of MC DIF emerged in the ELL model. And, while DIF analysis in both the ELL and SLD model hint there may be an interaction between *Item Format* and group membership, we do not ultimately know what effect the DIF items have on overall student proficiencies; thus, the possibility of an interaction effect will be further investigated in the multidimensional analysis. A few patterns are revealed when DIF is organized by the Reading Framework components. In the ELL model, Context: Informational items favor ELLs, but in the SLD model Context equally favors both groups. In both models items from the Aspect: Interpretation favor the focal group while those items from the Aspect: Content and Structure favor the reference. With little variability in difficulty between the subcategories, it is difficult to interpret these findings.

Model 4: Multidimensionality

The NAEP assessment of Reading Comprehension is a unidimensional test that is comprised of three *Item Formats* designed to assess a subset of reading *Aspects* across two different *Contexts.* The purpose of this section is to test the basic unidimensional assumption by fitting to multidimensional model to the data in order to account for an effect of the various assessment characteristics. The multidimensional approach challenges the premise that all persons and items belong on a single continuum of latent ability that together estimate an individual's overall propensity to answer an item correctly and an items' overall probability of being answered correctly. The first series of multidimensional models examines whether it is statistically useful to model *Item Format,* the focal assessment variable, as separate latent dimensions.

Item Format. Employing a model developed by Adams, Wilson, and Wang (1997) called a *multidimensional between item model,* I will model the subscales of *Item Format* as distinct, latent dimensions with each item belonging to only one subscale (Wu, 1997). As opposed to modeling several consecutive unidimensional models, the benefit of the single multidimensional model is that the interrelationship between the dimensions is calculated as a correlation. An additional benefit of modeling the data multidimensionally is that reliabilities similar to the unidimensional model are sustained; as opposed to modeling *Item Formats* separately which causes reliabilities to drop substantially due to the loss of information (Briggs & Wilson, 2003). Following Multidimensional analysis of *Item Format,* for exploratory purposes, I will also model a second and third series of multidimensional models to investigate the presence of independent dimensions among the subdomains of *Context* and *Aspect.*

Multidimensional analysis of *Item Format* unfolds in three stages. The most parsimonious multidimensional model is two-dimensional and considers ability as assessed by the multiple-choice items as one latent outcome and ability as assessed by the constructed-response items as a second latent outcome. The constructed-response outcome pairs both short-constructed response items which generally are single word to single sentence responses compared with the extended-constructed response items that require a more elaborate written response. Subsequently, I will analyze a three-dimensional model that in addition to calculating ability assessed by multiple-choice items as a latent outcome, will separate ability as assessed by short-constructed response and ability as assessed by extended-constructed response into two separate dimensions. A third model will investigate a three-dimensional model separating MC, SCR, and ECR items that will regress the three dimensions onto two examinee background variables: ELL and SLD status; this is the most complex of the three models as it will compare latent outcomes between the reference and focal groups for each of the dimensions indicating, to some extent, the impact of item format on student reading comprehension proficiencies. All analyses use a partial credit model because items are a mixture of dichotomous and polytomous items.

Two-dimensional model. Table 15 displays the relative fit of the Multidimensional Format models. The Two-Dimensional Format model is better-fitting than the Unidimensional Baseline model.

Table 15
Relative Model Fit: Multidimensional Format versus Baseline Model

Model	Final Deviance	Estimated Parameters	AIC	BIC
Baseline	4,875,366	136	4,875,638	4,877,020
Format –Two Dimensional	4,871,485	138	4,871,761	4,873,163
Format - Three Dimensional	4,870,058	141	4,870,340	4,871,772
Format - Three Dimensional with Regression	5,532,770	147	5,533,064	5,534,557

Note. Format Two Dimensional without Regression compared to Baseline model (AIC): 4,871,761 < 4,875,638 and (BIC): 4,873,163 < 4,877,020.
Format Three Dimensional without Regression compared to Baseline model (AIC): 4,870,340 < 4,875,638 and (BIC): 4,871,772 < 4,877,020.
Format Three Dimensional with Regression compared to Baseline model (AIC): 5,533,064 > 4,875,638 and (BIC): 5,534,557 > 4,877,020.

As indicated in Table 16, there is a high correlation between the MC dimension and the CR dimensions with 93% of the variance shared, thus indicating that it is not necessary to treat the two dimensions separately as there is very little practical difference between them. The examination of correlations between dimensions essentially signals the level of consistency in student proficiency across each dimension; it suggests that people who did best on the MC dimension also did best on the CR dimension, while people who did poorly on one item format also did poorly on the other item format.

Table 16
Correlations between Item Format Dimensions

Model	Correlation of Dimensions		
Format: Two-Dimensional		**MC**	
	CR	93%	
Format: Three-Dimensional		**MC**	**SCR**
	SCR	92%	
	ECR	92%	95%
Format: Three-Dimensional with Regression		**MC**	**SCR**
	SCR	88%	
	ECR	91%	90%

The estimated mean ability of the MC dimension is 1.045 (1.259) logits and the CR dimension is -0.053 (1.075) logits. The analysis of each dimension is centered on 0, thus we cannot compare abilities between the two dimensions because we do not know if the drop in latent ability for CR items is related to greater item difficulty in this dimension or a drop in person proficiency, or perhaps a little of both. But, a general statement can be made that an average student did better on an average item in the MC dimension than in the CR dimension. The variability of student abilities as assessed by the test items was larger in the MC dimension than the CR dimension.

Table 17

Parameter Estimates: Unidimensional, Baseline versus Multidimensional Format Model

Model	Population Mean (SE) Of Dimensions			SD		
Baseline	0.549 (0.003)			1.140		
Format: Two Dimensional	**MC**	**CR**		**MC**	**CR**	
All Ss	1.045 (0.003)	-0.053 (0.002)		1.258	1.075	
Format: Three Dimensional	**MC**	**SCR**	**ECR**	**MC**	**SCR**	**ECR**
All Ss	1.046 (0.003)	0.284 (0.002)	-0.158 (0.002)	1.259	1.298	1.035
Format: Three Dimensional	**MC**	**SCR**	**ECR**	**MC**	**SCR**	**ECR**
NE & Non-SLD	1.189 (0.003)	0.436 (0.003)	-0.031 (0.002)			
ELL diff ELL effect size	-1.102 (0.010) 0.876	-1.060 (0.010) 0.817	-0.954 (0.008) 0.921	1.190	1.261	1.035
SLD SLD effect size	-1.252 (0.013) 0.995	-1.474 (0.014) 1.136	-1.180 (0.011) 1.140			

Three-dimensional model. The Three-Dimensional Format model treats each of the three item formats as a separate latent outcome. This model is also better-fitting than the Unidimensional Baseline model (see Table 15). There is a high correlation between all dimensions ranging from 92-95% (see Table 16) again suggesting that there is little practical difference between the dimensions, thus a unidimensional approach of collectively calculating student proficiencies across all three item formats is justified.

The estimate of mean ability for the MC dimension is 1.046 (1.259), SCR dimension is 0.284 (1.298), and the CR dimension is -0.158 (1.035) logits (see Table 17). These proficiencies are very similar to the proficiencies calculated in each of the Consecutive, Unidimensional models of *Item Format* overviewed in Table 5. As expected, the mean proficiencies drop from MC<SCR<ECR indicating that they are moving in the expected direction given previous analyses of item difficulty. Yet, we cannot compare logit differences between the dimensions; the estimates are not mean performance differences because calculations may also be affected by variability in item difficulty between the separate dimensions. Variability of student abilities as assessed by the test items was largest in the SCR dimension, second largest in MC, and smallest in the ECR dimension.

Additionally, we see that the multidimensional model enhances reliability (see Table 18). In the multidimensional model, the reliability for each *Item Format* dimension comes closer to the baseline reliability of 0.857 (which is the standard reliability because it incorporates all the items possible in the scale) than each of the reliabilities calculated via the three consecutive models.

Table 18
Reliability Estimates for Baseline, Consecutive, and Multidimensional Item Format Models

		Item Format		
	All Items	**MC**	**SCR**	**ECR**
Baseline Model	0.857			
MC Consecutive Model		0.721		
SCR Consecutive Model			0.311	
ECR Consecutive Model				0.730
Multidimensional Models				
Two-Dimensional Format without Regression		0.863	0.881	
Three-Dimensional Format without Regression		0.879	0.886	0.900
Three-Dimensional Format with Regression		0.975	0.964	1.02

The most notable difference is for the SCR dimension which only has 11 items. The SCR consecutive model reliability (31%) is considerably enhanced in the multidimensional model (89% & 96%, without and with regression models respectively) because student responses are correlated with all the information available within the assessment. The MC dimension which has the greatest number of items has a smaller increase in reliability from the consecutive approach of 72% to the multidimensional model of 88% and 98% in the without and with regression models respectively; yet this is still a substantial increase in reliability because MC items account for only a little over half of the items.

Three-dimensional model with regression. In contrast to the previous multidimensional models, the Three-Dimensional Format with regression model is not better-fitting than the Unidimensional Baseline model (see Table 15). Again, the outcome of fit does not affect the validity of the multidimensional regression model. Of primary interest is the effect of group membership on *Item Format* proficiencies. The *Item Format* latent outcomes were regressed on two background variables: ELL and SLD. All three dimensions have similar and moderately-high correlations that range from 88-90% which suggests that even with the inclusion of regression variables, there is still no practical difference in separating the item formats; individuals who do well on one format tend to do well on the others, and those who do poorly, do poorly on all formats.

The estimate of mean ability for the MC dimension is 1.189 (1.191), SCR dimension is 0.436 (1.262), and the CR dimension is -0.031 (0.978) logits (see Table 17). This model calculates a large effect of group membership on *Item Format* proficiencies. Both subgroups of students do more poorly than their peers across the three dimensions. ELL students underperform on all three dimensions by approximately 1 logit and SLDs underperform by 1.2 to 1.5 logits (see Table 17). These differences are very large as they are around 1 standard deviation. Effect sizes calculated for subgroup differences demonstrates that for both ELL and SLD students the ECR dimension has a slightly greater logit difference: ELL effect size = 0.921 and SLD effect size =1.140.

Summary. The Two-Dimensional and Three-Dimensional Format models fit the data better than the unidimensional model, although the Multidimensional regression model does not. Estimations of the separate latent dimensions in all three Multidimensional *Item Format* models are so highly correlated that there is no practical difference (i.e. no loss of information) in reporting a single student proficiency. The Multidimensional Regression model indicates that there is a large effect of group membership on *Item Format* proficiencies, yet there is no

interaction between group membership and *Item Format* because groups perform similar across the various dimensions.

Table 19
Relative Model Fit: Multidimensional Context & Aspect versus Baseline Model

Model	Final Deviance	Estimated Parameters	AIC	BIC
Baseline	4,875,366	136	4,875,638	4,877,020
Context Two-Dimensional	5,322,744	138	5,323,020	5,324,422
Context Two-Dimensional with Regression	5,303,653	142	5,303,937	5,305,380
Aspect Four-Dimensional	4,949,670	145	4,949,960	4,951,433
Aspect Four-Dimensional with Regression	5,191,024	153	5,191,330	5,192,885

Note. Context Two Dimensional compared to Baseline model
(AIC): 5,323,020 > 4,875,638 and (BIC): 5,324,422 > 4,877,020.
Context Two Dimensional with Regression compared to Baseline model
(AIC): 5,303,937 > 4,875,638 and (BIC): 5,305,380 > 4,877,020.
Aspect Four Dimensional compared to Baseline model
(AIC): 4,949,960> 4,875,638 and (BIC): 4,951,433> 4,877,020.
Aspect Four Dimensional with Regression compared to Baseline model
(AIC): 5,191,330 > 4,875,638 and (BIC): 5,192,885 > 4,877,020.

Context. The second series of multidimensional models investigates the independency of Literary and Informational subdomains of *Context*. The data is fitted to two models, one with and one without regression. On the basis of AIC and BIC values, neither *Context* Multidimensional model fits the data as well as the Baseline model (see Table 19). In both *Context* models there are moderate correlations between the two dimensions: 77% without regression and 74% with regression (see Table 21). With such a large percent of the variance shared, the Literary and Informational dimensions are similar enough to be reported as a unidimensional score. It is worth noting, however, that there is a greater distinction between the dimensions of *Context* than the dimensions of *Item Format* which were highly correlated at 88-90%. In both models, variability of student abilities as assessed by the test items was slightly larger in the Literary dimension.

Table 20
Reliability Estimates of Context and Aspect Multidimensional Models

	All Items	Dimension 1	Dimension 2	Dimension 3	Dimension 4
Baseline Model	0.857				
Multidimensional Models					
Context without Regression		0.791	0.792		
Context with Regression		0.883	0.898		
Aspect with Regression		0.716	0.957	0.744	0.792

Table 21
Correlations between Context and Aspect Dimensions

Model		Correlation of Dimensions		
Context: Two Dimensional		**Lit**		
	Inf	77%		
Context: Two Dimensional with Regression		**Lit**		
	Inf	74%		
Aspect: Four Dimensional		**GU**	**I**	**C**
	I	70%		
	C	54%	71%	
	C&S	62%	75%	62%
Aspect: Four Dimensional with Regression		**GU**	**I**	**C**
	I	66%		
	C	50%	68%	
	C&S	56%	74%	60%
		GU	**I**	**C**

Note. Lit =Literary; Inf=Informational; GU=General Understanding; I=Interpretation; C=Connections, and C&S=Content and Structure

Context without regression. There is slight lose in reliability to 79% for each of the dimensions as opposed to a reliability of 86% in the Baseline model (see Table 20). When the latent variables of *Context* are modeled separately, the estimated mean proficiency for all students is 0.780 (1.227) logits for Literary text and 0.326 (1.133) logits for Informational text. The standard effect size shows the Literary text has greater effect on overall student proficiencies: $0.636 > 0.288$. While we do not know if the estimated difference is attributed to item difficulty or student proficiencies, it is clear that there is a notable difference between the two dimensions.

Table 22
Parameter Estimates: Unidimensional, Baseline versus Two-Dimensional Context Model

Model		Population Mean (SE) Of Dimensions		SD	
Baseline		0.549 (0.003)		1.140	
Context		**Literary**	**Informational**	**Literary**	**Informational**
	All Ss	0.780 (0.003)	0.326 (0.003)	1.227	1.133
	NE & Non-SLD	0.920 (0.003)	0.463 (0.003)		
	ELL diff **ELL effect size**	-1.032 (0.010) -0.841	-1.047 (0.009) -0.924	1.160	1.067
	SLD diff **SLD effect size**	-1.301 (0.013) -1.060	-1.204 (0.012) -1.063		

Context with regression. The slight lose in reliability in the previous *Context* model is recovered in the *Context* with regression: reliability increases from the Baseline of 86% to 88% for the Literary and 90% for the Informational dimensions (see Table 20). The estimate of mean ability for NE and Non-SLD students on the Literary dimension is 0.92 (1.160) logits and the Informational dimension is 0.463 (1.067) logits (see Table 22).

The latent outcomes of Literary and Informational ability were regressed on two background variables: ELL and SLD. This model calculates a large effect of group membership on *Context* proficiencies. Both subgroups of students do more poorly than their peers on both Literary and Informational dimensions. ELL students underperform by approximately 1 logit and SLDs underperform by 1.2 to 1.3 logits (see Table 22). These differences are very large as they are around 1 standard deviation. Effect sizes calculated for subgroup differences demonstrate that for both ELL and SLD students the Informational dimension has a slightly greater logit difference: ELLs Informational 0.924 > Literary 0.841 and SLDs Informational 1.063 > Literary 1.060.

Summary. Neither *Context* model fits the data as well as the Baseline model. The Literary and Informational dimensions in the models are correlated high enough to justify reporting student proficiencies as one score. The *Context* regression model demonstrates a large effect of group membership with both focal groups underperforming in comparison to their peers by one logit or more. Just as with *Item Format*, there does not appear to be an interaction between group membership and *Context* because effect size differences across the dimensions are very minimal.

Aspect. The third series of multidimensional models investigates the independency of the four subdomains of *Aspect*: General Understanding, Interpretation, Connections, and Content & Structure. The data is fitted to two models, one with and one without regression. On the basis of AIC and BIC values, neither Aspect model fits the data as well as the Baseline model (see Table 19). There is a low to moderate correlation among the four dimensions. Variability of student abilities as assessed by the test items was largest in the Connections dimension.

Aspect without regression. The category of General Understanding and Connections had the lowest correlation at 54% and Interpretation and Content & Structure had the highest correlation at 75%. The estimated mean ability of all students is 1.033 (1.808) logits for General

Understanding, 0.700 (1.408) logits for Interpretation, -0.256 (2.203) logits for Connections, and 0.410(2.139) logits for Content and Structure.

Table 23
Parameter Estimates: Unidimensional, Baseline versus Four-Dimensional Aspect Model

Model		Population Mean (SE) Of Dimensions				SD			
Baseline		0.549 (0.003)				1.140			
		GU	**I**	**C**	**C&S**	**GU**	**I**	**C**	**C&S**
Aspect									
	All Ss	1.033 (0.003)	0.700 (0.003)	-0.256 (0.003)	0.410 (0.003)	1.345	1.187	1.484	1.463
	NE & Non-SLD	1.143 (0.003)	0.833 (0.003)	-0.103 (0.003)	0.548 (0.003)				
	ELL diff	-0.918 (0.011)	-1.092 (0.009)	-1.021 (0.012)	-1.128 (0.011)	1.267	1.132	1.398	1.365
	ELL effect size	0.501	0.776	0.464	0.527				
	SLD diff	-1.336 (0.014)	-1.303 (0.013)	-1.274 (0.016)	-1.251 (0.015)				
	SLD effect size	0.739	0.925	0.578	0.585				

Note. GU=General Understanding; I=Interpretation; C=Connections, and C&S=Content and Structure

Aspect with regression. As expected we see a lose in reliability across three of the four dimensions when the data is modeled Multidimensionally for *Aspect*. This is likely due to the low number of items in three of the dimensions. The Interpretation dimension with increased reliability to 96%, however, is the only exception, but this dimension also comprised a large proportion of the NAEP items. There is a low to moderate correlation between the four dimensions (see Table 21). General Understanding and Connections had the lowest correlation at 50% and Interpretation and Content & Structure had the highest correlation at 74%. There is a much greater distinction between the dimensions of *Aspect* than the dimensions of *Item Format.*

The estimated mean ability of NE and Non-SLD students is 1.143 (1.267) logits for General Understanding, 0.833 (1.132) logits for Interpretation, -0.103 (1.399) logits for Connections, and 0.548(1.365) logits for Content and Structure. Just as with the Multidimensional *Item Format* and *Context* models, the *Aspect* model calculates a large effect of group membership on student proficiencies. Both subgroups of students do more poorly than their peers on all four dimensions. ELL students underperform by 0.918 to 1.128 logits and SLDs underperform by 1.251 to 1.336 logits (see Table 23). These differences are very large as they are around 1 standard deviation. Standardized effect sizes demonstrate a greater effect of the category of Interpretation for ELLs and for SLDs, but we do not know if the effect is related to a difference in item difficulty or a difference in student proficiencies across the dimensions.

Summary. Similar results to previous Multidimensional models emerge in the exploratory analysis of a Multidimensional *Aspect* model. The regression model does not fit as well as the Baseline model. There is an effect of group membership on *Aspect* proficiencies, however, there appears to be no interaction effect between group membership and *Aspect* as students perform similarly across the four dimensions. A distinct difference in the *Aspect* model, however, is that the dimensions have a much lower correlation than the Context and *Item Format* models.

Chapter 4: Discussion

Reading Comprehension is an essential and lifelong skill that for many children is acquired in school. It is, therefore, of great national interest how American students are faring. For over four decades the NAEP Reading Comprehension assessment has served as a register of that skill –and the news is not great: significant numbers of students continue to perform below what NAEP classifies as *Basic* levels. Typically, the lower end of the performance continuum is populated by Special Education students, English language learners, and students of poverty. It is not surprising that these populations underperform on assessments of Reading Comprehension; many of their academic struggles are related to the mastery of English language conventions. These students are of particular interest to educators and policymakers alike because they are the most vulnerable to low academic performance and high school drop-outs (Pellegrino, Jones, & Mitchell, 1999).

The primary goal of the NAEP assessment is to detect and report the status of student achievement and to track trends over time. This comprehensive set of national achievement data assists educators and legislators alike in the future crafting of efficacious educational agendas. Given the level of importance of the NAEP assessments, it is of the utmost importance that we better understand the nature of student competencies measured and the degree to which valid and relevant inferences about performance can be drawn. Increasing federal mandates to include greater numbers of English language learners and Special Education populations in large-scale assessment raises questions about the degree to which achievement scores reflect content knowledge versus English language proficiency (Mahoney, 2008; Snetzler & Qualls, 2000*).*

There is sufficient reason to believe that low levels of language proficiency will have a detrimental effect on the performance of CR items because they require students to produce a written response. Investigating such an effect is important because any performance differences attributable to skills outside of the target construct, (i.e., what psychometricians refer to as construct-irrelevant variance), can introduce item bias. Additional findings in *Item Format* effects are likely to have implications for the design of future assessments, including reading assessment. This study was an analysis of the 2007 NAEP reading achievement data via four different measurement models using the computer program ConQuest (Wu et al., 2007). This chapter provides an elaborated discussion of the results, possible explanations and implications of those results, limitations of the study, and tentative recommendations for future test development.

The preliminary Rasch model and traditional item analysis illuminated few measurement issues across the 100 NAEP Reading Comprehension items. Overall, the NAEP assessment demonstrated strong reliability, with only a few items falling slightly outside the expected fit index. In general items were acting within an acceptable range of fit, strengthening confidence in interpretations emerging from the NAEP dataset.

The literature on *Item Format* differences suggests that mixed item formats are beneficial in that they balance item efficiency with comprehensive content coverage, thus increasing the validity of the assessment. The Wright map produced in the Baseline analysis of the 4^{th} grade data demonstrated a wide range of item difficulty from -2.5 to 4 logits. This suggests that the practice of employing mixed item formats (i.e. dichotomous and polytomous items) on the NAEP assessment provides fairly comprehensive coverage across student proficiencies and demonstrates that the incorporation of CR items helps avoid ceiling effects given that the difficulty of two items exceed even the highest student proficiencies estimated at 3.5 logits. At

the lower end of difficulty, however, we see a floor effect where the two easiest items are plotted 0.5 logits above the ability of a group of students. This finding is the sort of evidence that has prompted the NAEP Validity Panel to undertake a line of work to create and evaluate a set of "easy reading blocks" in the fourth grade assessment; the hope is that these easier blocks will allow low-achieving students to contribute to the "information value" of the assessment by demonstrating the sorts of tasks that low achievers can actually perform. This illustration of item format hierarchy aligns with earlier research documenting the unique contributions of different *Item Format* levels. In general, MC items are on average easier (Rauch & Hartig, 2010) and contribute to efficiency (Lukehele et al., 1994), while polytomous items contribute to information over and above that of MC items even when collapsed into dichotomous categories (Donoghue, 1994; Ercikan et al., 1998; Hastedt, 2004).

The first research question addressed whether *Item Format* significantly influenced NAEP reading performance. When item difficulties on Wright map were organized by *Context* and *Aspect* there appeared to be no patterns of hierarchy between the subdomains. This is an interesting finding because these two assessment characteristics comprise NAEP's Reading Comprehension framework. Because there is little variability in difficulty across levels of *Context* (reading a. Literary and b. Informational texts) or *Aspect* (reading to a. identify a General Understanding, b. make Interpretations, c. make Connections, and d. identify Content and Structure), these elements appear to be parallel component skills of reading comprehension. This suggests that students should perform equally well or equally poorly across all elements, and any change in student proficiency would produce a similar change across all of the elements.

On the other hand, the Wright map suggested a very distinct hierarchy of *Item Format* difficulty across all students. Based on the range of difficulty, MC items rank as the easiest items capping out at 1.5 logits. SCR items rank second with a range up to 2.5 logits. And, the polytomous items (i.e. ECRs) are the most difficult capping out at almost 4.0 logits. Hence, we would expect to see lower proficiency students, for example, getting more MC items correct as compared to other item formats while conversely see the same students having more difficulty with ECR than the other item formats.

The findings from the baseline analysis provide evidence about how *Item Format* likely affects student reading proficiencies for all 4th graders, in particular with respect to the differences in the probability of correctly answering items in different formats. Overall, the probability of answering an item correctly greatly decreases from MC to SCR to ECR. The difference, for example, between the most difficult MC item and most difficult ECR item is a very large logit difference of 2.5 logits that translates into extremely different student probabilities of answering the MC item correctly than the ECR item at a particular score level. For example, an individual whose proficiency is aligned at 1.5 logits alongside the hardest MC item has a correct response probability of 50%, but for this same student, the probability of answering the most difficult ECR item at the highest level drops to just 2%. The great difference in probability is related to the large estimated logit difference between the items. But, for an individual whose ability is aligned with the most difficult ECR item, the probability of answering the ECR item at its highest level is 50% (rather than 2%) and their probability of answering the hardest MC item correctly is ~92% (rather than 50%) because it is 2.5 logits below their estimated ability.

The large logit difference between the most difficult ECR item and the most difficult MC item raise the possibility that CR items are being employed for more difficult subject matter (or perhaps more difficult ideas and relationships in the text) as suggested in the literature. But,

without further investigation, we cannot be certain whether *Item Format* differences in difficulty are driven by a) how they are paired with reading content; (b) differences in task demands (i.e. recognizing an answer in MC items versus producing a written response in CR items); or, (c) some other as yet unidentified aspects of reading comprehension that are not currently addressed in the Framework (e.g. see Hancock, 1994 on Bloom's taxonomy). Baseline analysis suggests that the hierarchy in *Item Format* difficulty is not attributed to, or directly related to, differences in *Context* or *Aspect.* The variability in item difficulty across *Item Format* that is not present in the other assessment characteristics- suggests that there is something specific to *Item Format* that is contributing to difficulty beyond the scope of the NAEP Reading Comprehension framework.

The second stage of this study used latent regression to estimate the overall mean performance differences between the ELL and SLD focal groups on all items. All four alternative models confirmed the presence of significant and detrimental effects for both ELL and SLD status. The effect of ELL or SLD status was estimated at a drop in proficiency of over one logit, with SLD student proficiencies impacted the most. And, for the small population of students who belonged to both groups (i.e., ELLs who were also classified as SLD), the interaction effect reduced proficiency by just under 2.0 logits. These findings align with previous research illustrating significant differences in student performance across subgroups, although much of the research has focused on differences related to gender (Mazzeo & Yamamoto, 1993; Garner & Engelhard, 1999; Hastedt & Sibberns, 2005; Le, 2009;) or ethnic membership (Rock & Kaling, 1988; Zwick & Ercikan, 1989, Scheuneman & Gerritz, 1990; O'Neal & Paek, 1993; Snetzler & Qualls, 2000; Taylor & Lee, 2011).

In the first steps of this dissertation, analysis demonstrated an effect for *Item Format* and a separate effect of group membership on student proficiencies. The third stage examined whether there is an interaction effect between the assessment characteristics (e.g. such as *Item Format)* and group membership variables by investigating the presence and directionality of significant DIF in the ELL and SLD models. The investigation of DIF is not only beneficial in assessment design and evaluation (i.e., it pinpoints items that might exhibit systematic bias), but is a useful tool for assessing differential strengths and weaknesses in performance characteristics of population subgroups; DIF analyses may reveal facets of items or groups of items that point to patterns for particular subgroups that might have been previously overlooked (Scheuneman & Gerritz, 1990). In some cases DIF denotes differences that are relevant to the construct, but it can also spotlight item bias when differences between groups can be attributed to construct-irrelevant characteristics (Taylor & Lee, 2011), such as the item requirement to write a response that may ultimately cloud our judgment about what a student understood while reading.

Consistent with the regression, DIF analysis demonstrated that group membership affected mean reading comprehension proficiency by more than one logit. A significant number of items demonstrated DIF in both models: twelve percent of items were flagged in the ELL model and 16 percent flagged in the SLD model. To address construct irrelevant DIF, test developers can either remove the flagged items from the pool, or equally select items so that DIF is balanced across focal (the population of interest—in this case ELL or SLD) and reference (usually the general population) groups. For the 2007 NAEP reading comprehension, directionality of DIF items were fairly evenly distributed across groups in both ELL and SLD models.

The range of item difficulties for DIF items favoring the focal groups was much lower than those favoring the reference group in both models. On one hand, this means that the focal groups are performing better than expected on easy items while the reference group is

performing worse than expected on them. Conversely, the reference group is scoring higher than expected on the harder items, and the focal groups are scoring lower than expected on them. This signals that some characteristic of easier items (which are more predominantly MC items) is favoring focal groups and something about more difficult items (which are predominantly ECR items) is favoring the reference group. We can only guess as to why, but perhaps the focal group tries harder with easier items because they seem accessible, while the reference group pays less attention because the items are so easy. The only way, however, to get at the heart of this is to employ a method such as ThinkAlouds that asks students to explain their thinking when answering items (Pressley & Afflerbach, 1995).

While we see an interesting pattern of DIF difficulty emerge across group membership, these patterns are not explained by patterns of DIF across *Item Format*. Overall, the patterns of DIF across *Item Format* suggest that there may be some interaction with group membership. When DIF items were grouped by *Item Format* categories, the great majority of flagged items in both models were MC items. There was a hint of imbalance in the SLD model: more flagged MC DIF items favored the reference group. Even though MC items on average are easier, there is something about MC items that favors the reference group, and, in turn, disadvantages the ELL and SLD focal groups. These results are similar to the findings in Taylor and Lee (2011) where a state-criterion referenced reading assessment demonstrated that MC items favored Whites and ECR items favored ethnic minorities across multiple testing years and grades.

A growing body of research has focused our attention on items that appear to put various subgroups (especially racial, ethnic, linguistic, or intellectual minority groups) at a disadvantage, but a deeper and more elaborate analyses of these items may also be instructive in identifying important differences between groups in test taking skills, cognitive styles, and testwiseness (O'Neill & McPeek, 1993). Since we are not able to review NAEP's "live" reading items (because of security matters), we can only conjecture about the origin of influence. There is sufficient research documenting that guessing may account for higher scores on MC items (e.g. Hastedt, 2004); it may be that the reference group is more test savvy in eliminating distracters or perhaps, reciprocally, the SLD group is more attracted by the appeal of them.

In both focal group models, the few SCR and ECR DIF items all favored the focal groups. My hypothesis was that the focal students would do more poorly on the CR items given the added demand of producing a written answer. To the contrary, a smaller percentage of CR items were flagged and, moreover, of those that were flagged they all favored the focal groups. Again, I can only conjecture about the causes of this counter-intuitive effect: perhaps CR items allow struggling students entry by the open-ended nature of the response, or perhaps lower-achieving students do better on items that do not attract them with intellectually seductive foils. One intriguing hypothesis is that the relative advantage may stem from the accommodations available to focal groups on NAEP. Out of the 191,040 students assessed, 13,977 (7%) were allowed an accommodation of extended time. We have evidence from studies such as Lukehele, Thissen, and Wainer (1994) demonstrating that MC and CR items differ greatly in average completion time. Thus, it is certainly reasonable to assume that when it comes to producing a written response on a test, extra time, which is the most prevalent accommodation, can make a large difference in how well students perform.

These are interesting findings since we know that: (a) CR items are increasingly more difficult than MC items; (b) the two focal groups are underperforming on all items in relationship to their peers; and (c) coupled with the fact that ELL and SLD students experience linguistic challenges with the English language. Yet, in this study, we don't see any evidence of a

detrimental effect of CR items on group membership; rather, to some degree we see the opposite effect where the flagged CR items work in their favor. The caveat is that very few of the forty-three CR items were flagged for DIF: 3 in the ELL model and 1 in the SLD model. With so few items, this finding is interpreted with caution and further investigation is needed since the items could merely be a result of random variability.

Varied results emerged when DIF analysis was applied to the components (i.e., the three levels of Context and the four Aspects) of the Reading Framework. In the ELL model, *Informational Context* items favored ELLs, but in the SLD model, there were no differences in levels of *Context*. In both the ELL and SLD model, items from the *Interpretation Aspect* favored the focal group while items from the *Context and Structure Aspect* favor the reference group. With little variability in difficulty between the subcategories, it is difficult to interpret why one subcategory may favor one group of students over another. What is notable here is that no clear pattern of *Item Format* by *Context* by *Aspect* interactions emerged, suggesting that with the minor tendencies noted, *Item Format* differences function similarly across levels of *Aspect* and *Context* within the Reading Comprehension Framework.

Multidimensional analysis of *Item Format* demonstrated high correlations for all models, indicating that neither the MC-CR dichotomy or the MC, SCR, and ECR dimensions exhibited enough independent variance to be calculated separately, therefore, it is justifiable to report NAEP reading comprehension scores unidimensionally because of the minor loss in information resulting from composite scores. These correlations were slightly higher than those reported in Rodriguez's (2003) meta-analysis that compared non-stem and stem equivalent item formats where he found an 87% correlation in content equivalent designs and 82% in non-content equivalent designs. Regression analysis indicates that student proficiencies for both focal groups dropped across all three dimensions. The largest logit difference for the NE and Non-SLDs was in the MC dimension and for ELLs and SLDs was in the ECR dimension. These results are consistent with the directionality of DIF results.

Multidimensional analysis of *Context* demonstrated moderate correlations between dimensions for the model with and without regression. The fact that these correlations were lower than that of Multidimensional *Item Format* suggests that there is a greater distinction between the dimensions of Literary and Informational than between the three *Item Formats,* yet, not enough of a difference to justify separating the dimensions. ELLs demonstrated a slightly larger logit decline in proficiency on items assessing the Informational Context, while SLDs had a larger drop in the Literary Context. In revisiting the Baseline model, there is little difference in difficulty between Contexts.[1]

Conclusion

This dissertation supports the findings of researchers such as Ercikan et al. (1998) who assert there are important differences in *Item Format* that warrant further study. *Item Format* in this study demonstrates a difference in difficulty with MC items comprising the lowest in difficulty and CR items comprising the highest. Multidimensional analysis suggests that despite the differences between *Item Format,* they are similar enough to be calculated as one score of Reading Comprehension. Aside from the unanswered question of what is driving *Item Format* differences, there is no evidence to suggest that the integration of Constructed-Response items

[1] It should be noted that NAEP continues to report separate scales for different *Contexts—Literary and Informational,* even though the correlations of scores between levels of *Context* are as only slightly lower than those between levels of *Aspect* or *Item Format*.

disadvantage English-language learners or individuals diagnosed with Specific Learning Disabilities, as these subgroups who underperform in relationship to their peers, do equally poorly across all three item formats.

Furthermore, the findings suggest that the incorporation of CR items allows for the assessment of a wider range of student abilities, thus increasing assessment reliability. There is also no evidence to contradict the many advantages of using MC items, such as the benefits in linking tests with CR items across multiple years (Ercikan et al., 1998; Campbell, 1999). In addition, there is great efficiency of MC items in breadth of content coverage and ease of scoring, as well as the objective nature that aids in scoring reliability.

An appraisal of the literature suggests that CR items are beneficial in their unique ability to assess higher-order content not obtainable through the sole use of multiple-choice items. But this does not appear to be the purpose—or more accurately, the outcome—in the use of CR items on the NAEP assessment. Given that there was little hierarchy in item difficulty across the subcategories of reading genre (i.e. *Context*) or domains of reading (i.e. *Aspect*), the estimated differences in *Item Format* difficulty appear to be unrelated to these framework components. As a result, it is hard to ascertain how the *Context* and *Aspect* elements of the Reading Comprehension framework are systematically related to item format on the NAEP assessment or what the decision process may be for developing items into a multiple-choice versus a constructed-response format other than meeting a standard ratio.

It is possible that the differences we are witnessing in *Item Format* difficulty capture an implicit structure of reading comprehension not captured in the Framework, which—if it existed—would be useful in deepening our understanding of student reading comprehension strengths and weaknesses. But, it seems more likely that differences are attributable to unique cognitive demands of *Item Format*. If the cognitive demand comes down to a distinction between recognition versus production, then the important question is whether any *Item Format* is producing item bias. In this dissertation, there is no evidence of such. Analysis indicates that there is an effect for *Item Format* on estimated student proficiencies. And, while DIF analysis in both the ELL and SLD model suggests patterns of interaction between *Item Format* and group membership, we do not ultimately know what effect DIF items have on overall student proficiencies. The multidimensional regression model settles this question: estimated differences in the student proficiencies are virtually the same across MC, SCR, and ECR dimensions for all 4^{th} grade students suggesting that there is not an interaction effect between *Item Format* and group membership.

Future Directions

In the words of Hancock (1994), "The items that we use in our assessments must first be true to those objectives we seek to teach and must not compromise those objectives for the sake of administrational convenience" (pg.155). Rauch and Hartig (2010) claim that the confounding of format and cognitive processes could be avoided if reading items were systematically constructed to keep them independent; for example, by explicitly instructing item writers to assess abilities necessary to master higher reading processes with open ended as well as with closed response formats. But, it is possible that in the mind of item writers, it is more difficult to construct MC items for higher-order content as asserted in the literature, and is equally possible that it is just as difficult to design CR items for lower-level content. Given the ubiquitous nature of mixed item formats, we need more information about whether there are limited applications of *Item Format.* In future research, it may be useful to construct a special NAEP study that designs paired item formats where both MC and CR items are created to assess identical item content.

Such a study as such may give insight into potential limitations of developing *Item Formats* across content to aid in the future design of large-scale assessment. If differences in student proficiencies were followed-up with Think Aloud protocols, this study could lend useful insight into cognitive differences between *Item Formats*. These differences may come down to actual differences in cognition, but may also be related to environmental influences such as instructional, curricular, and/or experiential factors. If item formats are suited to a particular content or level of knowing, explicating a construct would be useful to item writers so that format can be chosen accordingly. In addition, to aid in future assessment design, it is useful to understand the ideal ratio of MC to CR items in mixed item formats that would utilize all of their desired benefits while also producing fair results (Hastedt & Sibberns, 2005).

Limitations

NAEP's wide reach across the U.S. student population makes it an ideal resource for analyzing issues in test design and interpretation. The secure dataset is available for public consumption through an application process that serves to protect not only the privacy of individual student achievement data but NAEP content still in use. The diversity and quantity of students assessed allows for the application of sophisticated programs of analysis such as ConQuest that extend beyond traditional methods of data analysis. An additional benefit is NCES' extensive process of item development and evaluation through 3PL IRM which lends strong validity to the assessment design. The application of Item Response Modeling (IRM) to the question of *Item Format* effect allows data to be addressed in ways not possible with earlier methods of factor analysis and the like. IRM's ability to isolate person proficiencies and item difficulties can take into the account various assessment format effects on item difficulty.

That said, it is important to realize that possible generalizations of my findings are limited to 4^{th} grade data, reading comprehension assessments, and *Item Format* effects related to non-stem equivalent multiple-choice and open-ended items. The dataset used to evaluate *Item Format* effect is based on items developed for Reading Comprehension hence is possible that similar analysis in other content areas or with different items sets would deliver varying results. For that reason, it is important to replicate this work in other content areas. However, there is a greater preponderance of research in the field of mathematics likely due to the obvious construct differences between mathematics and writing, while reading and writing are often viewed as sibling skills. Also, while findings suggest that there were minimal influences of *Context* and *Aspect* on the effect of *Item Format*, the development of paired items is another means for controlling these variables and may produce different results than documented here. And, a final limitation of this study is NAEPs use of the BIB design means that on average each student only completes one fifth of the item pool. Therefore, while Item Response Modeling can statistically estimate item difficulties and student proficiencies of latent reading comprehension ability, results may differ in an assessment with fewer missing data points.

References

Abedi, J. (2002). Standardized assessment tests and English language learners: Psychometric issues. *Educational Assessment*, 8(3), 231-257.

Adams, R., Wilson, M.R., & Wang, W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: B. N. Petrov and F. Csaki (Eds). Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267–281.

Barnett, Foster, D. & Nagy, P. (1996). Undergraduate student response strategies to test questions of varying format. *Higher* Education, 32(2), 177-198.

Beller, M, & Gafni, N. (2000). Can item format (multiple-choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, 42(1/2), 1-21.

Bennett, R.E., Rock, D.A., & Wang, M, (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, 28(1), 77-92.

Bennett, R. & Ward, W. (Eds.). (1993). *Constructing versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, 29(3), 253-271.

Bridgeman, B., & Rock, D.A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement*, 30, 313-329.

Briggs, D. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4 (1),* 87-100.

Campbell, J.R. (1999). Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension (UMI No. 9938651).

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment*. NY: Routledge.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement,* 31(4), 295-311.

Downing, S. (2006). Selected-response item formats in test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development.* London: Lawrence Erlbaum Associates.

Embretson, S. E. & Reise, S.P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Ercikan, K. & Schwarz, R. (1995). Dimensionality of multiple-choice and constructed-response tests for different ability groups. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Ercikan, K., Schwarz, R.D., Julian, M.W., Burket, G.R., Weber, M.M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement,* 35(2), 137-154.

Garner, M., & Engelhard, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12,* 29-51.

Haladyna, T.M. (1997). Writing test items to measure higher level thinking. Needham Heights, MA: Allyn & Bacon.
Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333.
Hambleton, R.K. & Jirka, S. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. Downing & T. Haladyna (Eds.), Handbook of test development (pp. 399-420). Mahwah, NJ: Lawrence Erlbaum Publishers.
Hambleton, R.K. & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education*, 5(1), 1-16.
Hancock, G. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, 62(2), 143-157.
Hastedt, D. (2004). Differences between multiple-choice and constructed response items in PIRLS 2001. In: C. Papanastasiou (Ed.), *In PIRLS 2001, Proceedings of the IRC-2004 PIRLS*, Cyprus University Press, Nicosia, Cyprus.
Hastedt, D. & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation*, 31, 145-161.
Hollingworth, L., Beard, J. & Proctor, T. (2007). Item type. *Practical Assessment, Research, & Evaluation*, 12(18), 1-13.
Individuals with Disabilities Education Improvement Act, P.L. 108-466 (2004, 2005). 34 C.F.R. 300 (Proposed Regulations).
Katz, S., Bennett, R.E., & Berger, A.E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement*, 37(1), 39-57.
Le, L.T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, 9(2), 122-133.
Longford, N.T., Holland, P.W., & Thayer, D.T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. *Holland* & H. Wainer (Eds.), *Differential item functioning (pp. 171-196). Hills-dale, NJ: Lawrence Erlbaum.*
Lukehele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31(3), 234-250.
Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. *International Journal of Testing*, 8(1), 14-33.
Manhart, J. (1996). Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct. *National Council of Measurement in Education*, 1-48.
Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2), 333-368.
Martinez, M.E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34, 207-218.
Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
Mazzeo, J., K. & Yamamoto, et al. (1993). Extended constructed-response items in the 1992 NAEP: Psychometrically speaking, were they worth the price? National Council on Measurement in Education. Atlanta, Georgia.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed, pp.13-104). NY: American Council on Education & MacMillan.
Meulders, M. & Xie, Y. (2004). Person-by-item predictors. In P. D. Boeck & M.R. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp.213-240). New York: Springer.
Moore, S. (1996). Estimating differential item functioning in the polytomous case with random coefficients multinomial logit (RCML) model. In E. Inglehard & M. Wilson (Eds.), *Objective measurement: Theory into practice*, Volume 3. Norwood, NJ: Ablex Publishing.
National Assessment Governing Board (2007). Reading framework for the 2007 national assessment of educational progress. Washington, D.C. Retrieved from Institute of Educational Services online: http://nces.ed.gov/nationsreportcard/reading/whatmeasure.asp#sec4.
National Assessment of Educational Progress (NAEP) Validity Panel (2007). Scope of Work: Study of the Feasibility of a NAEP "Easy Booklet Alternative".
National Center of Educational Statistics (2007, September). *The Nation's Report Card: Reading 2007* (Publication No. NCES 2007496). Retrieved from National Center for Educational Statistics online: http://nces.ed.gov/pubsearch/getpubcats.asp?sid=031#.
National Research Council. (2000, 2002). *Scientific research in education.* Washington, DC: National Academy Press.
Nitko, A.J. (2004*). Educational assessment of children (4th ed.).* Upper Saddle River, NJ: Pearson.
No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2002). Available from http://www.ed/gov/policy/elsec/leg/esea02/index.html
Office of Special Education Programs [OSEP] (2003). Twenty-fifth annual report to congress on the implementation of the individuals with disabilities education act: Section 1, the national picture. Retrieved on November 2, 2007, from http://www.ed.gov/about/reports/annual/osep/2003/index.html.
O'Neill, K.A., & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Erlbaum.
Paek, I. (2002). Investigations of differential item functioning: Comparisons among approaches, and extensions to a multidimensional context. Unpublished doctoral dissertation, University of California, Berkeley.
Pearson, P.D. & Garavaglia, D.R. (1997). Improving the information value of performance items in large scale assessments. Commissioned by the NAEP Validity Studies Panel. Unpublished paper.
Pelligrino, J. W., Jones, L.R., & Mitchell, K. J. (1999). Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress. *Committee on the Evaluation of National and State Assessments of Educational Progress, National Research Council. Washington, D.C.: National Research Council.*
Pressley, M. & Afflerbach, P. (1995). Verbal protocols of reading: The nature of constructively responsive reading. Hillsdale, NJ, USA: Erlbaum.
Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press. (Original work published 1960 by the Danish Institute for Educational Research).

Rauch, D. & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, 52(4), 354-379.
Rock, D. & Kaling, C. (1988). Differential item functioning analysis of Math performance of Hispanic, Asian, and White NAEP respondents. Princeton, NJ, National Assessment of Educaitonal Progress, Research Report, RR 142.
Rodriguez, M.C. (2003). Construct equivalence in multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Education Measurement*, 40(2), 163-184.
Samejima (1976). A use of the information function in tailored testing. *Applied Psychological Measurement*, 1(2), 233-247.
Scheuneman, J.D. & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, 27(2), 109-131.
Schwartz G (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1, 147-170.
Snetzler, S. & Quall, A. (2000). Examination of differential item functioning on a standard achievement battery with limited English proficient students. *Educational and Psychological Measurement*, 60(4), 564-577.
Stedman, L. C. (2009). The NAEP long-term trend assessment: A review of its transformation, use, and findings. Washington, D.C.: National Assessment Governing Board. http://www.nagb.org/who-we-are/20-anniversary/stedman-long-term-formatted.pdf
Sykes, R. & Yen, W. (2000). The scaling of mixed-item format tests with the one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, 37(3), 221-244.
Taylor, C. & Lee, Y. (2011). Ethnic DIF in reading tests with mixed item formats. *Educational Assessment*, 16(1), 35-68.
Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests: An analysis of two tests. *Journal of Educational Measurement*, 31(2), 113-123.
Traub, R. E. & Fisher, C. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, 1(3), 355-369.
Traub, R. E. & MacRury, MacRury, K. (1990). Multiple-choice vs. free-response in the testing of scholastic achievement. Unpublished manuscript. The Ontario Institute for Studies in Education. Published in German in K. Ingenkamp & R.S. Jager (Eds.) *Tests und trends 8* (pp128-159). Weinheim & Basel: Beltz Verlag.
United States General Accounting Office. (2003). Title 1: Characteristics of tests will influence expenses; Information sharing may help States realize efficiencies. Retrieved May 06, 2009, from http://www.gao.gov/new.items/d03389.pdf.
Van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement*, 14, 1-12.
Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, 6(2), 103-118.

Ward, W., Dupree, D., & Carlson, S.B. (1987). A comparison of free-response and multiple-choice questions in the assessment of reading comprehension. Princeton, N.J., Educational Testing Service (87-20).
Wilson, M. (2005). Constructing Measures: An Item Response Modeling Approach. Mahwah, NJ: Erlbaum.
Wright, B.D. & Masters G.N. (1982). Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press.
Wright, B.D. & Stone, M. (1980). Best Test Design. Rasch Measurement. Chicago: MESA Press.
Wu, M.L., Adams, R.J., & Wilson, M. (2007). ACER ConQuest 2.0 [computer program]. Hawthorn, Australia: ACER Press.
Zwick, R, Donoghue, J.R, & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.
Zwick, R. & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26(1), 55-66.

Appendices

Appendix A. Item Properties by Format, Context, and Aspect.

Characterization of individual items

	Item Property								
	Item Format			Context		Aspect			
	MC	SCR	ECR	Lit	Inf	GU	I	MC	C/S
1.	X			X		X			
2.	X			X		X			
3.	X			X		X			
4.	X			X			X		
5.	X			X			X		
6.	X			X			X		
7.	X			X			X		
8.	X			X			X		
9.	X			X			X		
10.	X			X			X		
11.	X			X			X		
12.	X			X			X		
13.	X			X			X		
14.	X			X			X		
15.	X			X			X		
16.	X			X			X		
17.	X			X			X		
18.	X			X			X		
19.	X			X			X		
20.	X			X			X		
21.	X			X			X		
22.	X			X			X		
23.	X			X			X		
24.	X			X			X		
25.	X			X				X	
26.	X			X					X
27.	X			X					X
28.	X			X					X
29.	X			X					X
30.	X			X					X
31.	X				X	X			
32.	X				X	X			
33.	X				X		X		
34.	X				X		X		
35.	X				X		X		
36.	X				X		X		
37.	X				X		X		
38.	X				X		X		
39.	X				X		X		
40.	X				X		X		
41.	X				X		X		
42.	X				X		X		
43.	X				X		X		
44.	X				X		X		
45.	X				X		X		
46.	X				X		X		
47.	X				X		X		
48.	X				X		X		
49.	X				X		X		
50.	X				X		X		
51.	X				X		X		
52.	X				X				X
53.	X				X				X
54.	X				X				X
55.	X				X				X
56.	X				X				X
57.	X				X				X
58.		X		X		X			
59.		X		X			X		
60.		X		X			X		
61.		X		X			X		
62.		X		X			X		

63.	X		X				X			
64.	X		X				X			
65.	X		X						X	
66.	X			X	X					
67.	X			X	X					
68.	X			X			X			
69.		X	X			X				
70.		X	X			X				
71.		X	X			X				
72.		X	X			X				
73.		X	X			X				
74.		X	X			X				
75.		X	X			X				
76.		X	X			X				
77.		X	X			X				
78.		X	X				X			
79.		X	X						X	
80.		X	X						X	
81.		X	X						X	
82.		X		X	X					
83.		X		X		X				
84.		X		X		X				
85.		X		X		X				
86.		X		X		X				
87.		X		X		X				
88.		X		X		X				
89.		X		X		X				
90.		X		X		X				
91.		X		X		X				
92.		X		X		X				
93.		X		X		X				
94.		X		X		X				
95.		X		X		X				
96.		X		X		X				
97.		X		X		X				
98.		X		X			X			
99.		X		X			X			
100.		X		X			X			

Note. An "X" in the cell indicates that the property applies to the item; MC = Multiple-choice; SCR=Short-constructed response; ECR=Extended-constructed response; Lit=Literary; Inf=Informational; GU=General Understanding; I=Interpretation; MC=Making Connections; C/S= Content and Structure.

Appendix B. Crosstabulation of Item Number by *Item Format, Context,* and *Aspect*

	MC		SCR		ECR		Total Aspect
	Liter	**Infor**	**Liter**	**Infor**	**Liter**	**Infor**	
General Understanding	1 2 3	31 32	58	66 67		82	**9**
Interpretation	4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24	33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51	59 60 61 62		69 70 71 72 73 74 75 76 77	83 84 85 86 87 88 89 90 91 92 93 94 95 96 97	**68**
Connections	25		63 64	68	78	98 99 100	**7**
Content and Structure	26 27 28 29 30	52 53 54 55 56 57	65		79 80 81		**15**
Total Context	**30**	**27**	**8**	**3**	**13**	**19**	
Total Item Format	**57 MC**		**11 SCR**		**32 ECR**		**100**

Note: Underlined items are from the released passage

Appendix C: NAEP Reading Comprehension Assessment Variables

Variable 1. Student reporting sample (n=191,040)
Variable 2. English language learners (ELL; n=15,784)
Variable 3. Students classified with Specific Learning Disability (SLD; n=8,244)
Variable 4. Interaction variable of ELL*SLD (n=858)
Variable 5-104. NAEP's 100 Reading Comprehension items categorized by:

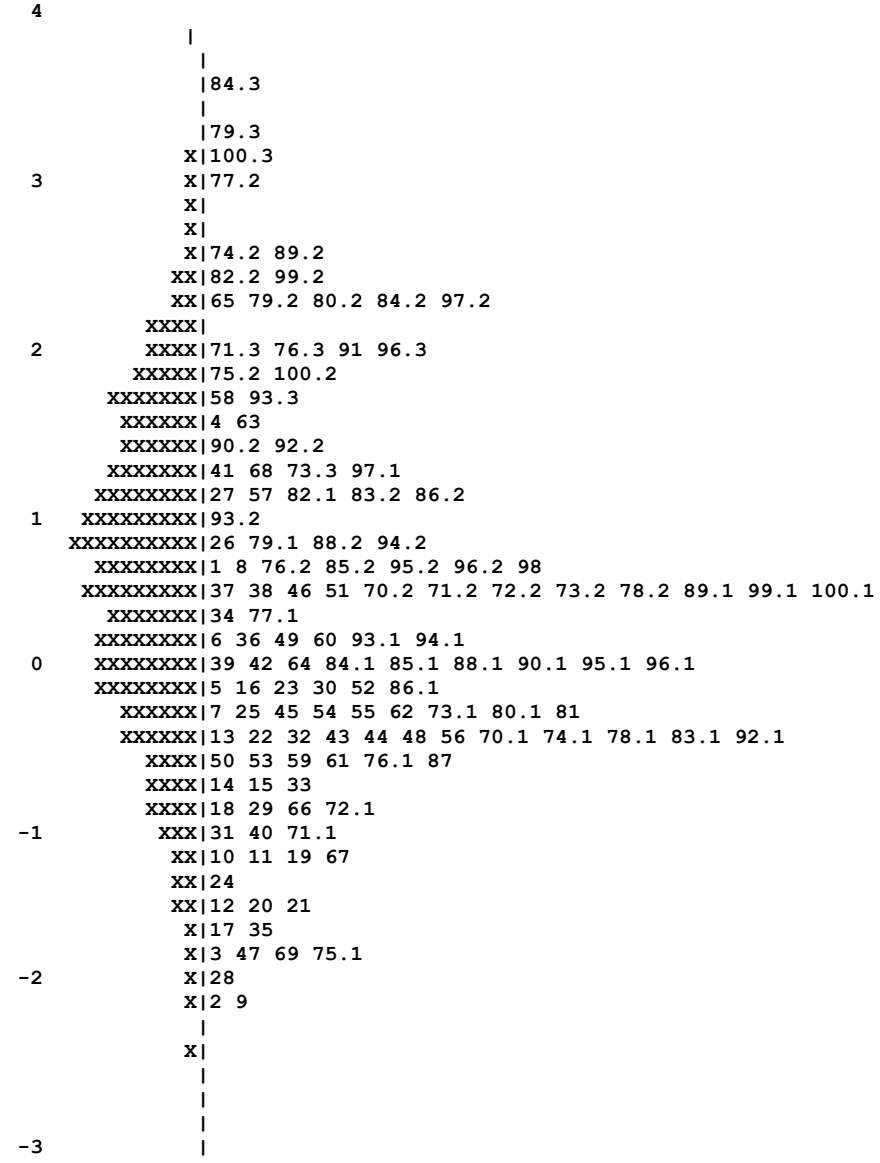
- *Item Format*
    - MC (57 items),
    - SCR (11 items), and
    - ECR (32 items);
- *Context*
    - Literary (51 items), and
    - Informational (49 items);
- *Reading Aspect*
    - General Understanding (9 items),
    - Interpretation (68 items),
    - Connections (7 items), and
    - Content & Structure (15 items).

Appendix D. Released Passage Item Details

Item 8: What does the word "pleading" mean? (MC, Lit, I)
Item 9: Why does Rosa return to the school yard? (MC, Lit, I)
Item 10: When Rosa tiptoes to the ducks and whispers… (MC, Lit, I)
Item 11: Which lesson is most important to the story? (MC, Lit, I)
Item 12: When worker speaks to her why Rosa feels proud? (MC, Lit, I)
Item 28: When Rosa plays baseball what does it show? (MC, Lit, C/S)
Item 59: Why is the gym teacher important? (SCR, Lit, GU)
Item 69: Explain why Rosa visits the ducks. (ECR, Lit, I)
Item 78: Describe how Rosa is like someone you know. (ECR, Lit, C)
Item 80: Rosa's Creek better title? (ECR, Lit, C/S)

Note: Released items are written as denoted in the NAEP 2007 database of Reading Comprehension code book; MC=Multiple-choice; SCR=Short-constructed Response; ECR=Extended-constructed response; Lit=Literary; I=Interpretation; C/S=Content & Structure; GU=General Understanding; C=Connections.

Appendix E. Wright Map of Baseline Model Including Steps



Appendix F. Baseline Model: Average Fit Analysis.

Note: an infit mean square < 0.75 is an indicator of less randomness than expected while an infit mean square > 1.33 is an indicator of more randomness than expected

```
 ================================================================================
Partial Credit Model: baseline, all items                  Sun Apr 03 14:50 2011
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
================================================================================
TERM 1: item
-------------------------------------------------------------------------------------------
-
   VARIABLES                              UNWEIGHTED FIT            WEIGHTED FIT
---------------                      ----------------------   ----------------------
     item           ESTIMATE  ERROR    MNSQ       CI        T    MNSQ       CI        T
-------------------------------------------------------------------------------------------
-
 1   R012610           0.564   0.012    0.96 ( 0.99, 1.01) -6.1   0.97 ( 0.99, 1.01) -7.6
 2   R052901          -2.178   0.019    0.75 ( 0.99, 1.01)-38.4   0.92 ( 0.98, 1.02) -6.2
 3   R053001          -1.872   0.017    1.16 ( 0.99, 1.01) 20.8   1.04 ( 0.98, 1.02)  3.9
 4   R012602           1.545   0.012    1.29 ( 0.99, 1.01) 37.0   1.10 ( 0.99, 1.01) 18.3
 5   R012603          -0.152   0.012    0.93 ( 0.99, 1.01) -9.6   0.96 ( 0.99, 1.01) -8.1
 6   R012606           0.110   0.012    0.96 ( 0.99, 1.01) -6.0   0.98 ( 0.99, 1.01) -4.4
 7   R012608          -0.234   0.012    1.05 ( 0.99, 1.01)  7.0   1.03 ( 0.99, 1.01)  6.5
 8   R020501           0.577   0.012    1.00 ( 0.99, 1.01) -0.7   0.99 ( 0.99, 1.01) -2.5
 9   R020601          -2.133   0.019    0.68 ( 0.99, 1.01)-50.8   0.88 ( 0.98, 1.02) -9.8
10   R021101          -1.172   0.014    1.25 ( 0.99, 1.01) 32.3   1.07 ( 0.98, 1.02)  8.1
11   R021601          -1.259   0.015    1.21 ( 0.99, 1.01) 27.5   1.09 ( 0.98, 1.02) 10.1
12   R022101          -1.533   0.016    0.64 ( 0.99, 1.01)-57.9   0.82 ( 0.98, 1.02)-19.9
13   R052902          -0.500   0.013    1.06 ( 0.99, 1.01)  8.4   1.05 ( 0.99, 1.01)  7.7
14   R052903          -0.672   0.013    0.83 ( 0.99, 1.01)-25.0   0.91 ( 0.99, 1.01)-13.5
15   R052905          -0.695   0.013    0.82 ( 0.99, 1.01)-25.8   0.90 ( 0.99, 1.01)-15.6
16   R052910          -0.177   0.012    0.89 ( 0.99, 1.01)-15.6   0.93 ( 0.99, 1.01)-13.2
17   R053002          -1.720   0.016    0.77 ( 0.99, 1.01)-35.1   0.93 ( 0.98, 1.02) -6.9
18   R053005          -0.952   0.014    1.08 ( 0.99, 1.01) 10.6   1.02 ( 0.99, 1.01)  2.6
19   R053007          -1.163   0.014    0.84 ( 0.99, 1.01)-23.1   0.92 ( 0.98, 1.02)-10.8
20   R053701          -1.582   0.016    0.97 ( 0.99, 1.01) -4.8   0.98 ( 0.98, 1.02) -2.5
21   R053801          -1.452   0.015    0.91 ( 0.99, 1.01)-12.9   0.97 ( 0.98, 1.02) -3.2
22   R054401          -0.486   0.013    0.88 ( 0.99, 1.01)-16.9   0.93 ( 0.99, 1.01)-12.4
23   R054501          -0.139   0.012    1.05 ( 0.99, 1.01)  7.0   1.05 ( 0.99, 1.01)  8.8
24   R053601          -1.425   0.015    0.77 ( 0.99, 1.01)-35.5   0.91 ( 0.98, 1.02)-10.1
25   R052908          -0.352   0.013    0.80 ( 0.99, 1.01)-30.4   0.87 ( 0.99, 1.01)-24.1
26   R012605           0.807   0.012    1.20 ( 0.99, 1.01) 25.6   1.13 ( 0.99, 1.01) 28.4
27   R012609           1.094   0.012    1.07 ( 0.99, 1.01)  9.8   1.04 ( 0.99, 1.01)  9.2
28   R022201          -1.902   0.017    0.63 ( 0.99, 1.01)-58.8   0.85 ( 0.98, 1.02)-13.8
29   R053008          -0.893   0.013    0.85 ( 0.99, 1.01)-21.7   0.92 ( 0.99, 1.01)-11.7
30   R054101          -0.138   0.012    1.09 ( 0.99, 1.01) 11.6   1.07 ( 0.99, 1.01) 12.1
31   R023301          -0.999   0.014    0.96 ( 0.99, 1.01) -6.3   0.98 ( 0.99, 1.01) -2.2
32   R054601          -0.450   0.013    1.08 ( 0.99, 1.01) 11.1   1.07 ( 0.99, 1.01) 11.1
33   R017302          -0.806   0.013    0.80 ( 0.99, 1.01)-29.0   0.91 ( 0.99, 1.01)-13.2
34   R017308           0.272   0.012    0.96 ( 0.99, 1.01) -5.7   0.97 ( 0.99, 1.01) -7.1
35   R022401          -1.582   0.016    0.86 ( 0.99, 1.01)-20.0   0.98 ( 0.98, 1.02) -1.8
36   R023101           0.097   0.012    0.86 ( 0.99, 1.01)-20.6   0.90 ( 0.99, 1.01)-21.5
37   R024101           0.440   0.012    1.00 ( 0.99, 1.01) -0.5   0.99 ( 0.99, 1.01) -2.8
38   R034501           0.510   0.012    1.15 ( 0.99, 1.01) 20.0   1.10 ( 0.99, 1.01) 21.1
39   R034701          -0.034   0.012    1.28 ( 0.99, 1.01) 35.7   1.18 ( 0.99, 1.01) 35.3
40   R034801          -1.021   0.014    0.82 ( 0.99, 1.01)-27.0   0.91 ( 0.99, 1.01)-11.8
41   R035101           1.185   0.012    1.14 ( 0.99, 1.01) 18.3   1.08 ( 0.99, 1.01) 15.8
42   R035401           0.030   0.012    1.06 ( 0.99, 1.01)  8.3   1.05 ( 0.99, 1.01) 10.6
43   R053102          -0.427   0.012    0.94 ( 0.99, 1.01) -7.8   0.99 ( 0.99, 1.01) -1.2
44   R053103          -0.516   0.013    0.82 ( 0.99, 1.01)-26.2   0.91 ( 0.99, 1.01)-16.2
45   R053104          -0.316   0.012    1.11 ( 0.99, 1.01) 15.1   1.07 ( 0.99, 1.01) 12.3
46   R053109           0.463   0.012    0.91 ( 0.99, 1.01)-13.4   0.92 ( 0.99, 1.01)-17.3
47   R054701          -1.805   0.017    0.82 ( 0.99, 1.01)-26.8   0.96 ( 0.98, 1.02) -4.0
48   R054901          -0.496   0.013    0.90 ( 0.99, 1.01)-14.1   0.96 ( 0.99, 1.01) -6.5
49   R055001           0.216   0.012    0.89 ( 0.99, 1.01)-15.1   0.93 ( 0.99, 1.01)-16.0
50   R055201          -0.636   0.013    0.86 ( 0.99, 1.01)-20.5   0.93 ( 0.99, 1.01)-11.2
51   R057501           0.501   0.012    1.02 ( 0.99, 1.01)  3.1   1.01 ( 0.99, 1.01)  2.2
```

52	R017304	-0.077	0.012	0.97 (0.99, 1.01)	-3.9	0.99 (0.99, 1.01) -2.4
53	R017306	-0.562	0.013	0.86 (0.99, 1.01)	-21.0	0.91 (0.99, 1.01) -14.9
54	R022901	-0.356	0.012	0.85 (0.99, 1.01)	-21.6	0.91 (0.99, 1.01) -17.2
55	R023801	-0.242	0.012	0.98 (0.99, 1.01)	-2.5	1.00 (0.99, 1.01) -0.1
56	R053107	-0.479	0.013	0.79 (0.99, 1.01)	-30.6	0.88 (0.99, 1.01) -21.5
57	R053110	1.121	0.012	1.20 (0.99, 1.01)	25.7	1.10 (0.99, 1.01) 20.9
58	R012601	1.734	0.013	1.12 (0.99, 1.01)	15.9	1.03 (0.99, 1.01) 4.9
59	R020701	-0.666	0.013	0.89 (0.99, 1.01)	-15.1	0.92 (0.99, 1.01) -13.2
60	R052906	0.108	0.012	0.99 (0.99, 1.01)	-1.5	0.99 (0.99, 1.01) -2.8
61	R053003	-0.655	0.013	0.90 (0.99, 1.01)	-13.8	0.94 (0.99, 1.01) -9.4
62	R053009	-0.311	0.012	0.88 (0.99, 1.01)	-17.4	0.91 (0.99, 1.01) -16.6
63	R012612	1.532	0.012	0.96 (0.99, 1.01)	-4.9	1.00 (0.99, 1.01) -0.4
64	R052909	-0.024	0.012	0.86 (0.99, 1.01)	-20.5	0.89 (0.99, 1.01) -21.4
65	R012604	2.226	0.014	0.88 (0.99, 1.01)	-16.6	0.97 (0.99, 1.01) -4.5
66	R017301	-0.872	0.013	1.06 (0.99, 1.01)	8.3	1.01 (0.99, 1.01) 1.1
67	R053101	-1.170	0.014	1.13 (0.99, 1.01)	17.3	1.05 (0.98, 1.02) 6.6
68	R023201	1.199	0.012	1.01 (0.99, 1.01)	1.7	1.02 (0.99, 1.01) 3.7
69	R020401	-1.751	0.017	0.90 (0.99, 1.01)	-14.8	0.95 (0.98, 1.02) -4.8
70	R052904	-0.035	0.008	1.19 (0.99, 1.01)	25.3	1.12 (0.99, 1.01) 16.4
71	R052907	0.462	0.008	1.01 (0.99, 1.01)	1.0	1.01 (0.99, 1.01) 1.3
72	R053004	-0.242	0.009	1.16 (0.99, 1.01)	21.2	1.11 (0.99, 1.01) 16.2
73	R053006	0.444	0.007	1.21 (0.99, 1.01)	26.8	1.16 (0.99, 1.01) 22.3
74	R053901	1.055	0.010	1.01 (0.99, 1.01)	1.2	1.01 (0.99, 1.01) 2.1
75	R054001	0.010	0.011	1.02 (0.99, 1.01)	2.6	1.02 (0.99, 1.01) 2.7
76	R054201	0.647	0.007	1.04 (0.99, 1.01)	5.4	1.05 (0.99, 1.01) 6.8
77	R054301	1.688	0.011	0.95 (0.99, 1.01)	-6.4	0.98 (0.99, 1.01) -2.8
78	R021201	0.003	0.008	1.29 (0.99, 1.01)	36.1	1.19 (0.99, 1.01) 26.2
79	R012607	2.175	0.012	0.91 (0.99, 1.01)	-12.8	0.95 (0.98, 1.02) -5.7
80	R021701	1.032	0.009	1.10 (0.99, 1.01)	12.8	1.10 (0.99, 1.01) 14.8
81	R053010	-0.285	0.012	0.94 (0.99, 1.01)	-9.0	0.95 (0.99, 1.01) -8.6
82	R055301	1.790	0.010	1.03 (0.99, 1.01)	3.7	1.03 (0.99, 1.01) 4.0
83	R017303	0.297	0.008	1.08 (0.99, 1.01)	10.8	1.07 (0.99, 1.01) 11.1
84	R017307	1.962	0.012	0.96 (0.99, 1.01)	-5.1	0.98 (0.99, 1.01) -3.3
85	R017310	0.366	0.008	1.13 (0.99, 1.01)	17.0	1.09 (0.99, 1.01) 13.1
86	R023501	0.476	0.008	1.34 (0.99, 1.01)	42.5	1.25 (0.99, 1.01) 36.4
87	R023601	-0.624	0.013	0.92 (0.99, 1.01)	-10.7	0.95 (0.99, 1.01) -8.5
88	R034601	0.470	0.008	1.43 (0.99, 1.01)	52.4	1.27 (0.99, 1.01) 39.5
89	R034901	1.527	0.010	1.08 (0.99, 1.01)	10.2	1.06 (0.99, 1.01) 8.7
90	R035001	0.753	0.008	1.02 (0.99, 1.01)	2.4	1.02 (0.99, 1.01) 3.8
91	R035201	1.954	0.013	0.74 (0.99, 1.01)	-38.9	0.88 (0.99, 1.01) -19.5
92	R053105	0.471	0.009	0.95 (0.99, 1.01)	-7.6	0.95 (0.99, 1.01) -7.9
93	R053106	0.933	0.007	1.18 (0.99, 1.01)	24.0	1.14 (0.99, 1.01) 19.8
94	R053108	0.416	0.008	1.12 (0.99, 1.01)	16.3	1.09 (0.99, 1.01) 13.9
95	R054801	0.294	0.008	0.97 (0.99, 1.01)	-3.9	0.98 (0.99, 1.01) -3.4
96	R055101	0.881	0.007	1.06 (0.99, 1.01)	8.6	1.05 (0.99, 1.01) 7.2
97	R055401	1.730	0.009	1.17 (0.99, 1.01)	22.5	1.10 (0.99, 1.01) 12.8
98	R017309	0.698	0.012	0.97 (0.99, 1.01)	-4.1	0.98 (0.99, 1.01) -3.6
99	R024001	1.495	0.010	1.03 (0.99, 1.01)	3.6	1.03 (0.99, 1.01) 4.0
100	R035301	1.860*		0.99 (0.99, 1.01)	-1.5	1.01 (0.99, 1.01) 2.0

An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability = 1.000
Chi-square test of parameter equality = 669815.28, df = 99, Sig Level = 0.000

Appendix G. DIF Analysis of the ELL Model

Item #	Format	Context	Aspect	Group	Coefficient	DIF Magnitude	DIF Level
100 R035301 1 0	ECR	2	3	NE	0.002* 0.061	<.426	NO DIF
71 R052907 1 0	ECR	1	2	NE	0.011 0.005	<.426	NO DIF
90 R035001 1 0	ECR	2	2	NE	0.011 0.005	<.426	NO DIF
5 R012603 1 0	MC	1	2	NE	0.013 0.006	<.426	NO DIF
46 R053109 1 0	MC	2	2	NE	0.018 0.006	<.426	NO DIF
85 R017310 1 0	ECR	2	2	NE	0.019 0.005	<.426	NO DIF
31 R023301 1 0	MC	2	1	NE	0.019 0.007	<.426	NO DIF
79 R012607 1 0	ECR	1	4	NE	0.020 0.005	<.426	NO DIF
69 R020401 1 0	ECR	1	2	NE	0.020 0.007	<.426	NO DIF
83 R017303 1 0	ECR	2	2	NE	0.033 0.005	<.426	NO DIF
82 R055301 1 0	ECR	2	1	NE	0.033 0.006	<.426	NO DIF
51 R057501 1 0	MC	2	2	NE	0.034 0.006	<.426	NO DIF
35 R022401 1 0	MC	2	2	NE	0.034 0.007	<.426	NO DIF
42 R035401 1 0	MC	2	2	NE	0.035 0.006	<.426	NO DIF
18 R053005 1 0	MC	1	2	NE	0.036 0.006	<.426	NO DIF
43 R053102 1 0	MC	2	2	NE	0.036 0.006	<.426	NO DIF
76 R054201 1 0	ECR	1	2	NE	0.037 0.005	<.426	NO DIF
72 R053004 1 0	ECR	1	2	NE	0.048 0.006	<.426	NO DIF
81 R053010 1 0	ECR	1	4	NE	0.048 0.006	<.426	NO DIF
88 R034601 1 0	ECR	2	2	NE	0.052 0.005	<.426	NO DIF
30 R054101 1 0	MC	1	4	NE	0.053 0.006	<.426	NO DIF
73 R053006 1 0	ECR	1	2	NE	0.055 0.005	<.426	NO DIF
48 R054901 1 0	MC	2	2	NE	0.055 0.006	<.426	NO DIF
98 R017309 1 0	ECR	2	3	NE	0.061 0.006	<.426	NO DIF
6 R012606 1 0	MC	1	2	NE	0.063 0.006	<.426	NO DIF
1 R012610 1 0	MC	1	1	NE	0.073 0.006	<.426	NO DIF
86 R023501 1 0	ECR	2	2	NE	0.074 0.005	<.426	NO DIF
68 R023201 1 0	SCR	2	3	NE	0.084 0.006	<.426	NO DIF
99 R024001 1 0	ECR	2	3	NE	0.088 0.006	<.426	NO DIF
94 R053108 1 0	ECR	2	2	NE	0.089 0.005	<.426	NO DIF
11 R021601 1 0	MC	1	2	NE	0.091 0.007	<.426	NO DIF
38 R034501 1 0	MC	2	2	NE	0.092 0.006	<.426	NO DIF
45 R053104 1 0	MC	2	2	NE	0.096 0.006	<.426	NO DIF
97 R055401 1 0	ECR	2	2	NE	0.102 0.006	<.426	NO DIF
66 R017301 1 0	SCR	2	1	NE	0.104 0.006	<.426	NO DIF

67	R053101	1	0	SCR	2	1	NE	0.113	0.007	<.426	NO DIF
80	R021701	1	0	ECR	1	4	NE	0.115	0.006	<.426	NO DIF
23	R054501	1	0	MC	1	2	NE	0.125	0.006	<.426	NO DIF
10	R021101	1	0	MC	1	2	NE	0.125	0.007	<.426	NO DIF
32	R054601	1	0	MC	2	1	NE	0.126	0.006	<.426	NO DIF
63	R012612	1	0	SCR	1	3	NE	0.132	0.006	<.426	NO DIF
41	R035101	1	0	MC	2	2	NE	0.162	0.006	<.426	NO DIF
58	R012601	1	0	SCR	1	1	NE	0.180	0.006	<.426	NO DIF
78	R021201	1	0	ECR	1	3	NE	0.183	0.005	<.426	NO DIF
3	R053001	1	0	MC	1	1	NE	0.199	0.007	<.426	NO DIF
39	R034701	1	0	MC	2	2	NE	0.211	0.006	<.426	NO DIF
7	R012608	1	0	MC	1	2	NE	0.223	0.006	=.446	SLI TO MOD
57	R053110	1	0	MC	2	4	NE	0.278	0.006	=.426-.638	SLI TO MOD
27	R012609	1	0	MC	1	4	NE	0.291	0.006	=.426-.638	SLI TO MOD
4	R012602	1	0	MC	1	2	NE	0.374	0.006	=.748	MOD TO SEV
26	R012605	1	0	MC	1	4	NE	0.374	0.006	>.638	MOD TO SEV
9	R020601	1	0	MC	1	2	ELL	-0.006	0.007	<.426	NO DIF
93	R053106	1	0	ECR	2	2	ELL	-0.008	0.005	<.426	NO DIF
89	R034901	1	0	ECR	2	2	ELL	-0.014	0.006	<.426	NO DIF
96	R055101	1	0	ECR	2	2	ELL	-0.018	0.005	<.426	NO DIF
60	R052906	1	0	SCR	1	2	ELL	-0.020	0.006	<.426	NO DIF
22	R054401	1	0	MC	1	2	ELL	-0.021	0.006	<.426	NO DIF
87	R023601	1	0	ECR	2	2	ELL	-0.021	0.006	<.426	NO DIF
21	R053801	1	0	MC	1	2	ELL	-0.026	0.007	<.426	NO DIF
75	R054001	1	0	ECR	1	2	ELL	-0.027	0.006	<.426	NO DIF
47	R054701	1	0	MC	2	2	ELL	-0.028	0.007	<.426	NO DIF
55	R023801	1	0	MC	2	4	ELL	-0.030	0.006	<.426	NO DIF
59	R020701	1	0	SCR	1	2	ELL	-0.035	0.006	<.426	NO DIF
84	R017307	1	0	ECR	2	2	ELL	-0.035	0.006	<.426	NO DIF
61	R053003	1	0	SCR	1	2	ELL	-0.038	0.006	<.426	NO DIF
77	R054301	1	0	ECR	1	2	ELL	-0.041	0.006	<.426	NO DIF
65	R012604	1	0	SCR	1	4	ELL	-0.042	0.007	<.426	NO DIF
34	R017308	1	0	MC	2	2	ELL	-0.044	0.006	<.426	NO DIF
37	R024101	1	0	MC	2	2	ELL	-0.048	0.006	<.426	NO DIF
49	R055001	1	0	MC	2	2	ELL	-0.048	0.006	<.426	NO DIF
29	R053008	1	0	MC	1	4	ELL	-0.049	0.006	<.426	NO DIF
70	R052904	1	0	ECR	1	2	ELL	-0.056	0.005	<.426	NO DIF

28	R022201	1	0	MC	1	4	ELL	-0.063	0.007	<.426	NO DIF
62	R053009	1	0	SCR	1	2	ELL	-0.070	0.006	<.426	NO DIF
50	R055201	1	0	MC	2	2	ELL	-0.072	0.006	<.426	NO DIF
53	R017306	1	0	MC	2	4	ELL	-0.083	0.006	<.426	NO DIF
25	R052908	1	0	MC	1	3	ELL	-0.084	0.006	<.426	NO DIF
74	R053901	1	0	ECR	1	2	ELL	-0.084	0.006	<.426	NO DIF
2	R052901	1	0	MC	1	1	ELL	-0.085	0.007	<.426	NO DIF
19	R053007	1	0	MC	1	2	ELL	-0.093	0.007	<.426	NO DIF
95	R054801	1	0	ECR	2	2	ELL	-0.094	0.005	<.426	NO DIF
64	R052909	1	0	SCR	1	3	ELL	-0.094	0.006	<.426	NO DIF
17	R053002	1	0	MC	1	2	ELL	-0.109	0.007	<.426	NO DIF
16	R052910	1	0	MC	1	2	ELL	-0.112	0.006	<.426	NO DIF
13	R052902	1	0	MC	1	2	ELL	-0.117	0.006	<.426	NO DIF
33	R017302	1	0	MC	2	2	ELL	-0.133	0.006	<.426	NO DIF
52	R017304	1	0	MC	2	4	ELL	-0.136	0.006	<.426	NO DIF
8	R020501	1	0	MC	1	2	ELL	-0.163	0.006	<.426	NO DIF
36	R023101	1	0	MC	2	2	ELL	-0.183	0.006	<.426	NO DIF
54	R022901	1	0	MC	2	4	ELL	-0.187	0.006	<.426	NO DIF
92	R053105	1	0	ECR	2	2	ELL	-0.190	0.006	<.426	NO DIF
12	R022101	1	0	MC	1	2	ELL	-0.192	0.007	<.426	NO DIF
24	R053601	1	0	MC	1	2	ELL	-0.192	0.007	<.426	NO DIF
40	R034801	1	0	MC	2	2	ELL	-0.214	0.007	=.428	SLI TO MOD
44	R053103	1	0	MC	2	2	ELL	-0.215	0.006	=.426-.638	SLI TO MOD
20	R053701	1	0	MC	1	2	ELL	-0.217	0.007	=.426-.638	SLI TO MOD
56	R053107	1	0	MC	2	4	ELL	-0.248	0.006	=.426-.638	SLI TO MOD
14	R052903	1	0	MC	1	2	ELL	-0.267	0.006	=.426-.638	SLI TO MOD
15	R052905	1	0	MC	1	2	ELL	-0.298	0.006	=.426-.638	SLI TO MOD
91	R035201	1	0	ECR	2	2	ELL	-0.306	0.006	=.426-.638	SLI TO MOD

Appendix H. DIF Analysis for the SLD Model

Item #	Format	Context	Aspect	Group	Coefficient	DIF Magnitude	DIF Level
25 R052908 1 0	MC	1	3	Non	0.005 0.006	<.426	NO DIF
90 R035001 1 0	ECR	2	2	Non	0.013 0.005	<.426	NO DIF
43 R053102 1 0	MC	2	2	Non	0.016 0.006	<.426	NO DIF
39 R034701 1 0	MC	2	2	Non	0.017 0.006	<.426	NO DIF
99 R024001 1 0	ECR	2	3	Non	0.023 0.006	<.426	NO DIF
18 R053005 1 0	MC	1	2	Non	0.025 0.006	<.426	NO DIF
63 R012612 1 0	SCR	1	3	Non	0.025 0.006	<.426	NO DIF
96 R055101 1 0	ECR	2	2	Non	0.026 0.005	<.426	NO DIF
19 R053007 1 0	MC	1	2	Non	0.029 0.007	<.426	NO DIF
89 R034901 1 0	ECR	2	2	Non	0.031 0.006	<.426	NO DIF
73 R053006 1 0	ECR	1	2	Non	0.034 0.005	<.426	NO DIF
86 R023501 1 0	ECR	2	2	Non	0.041 0.005	<.426	NO DIF
3 R053001 1 0	MC	1	1	Non	0.043 0.007	<.426	NO DIF
16 R052910 1 0	MC	1	2	Non	0.045 0.006	<.426	NO DIF
30 R054101 1 0	MC	1	4	Non	0.047 0.006	<.426	NO DIF
78 R021201 1 0	ECR	1	3	Non	0.051 0.005	<.426	NO DIF
45 R053104 1 0	MC	2	2	Non	0.056 0.006	<.426	NO DIF
29 R053008 1 0	MC	1	4	Non	0.064 0.006	<.426	NO DIF
68 R023201 1 0	SCR	2	3	Non	0.067 0.006	<.426	NO DIF
81 R053010 1 0	ECR	1	4	Non	0.067 0.006	<.426	NO DIF
85 R017310 1 0	ECR	2	2	Non	0.070 0.005	<.426	NO DIF
93 R053106 1 0	ECR	2	2	Non	0.071 0.005	<.426	NO DIF
82 R055301 1 0	ECR	2	1	Non	0.086 0.006	<.426	NO DIF
84 R017307 1 0	ECR	2	2	Non	0.086 0.006	<.426	NO DIF
55 R023801 1 0	MC	2	4	Non	0.089 0.006	<.426	NO DIF
100 R035301 1 0	ECR	2	3	Non	0.095* 0.061	<.426	NO DIF
6 R012606 1 0	MC	1	2	Non	0.116 0.006	<.426	NO DIF
88 R034601 1 0	ECR	2	2	Non	0.121 0.005	<.426	NO DIF
98 R017309 1 0	ECR	2	3	Non	0.129 0.006	<.426	NO DIF
58 R012601 1 0	SCR	1	1	Non	0.135 0.006	<.426	NO DIF
34 R017308 1 0	MC	2	2	Non	0.139 0.006	<.426	NO DIF
37 R024101 1 0	MC	2	2	Non	0.141 0.006	<.426	NO DIF
97 R055401 1 0	ECR	2	2	Non	0.154 0.006	<.426	NO DIF
94 R053108 1 0	ECR	2	2	Non	0.162 0.005	<.426	NO DIF
41 R035101 1 0	MC	2	2	Non	0.169 0.006	<.426	NO DIF
52 R017304 1 0	MC	2	4	Non	0.193 0.006	<.426	NO DIF
23 R054501 1 0	MC	1	2	Non	0.198 0.006	<.426	NO DIF
46 R053109 1 0	MC	2	2	Non	0.203 0.006	<.426	NO DIF
38 R034501 1 0	MC	2	2	Non	0.291 0.006	=.426-.638	SLI TO MOD
1 R012610 1 0	MC	1	1	Non	0.308 0.006	=.426-.638	SLI TO MOD
7 R012608 1 0	MC	1	2	Non	0.335 0.006	=.670	MOD TO SEV
26 R012605 1 0	MC	1	4	Non	0.339 0.006	>.638	MOD TO SEV
42 R035401 1 0	MC	2	2	Non	0.354 0.006	>.638	MOD TO SEV
27 R012609 1 0	MC	1	4	Non	0.381 0.006	>.638	MOD TO SEV
4 R012602 1 0	MC	1	2	Non	0.407 0.006	>.638	MOD TO SEV

57	R053110	1	0	MC	2	4	Non	0.436	0.006	>.638	MOD TO SEV
15	R052905	1	0	MC	1	2	SLD-	-0.123	0.006	<.426	NO DIF
5	R012603	1	0	MC	1	2	SLD -	-0.004	0.006	<.426	NO DIF
22	R054401	1	0	MC	1	2	SLD -	-0.007	0.006	<.426	NO DIF
48	R054901	1	0	MC	2	2	SLD -	-0.015	0.006	<.426	NO DIF
11	R021601	1	0	MC	1	2	SLD -	-0.019	0.007	<.426	NO DIF
83	R017303	1	0	ECR	2	2	SLD -	-0.023	0.005	<.426	NO DIF
79	R012607	1	0	ECR	1	4	SLD -	-0.025	0.005	<.426	NO DIF
10	R021101	1	0	MC	1	2	SLD -	-0.027	0.007	<.426	NO DIF
72	R053004	1	0	ECR	1	2	SLD -	-0.029	0.006	<.426	NO DIF
71	R052907	1	0	ECR	1	2	SLD -	-0.034	0.005	<.426	NO DIF
77	R054301	1	0	ECR	1	2	SLD -	-0.036	0.006	<.426	NO DIF
50	R055201	1	0	MC	2	2	SLD -	-0.040	0.006	<.426	NO DIF
51	R057501	1	0	MC	2	2	SLD -	-0.040	0.006	<.426	NO DIF
32	R054601	1	0	MC	2	1	SLD -	-0.041	0.006	<.426	NO DIF
13	R052902	1	0	MC	1	2	SLD -	-0.045	0.006	<.426	NO DIF
62	R053009	1	0	SCR	1	2	SLD -	-0.056	0.006	<.426	NO DIF
20	R053701	1	0	MC	1	2	SLD -	-0.058	0.007	<.426	NO DIF
65	R012604	1	0	SCR	1	4	SLD -	-0.061	0.007	<.426	NO DIF
67	R053101	1	0	SCR	2	1	SLD -	-0.064	0.007	<.426	NO DIF
76	R054201	1	0	ECR	1	2	SLD -	-0.065	0.005	<.426	NO DIF
70	R052904	1	0	ECR	1	2	SLD -	-0.066	0.005	<.426	NO DIF
87	R023601	1	0	ECR	2	2	SLD -	-0.067	0.006	<.426	NO DIF
80	R021701	1	0	ECR	1	4	SLD -	-0.070	0.006	<.426	NO DIF
36	R023101	1	0	MC	2	2	SLD -	-0.071	0.006	<.426	NO DIF
95	R054801	1	0	ECR	2	2	SLD -	-0.075	0.005	<.426	NO DIF
56	R053107	1	0	MC	2	4	SLD -	-0.075	0.006	<.426	NO DIF
40	R034801	1	0	MC	2	2	SLD -	-0.075	0.007	<.426	NO DIF
8	R020501	1	0	MC	1	2	SLD -	-0.078	0.006	<.426	NO DIF
64	R052909	1	0	SCR	1	3	SLD -	-0.080	0.006	<.426	NO DIF
92	R053105	1	0	ECR	2	2	SLD -	-0.082	0.006	<.426	NO DIF
75	R054001	1	0	ECR	1	2	SLD -	-0.090	0.006	<.426	NO DIF
35	R022401	1	0	MC	2	2	SLD -	-0.097	0.007	<.426	NO DIF
28	R022201	1	0	MC	1	4	SLD -	-0.098	0.007	<.426	NO DIF
21	R053801	1	0	MC	1	2	SLD -	-0.118	0.007	<.426	NO DIF
60	R052906	1	0	SCR	1	2	SLD -	-0.141	0.006	<.426	NO DIF
61	R053003	1	0	SCR	1	2	SLD -	-0.141	0.006	<.426	NO DIF
74	R053901	1	0	ECR	1	2	SLD -	-0.144	0.006	<.426	NO DIF
31	R023301	1	0	MC	2	1	SLD -	-0.158	0.007	<.426	NO DIF
66	R017301	1	0	SCR	2	1	SLD -	-0.159	0.006	<.426	NO DIF
53	R017306	1	0	MC	2	4	SLD -	-0.162	0.006	<.426	NO DIF
49	R055001	1	0	MC	2	2	SLD -	-0.165	0.006	<.426	NO DIF
44	R053103	1	0	MC	2	2	SLD -	-0.168	0.006	<.426	NO DIF
54	R022901	1	0	MC	2	4	SLD -	-0.173	0.006	<.426	NO DIF
24	R053601	1	0	MC	1	2	SLD -	-0.173	0.007	<.426	NO DIF
2	R052901	1	0	MC	1	1	SLD -	-0.182	0.007	<.426	NO DIF
47	R054701	1	0	MC	2	2	SLD -	-0.212	0.007	<.426	NO DIF
59	R020701	1	0	SCR	1	2	SLD -	-0.223	0.006	=.446	SLI TO MOD
14	R052903	1	0	MC	1	2	SLD -	-0.226	0.006	=.426-.638	SLI TO MOD
17	R053002	1	0	MC	1	2	SLD -	-0.234	0.007	=.426-.638	SLI TO MOD

69	R020401	1	0	ECR	1	2	SLD -	-0.242	0.007	=.426-.638 SLI TO MOD
9	R020601	1	0	MC	1	2	SLD -	-0.245	0.007	=.426-.638 SLI TO MOD
12	R022101	1	0	MC	1	2	SLD -	-0.246	0.007	=.426-.638 SLI TO MOD
33	R017302	1	0	MC	2	2	SLD -	-0.291	0.006	=.426-.638 SLI TO MOD
91	R035201	1	0	ECR	2	2	SLD -	-0.292	0.006	=.426-.638 SLI TO MOD

Appendix I. Conquest Control Files

BASELINE

```
Title Baseline Model, all items;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format ELL3 1 rptsamp 2 LEP 3 XS00301 4 responses 5-104;
labels << C:\Users\diss master\Desktop\ConQuest\input\\NAEP07R4_all_conquest_042010_label.txt;
codes 0, 1, 2, 3;
model item+item*step;
estimate!iterations=9999,stderr=full;
export par>>baseline.prm;
export reg>>baseline.reg;
export cov>>baseline.cov;
show !estimates=latent>> baseline_cc.shw;
show >> baseline2_cc.shw;
reset;
quit;
```

REGRESSION

```
Title Regression Model: all items, LEP;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format ELL3 1 rptsamp 2 LEP 3 XS00301 4 responses 5-104 InteractionLEP 106;
labels << C:\Users\diss master\Desktop\ConQuest\input\\NAEP07R4_all_conquest_042010_label.txt;
codes 0, 1, 2, 3;
regression LEP;
model item+item*step;
estimate!iterations=9999;
export par>>LEP_reg.prm;
export reg>>LEP_reg.reg;
export cov>>LEP_reg.cov;
show >> LEP_reg_cc.shw;
reset;

Title Regression Model: all items, SLD;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format ELL3 1 rptsamp 2 LEP 3 XS00301 4 responses 5-104 InteractionLEP 106;
labels << C:\Users\diss master\Desktop\ConQuest\input\\NAEP07R4_all_conquest_042010_label.txt;
codes 0, 1, 2, 3;
regression xs00301;
model item+item*step;
estimate!iterations=9999;
export par>>SLD_reg.prm;
export reg>>SLD_reg.reg;
export cov>>SLD_reg.cov;
show >> SLD_reg_cc.shw;
reset;
```

```
 Title Regression Model: all items, LEP & SLD;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format ELL3 1 rptsamp 2 LEP 3 XS00301 4 responses 5-104 InteractionLEP 106;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
codes 0, 1, 2, 3;
regression LEP XS00301;
model item+item*step;
estimate!iterations=9999;
export par>>LEP_SLD_reg.prm;
export reg>>LEP_SLD_reg.reg;
export cov>>LEP_SLD_reg.cov;
show >> LEP_SLD_reg_cc.shw;
reset;

Title Regression Model: all items, LEP, SLD, LEP*SLD;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format ELL3 1 rptsamp 2 LEP 3 XS00301 4 responses 5-104 InteractionLEP 106;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
codes 0, 1, 2, 3;
regression LEP xs00301 InteractionLEP;
model item+item*step;
estimate!iterations=9999;
show!estimates=latent>>LEPinterSLD_reg_cc.shw;
reset;
quit;
```

DIFFERENTIAL ITEM FUNCTIONING

```
Title Differential Item Functioning: all items, LEP simple;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format rptsamp 2 LEP 3 XS00301 4 responses 5-104 interactionLEP 106;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
codes 0, 1, 2, 3, 4;
model item-LEP+item*LEP+item*step;
estimate!iterations=9999;
export par>>DIF_LEP.prm;
export reg>>DIF_LEP.reg;
export cov>>DIF_LEP.cov;
show>>DIF_LEP_cc.shw;
reset;

Title Differential Item Functioning: all items, SLD simple;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format rptsamp 2 LEP 3 XS00301 4 responses 5-104 interactionLEP 106;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
codes 0, 1, 2, 3, 4;
model item-xs00301+item*xs00301+item*step;
estimate!iterations=9999;
```

```
export par>>DIF_SLD.prm;
export reg>>DIF_SLD.reg;
export cov>>DIF_SLD.cov;
show>>DIF_SLD_cc.shw;
reset;
quit;
```

MULTIDIMENSIONAL ITEM FORMAT

```
Title Multidimensional Latent Regression: MC vs. CR, no regression;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format responses 5-104;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
model item+item*step;
score (0,1,2,3) (0,1,2,3) ()!items(1-57);
score (0,1,2,3) () (0,1,2,3)!items(58-100);
import anchor_parameters>>format2dimen.prm;
import anchor_reg_coefficients>>format2dimen.reg;
import anchor_covariance>>format2dimen.cov;
set update=yes, warnings=no;
codes 0, 1, 2, 3;
estimate!method=quadrature, nodes=30, minnode=-8, maxnode=8, iterations=9999;
show!estimates=latent>>format2dimen_redo.shw;
reset;

Title Multidimensional Latent Regression: MC, SCR, ECR, no regression;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format responses 5-104;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
model item+item*step;
score (0,1,2,3) (0,1,2,3) () ()!items(1-57);
score (0,1,2,3) () (0,1,2,3) ()!items(58-68);
score (0,1,2,3) () () (0,1,2,3)!items(69-100);
import anchor_parameters<<format3dimen.prm;
import anchor_reg_coefficients<<format3dimen.reg;
import anchor_covariance<<format3dimen.cov;
set update=yes, warnings=no;
codes 0, 1, 2, 3;
estimate!method=quadrature, nodes=30, minnode=-8, maxnode=8, iterations=9999;
show!estimates=latent>>format3dimen_redo.shw;
reset;

Title Multidimensional Latent Regression: MC, SCR, ECR, regression;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format LEP 3 xs00301 4 responses 5-104;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
model item+item*step;
score (0,1,2,3) (0,1,2,3) () ()!items(1-57);
```

```
score (0,1,2,3) () (0,1,2,3) ()!items(58-68);
score (0,1,2,3) () () (0,1,2,3)!items(69-100);
regression LEP, xs00301;
import anchor_parameters<<format3dimen_reg.prm;
import anchor_reg_coefficients<<format3dimen_reg.reg;
import anchor_covariance<<format3dimen_reg.cov;
set update=yes, warnings=no;
codes 0, 1, 2, 3;
estimate!method=quadrature, nodes=30, minnode=-8, maxnode=8, iterations=9999;
show!estimates=latent>>format3dimen_reg_redo2.shw;
reset;
quit;
```

MULTIDIMENSIONAL CONTEXT

```
Title Multidimensional Latent Regression: context, no regression;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format responses 5-104;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
model item+item*step;
score (0,1,2,3) (0,1,2,3) ()!items(1-30, 58-65, 69-81);
score (0,1,2,3) () (0,1,2,3)!items(31-57, 66-68, 82-100);
export parameters>>Context2dimen.prm;
export reg>>Context2dimen.reg;
export cov>>Context2dimen.cov;
set update=yes, warnings=no;
codes 0, 1, 2, 3;
estimate!method=quadrature, nodes=30, minnode=-8, maxnode=8, iterations=9999;
show!estimates=latent>>Context_cc.shw;
reset;

Title Multidimensional Latent Regression: context, regression;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format LEP 3 xs00301 4 responses 5-104;
labels << C:\Users\diss master\Desktop\ConQuest\input\NAEP07R4_all_conquest_042010_label.txt;
model item+item*step;
score (0,1,2,3) (0,1,2,3) ()!items(1-30, 58-65, 69-81);
score (0,1,2,3) () (0,1,2,3)!items(31-57, 66-68, 82-100);
regression LEP, xs00301;
export parameters>>Context_reg.prm;
export reg>>Context_reg.reg;
export cov>>Context_reg.cov;
set update=yes, warnings=no;
codes 0, 1, 2, 3;
estimate!method=quadrature, nodes=30, minnode=-8, maxnode=8, iterations=9999;
show!estimates=latent>>Context_reg_cc.shw;
reset;
quit;
```

MULTIDIMENSIONAL ASPECT

```
Title Multidimensional Latent Regression: aspect, regression;
datafile  C:\Users\diss
master\Desktop\ConQuest\input\\NAEP07R4_all_recode_filter_interaction2_012611.txt;
format LEP 3 xs00301 4 responses 5-104;
labels << C:\Users\diss master\Desktop\ConQuest\input\\NAEP07R4_all_conquest_042010_label.txt;
model item+item*step;
score (0,1,2,3) (0,1,2,3) () () ()!items(1-3,31-32,58,66-67,82);
score (0,1,2,3) () (0,1,2,3) () ()!items(4-24,33-51,59-62,69-77,83-97);
score (0,1,2,3) () () (0,1,2,3) ()!items(25,63-64,68,78,98-100);
score (0,1,2,3) () () () (0,1,2,3)!items(26-30,52-57,65,79-81);
regression LEP, xs00301;
export parameters>>Aspect_reg.prm;
export reg>>Aspect_reg.reg;
export cov>>Aspect_reg.cov;
set update=yes, warnings=no;
codes 0, 1, 2, 3;
estimate!method=quadrature, nodes=20, minnode=-8, maxnode=8, conv=.1,iterations=9999;
show!estimates=latent>>aspect_reg_cc.shw;
reset;
quit;
```