Differences in K-12 science standards and Differential Item Functioning (DIF) in NAEP Science Assessments

By
© 2020

Rabia Esma Sipahi Akbas
M.Sc., Texas A&M University, 2012
B.Sc., Ankara University, 2008

Submitted to the graduate degree program in Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chair: John P. Poggio

Vicki Peyton

Meagan Patterson

David Hansen

Eve Levin

Date Defended: 11 May 2020

The dissertation committee for Rabia Esma Sipahi Akbas certifies that this is the approved version of the following dissertation:

Differences in K-12 science standards and Differential Item Functioning (DIF) in NAEP Science Assessments

Chair: John P. Poggio

Date Approved: 06 June 2020

Abstract

This study examined the role of differences in K-12 science standards across US states on measurement comparability, as indicated by differential item functioning (DIF), on the National Assessment of Educational Progress (NAEP) science assessments at grades 4, 8, and 12. Restricted data sets as offered and maintained by NCES were used in this investigation. Extant data was obtained from Institute for Education Sciences (IES) and includes item-level responses of students along with student characteristics, school district and state identification. The report of Gross et al 2013, which evaluated state science standards against Next Generation Science Standards, was used to categorize states' science standards as superior or inferior. The Mantel-Haenszel and Logistic regression analysis procedures were used to conduct DIF analysis. The findings evidenced that there is DIF based on differences in state science standards in three (3) test questions out of the 126 test questions asked at 12th grade. However, no DIF based on differences in state science standards have been found on the 4th and 8th grade examination items. In these DIF detected items on the 12th grade exams, the odds of students from superior states responding correctly to these items are 1.537 to 1.664 times higher than that of students from inferior standard states. These items were from physical science and life science content areas and favor students from superior science standard states over students from inferior science standard states. However, no DIF items were detected in earth and space science content area. Additional analyses have shown that the results are robust (i.e., no DIF detected) when controlled for the gender and the race of students. Further, examinee ability was measured at the content level and the results evidence that using this alternative measure of ability had no effect on the main findings. Overall, the findings strongly suggest that differences in state K-12 science standards is associated with the degree to which NAEP science assessments measure what they claim to measure: the true science ability of students and become an important factor when assessing measurement validity.

Keywords: differential item functioning, NAEP science assessments, K-12 state science standards, curriculum

Acknowledgements

I would like to express my sincere gratitude to my committee chair and academic advisor, Professor John P. Poggio, for his wonderful mentorship and guidance throughout my graduate years at the University of Kansas. He always encouraged and supported me to pursue and explore my own my own interests while providing guidance throughout all challenges. Without him, the completion of this work would not be possible. Thank you for your commitment to me and many hours you have dedicated to my education.

Thank you to all my professors in the REMS program at the University of Kansas, and to my dedicated committee members, Dr. Vicki Peyton, Dr. Meagan Patterson, and Dr. David Hansen, and Dr. Eve Levin for their guidance, suggestions and helps at the proposal meeting and support at the defense meeting.

My special thanks to the Institute of Education Sciences for their support in providing me the data and answering all my questions very patiently that made this thesis possible.

My success in this degree program would not have been possible without the love, support, and guidance I received from my family. Thank you to my parents Nazik Sipahi and Yusuf Sipahi and in-laws Esvet Akbas and Mehmet Akbas for their endless support, encouragement, and prayers throughout this entire journey, also coming overseas to take care of my son and me while I was working on comprehensive exams. I would like to extend my gratitude to my brother, Ali Sipahi, sister Esra Sipahi, and sister in-law, Neslihan Sipahi for cheering me up when I feel lost and desperate during the frustrating process of research and writing my thesis.

Finally, my deepest gratitude goes to my husband, Ferhat Akbas, for his understanding, patience, limitless support, and unconditional love. I am so blessed to have such a smart and kind person in my life who is always available to listen, discuss and pushes me along every steps of my PhD education. My very special thanks to my one and only son Furkan Akbas, the light and hope of my life. With his limitless love, kisses, and hugs, I was able to overcome every trial that comes to my way.

Table of Contents

List of Tables

Chapter 1: Introduction

In the US, individual states have the opportunity and flexibility to develop their own curriculum, their own assessments, and set their own performance standards for proficiency in science for K-12 students (Wenning, et al., 2003). However, US students at grades 4, 8, and 12 take the ***identical*** National Assessment of Educational Progress (NAEP) science assessments, even though the state science standards embraced are not the same. Indeed, there are very real and crucial differences in K-12 science standards across US states when the Next Generation Science Standards (NGSS) are used as a benchmark (Gross et al., 2013). Interestingly, despite the meaningfulness of differences in state standards, which may lead to differential instruction and thus item response patterns, difficulty levels and potentially impact measurement invariance assumption in science assessments, there have not been in-depth evaluation studies examining this issue.

Background

There is a long history on detecting differential item functioning (DIF) in large-scale assessments and importance of removing bias, unfairness and insensitivity from assessments. DIF occurs when a test item functions differently for different individuals from different groups but with the same ability level. DIF procedures compare performances of groups or subgroups on each item for examinees having the same level of performance, usually measured by total test score. As pointed out by Zieky (1993) the fairness of an item depends directly on the purpose for which a test is being used and the DIF judgment is required to determine whether or not the difference in difficulty is unfairly related to group membership. The judgment of fairness is based on whether the difference in difficulty is believed to be related to the construct being measured. Yet, our knowledge and understanding of how DIF occurs are far from complete.

Understanding the fairness and validity of NAEP science exams is extremely important since the National Assessment of Education Progress (NAEP) is the most-often used indicator of student learning in science. Results from NAEP are used to compare student achievement across states and to judge states' student proficiency levels. Hence, constructing a test that would minimize the differences in students' exposure to various contents in NAEP tests is vital for healthy inferences from these tests. An important way to construct a fair test is to examine all the dimensions that has a potential to lead to DIF in test items. Yet, the role of science curriculum has been examined in other studies and found crucial, but it has not been controlled within a DIF context with the NAEP Science Assessments.

Statement of Problem

NAEP tests are designed to measure the proficiency of students in a subject but the standards used to educate/prepare these students are different. While NAEP has its own standards when preparing the exam questions, states have maintained their own standards and judging students from different states with various levels of science standards may not be fair. In this manner, besides the overall student ability, differences in science standards might affect the individual's success in answering different questions in NAEP assessments and create inequality. Detecting this inequality using DIF analysis is an important criterion when making claims regarding whether an item should be included in an assessment or whether modification is required in order to reduce or eliminate construct-irrelevant variance across comparison groups.

Differences in state science standards might lead to DIF in NAEP science exams because they generate differential curriculum coverage, in content, depth, the teaching techniques, or the order of the material covered, across different states. Curriculum differences can result in varying degrees of student exposure to the content and processes required to answer the items correctly which leads to differential response patterns and difficulty levels. This view is supported by curricular and testing professionals such as Mehrens & Phillips, 1986; Miller & Linn, 1988; Ercikan, et al., 2004; and Emenogu & Childs, 2005. In this manner, if NAEP science exams have more items that are appropriate for some states than others, it brings into question the adequacy of the test items for cross-state comparisons.

Purpose of Study

The purpose of this investigation was to examine and evaluate whether differences in K-12 science standards across states lead to differential item functioning (DIF) in NAEP science assessments test items. Due to differences in instruction, curriculum, students in different states potentially are exposed to different subject matters and teaching methods. Different states also place different weights on different topics and the order in which topics are introduced and subsequent instructions provided to students are different as well. Therefore, if some subjects are omitted or teaching standards for topics are shortsighted in a state, the order in which topics are introduced are different, the method of the topics covered is outdated, or not proper for science educations, etc., then students can be expected to do poorly in some items relative to their performance on the rest of the test.

Therefore, and in summary, this study examined whether the differences in K-12 science standards across states generate differential item functioning (DIF) in NAEP science assessments test items in 4th, 8th, and 12th grade students. If differences in state science standards lead to DIF in some of the NAEP science assessment questions due to similarity of science standards in NAEP and some states that have closer standards to NAEP, then it is expected to find that students who are educated in these closer standard states should perform better in these questions even if they have a similar ability in the remaining science items. After carefully examining this first research question, this study further examined two additional research questions to assess the robustness of first research hypothesis findings and to distinguish the findings from the well-known determinants of DIF. As a second research question, this study examined whether controlling for gender and race, which are documented to be the most prominent sources of DIF, affect the findings in the first research question. For the third research hypothesis, this study introduced the idea of measuring more exacting examinee ability using subscale scores within each broad content area instead of the test total score and repeated the main analyses using subscale scores. A more detailed explanation of the research questions, their hypothesis development, and statistical methods employed are presented in Chapter 3.

Chapter 2: Literature Review

Development of State Science Standards in the US

Starting by the mid-1980s, most of the American public and policymakers accepted the idea that the United States had an escalating educational crisis. The National Commission on Excellence in Education (NCEE) panel, which included several people from various backgrounds and political views, produced a unanimous and very influential report, A Nation at Risk: The Imperative for Education Reform[1]. The commission, in an open letter, argued that the American education is in an inacceptable state and emphasized that US dominance in commerce, industry, science, and technological innovation is under the threat. One of the most important messages from the report is that to reverse the decline in quality of American education, "state and local high school graduation course requirements should be strengthened, higher academic standards be established, more time be spent in school, the preparation of teachers be improved,

[1] https://www2.ed.gov/pubs/NatAtRisk/index.html

and that elected officials across the nation be held accountable for making the necessary improvements".

This improvement required an understanding of relative state and local achievements. However, while The National Assessment of Educational Progress (NAEP) had been in place since 1969, strong opposition from state officials and certain education associations had prevented the reporting of those results at a state level. Following the highly influential report of A Nation at Risk, Secretary of Education Terrel Bell instituted "wall chart" in 1984, which allowed for ranking the states by their educational achievements. Following a significant media attention, state level NAEP data became even more important in monitoring the progress of student achievement in the forthcoming national education goal.

When improving education at the national and state levels became an important issue for the American public, at the December 1988 meeting of the president-elect and the governors, the idea of setting long-range education goals was proposed, and both sides agreed to pursue it further. This idea has been more materialized in 1989 Charlottesville education summit in which national education goals for the year of 2000 has been declared by the governors. After several rounds of federal legislation, every state ultimately started to develop its own academic standards in the core subjects of the K–12 curriculum.[2]

The purpose of K–12 standards is to communicate the knowledge, skills, and capabilities that students should gain along the K–12 path. State standards set clear academic standards for the schools by conveying the critical science content students need to learn and properly sequence and prioritize that content. In other words, these standards are the foundations of the

[2] More detailed information about this summit is available from https://govinfo.library.unt.edu/negp/reports/negp30.pdf

curricular and instructional materials and assessments in each state and they provide the necessary foundation for local decisions around curriculum, assessments, and instruction.

A Brief History of NAEP Assessments

The National Assessment of Educational Progress (NAEP) is a congressionally mandated and largest assessment of students' knowledge and ability in various content areas across the nation, states, and in some urban districts. It is administered by the National Center for Education Statistics (NCES), within the Institute of Education Sciences (IES) of the U.S. Department of Education.

The first administration of NAEP was in 1969. Since then, NAEP has gathered information on student achievement in selected academic subjects. In the early periods, the results were reported on an item-by-item basis for the nation, regions of the country, and certain demographic groups. This item-by-item reporting limited the attention from policymakers and the general public. Starting 1983, collaborating with ETS, results in a content area were reported and NAEP became known as the "Nation's Report Card. In 1990, for the first time, results were reported state-by-state and in terms of achievement levels by various subgroups of students, defined by demographic conditions related to geographical, racial, ethnic, sociological, and poverty markers.

Currently, NCES administers NAEP assessments in a variety of subjects including Arts, Civics, Economics, Geography, Mathematics, Reading, Science, Technology and Engineering Literacy (TEL), History, and Writing. Among these subject areas, while national results are available for all subjects assessed by NAEP, state and selected urban district results are available for mathematics, reading, and (in some assessment years) science and writing.

In NAEP assessments, a representative sample of the nation's students' academic performance in various subjects are measured at grades 4, 8 and 12. In order to ensure that the results reflect a complete representation of the nation's performance, the tests are administrated in a sample of schools whose students reflect the varying demographics of a specific area. In each school, the students are chosen at random to participate in NAEP and regardless of race/ethnicity, socioeconomic status, disability, or any other factor, every student has the same chance of being chosen.

NAEP results are currently used for three major purposes: monitoring trends in student achievement; providing evaluative statements regarding the level of student achievement; and making state-by-state comparisons. NAEP exam results have important consequences since teachers, principals, parents, policymakers, and researchers all use NAEP results to compare their progress with other states and develop ways to improve their education. The results of NAEP exams also get significant publicity which further creates a pressure on state officials to improve their education in various content areas. In particular, state and district level decision makers started to apply the results, sometimes inappropriately, to policies and program planning which makes the assessments important for all stake holders.

In science content area, the first assessment, whose results were available at the state level, is administrated in 1996. From then, NAEP science assessments are given in 2000, 2005, 2009, 2011, and 2015. NAEP introduced a new framework in 2009 which replaced the one used for the 1996, 2000, and 2005 science assessments. The new framework is developed to more fairly assess students due to the advances in both science and cognitive research, the growth in national and international science assessments, and advances in innovative assessment approaches.

According to the science framework, student's science knowledge and skills are measured in three broad areas: physical science, life science, and Earth and space sciences. As a second dimension of the framework, four main science practices, including identifying science principles, using science principles, using scientific inquiry, and using technological design, are identified to assess student science knowledge. Based on the NAEP results, examinees are categorized as Basic, Proficient and Advanced levels.

Differential Item Functioning

As discussed above, an important purpose of NAEP is to provide a fair and accurate measurement of student academic achievement and reporting of trends in such achievement. NAEP reports are increasingly used for monitoring the state of education in the subjects that are assessed, as models for designing other large-scale assessments, and for secondary research purposes. Therefore, validity and fairness investigations are becoming important part of NAEP assessments.

An essential part of validity and fairness investigations is the comparability of measurement at the item and test levels (Ercikan, 2006). When the probability of successfully answering an item is affected by construct-irrelevant factors such as differential familiarity with item types, formats, or vocabulary knowledge for one or more of the comparison groups then differential item functioning (DIF) is signaled for that item. (Gierl et al., 1999; Shepard, Camilli, & Averill, 1981, Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Ercikan & Lyons-Thomas, 2013; Oliveri & Ercikan, 2011). Accurate DIF detection is central to making claims regarding whether an item should be used in an assessment or whether modification is required in order to reduce or eliminate construct-irrelevant variance across comparison groups. When a test includes a high number of DIF items, the cross-state or cross-group comparability is at risk. The interpretation of a scale can be severely biased due to unstable item characteristics from one state/group to another. In this regard, DIF analysis becomes an important part of assessing the fairness and validity of NAEP assessments.

Another important part of DIF investigations is to understand whether the DIF is uniform or nonuniform across different ability groups. Uniform DIF indicates that one group is systematically at a disadvantage when responding to the item. Non-uniform DIF means that one group has an advantage for some proficiency levels but is at a disadvantage (or loses the advantage) at other proficiency levels. For example, assume that differences in state science standards favor students in superior standard states in one item. This finding indicates that there is DIF for that item. The next in DIF analysis would be examining whether the DIF is constant across ability levels (uniform DIF) or varying across ability level (non-uniform DIF). Finding a uniform DIF indicates that, being in a superior standard state is advantageous for students regardless of their ability level. On the other hand, finding a non-uniform DIF indicates that being in a superior standard state is advantageous only for students with certain ability levels and is not advantageous (or even disadvantageous) for other ability levels.

In context of state science standards, detecting the type of DIF, uniform or non-uniform, is important since ability level might change the marginal effect of difference in state standards on students' probability to answer the questions correctly. For example, it might be argued that high ability students might be less affected by differences in state science standards compared to low ability students since they already know the subjects well and can close the gap with their high level of ability. In this case, one would find that differences in state science standards would result in DIF, only among low ability students but not among high ability students.

A number of different statistical methods are used for DIF detection. The most prominent ones include Mantel-Haenszel (MH) method (Holland & Thayer, 1988), the standardized p-difference index (Dorans & Holland, 1993), logistic regression (Swaminathan & Rogers, 1990), IRT approaches Lord (1980), Raju (1988) and Holland and Wainer (1993), Raju‛s area measures (Cohen & Kim, 1993), SIBTEST (Shealy & Stout, 1993), and Rasch-based random coefficient multinomial logit model (RCMLM) for DIF detection (Meulders & Xie, 2004).

Each of these DIF detection methods has its own advantages and disadvantages (Millsap & Everson, 1993). One important difference between these methods is that, while the MH procedure, standardized p-difference index, and logistic regression method are based on observed scores, Raju‛s area measures and Rasch-based logit models assume the unobserved latent variable underlying the assessed performance. The use of the total score as a proxy of the latent trait encounters problems when the responses follow complex IRT models. However, since NAEP uses imbalanced booklet method for all the assessment, in which students answer only part of the exam questions, matching students based on their total scores is more proper for examining DIF in NAEP exams.

For this research, logistic regression (Swaminathan & Rogers, 1990) and Mantel-Haenszel (Holland & Thayer, 1988) methods were used to detect DIF. The Mantel-Haenszel (MH) statistic is one of the most widely used methods in detecting item-level measurement bias, largely because it is conceptually simple, relatively easy to use, and provides a chi-square test significance. NAEP also uses MH to perform gender and race based DIF analysis. Moreover, besides a test of the null hypothesis, it also estimates the size of DIF in an item. However, this method cannot detect non-uniform DIF which indicates the degree of difference in item bias for examinees with low and high total scores. The logistic regression method can detect both uniform and nonuniform DIF. Indeed, logistic regression has been found to be more powerful than an IRT based analysis of variance method at detecting (nonuniform) DIF (Whitmore and Schumacker, 1999). It can be extended to multiple examinee groups and one can include various other controls in the specification. Furthermore, these two methods are selected because they are examples of most used DIF detection methods, and they can be automatized in commonly used statistical software. Unlike IRT models, they do not require large sample size, which would become a potential problem particularly for the analysis of 12th grade. Using the two methods also assesses the robustness of my findings. SAS software is used to conduct analyses.

The topic of DIF has been researched in large scale assessments as it indicates potential item bias against a comparison group. Studies have compared the functioning of items for females and males, for students of different ethnicities or cultural backgrounds, and for students taking tests in different languages (e.g., Allalouf, Hambleton, & Sireci, 1999; Gierl & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999; Swanson, Clauser, Case, Nungester, & Featherman, 2002; Zwick, Thayer, & Lewis 2000).

As an important basis for this study, curriculum and the match between the curriculum and the content is shown to be an important part of the DIF investigations. For example, Harnish and Linn (1981), Lawson, Bordignon, and Nagy (2002), Leinhardt and Seewald (1981), Mehrens and Phillips (1986), and Muthén, Kao, and Burstein (1991) examined the effects of differences in instructional experiences of students on the resulting achievement estimates and observed item difficulties. Furthermore, Mehrens & Phillips (1986) and Miller & Linn (1988) have suggested that the degree of match between an assessment and the curriculum can have a large impact on achievement test scores. This study complements these studies by examining the role of standard differences state science standards which would affect the curriculum and the match between the curriculum and the content in the context of DIF analysis.

Differential Item Functioning in NAEP Assessments

In order to conduct DIF analysis in NAEP science assessments, one needs to use restricted NAEP data which provides examinee level information. However, potentially due to difficulty of obtaining the data and the length of the bureaucratic procedures, the use of restricted data for DIF analysis is not very common in the literature. As an earlier example, Zwick and Ercikan (1989) used data from the 1986 United States history assessment from the National Assessment of Educational Progress (NAEP) to examine the effects of using more homogeneous (e.g., dissected gender) groups as compared to larger (heterogeneous gender) groups on the number of DIF items. It is worth to note that the data is restricted and obtaining and working on the data is subject to severe restrictions, which might be one of the reasons why we do not have many studies focusing NAEP restricted data to conduct DIF analysis. Moreover, Educational Testing Service (ETS) and National Center for Education Statistics (NCES) regularly conduct DIF analysis for large scale assessments including NAEP and report the results.

As reported in NCES national report card,[3] while several procedures have been used to identify differential item functioning (DIF) in NAEP science tests, analyses for each assessment involved three comparisons: male vs. female students, White vs. Black students, and White vs. Hispanic students. However, the potential effect of differences in state science standards have been ignored in DIF analysis. Taking curriculum differences across states into account is very important since there are important implications of ignoring potential DIF sources. Potential development and administration of biased tests due to under-detection, cancellation, and

[3] http://nces.ed.gov/nationsreportcard/naepdata.

inaccurate interpretation of DIF would be good examples of such implications (Ercikan & Oliveri, 2013). It might also contribute to challenges in identifying sources of DIF documented in previous research (Ercikan et al., 2010; Oliveri & Ercikan, 2011).

Differences in State and NAEP Standards in Science Assessments

Due to NAEP's increasingly important role as a powerful policy benchmark, Congress called for an independent evaluation of NAEP in 2002. The purpose of this congressional mandate is to investigate whether the assessment program follows professional standards, with emphasis given to the achievement levels, sampling procedures, and fairness issues. For example, Buckendahl et. al. (2009) is one of these reports which assess the quality and consequences of NAEP assessments. In particular, the report emphasizes that when making comparisons of achievement among states using NAEP, a critical issue is the degree of alignment between the assessment (i.e. the NAEP assessment framework and questions) and states' education systems characterized in their content standards, curricula, instructional practices, and assessments. The report also mentions the importance of validity and fairness of NAEP exams and encourages to conduct more differential item functioning to eliminate any bias from these exams.

An important issue that might affect the validity and fairness of the NAEP assessments is that NAEP prepares the exams based on their own standards and does not balance the coverage of content and process to accommodate differences across different states. NAEP frameworks are prepared by The Governing Board that works with a committee of subject matter experts, practitioners, and members of the general public—including researchers, educators, business leaders, and policymakers—to develop a rich and rigorous set of standards that define what students should know and be able to do in a particular subject.[4] Survey questionnaires, administered to students, teachers, and school administrators who participate in a science assessment, are also used to collect and report contextual information about students' learning experience in and out of the classroom. However, while PISA has already set a goal to make the exam in a way that would minimize the role of curriculum in measuring students overall scientific literacy (Huang 2009), NAEP lacks such a goal and continues to prepare the exams based on their own standards.

States also develop their own standards and decide on what framework to use in their education. In particular, in attempt to increase the science standards for K-12 education, 26 states[5], and important education institutions such as the National Research Council (NRC), the National Science Teachers Association (NSTA), the American Association for the Advancement of Science (AAAS) developed The Next Generation Science Standards (NGSS). NGSS are science standards for K-12 education and set the expectations for what students should know and able to do. Note that, most of the previous state standards were based on the Benchmarks for Science Literacy (1993) and the National Science Education Standards (1996). According to National Science Technology Association, currently 20 states and DC (representing over 36% of U.S. students) have adopted the Next Generation Science Standards (NGSS).[6]

NGSS have been developed following a two-step process. In the first step, framework for K–12 science education is developed which identified the science all K–12 students should

[4] https://nces.ed.gov/

[5] In developing NGSS, a total of 26 states including Arizona, Arkansas, California, Delaware, Georgia, Illinois, Iowa, Kansas, Kentucky, Maine, Maryland, Massachusetts, Michigan, Minnesota, Montana, New Jersey, New York, North Carolina, Ohio, Oregon, Rhode Island, South Dakota, Tennessee, Vermont, Washington, and West Virginia were involved. Currently, around 40 states have shown interest in the standards.

[6] The 20 states are Arkansas, California, Connecticut, Delaware, Hawaii, Illinois, Iowa, Kansas, Kentucky, Maine, Maryland, Michigan, Nevada, New Hampshire, New Jersey, New Mexico, Oregon, Rhode Island, Vermont and Washington.

know. In the second stage, NGSS are developed to prepared students for college and careers. NGSS aims to improve state science standards in several dimensions. They provide a set of performance expectations that integrate practices, fundamental ideas, and crosscutting concepts to prepare all students for college, career, and citizenship. Also, rather than focusing on multiple choice questions that emphasized definitions, they are designed to be assess in real world contexts. For example, unlike the existing state standards before NGSS, engineering is aimed to be integrated with science and explicit connections to mathematics and English language arts included.

It is important to note that, although states attempt to increase their science education by adopting higher standards, the process of generating state science standards varies from state to state. Usually, in each state, the department of education assembles a committee of scientists, teachers, parents, and others to write and/or revise earlier standards. In some cases, state school boards get involved in the process, which has the adverse effect of placing elected officials, with little or no knowledge or expertise in the fields whose curricula they govern, responsible for approving standards written by experts. Hence, how states adopt standards and use them in their curricula significantly varies and the final outcome of the standards might be different than their initial goal.

Gross et al (2013) evaluated The Next Generation Science Standards (NGSS) to assess their effectiveness in teaching science to K-12 students. For the purposes of this study, one of the most important conclusions from Gross et al. is that there are important differences in state science standards. According to Gross et al. some states exhibit strong weakness in setting appropriately clear, rigorous, and specific standards and clarifying what they expect of their schools, teachers, and students in science education. These clearly weak standard states are categorized as *inferior* standard states. On the other hand, some of the states very strongly set clear, rigorous, and specific standards and clarified their expectation. These clearly strong states are categorized as *superior* standard states.

Since these standards are the foundations of the curricular and instructional materials, they affect the curriculum and students' science learning experience in each state. Due to differences in science standards, there are significant differences in state science curriculums and students in different states are exposed to different subject matters and teaching methods. Furthermore, the emphasis and the sequence of the topics are also significantly different in various states.

In addition, Gross et. al 2013 emphasize that NAEP science exams are constructed without taking the differences in state science standards into account and the questions are prepared using similar standards compared to the superior standard states.[7] They argue that NAEP standards include the necessary ground with neither critical omissions nor trivialities and takes a score of 9 out of 10 for its standards that are comparable to the ones adopted in superior standard states. In other words, while NAEP science exam is taken by students in various states with varying degrees of curricular standards, the test questions are prepared using standards which are closer to the ones in superior standard states. This similarity in standards would increase similarity in curriculum and familiarity with the terminology, the concept and the logic of the questions compared to the students in inferior standard states.

[7] More detailed information on NAEP science framework can be obtained from https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/science/2015-science-framework.pdf.

Chapter 3. Hypothesis Development

Research Question 1: Differences in Science Standards as a Source of DIF

This study argues that differences in science standards across states have a significant potential to lead to DIF in NAEP science exam for at least three important reasons. First, there is a significant difference in the materials covered between inferior and superior standard states and some essential content was omitted from curriculum in inferior standard states. According to the literature, the match between the curriculum and the content of the test is important for DIF analysis. For example, Harnish and Linn (1981), Lawson, Bordignon, and Nagy (2002), Leinhardt and Seewald (1981), Mehrens and Phillips (1986), and Muthén, Kao, and Burstein (1991) examined the roles of differences in instructional experiences and curriculum on student achievements and found that the degree of match between an assessment and the curriculum can have a large impact on achievement test scores.

For example, Alabama is categorized an inferior science standard state and evolution, which should be an essential element of the genetics content, is almost entirely missing from science education (Gross et. al 2013). A similar criticism also applies to many other inferior standard states. On the other hand, virtually all critical life science content including evolution, is included and well developed in curriculum of superior standard states. Indeed, Bowman (2008) finds that students in weak standard states are three times as likely as those in strong standard states to receive instruction that evolution is not scientifically credible.

Hence, even if we assume that all other materials are equally covered in inferior vs superior standard states, since NAEP has superior science standards, it is likely that a question about evolution might be asked in the exam. Hence, students in inferior standard states would do poorly on these questions even if they perform similar to superior standard states in other questions. Another example of omitted subject is carbon chemistry. According to Gross et. al., topics such as hydrogen bonding, Lewis dot structures, carbon chemistry, molecular shapes and polarities are generally not well covered in high schools, the problem mainly present more widely in inferior state standards. Hence, these two examples would give an idea of how missing or omitted content might affect examinee responses in NAEP science exams.

Second, there are significant differences in the sequence of courses, or the time spent on topics. According to Sireci and Swaminathan (1996) these differences might lead to DIF. This is because sequential courses can be considered as building blocks of an efficient learning for certain subject and wrong sequences or improper focus on certain topics would impact student learnings in certain subjects. For example, in inferior standard states evaluation of grade-to-grade progress is rather weak and some content that was never explicitly stated in earlier grades was nevertheless assumed in later grades. This would cause some of the objects in the higher grades in inferior standard states not to be covered in depth and students' overall learning of this material in higher grades would be significantly affected. Furthermore, the standards differ in math requirements which is essential to the learning of physics and chemistry at the high school level. Hence, students who lack the ability of math from earlier grades and who are not well educated in math during high school education would have difficulty in learning subjects which necessitates a significant level of knowledge in math. In this case, probability to correctly answer the questions in NAEP science exams, which are designed to measure students' knowledge and skill in science, would be affected by examinees' math knowledge.

As a third point, differential availability of textbooks, teaching practices, expectations from students, the use of vocabulary, and other materials across different states might be important sources of DIF in exams (Huang, 2010). For example, NAEP identifies four science practices including identifying science principles, using science principles, using scientific inquiry, and using technological design to prepare the exam questions. Students are expected to apply the content they learn in a design or engineering problem. However, it is very unlikely that these practices are covered in inferior standards states in a way they are covered in superior standard states.

For example, inquiry-based learning standards, which means helping students learn scientific content through discovery, as opposed to through direct instruction of specific content, are vague to the point of uselessness in inferior states. In Idaho (inferior standard state), for instance, students are merely asked to "make observations" or to "use cooperation and interaction skills. Again, any question related to scientific inquiry has potential to create differences in response patterns between inferior and superior standard states. Another example can be from the use of vocabulary in science education. For example, standards mention list of technical vocabulary words that students should learn, like convergent or divergent plate boundary and atmospheric layers. However, if this terminology is not adequately explained in inferior standard states, then any questions using these words, although the central theme of the question is not about the particular word, would favor superior standard states.

Overall, differences in state science standards have a significant potential to result in curricular differences in favor of students from superior standard states. This difference ultimately could lead to a differential responding pattern across inferior and superior standard states. It is important to note that, while differences in curriculum might result in DIF, whether these items should be excluded from the exam is a part of the DIF investigation after DIF detection. Even if an item is flagged as exhibiting DIF, a more comprehensive and careful investigation is further need to determine the sources of DIF. In most practices, DIF flagged items are later reviewed by professionals to decide on whether the sources of the DIF is related to latent treat the exam is aiming to measure. If the source of DIF is related to the main latent trait then the item is fair and should not be excluded. Nonetheless, the act of identifying these gaps in conceptual understanding can inform teaching and, subsequently, help educators and policy makers to reduce such gaps in NAEP exams. Unfortunately, for this study since the 2015 NAEP science questions have not been released to the public and are still secure for use in NAEP assessments, examining these questions is not be possible for an aftermath investigation of DIF detected questions.

Therefore, the first research question for this study is as following:

1-) Do differences in state science standards result in differential performance in NAEP science assessment test items at grades 4, 8, and 12?

Research Question 2: Controlling for Gender and Race of Examinees

It is well established that gender and race are among the most important and well-established sources of DIF (Oliveri, 2012). For example, Bolger & Kellaghan (1990), Hamilton, (1999), and Zenisky, Hambleton, & Robin (2003) suggested that multiple-choice items seem to benefit males, while open-ended items are more biased for females. On the other hand, Becker, 1989; Burkam, Lee & Smerdon, 1997; Jovanovic, Solano-Flores, & Shavelson, 1994; Young & Fraser, 1994 studied the effect of item contents and find that males seem to outperform females on physical, earth, and space science items. Consistently, items requiring spatial reasoning or visual content favored males (Halpern, 1992) and test characteristics and culture also contribute to gender-based DIF. On the other hand, several researchers also studied potential causes of DIF regarding race (Holland & Wainer, 1993; Hough, Oswald, & Ployhart, 2001).

As a part of the fairness investigation, for each of their assessments, NAEP conducts DIF analysis. In the analysis, NAEP mainly focuses on two well established sources of DIF: race and gender. NAEP provides three comparisons: scores of male students versus scores of female students, scores of White students versus scores of Black students, and scores of White students versus scores of Hispanic students. The fact that NAEP provides DIF analysis based on these two student characteristics raises the issue that any DIF detection based on differences in state science standards should not be an artifact of differences in gender or race. Hence, besides the state science standards, controlling for these two characteristics also ensures that any DIF detection based on state science standards is not an artifact of differences in race and gender between superior and inferior standard states. In other words, while the first question focuses on comparing the inferior standard states examines with superior standard state examines, it is also important to take within group heterogeneity based on gender and race into account (Ercikan & Oliveri, 2013) and perform the analysis accordingly. This idea yields the following research question.

2. *Does controlling for gender and race of the student impact DIF detection on NAEP science assessment test items at grades 4, 8, and 12?*

Research Question 3: An Alternative Measure of ability

DIF analyses have primarily been conducted by comparing performances of subgroups on each item within sets of examinees having the same level of skill. Since examinees' true level of ability is unobservable, measuring the skill of examinees becomes the key decision for the accuracy of the DIF analyses. As suggested by the literature, I used the total score of examinees on the entire exam to measure ability. However, controlling for examinees' skills in each content sub-area of a test might better capture their ability and yield more accurate DIF results compared to using the total score of examinees in the test. In the literature, one noticeable study by Sipahi and Poggio (2020) examines this issue. Using 2015 NAEP science exam, they examined the effect of content area ability on gender and race based DIF items. They find that in 4th, 8th, and 12th grades, controlling for content ability reduces the number of DIF items by at least 20%.

DIF analyses primarily have been conducted by comparing performances of subgroups on each item within sets of examinees having the same level of skill in the content area being tested. Since examinees' true level of ability is unobservable, how the researcher measures the skill of examinees becomes the key decision for the accuracy of the DIF analyses. In most of the cases, the literature suggests using the total score of examinees on the entire exam when we assume a unidimensional trait is being assessed. The assumption is to use examinee's total test score to create homogenous groupings, which is essential for DIF analysis (Oliveri, 2012), with respect to ability across examinees then other factors as gender, race, cultural grouping, income, etc., would be a potential source of DIF. MH and logistic regression analysis also matches students based on their total score and all the inferences are based on this total score measure of ability.

However, NAEP measures students' learning in three different content areas: physical science, life science, and Earth and space sciences. Therefore, students' distinct sub-score, standard or indicator content ability as measured by their subscale scores within each content area instead of total test score might better reflect their specific trait ability. The following example better illustrates this issue. Consider two otherwise identical students, A and B, who have a total score of 70 out of 100 in a science exam with two content areas S1 and S2. According to the traditional DIF analysis, we treat these two students to have equal ability in this science exam. However, if student A gets 50 from content area S1 and student B gets 50 from content area S2, then comparing these students for the same question would not be accurate. Apparently, while student A has more ability in content area S1, student B would have more ability in content area S2 but this information is lost when we simply use the total score of 70 in the analysis. Hence, instead of the total test score of 70, one should better dissect this total score into more refined content area scores (50 and 20) and control for them separately in the DIF analysis. Indeed, using 2015 NAEP science exam, Sipahi and Poggio (2020) examined the effect of content area ability on gender and race based DIF items. They find that in 4th, 8th, and 12th grades, controlling for content ability reduces the number of DIF items by at least 20%.

Several factors might contribute to the differences in students, ability/success in different content areas. First, the ability of a student to solve problems might be different for different content areas of the test. While a student might be superior in solving items in physical science, another one might be better in life science when taking a science test. Students' interests in the subject due to their geographical or personal differences might affect this outcome. Second, the sources of DIF such as language or cultural differences might affect the success of students differently in different content areas. For example, the vocabulary used or the presentation of questions such as including highly visual stimulus items that are shown to create DIF is not necessarily the same across different content areas. While visual items might be used in one area, more language skills might be required in other content areas. Finally, due to curriculum differences across states, students might excel in one area while ignoring the other areas. From the research hypothesis perspective, operationally, the degree of relationship between subskill scores operationally sets the criterion as to when to use subscale (defined by subskill, standard, indicator, or objective) scores rather than test total score. This argument yields the following third research question.

3. Does controlling for separate scores on the three distinct content areas of physical science, life science, and earth and space sciences affect DIF detection based on state science standards in NAEP science assessments test items at grades 4, 8, and 12?

Chapter 4: Research Design and Methods

Datasets

This study used 2015 NAEP science assessments for grades 4, 8 and 12. Restricted respondent-level data as maintained by NCES is used in all the analysis. Responses of students along with student characteristics, school district and state identification were included in the data. More information on accessing and using restricted data is available from NCES's web site.[8] NAEP data was chosen for this study because as stated by NCES, NAEP results provide a national sample using uniform questions and serve as a common metric for all states and selected urban districts. Participating schools and students were selected to be representative of all schools nationally.

The restricted dataset was provided in CDs and the data analysis can be conducted only in a location approved by the NCES. Only approved people have access to the datasets and all publications and dissemination of analysis using the restricted data is subject to NCES' approval.

In Table 1, the information about the number of examinees and number of questions are provided. The results from the 2015 science assessment at grades 4, 8, and 12 were based on a representative sample of 115,400 fourth graders from 7,650 schools, 110,900 eighth graders from 6,050 schools, and 11,000 twelfth graders from 730 schools.[9]

[8] https://nces.ed.gov/statprog/instruct_access_faq.asp

[9] Due to confidentiality requirements of NAEP, the number of students is rounded to the nearest hundred. Alaska, Colorado, Louisiana, Pennsylvania, and the District of Columbia did not participate in the 2015 NAEP science assessment.

There are 99, 111, and 126 multiple choice questions in 4th, 8th and 12th grade exams, respectively. As evident from Panel B of Table 1, while Physical Science, Life Science, and Earth and Space Sciences gain equal weight at grade 4; more emphasis on Earth and Space Sciences is given at grade 8; and a shift to more emphasis on Physical Science and Life Science is present at grade 12.

Each student in the exam was given two blocks of questions and therefore, the number of questions given to answer was a minimum of 29 questions. However, it is important to note that due to this method, for each question only a part of the entire student population in each grade level answered each question. Panel C of Table 1 provides a summary of number of students in inferior and superior standard states answering the questions. According to Panel C of Table 1, on average, 6900, 6200, 330 students answered each question in 4th, 8th, and 12th grade students in inferior standard states and 5700, 5400, 750 students answered each question in 4th, 8th, and 12th grade students in superior standard states.[10]

According to Zwick (2012), 500 and 200 examinees in reference and focal groups, respectively, are needed to conduct a robust DIF analysis for a 50-item exam. Hence, the number of students in each grade level suggested that there was enough number of observations for each question to conduct DIF analysis based on the state science standards.

[10] Due to confidentiality requirements of NAEP, the number of students is rounded to the nearest hundred or tenth.

Table 1: 2015 NAEP Science Exam and Population Characteristics

	Panel A: Total Number of Examinees		
	4th Grade	8th Grade	12th Grade
	115,400	110,900	11,000

	Panel B: Number of Multiple-Choice Questions		
	4th Grade	8th Grade	12th Grade
Physical Science	31	37	44
Earth and Space Science	34	44	29
Life Science	34	30	53
Total	99	111	126

	Panel C: Average Number of Examinee Responses		
	4th Grade	8th Grade	12th Grade
Inferior Standard	6900	6200	330
Superior Standard	5700	5400	750

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Determining State Science Standards Rankings

An important part this study is to categorize states using their science standards. This necessitates the evaluation and comparison of every states and NAEP's science standards by professionals in various areas. For example, while educators in one state might know strengths and weaknesses of their own science standards, unless they also examine NAEP and other states' standards, they would not know how their own standards stand relative to other standards.

In the literature, state science standards have been formally reviewed by Braden et al. (2000), Gross et al. (2005), Gross et al. (2013). These studies are all funded by the Fordham Institute and several professional with different backgrounds prepared these reports. There have been some other studies, including Lerner (2000) and Mead (2009), which also examined state science standards but their focused were mainly limited to how the state curriculum covers the topic evolution. It is also important to note that this study examined NAEP 2015 science exam and state science standards are dynamically changing over time. Hence, one needs the evaluation of most recent science standards to have the best comparison of state science standards.

Most recently, Gross et al. (2013) report the evaluation of each state's K-12 science standards against Next Generation Science Standards (NGSS), whose aim is to create standards for science teaching in US schools. Their study is the most comprehensive report on relative state science standards and therefore best meets the needs of this study to rank states relative to each other's and NAEP science standards.

In the report, both NGSS and individual state standards are evaluated against a grading metric to determine how clearly and carefully states cover important content in four areas: physical science, life science, earth and space science, and scientific inquiry and methodology. The report indicates that, to evaluate the most recently adopted standards, they searched the websites of state education departments and collected the most recent version of each state's science standards from its department of education website. For example, for state of Kansas science standards was downloaded from Kansas state department of education website.[11] The team also examined assessment frameworks and curriculum guides if these were characterized as key documents by the state. Next, the science standards coordinator(s) for each state has been contacted to confirm the accuracy of the documents and advise of a team of content experts has been asked to apply a set of criteria to them.

For each grade level, experts, who wrote the report, developed criteria that outlined the essential content that should be included in K-12 science standards. As an example, in Physical

[11] https://community.ksde.org/Default.aspx?tabid=5785

Science, the report stated that some of the general expectations for learnings through grade eight are to know the common forms and states of matter, including solids, liquids, and gases, elements, compounds, and mixture, to know how to use the standard units of measurement (SI, to define “gravity”, to understand kinetic and potential energy, and their transformations, to know matter is made of atoms etc. Similarly, for earth and space science, students are expected to describe the organization of matter in the universe into stars and galaxies, recognize Earth as one planet among its solar system neighbors, describe the motions of planets in the solar system and recognize our star as one of a multitude in the Milky Way, identify the sun as the major source of energy for processes on Earth’s surface, describe the hydrologic (water) cycle and etc.

Following the evaluation of a state’s standards are evaluated against the science content criteria, the standards are judged against a grading metric. The grading metric focuses on two dimensions: content and rigor vs clarity and specificity.

For the content and rigor dimension:

- the standards are examined on their content comprehensiveness and their efforts to distinguish between more important and less important content and skills
- the way the content is communicated both to the teachers and to the students
- the appropriateness of the level of rigor for the targeted grade levels
- the effectiveness of teaching methods

For the clarity and specificity dimension:

- the standards are examined to evaluate whether both scope and sequencing of the material are apparent and reasonable and whether they provide guidance to students, parents, teachers, curriculum directors, test developers, textbook writers on the expectations and goals of the standards.

Based on this careful evaluation, each state's and NAEP's science standards are scored to take a value between 1 and 10, where 10 represents the highest science standards. According to the report, while 13 states are classified as having clearly *superior* standards, 16 states are being classified as having clearly *inferior* standards. Table 2 shows the states based on their relative superiority or inferiority. It is also important to note that NAEP and TIMMS frameworks are also examined in this report and have been classified as clearly superior. The remaining states have relatively medium level of science standards and their relative superiority/inferiority is "too close to call." Hence, these states are excluded from the analysis to have a better comparison of science standards between clearly inferior and superior states.

Table 2: States by Their Relative K-12 Science Standards

Clearly superior	Medium Level	Clearly inferior
Arkansas	Alabama	Alaska*
California	Arizona	Colorado*
DC*[12]	Connecticut	Idaho
Indiana	Delaware	Iowa
Kansas	Florida	Kentucky
Louisiana*	Georgia	Montana
Maryland	Hawaii	Nebraska
Massachusetts	Maine	Nevada
New York	Michigan	New Jersey
Ohio	Minnesota	North Dakota
South Carolina	Mississippi	Oklahoma
Utah	Missouri	Oregon
Virginia	New Mexico	Pennsylvania*
NAEP	New Hampshire	South Dakota
	North Carolina	Wisconsin
	Rhode Island	Wyoming
	Tennessee	
	Texas	
	Vermont	
	Washington	
	West Virginia	

*[12] These states did not participate to NAEP 2015 science exam.

SOURCE: Gross, P.R. (2013). Thomas B. Fordham Institute Final Evaluation of the Next Generation Science Standards. Retrieved from: http://edexcellence.net/publications/final-evaluation-of-NGSS.html

Differential Item Functioning Procedures

Mantel-Haenszel

The MH DIF procedure compares dichotomous item performance between two groups after matching examinees on overall scores. The MH-test statistic is computed by comparing the observed frequency of correct and incorrect answers split out by group membership and ability level, to the expected frequency if there were no DIF. Examinees in the focal and reference groups were matched on total test or questionnaire scores by dividing examinees in both groups into defined strata on those scores. Since, Institute of Education Science (IES) uses balanced incomplete block (BIB) or partially balance incomplete block (pBIB) design to perform the test, there is no single set of items common to all examinees. Therefore, for each student, the measure of proficiency used was the total item score on the entire booklet. These scores were then pooled across booklets for each analysis.

Estimates of the odds ratio for a given item, across the strata of the matching variable, can be computed from a 2 x 2 x *K* contingency table with *k* denoting the *k*-th stratum, (k = 1, 2,…*K*). The following table shows the 2 x 2 contingency table for the *k*-th stratum of an item.

	Response (1)	Response (0)	Total
Reference group	a_k	b_k	
Focal group	c_k	d_k	
Total			N_k

The letters, a_k, b_k, c_k, d_k, in each cell represent the numbers of responders in the cells. N_k denotes the number of responders in the *k*-the stratum. Some DIF occurs if the odd ratio for an item is greater than 1 or less than 1. The common odds-ratios formula is:

$$OR_{MH} = \frac{\sum a_k d_k / N_k}{\sum b_k c_k \, / N_k}$$

In the SAS system, the Cochran-Mantel-Haenszel statistic (Landis, Heyman, & Koch, 1978) can be generated using the FREQ procedure. The CMH option in the TABLE statement requests this statistic. Responders were stratified on total scores using PROC RANK. The CMH statistics were separately obtained for each item. For any given item, the null hypothesis is that there is no association between the defined groups (state science standards in this project) and the item responses across strata.

Logistic Regression

Logistic regression as a test of DIF was proposed by Swaminathan and Rogers (1990) and Rogers and Swaminathan (1993). Logistic regression is a generalized linear model to calculate the probability of giving a correct answer to a dichotomous item given the ability and group membership. The probability of a positive response to an item is modeled as a function of total scores (Ability), group membership (G), and the interaction between total score (Ability) and

$$\mathrm{P}\ (\mathrm{Y} \ = \ 1|\mathrm{Ability}, \mathrm{G}) = \frac{\exp(\beta0 + \beta1 * \mathrm{Ability} + \beta2 * \mathrm{G} + \beta3 * \mathrm{Ability} * \mathrm{G}}{1 + \exp(\beta0 + \beta1 * \mathrm{Ability} + \beta2 * \mathrm{G} + \beta3 * \mathrm{Ability} * \mathrm{G})}$$

For each item, three models with increasing numbers of predictors are used. In the first model, Ability is included, in the second model both ability and group membership are included, and in the third model both ability, group membership and their interactions are included. A logit transformation is applied to the probability equations and the following models are obtained.

$$\text{Model 1: logit (P)} = \beta_0 + \beta_1 * \text{Ability}$$

$$\text{Model 2: logit (P)} = \beta_0 + \beta_1 * \text{ Ability} + \beta_2 * \textit{ Superior}$$

$$\text{Model 3: logit (P)} = \beta_0 + \beta_1 * \text{ Ability} + \beta_2 * \textit{ Superior} + \beta_3 * \text{ Ability } * \textit{ Superior}$$

where Ability denotes the value of the responder on the trait, and *Superior* denotes group membership based on science standards and P denotes the logit of the probability of responders answering positively or correctly. Like the MH procedure, responders' values on the trait being measured are represented by their total scores.

In order to determine the presence of DIF, we want to know whether β2 and β3 are significantly different from 0. β2 is different than zero when examinees in one group score higher on the item than examinees in the other group, conditional on ability level (uniform DIF). β3 is different than zero when there is an interaction effect between group membership and total test score (nonuniform DIF).

Alternatively, the likelihood ratio test (LRT) is used to compare the likelihood of two models. The model with the smaller -2logL has better fit to the data. The LRT statistic is calculated by:

$$G^2 = [-2lnL(model\_r)] - [-2lnL(model\_f)] \sim \chi^2{}_{(d)}$$

where model_r denotes reduced models and model denote full models. G^2 follows the chi square distribution and d is the difference in numbers of parameter between the reduced and full models.

The null hypothesis is that item parameters between reference and focal group do not differ. Uniform DIF can be identified by comparing the LRT statistic between Models 1 and 2, with degree of freedom (df) = 1. Nonuniform DIF is tested by comparing Models 2 and 3, with df = 1. An overall test of DIF can be conducted by comparing Models 1 and 3, with df = 2. Several SAS procedures can be used to carry out logistic regression analysis. I use the Logistic procedure to detect DIF for dichotomous items. The Logistic procedure fits linear logistic regression models for dichotomous response categories using Fisher's method to maximize the likelihood (ML) function.

It is important to note that, by using both MH and Logistic regression procedures, the robustness of the results can be easily assessed. However, only logistic regression analysis is useful to detect whether any superior standards based DIF is similar across different ability levels (i.e. uniform or non-uniform DIF). If differences in science standards lead to DIF similarly within high ability vs low ability students, then the coefficient on β_3, which capture the effect of differences in state science standards on probability of answering question correctly in different ability groups, should be significant (similarly, in likelihood ratio test (LRT), LRT statistic between Models2 and 3 should be significant).

DIF Magnitude

An important part of the DIF investigations is to determine the magnitude of the DIF effect and categorize questions based on not only statistical significance but how much it favors students in the focal group. For example, in case of state science standards, while it is important to detect statistical significance, how much differences in state science standards affect the probability of answering the questions correctly are also an important part of the DIF investigation.

The first DIF classification for the magnitude of DIF is developed in 1987 by ETS and further developed by Petersen 1988. Over the years, only with minor changes (see Zwick, Thayer, and Lewis (2000) and Zwick, R., Ye, L., & Isham, S.) most of the papers in the literature and NAEP uses this formulation by the ETS to assign items to these categories. ETS formulation uses the following MH D-DIF index, which was developed by Holland and Thayer (1988)

$$\text{MH D-DIF} = -2.35 \ln (\alpha_{MH})$$

where α_{MH} , represents Mantel-Haenszel common odds ratio. This MH D-DIF index is an estimate of the Mantel-Haenszel common odds ratio expressed on the delta scale for item difficulty. An MH D-DIF value of -1, for example, means that the item is estimated to be more difficult for the focal group than for the reference group by an average of one delta point, conditional on ability. An intuition behind these critical MH D-DIF values are given by Holland (2004) that the critical values of αMH can be thought as the 90th and the 95th percentile of the standard normal distribution (one tailed test), which are 1.282 and 1.645. Hence, if focal group is that much above or below that base group, then we conclude that there is DIF in an item. However, NAEP continues to use 1 and 1.5 (rounded values to the nearest half points) as the critical values, and therefore I also applied the same rule in this paper. This ensured that my results on state science standards and the ones presented by NAEP on gender and race are comparable.

In terms of odds ratios, an MH D-DIF statistic of -1 implies that $-2.35 \ln (\alpha_{MH}) = -1$, or $\alpha_{MH} = 1.5$. This means that the odds of answering correctly for the reference group are more than 50% higher than the odds of answering correctly for comparable members of the focal group. In our setting, this would mean that after matching the ability levels, the probability of correctly answering an item is 50% higher for students in superior standard states relative to inferior standard states. Similarly, an MH D-DIF of +1 means that the odds of answering correctly for the reference group are 1/1.530 = .653 times the odds of answering correctly for comparable members of the focal group.

A similar statistic is also calculated for logistic regression analysis as following.

$$\text{LR-D-DIF or } \Delta_{LR} = -2.35\ (\beta_2)$$

In this equation β_2 represents the coefficient on the variable of interest such as state standard level dummy in a logistic regression. The interpretation is similar to the one in MH analysis. Based these two statistics, below describes the categorization of DIF items NAEP uses in their own analysis.

Category A. Items with negligible or nonsignificant DIF. Defined by MH D-DIF (LR-D-DIF) not significantly (at 5% significance level) different from zero or absolute value less than 1.0.

Category B. items exhibiting a weak indication of DIF. Defined by MH D-DIF (LR-D-DIF) significantly (at 5% significance level) different from zero and absolute value of at least 1.0 and either less than 1.5.

Category C. items exhibiting a strong indication of DIF. Defined by absolute value of by MH D-DIF (LR-D-DIF) of at least 1.5 and significantly different than zero (at 5% significance level).

Briefly, the intuition behind categorizing the DIF items is that even if a factor, such as differences in state science standards, leads to statistically significant differences in odds of answering a question correctly, if the magnitude of this differences in not important, then we simply ignore the effect. Hence, the categorization highlights the importance of both statistical significance and magnitude of the effect and emphasizes to simultaneously have both effects to label an item as DIF item.

Analyses

Research Question 1: Detecting DIF based on differences in State Science Standards

In this first set of analyses, the first hypothesis of this study that whether the differences in state science standards result in differential item functioning for Dichotomous items in NAEP 2015 Science exam was tested.

The first set of analysis to detect state science standards based DIF was conducted using the Mantel Haenszel procedure as described above. For each question, students were ranked into deciles based on their ability which is computed as the examinees total score in other questions. Then a dummy variable, Superior, was created. This dummy variable equals to 1 if the state in which an examinee takes the exam belongs to the superior science standard, and zero otherwise. PROC FREQ procedure in SAS is used to conduct the analysis.

In the second set of analysis, logistic regression analysis was performed. In logistic regression analysis, the score of each student on each question (1 if the student answered the question correct and 0 otherwise) is regressed on student ability (their total score from the rest of the questions) and the state science standard dummy (Superior). Superior dummy variable takes a value of 1 if the state a student is taking the exam has relatively superior science standards and takes a value of 0 if the state has relatively inferior science standards.

In each of the MH and logistic regression analysis, whether an item exhibits DIF or not was determined using NAEP categorization as discussed above.

For logistic regression analysis, the type of DIF, whether uniform or non-uniform, also were examined. This analysis was conducted by focusing on the LRT statistic difference between model 2 and model 3. If any significance was detected, then it was concluded that there exists non-uniform DIF for that particular item.

Research Question 2: Controlling for Gender and Race in DIF analysis

The purpose of this second set of analysis was to examine whether controlling for gender and race differences among students affects the findings in the previous section. Since NAEP focuses on these two student characteristics on detecting DIF, it is important to investigate how controlling for race and gender alters the relation between student's probability of success in individual items and their state science standards.

To test this idea, the logistic regression analysis was repeated by controlling for gender and race of students. Note that, in this set of analysis, we controlled for more than one potential DIF source and therefore Mantel Haenszel analysis could not be performed to answer this question. The goal is to understand whether state science standard differences lead to DIF after controlling for gender and race of examinees. To control for the effect of gender and race in the analysis, three additional dummy variables were added to the second equation in the logistic regressions as following.

$$\text{logit (P)} = \beta_0 + \beta_1\text{*Ability} + \beta_2\text{*Superior} + \beta_3\text{*Gender} + \beta_4\text{*Hispanic} + \beta_5\text{*Black}$$

In this regression, Gender equals to 1 if a student is male and 0 if a student is female. Black and Hispanic are race dummies and they take value of 1 if a student is black or Hispanic, respectively, and zero otherwise. The coefficient on the state science standards dummy can be interpreted as how successfully answering an item depend on differences in state science standards after controlling for gender and race of the students. This analysis also addresses the concerns related to within group heterogeneity (Ercikan & Oliveri, 2013) and assess the robustness of the finding after controlling for two important potentially confounding student characteristics.

Research Question 3: Dissecting Ability by Content Area

In the third research question, the findings in the first research question were re-examined by using the separate subscale skill measures instead of the total score of students from the test. The analysis was conducted only using logistic regression model because of the necessity of controlling for three different ability levels in the same model. Specifically, the following logistic regression model is used.

$$\text{logit (P)}= \beta_0 + \beta_1\text{*Physical\_Ability} + \beta_2\text{*Life\_Ability} + \beta_3\text{*Earth\_Ability} + \beta_4\text{* Superior}$$

In this regression, Physical_Ability, Life_Ability, and Earth_Ability variables were calculated as each examinee's total score in each of the distinct content areas. The main variable of coefficient is β_4 . This coefficient captured the effect of differences in state science standards on examinees probability of correctly answering different items after controlling for student ability in each content area separately.

Chapter 5: Results

Research Question 1: Detecting DIF based on differences in State Science Standards

In this section, the results for the tests on whether the differences in state science standards lead to any differential item functioning for Dichotomous items in NAEP 2015 Science exam for 4th, 8th, and 12th grade students are presented. Two set of results, Mantel Haenszel and logistic regressions, are presented.

As the first set of results, the summary results of the Mantel-Haenszel procedure are presented in Tables, 3, 4 and 5. In each table, Cochran-Mantel-Haenszel Statistics (CMH_Statistics, Landis, Heyman, & Koch, 1978), the associated p_value, the odds ratio (Odds_Ratio) and the Delta Scale are shown. For each grade, while the analysis is conducted for all the questions, the results are presented only for the questions which attain a minimum statistical significance level of 5%.

According to the results in Table 3, among 4th grade students, 34 questions (out of 99) exhibit statistical significance based on the CMH statistics (i.e. p_value is smaller than 0.05 suggesting a significant at least at 5% level). Statistically significant CMH statistics for these 34 questions show that, the odds of answering a question is significantly higher for one group compared to the other. Since, some odds ratios are smaller and some are bigger than 1, the tests suggest that both inferior and superior standards favor examines. However, when magnitude significance of these statistical significances is further examined, the delta score (in absolute terms) based on the MH Adjusted Odds ratio never exceeds 1. In particular, the delta scale ranges from -0.582 to 0.548 and never exceeds the limits set by the NAEP to be considered to exhibit DIF. Hence, all the questions that exhibit statistical significance fall into DIF Category A which indicates a negligible or nonsignificant DIF. Therefore, based on the MH procedure, the differences in state science standards do not result in DIF in NAEP science tests for the 4th graders when I examine the entire 4th grade sample.

Similar analysis was repeated for the 8th grade students. According to the results in Table 4, 50 (out of 111) questions exhibits statistical significance in the CMH_Statistics. However, once again, when the effect size is examined, none of the questions exhibit any significant effect size. In particular, the delta scale ranges from -0.826 to 0.569 but never exceed the limits set by the NAEP to be considered to exhibit DIF and fall into DIF Category A which indicates a negligible or nonsignificant DIF. Hence, it is concluded that state science standards do not lead to DIF in NAEP science tests for the 8th grade student when the overall groups are examined.

Table 3. Summarized Results of MH method for 4th Grades

Item #	CMH_	P_	Odds_	Delta_	Ques.	CMH_	P_	Odds_	Delta_
	Statistics	VALUE	Ratio	Scale	#	Statistics	VALUE	Ratio	Scale
43	34.159	0	1.281	-0.582	28	7.433	0.006	0.906	0.233
31	24.091	0	1.208	-0.445	119	7.536	0.006	0.904	0.237
131	14.877	0	1.163	-0.354	70	7.64	0.006	0.884	0.288
72	16.23	0	1.156	-0.34	87	7.308	0.007	0.864	0.345
135	12.387	0	1.15	-0.329	19	6.667	0.01	0.897	0.256
13	14.191	0	0.866	0.339	85	6.364	0.012	1.106	-0.236
41	14.876	0	0.859	0.357	88	5.919	0.015	1.103	-0.23
23	22.562	0	0.838	0.416	120	5.658	0.017	1.092	-0.206
127	21.539	0	0.81	0.494	14	5.723	0.017	0.903	0.24
67	24.078	0	0.792	0.548	18	5.545	0.019	0.903	0.239
25	10.483	0.001	1.158	-0.345	77	5.29	0.021	0.871	0.324
138	11.089	0.001	1.134	-0.295	63	5.08	0.024	0.923	0.19
58	12.099	0.001	1.131	-0.29	27	5.009	0.025	0.848	0.387
71	9.851	0.002	0.84	0.408	117	4.83	0.028	1.088	-0.198
73	8.578	0.003	1.115	-0.255	94	4.786	0.029	0.823	0.457
80	9.03	0.003	0.887	0.283	10	4.396	0.036	0.917	0.203
16	7.947	0.005	0.826	0.448	113	4.232	0.04	1.075	-0.171

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Table 4. Summarized Results of MH method for 8th Grades

Item	CMH_	P_	Odds_	Delta_	Ques. #	CMH_	P_	Odds_	Delta_
#	Statistics	VALUE	Ratio	Scale		Statistics	VALUE	Ratio	Scale
103	68.773	0	1.421	-0.826	157	9.34	0.002	0.885	0.287
64	38.179	0	1.298	-0.612	123	9.399	0.002	0.883	0.292
93	28.913	0	1.252	-0.529	71	8.614	0.003	1.128	-0.283
91	22.866	0	1.221	-0.469	86	8.809	0.003	0.893	0.266
90	20.229	0	1.21	-0.447	2	9.058	0.003	0.89	0.274
149	18.273	0	1.2	-0.428	87	7.475	0.006	0.892	0.268
121	19.802	0	1.197	-0.424	143	7.386	0.007	1.138	-0.304
147	17.38	0	1.184	-0.398	23	6.416	0.011	0.904	0.238
120	15.868	0	1.18	-0.389	115	6.322	0.012	1.129	-0.286
68	14.24	0	1.174	-0.377	100	5.679	0.017	1.101	-0.226
82	12.866	0	1.173	-0.375	141	5.733	0.017	1.099	-0.221
119	12.938	0	1.157	-0.342	35	5.626	0.018	1.099	-0.222
85	14.85	0	0.861	0.353	60	5.536	0.019	1.098	-0.22
150	15.639	0	0.855	0.369	40	5.229	0.022	0.911	0.22
160	15.015	0	0.854	0.372	10	5.151	0.023	1.114	-0.254
105	20.577	0	0.827	0.447	34	5.085	0.024	1.106	-0.236
36	21.068	0	0.82	0.465	8	4.904	0.027	1.091	-0.206
45	24.029	0	0.818	0.472	83	4.717	0.03	1.094	-0.212
102	21.415	0	0.808	0.502	39	4.682	0.03	1.087	-0.195
63	11.417	0.001	1.153	-0.334	5	4.658	0.031	1.091	-0.204
57	10.1	0.001	1.131	-0.29	84	4.464	0.035	0.898	0.252
27	11.322	0.001	0.871	0.324	70	4.23	0.04	1.092	-0.207
44	11.406	0.001	0.869	0.329	32	4.224	0.04	0.922	0.191
25	9.937	0.002	1.188	-0.405	101	4.165	0.041	0.92	0.197
53	9.853	0.002	1.137	-0.301	50	3.948	0.047	1.096	-0.215

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Finally, MH results for 12th graders are presented Table 5. According to the results, 8(out of 126) items exhibit statistical significance at 5% level based on the CMH statistics. Turning to the magnitude of the DIF, the delta score (in absolute terms) based on the MH Adjusted Odds ratio exceeds 1 in three questions, 68, 101, 31. For question 68, the associated CHM statistics is

10.621 (significant at 1% level) and odds ratio is 1.664. The odds ratio suggests that the odds of students in superior states responding correctly to this item are 1.664 times higher than that of students in inferior standard states. More importantly, delta scale is -1.196 suggesting a weak DIF (Category B) based on the NAEP classification. Similarly, questions 101 and 31 exhibit DIF with 1.568 and 1.537 odd ratios, respectively, and both are significant at 1% level. Based on the delta scale values, -1.057 and -1.010, respectively, these items exhibit weak DIF (Category B).

Table 5. Summarized Results of MH method for 12th Grades

Item #	CMH_ Statistics	P_ VALUE	Odds_ Ratio	Delta_ Scale
68	**10.621**	**0.001**	**1.664**	**-1.196**
101	**8.899**	**0.003**	**1.568**	**-1.057**
36	5.279	0.022	1.421	-0.825
166	5.154	0.023	0.721	0.77
31	**5.029**	**0.025**	**1.537**	**-1.01**
73	4.905	0.027	0.709	0.809
96	4.343	0.037	1.347	-0.7
66	3.894	0.048	0.757	0.655

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Next, the results of logistic regression analysis are presented. In logistic regression analysis, the score of each student on each question (1 if the student answered the question correctly and 0 otherwise) is regressed on student ability (their total score from the rest of the questions) and the state science standard dummy (Superior). Superior dummy variable takes a value of 1 if the state a student is taking the exam has relatively superior science standards and takes a value of 0 if the state has relatively inferior science standards.

In Tables 6, 7, and 8, the summarized the results of logistic regression for the 4th ,8th, and 12th grades are presented. In particular, -2logL values for models from 1 to 3 and the coefficient on superior state dummy variable Superior, β_2, its p value, and the associated delta score are presented.

According to the results in Tables 6 and 7, similar to the Mantel Haenszel analysis, there are 34 and 48 items for the 4th and 8th grade students, respectively, that exhibit statistical significance at the 5% level. It is important to note that, for 4th and 8th grade students, some questions which are shown to exhibit statistical significance based on MH test lost their significances in the logistic regression tests and vice versa. For example, while there are 50 items in 8th graders that attains statistical significance, this number reduces to 48 in logistic regression. Also, while question 11 in 4th graders attains statistical significance based on logistic regression, that item show no significance for the MH tests. Hence, these two methods sometimes yield different results which further confirms the necessity of both methods to attain more robust findings. Nonetheless, for both grades, although the coefficient on β_2 is significant in these items, the absolute value of the delta score never exceeds 1. This suggests that these items exhibit negligible or no DIF based on differences in state science standards when the entire sample is used for these grades.

It is important to note that, in this study, all three models in the logistic regression analysis are performed and only a limited set of the parameter values are presented. However, to test whether any item exhibits non-uniform DIF, one also needs to compare the likelihood of three models using the likelihood ratio test, the LRT statistic. Although none of the items exhibited any DIF based on the delta scale based on β_2 parameter in model 2, these items might exhibit non-uniform DIF. In particular, for some questions, the log likelihood value in model 3 is significantly lower compared to models 1 and 2. For example, in Table 6 item 31, the LRT statistic G^2 (Model3 – Model2) is 11.5 and larger than chi square statistic χ^2 (1, 0.05) = 3.841. This result shows that item 31 might exhibit a nonuniform DIF.

As discussed before, to flag an item as a DIF item, we need to have both statistical significance and a large effect size (as measured by delta). However, when we examine the coefficient in question 31 for 4th grades, in model 2, while the coefficient on Superior is statistically significant, delta is smaller than 1 pointing to a negligible DIF. Turning to model 3 in untabulated results, I found that the coefficient on superior is negative (-0.3354) and significant at 5% level. However, note that this coefficient represents the difference of probability of answering the question correctly between students in superior vs inferior states only when the ability level is 0. While having a 0 ability is very unlikely, even if we assume there are many students like that, the delta scale for 0 ability students becomes 0.79(-2.35*-0.3354) and it is smaller than 1, pointing to a negligible DIF. On the other hand, the coefficient on beta3 (Ability*Superior) is found to be positive (0.027) and significant at 1% level. The positive and significant coefficient suggests that, for students who have high ability level, the effect of superior state dummy on correctly answering the question would be higher. However, once again one needs to find the delta scale value for the higher ability group level to assess the presence of DIF. For the highest ability level, the overall effect of superior state dummy is beta2+beta3*ability. This means the total coefficient becomes close to 0 for high ability levels (beta2 is negative and beta3 is positive so the effects cancel each other). In sum, for both high and low ability levels, we are left with a delta scale smaller than 1 and neither uniform nor non-uniform DIF is present for these items. This result applies to all the questions that shows a significant difference in -2logLs between model 2 and model 3. Hence, while the results show

DIF (and in some cases non-uniform DIF) based on statistical significance in many questions, the effect size is negligible and therefore, I conclude that there is no DIF in these items.

Table 6. Summarized Results of Logistic Regression for 4th Grades

	Model 2: β_2 Coefficients Summary Stats			-2logL		
Item #	**β_2**	**P_Value**	**Delta_Scale**	**Model1**	**Model2**	**Model3**
43	0.262	0	-0.616	13710.9	13672.7	13670.7
31	0.194	0	-0.457	15555.7	15530.2	15517.7
131	0.167	0	-0.393	15678.8	15660.3	15653.6
135	0.146	0	-0.344	15003.3	14989.8	14988.2
72	0.144	0	-0.339	17614.9	17598.7	17598.7
138	0.143	0	-0.337	16500	16485.4	16484.3
58	0.127	0	-0.299	18295	18282.1	18281.5
41	-0.139	0	0.326	15367.5	15355	15352.2
13	-0.15	0	0.352	15840.7	15825.4	15823.2
23	-0.177	0	0.416	18061.7	18039.3	18038.7
127	-0.21	0	0.493	11666.5	11645	11641.1
67	-0.229	0	0.538	11151.6	11128.5	11128.4
25	0.15	0.001	-0.352	12089.8	12079	12077.6
71	-0.172	0.002	0.405	8017	8007.4	8007.4
80	-0.118	0.003	0.278	14712.6	14703.8	14701.5
73	0.107	0.004	-0.251	16768.9	16760.7	16759.8
119	-0.106	0.004	0.249	16887.1	16878.7	16878.6
16	-0.19	0.005	0.447	6031.8	6023.9	6023.7
70	-0.123	0.006	0.29	12229.6	12221.9	12221.5
28	-0.097	0.007	0.228	17260.8	17253.7	17251.2
19	-0.112	0.008	0.262	13526.3	13519.3	13514.3
87	-0.144	0.008	0.339	9115.2	9108.2	9107.5
120	0.093	0.012	-0.218	16870.1	16863.8	16859.5
88	0.101	0.013	-0.236	14673.3	14667.1	14667
85	0.099	0.013	-0.233	14685.1	14678.9	14678.7
77	-0.141	0.019	0.33	7345.9	7340.4	7340.1
14	-0.099	0.021	0.233	13407.7	13402.4	13402.3
27	-0.17	0.022	0.398	4937.6	4932.3	4932.3
117	0.086	0.025	-0.201	16040.9	16035.9	16035.9
18	-0.097	0.025	0.227	12467.8	12462.8	12462.2
63	-0.08	0.026	0.188	17866.7	17861.7	17860.9
10	-0.087	0.034	0.205	13639.8	13635.3	13634.7
11	0.077	0.041	-0.181	16489.9	16485.7	16484.7
113	0.072	0.042	-0.169	18349.6	18345.4	18339.9

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Table 7. Summarized Results of Logistic Regression for 8th Grades

	Model 2: β_2 Coefficients Summary Stats			-2logL		
Item #	**β_2**	**P_Value**	**Delta_Scale**	**Model1**	**Model2**	**Model3**
103	0.376	0	-0.883	13261	13182.9	13182.8
64	0.266	0	-0.626	13328.9	13289.3	13288.3
93	0.236	0	-0.555	13421.8	13390.3	13390.3
91	0.219	0	-0.515	13482.4	13455.1	13449.7
90	0.207	0	-0.486	13193.9	13170.3	13170.3
149	0.2	0	-0.469	13023.4	13001.7	13000.1
121	0.181	0	-0.424	14159.5	14139.8	14139.4
147	0.177	0	-0.415	14031.9	14013.1	14011.6
120	0.172	0	-0.404	13637	13619.9	13619.9
68	0.167	0	-0.391	13076.9	13061.7	13059.8
82	0.165	0	-0.387	12302.3	12288.5	12283.7
119	0.156	0	-0.367	14145.3	14130.5	14127.9
63	0.148	0	-0.347	13274.7	13262.5	13262.2
150	-0.141	0	0.331	14658.1	14645.5	14642.7
27	-0.145	0	0.341	13843.5	13831.1	13829.9
85	-0.146	0	0.342	14935.4	14921.6	14915.8
160	-0.161	0	0.378	13981.4	13965.9	13963.7
105	-0.178	0	0.419	13267.5	13249.7	13246.1
36	-0.188	0	0.441	12631.9	12613.4	12613.1
45	-0.193	0	0.454	13753.9	13732	13731.8
102	-0.193	0	0.455	11210.5	11193.3	11190.8
25	0.19	0.001	-0.446	8453.7	8441.9	8438.3
53	0.142	0.001	-0.335	13967.8	13955.8	13943.4
57	0.125	0.001	-0.293	15089.6	15079.4	15078.8
157	-0.132	0.001	0.31	14447.2	14436.4	14434
123	-0.132	0.001	0.311	14056.4	14045.8	14044.5
44	-0.138	0.001	0.325	13490.2	13479.3	13479.2
71	0.123	0.003	-0.289	13871.2	13862.2	13861.3
10	0.137	0.004	-0.322	10817.4	10809.3	10808.4

115	0.137	0.005	-0.322	10516.7	10508.8	10507.9
86	-0.106	0.005	0.249	15507.9	15500.2	15494.4
35	0.106	0.008	-0.25	14558.4	14551.3	14551.2
2	-0.102	0.008	0.24	15230.2	15223.3	15223.2
34	0.117	0.009	-0.275	12167.3	12160.5	12159.9
143	0.123	0.01	-0.29	10900.2	10893.5	10893.2
87	-0.107	0.011	0.252	13383.9	13377.3	13375.6

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Table 7 (Continued). Summarized Results of Logistic Regression for 8th Grades

	Model 2: β_2 Coefficients Summary Stats			-2logL		
Item #	**β_2**	**P_Value**	**Delta_Scale**	**Model1**	**Model2**	**Model3**
100	0.101	0.012	-0.238	14215.4	14209.2	14209.1
8	0.1	0.012	-0.234	14756	14749.7	14749.6
60	0.098	0.014	-0.23	14562	14556	14555.9
5	0.099	0.015	-0.232	14335.1	14329.1	14326.4
39	0.091	0.018	-0.214	15334.5	15328.9	15325.5
23	-0.094	0.018	0.222	14469.6	14464	14463
83	0.097	0.019	-0.229	13647.9	13642.4	13642.4
32	-0.09	0.023	0.211	14732.1	14726.9	14726.8
50	0.104	0.025	-0.245	11427.6	11422.6	11422.1
70	0.093	0.031	-0.218	12908.3	12903.6	12898.8
141	0.084	0.033	-0.197	14831.4	14826.9	14826.1
40	-0.088	0.033	0.206	13929.1	13924.5	13909.8
101	-0.085	0.04	0.199	13799.2	13795	13794.4

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Turning to the logistic regression analysis for 12th grade students, the results are presented in Table 8. According to the results, similar to the MH results in Table 4, the coefficients on β_2 in model 2 are positive and significant at 5% significance level for items 68, 101 and 31. The positive and significant coefficient on β_2 indicate that the odds of performance on this item are significantly different between the superior and inferior science standard states and the questions favors the students in the superior states. In each case, the coefficient on the ability are significantly different from zero (not tabulated), which suggests that the log odds of correctly answering these items increases as total scores increase. More importantly, the absolute value of delta scale is higher than 1 for these three items and they exhibit weak DIF (category B).

Table 8. Summarized Results of Logistic Regression for 12th Grades

Model 2: β_2 Coefficients Summary Stats				-2logL		
Item #	**β_2**	**P_Value**	**Delta_Scale**	**Model1**	**Model2**	**Model3**
68	**0.532**	**0.001**	**-1.249**	1174	1162.5	1162
101	**0.481**	**0.002**	**-1.131**	1200.4	1190.8	1188.4
36	0.374	0.015	-0.879	1136	1130.2	1129.5
166	-0.353	0.016	0.829	1290.5	1284.7	1284.1
31	**0.462**	**0.017**	**-1.086**	892.6	886.6	886.6
96	0.317	0.027	-0.744	1365.7	1360.8	1360.1
66	-0.296	0.035	0.696	1359	1354.6	1354.4
73	-0.308	0.05	0.723	1103.8	1099.9	1099.9

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Next, I compare the likelihood of two models using the likelihood ratio test, the LRT statistic. For question 68, the LRT statistic G^2 (Model1 – Model2) is 11.5 and larger than chi square statistic χ^2 (1, 0.05) = 3.841. Therefore, I confirm that the second model which includes superior state science standards dummy variable fits the data better. On the other hand, G^2 (Model2 – Model3) is 0.5 and smaller than chi square statistic χ^2 (1, 0.05) = 3.841. Therefore, I conclude that the item does not show nonuniform DIF but have uniform DIF. I find similar results for questions 101 and 31 and conclude that neither of these questions exhibit nonuniform DIF but all exhibit a weak DIF based on NAEP criteria for DIF detection.

Overall, the findings for research question 1 suggest that differences in science standards lead to DIF in three items among 12th grade students. However, the results did not show any DIF among 4th and 8th grade students. According to the results, among 12th grade students, being in a high science standard state standard favors student in three of the items. Delta Scale analysis, which categorizes the DIF items based on the magnitude of the effect, shows that all these three items exhibit weak DIF. In other words, for examinees in superior standard states, the probability of correctly answering these three items are 1.537 to 1.664 times higher than that of students in inferior standard states. Moreover, logistic regression results show that the DIF in 12th grade students seem to have a uniform DIF. In other words, the effect of differences in state science standards among 12th grade students is not different (uniform) across different ability levels and for these three items, standards increase the probability of answering questions correctly at all ability levels in superior states.

On the other hand, no DIF is detected for the 4th and 8th grade students using their entire sample. A more detailed discussion on the differences in findings between 12th and 4-8th grades are provided in the discussion section.

Research Question 2: Controlling for Gender and Race

In this section, the results for the tests on whether controlling for gender and race alters the inferences from the first research question is presented. Two set of results, Mantel Haenszel and logistic regressions, are presented.

To test this idea, first the logistic regression analysis is repeated by controlling for gender and race of students. Three additional dummy variables are added to the second equation in the logistic regressions as following.

$$\text{logit (P)} = \beta_0 + \beta_1\text{*Ability} + \beta_2\text{*Superior} + \beta_3\text{*Gender} + \beta_4\text{*Hispanic} + \beta_5\text{*Black}$$

In this regression, Gender equals to 1 if a student is male and 0 if a student is female. Black and Hispanic are race dummies and they take value of 1 if a student is black or Hispanic, respectively, and zero otherwise.

In Tables 9, 10, and 12, the summarized results of logistic regression for the 4th ,8th, and 12th grades are presented. In particular, the coefficient on Superior, Gender, Hispanic*,* and Black along with their p value, and the associated delta score are presented. In each table, only the results for the questions in which the coefficient on Superior dummy is significant at least at 5% level (i.e. p_value <0.05) are presented.

According to the results in Table 9, among 4th graders, the coefficient on state standards dummy is significant in 31 questions. Note that, without controlling for gender and race, there were 34 questions attaining a significance level of 5% suggesting that gender and race were partially driving the significance on state standards dummy in these questions. However, as before, in none of the questions the absolute value of delta score on state standards dummy exceeds 1. Therefore, controlling for gender and race does not alter our main conclusion for 4th grade students that none of the questions exhibit state standards based DIF in NAEP science exam for 4th grade students. On the other hand, Table 9 also shows that two questions exhibit gender and race-based DIF. In particular, question 12 favor male students over female students and favors White students over Black students. The delta scores of -1.659 and 1.421, indicate that this question exhibits a strong gender based DIF and a weak race-based DIF. On the other hand, question 67 favors White students over Black students and there is a weak race based DIF in this item.

Table 9. Summarized Results of Logistic Regression for 4th Grades controlling for Gender and Race

	Superior			Gender			Hispanic			Black		
Item #	β	P_Val	Delta	β	P_Val	Delta	β	P_Val	Delta	β	P_Val	Delta
43	0.26	0.00	-0.60	0.11	0.01	-0.26	-0.15	0.01	0.36	0.09	0.17	-0.21
31	0.19	0.00	-0.44	0.07	0.09	-0.15	-0.11	0.06	0.25	0.10	0.14	-0.23
131	0.15	0.00	-0.34	0.01	0.86	-0.02	0.05	0.38	-0.12	0.22	0.00	-0.51
127	-0.17	0.00	0.40	-0.18	0.00	0.41	-0.34	0.00	0.80	-0.38	0.00	0.89
135	0.14	0.00	-0.33	0.28	0.00	-0.65	-0.11	0.06	0.27	0.10	0.14	-0.24
28	-0.12	0.00	0.29	0.12	0.00	-0.28	0.14	0.01	-0.33	0.21	0.00	-0.49
23	-0.12	0.00	0.29	0.18	0.00	-0.42	-0.27	0.00	0.64	-0.37	0.00	0.87
67	-0.16	0.00	0.37	0.18	0.00	-0.41	-0.26	0.00	0.62	**-0.44**	**0.00**	**1.03**
138	0.12	0.00	-0.29	-0.05	0.19	0.11	0.16	0.00	-0.36	0.15	0.02	-0.36
72	0.11	0.00	-0.27	-0.28	0.00	0.65	0.18	0.00	-0.43	0.28	0.00	-0.65
13	-0.12	0.00	0.28	0.14	0.00	-0.33	-0.10	0.06	0.23	-0.30	0.00	0.69
58	0.11	0.00	-0.25	-0.01	0.81	0.02	0.12	0.02	-0.28	0.16	0.01	-0.38
119	-0.10	0.01	0.25	-0.03	0.35	0.08	0.09	0.09	-0.21	-0.04	0.56	0.09
25	0.13	0.01	-0.31	0.14	0.00	-0.32	-0.01	0.90	0.02	0.17	0.02	-0.41
80	-0.10	0.01	0.24	-0.19	0.00	0.44	-0.03	0.56	0.08	-0.16	0.03	0.38
71	-0.14	0.01	0.33	-0.10	0.08	0.23	-0.21	0.01	0.48	-0.26	0.00	0.62
87	-0.14	0.01	0.33	0.06	0.27	-0.14	0.05	0.45	-0.13	-0.06	0.48	0.14
16	-0.17	0.02	0.40	0.40	0.00	-0.93	0.17	0.05	-0.41	-0.16	0.10	0.37
120	0.09	0.02	-0.21	0.28	0.00	-0.66	-0.07	0.22	0.15	0.07	0.25	-0.17
70	-0.11	0.02	0.26	-0.28	0.00	0.65	-0.15	0.01	0.35	-0.17	0.02	0.41
41	-0.09	0.02	0.22	0.33	0.00	-0.77	-0.29	0.00	0.69	-0.42	0.00	1.00
117	0.09	0.03	-0.20	0.35	0.00	-0.81	-0.21	0.00	0.50	0.01	0.90	-0.02
63	-0.08	0.03	0.19	-0.03	0.35	0.08	-0.01	0.92	0.01	0.00	0.96	-0.01
12	0.10	0.03	-0.24	**0.71**	**0.00**	**-1.66**	-0.19	0.00	0.44	**-0.61**	**0.00**	**1.42**
73	0.08	0.03	-0.19	-0.16	0.00	0.37	0.06	0.25	-0.14	0.27	0.00	-0.64
112	0.08	0.04	-0.18	-0.06	0.07	0.15	-0.25	0.00	0.58	-0.06	0.35	0.13
94	-0.19	0.04	0.45	0.02	0.86	-0.04	-0.05	0.66	0.11	0.13	0.31	-0.29
77	-0.13	0.04	0.30	0.11	0.06	-0.27	0.02	0.77	-0.05	-0.14	0.13	0.32
85	0.08	0.04	-0.19	0.00	0.94	0.01	-0.02	0.77	0.04	0.20	0.01	-0.46
32	0.08	0.05	-0.19	-0.15	0.00	0.36	0.06	0.34	-0.13	-0.11	0.11	0.27
89	-0.17	0.05	0.40	-0.14	0.10	0.33	0.19	0.08	-0.43	0.16	0.16	-0.38

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Turning to the results in Table 10, among 8th grade students, the coefficient on state standards dummy is significant in 31 questions. Similar to the 4th grade results, controlling for gender and race does not alter our main conclusion for 8th grade students that none of the questions exhibit state standards based DIF in NAEP science exam for 8th grade students.

Table 10. Summarized Results of Logistic Regression for 8th Grades controlling for Gender and Race

	Superior			Gender			Hispanic			Black		
Item #	**β**	**P_Val**	**Delta**	**β**	**P_Val**	**Delta**	**β**	**P_Val**	**Delta**	**β**	**P_Val**	**Delta**
103	0.37	0.00	-0.87	-0.02	0.73	0.04	0.00	0.98	0.00	0.06	0.36	-0.15
64	0.27	0.00	-0.63	0.25	0.00	-0.60	0.11	0.08	-0.25	-0.02	0.75	0.05
93	0.23	0.00	-0.53	0.41	0.00	-0.97	0.03	0.62	-0.07	0.10	0.15	-0.23
91	0.22	0.00	-0.51	0.04	0.40	-0.08	-0.04	0.48	0.10	0.02	0.82	-0.04
90	0.20	0.00	-0.48	0.00	0.96	0.01	0.06	0.30	-0.15	0.03	0.74	-0.06
45	-0.20	0.00	0.46	0.11	0.01	-0.25	-0.19	0.00	0.44	0.02	0.74	-0.05
149	0.19	0.00	-0.46	0.31	0.00	-0.72	0.06	0.34	-0.14	0.04	0.56	-0.10
121	0.18	0.00	-0.42	-0.02	0.56	0.06	0.03	0.62	-0.07	0.01	0.91	-0.02
85	-0.16	0.00	0.39	-0.35	0.00	0.82	0.12	0.03	-0.28	0.15	0.02	-0.35
68	0.18	0.00	-0.42	0.37	0.00	-0.87	-0.10	0.09	0.24	-0.07	0.29	0.17
147	0.16	0.00	-0.38	0.04	0.28	-0.10	0.03	0.59	-0.08	0.14	0.05	-0.34
36	-0.17	0.00	0.40	-0.07	0.14	0.15	0.07	0.27	-0.16	-0.14	0.04	0.34
120	0.16	0.00	-0.37	-0.17	0.00	0.40	0.16	0.01	-0.37	0.11	0.12	-0.25
57	0.15	0.00	-0.35	-0.07	0.07	0.16	-0.02	0.80	0.03	-0.19	0.00	0.45
25	0.20	0.00	-0.47	0.28	0.00	-0.65	-0.17	0.02	0.39	-0.07	0.37	0.17
160	-0.14	0.00	0.34	0.05	0.22	-0.12	-0.31	0.00	0.73	-0.15	0.04	0.35
157	-0.14	0.00	0.33	0.14	0.00	-0.34	0.20	0.00	-0.47	0.04	0.59	-0.09
82	0.16	0.00	-0.36	0.40	0.00	-0.95	-0.05	0.46	0.12	0.07	0.39	-0.17
102	-0.16	0.00	0.38	-0.17	0.00	0.40	0.04	0.55	-0.09	-0.21	0.00	0.49
150	-0.14	0.00	0.32	0.39	0.00	-0.91	-0.12	0.06	0.27	-0.07	0.31	0.16
63	0.14	0.00	-0.33	0.35	0.00	-0.82	-0.05	0.45	0.12	0.11	0.14	-0.26
53	0.13	0.00	-0.31	0.27	0.00	-0.63	0.11	0.06	-0.26	0.10	0.15	-0.24
119	0.12	0.00	-0.29	0.26	0.00	-0.61	0.07	0.24	-0.16	0.27	0.00	-0.65
123	-0.12	0.00	0.29	0.24	0.00	-0.57	-0.07	0.23	0.16	-0.09	0.19	0.21
86	-0.11	0.00	0.27	0.20	0.00	-0.46	0.16	0.00	-0.38	0.04	0.54	-0.09
100	0.12	0.00	-0.28	-0.11	0.01	0.25	-0.17	0.00	0.40	-0.15	0.03	0.36
105	-0.13	0.00	0.29	**0.69**	**0.00**	**-1.62**	-0.31	0.00	0.73	**-0.61**	**0.00**	**1.43**
50	0.14	0.01	-0.32	0.39	0.00	-0.91	-0.23	0.00	0.55	-0.24	0.00	0.57
2	-0.11	0.01	0.26	0.00	0.92	-0.01	0.06	0.28	-0.15	0.06	0.37	-0.14
71	0.11	0.01	-0.26	-0.06	0.13	0.15	0.06	0.34	-0.13	0.10	0.13	-0.24
27	-0.11	0.01	0.26	-0.09	0.03	0.21	-0.06	0.29	0.14	-0.28	0.00	0.65
83	0.11	0.01	-0.26	0.34	0.00	-0.81	-0.11	0.09	0.26	-0.17	0.03	0.39
87	-0.11	0.01	0.25	0.11	0.01	-0.27	-0.06	0.36	0.13	-0.01	0.88	0.03
5	0.10	0.02	-0.23	-0.13	0.00	0.31	-0.02	0.73	0.05	-0.01	0.90	0.02
44	-0.10	0.02	0.24	0.24	0.00	-0.57	-0.13	0.03	0.31	-0.35	0.00	0.81
115	0.12	0.02	-0.28	**0.55**	**0.00**	**-1.29**	0.05	0.44	-0.12	0.07	0.37	-0.16
40	-0.09	0.03	0.22	0.13	0.00	-0.32	-0.04	0.52	0.09	0.05	0.50	-0.12
8	0.09	0.03	-0.21	0.33	0.00	-0.77	0.12	0.04	-0.28	0.10	0.14	-0.23
34	0.10	0.03	-0.24	0.28	0.00	-0.65	-0.05	0.37	0.13	0.11	0.12	-0.26
84	-0.11	0.03	0.27	0.05	0.37	-0.11	-0.06	0.35	0.15	0.16	0.04	-0.38

60	0.09	0.03	-0.21	0.05	0.22	-0.12	0.05	0.44	-0.11	0.09	0.20	-0.20
10	0.10	0.04	-0.24	-0.07	0.14	0.17	-0.05	0.41	0.12	0.27	0.00	-0.62
88	-0.09	0.04	0.21	0.28	0.00	-0.67	-0.06	0.31	0.15	0.30	0.00	-0.71
70	0.09	0.04	-0.21	0.53	0.00	-1.24	0.01	0.94	-0.01	-0.05	0.54	0.11
69	-0.08	0.05	0.19	0.12	0.00	-0.27	0.04	0.48	-0.10	0.14	0.05	-0.32

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Finally, Table 11 presents the results for the 1 2th grade students. According to the results, in seven questions state science standards dummy, Superior, is significantly related to the probability of success in each question. Note that, without controlling for gender there are 8 questions that has a significant coefficient. Most importantly, the same 3 questions, 68,101, and 31, both have a significant coefficient and have an absolute value of delta bigger than 1 suggesting that race and gender control does not alter any of our inferences for DIF in 12th grades. Furthermore, in these questions, item #66 exhibits weak DIF based on gender and favors males over females and item 36 exhibits weak race based DIF and favors Hispanic students over White students. Hence, in questions that exhibits DIF based on state science standards do not exhibit DIF in the other two important DIF sources.

Table 11. Summarized Results of Logistic Regression for 12th Grades controlling for Gender and Race

	Superior			Gender			Hispanic			Black		
Item #	β	P_Val	Delta	β	P_Val	Delta	β	P_Val	Delta	β	P_Val	Delta
68	**0.56**	**0.00**	**-1.31**	-0.02	0.87	0.05	-0.12	0.50	0.28	-0.14	0.48	0.34
101	**0.49**	**0.00**	**-1.14**	0.35	0.01	-0.82	-0.01	0.98	0.01	0.07	0.72	-0.17
31	**0.51**	**0.01**	**-1.19**	0.17	0.33	-0.39	-0.45	0.05	1.06	-0.10	0.71	0.23
166	-0.37	0.01	0.88	0.04	0.74	-0.11	0.20	0.22	-0.47	0.09	0.66	-0.20
96	0.36	0.01	-0.86	-0.09	0.48	0.22	-0.30	0.08	0.70	-0.23	0.24	0.53
66	-0.31	0.03	0.74	**1.05**	**0.00**	**-2.46**	0.01	0.95	-0.03	-0.42	0.05	0.99
36	0.31	0.05	-0.73	-0.14	0.35	0.32	**0.43**	**0.02**	**-1.01**	0.10	0.62	-0.24

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Overall, the findings for research question 2 suggest that controlling for gender and race does not change the inferences from the 1 research question. According to the results, after controlling for gender and race, while differences in state science standards do not create DIF among 4th and 8th grade students, three items exhibit DIF among 12th grade students.

Research Question 3: Dissecting Ability by Content Area

This section presents the results for the third research question of this study that whether the finding that no items exhibit DIF in 4th and 8th grades and 3 items exhibit DIF in 12 grades might be altered when the ability is measured using sub content score of examines instead of their total exam scores. The following logistic regression model is used.

$$\text{logit (P)}= \beta_0 + \beta_1\text{*Physical\_Ability} + \beta_2\text{*Life\_Ability} + \beta_3\text{*Earth\_Ability} + \beta_4\text{* Superior}$$

In this regression, Physical_Ability, Life_Ability, and Earth_Ability variables are calculated as each examinee's total score in each of the distinct content areas. The main variable of coefficient is β_4 which captures the effect of differences in state science standards on examinees probability of correctly answering different items after controlling for student ability in each content area separately.

The results are presented in Tables 12, 13, and 14 for 4th, 8th and 12th grade students, respectively. According to the results in Table 12, among 4th grade students, after dissecting ability, there are 34 items that have a coefficient on the superior state standard dummy significant at 5% level. However, once again, based on the delta scale categorization, all of these items are categorized as items with negligible or nonsignificant DIF (Category A).

Table 12. Summarized Results of Logistic Regression for 4th Grades controlling for Separate Content Ability

Item #	β_4	P_Val	Delta	Item #	β_4	P_Val	Delta
43	0.257	<.0001	-0.603	19	-0.119	0.006	0.280
31	0.197	<.0001	-0.464	16	-0.188	0.006	0.441
127	-0.220	<.0001	0.517	87	-0.156	0.008	0.366
67	-0.226	<.0001	0.532	28	-0.098	0.008	0.229
23	-0.172	<.0001	0.405	120	0.099	0.008	-0.231
72	0.147	<.0001	-0.345	70	-0.117	0.009	0.274
131	0.167	<.0001	-0.393	85	0.098	0.014	-0.229
13	-0.153	<.0001	0.359	77	-0.143	0.019	0.335
58	0.131	0.000	-0.307	27	-0.170	0.023	0.399
138	0.139	0.001	-0.326	63	-0.078	0.030	0.184
25	0.154	0.001	-0.362	113	0.077	0.032	-0.180
119	-0.115	0.002	0.271	18	-0.093	0.032	0.219
41	-0.129	0.002	0.304	10	-0.089	0.033	0.210
135	0.130	0.003	-0.306	117	0.081	0.035	-0.191
71	-0.168	0.003	0.395	11	0.079	0.038	-0.185
80	-0.118	0.003	0.278	14	-0.090	0.039	0.211
73	0.110	0.003	-0.257	88	0.090	0.039	-0.212

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Table 13 shows the results for 8th grade students. According to the results, there are 46 items that have a coefficient on the Superior state standard dummy significant at 5% level. However, again, all these DIF items exhibits negligible DIF or nonsignificant DIF (Category A).

Table 13. Summarized Results of Logistic Regression for 8th Grades controlling for Separate Content Ability

Item #	β_4	P_Val	Delta	Item #	β_4	P_Val	Delta
103	0.381	<.0001	-0.894	71	0.118	0.005	-0.278
93	0.234	<.0001	-0.550	54	0.143	0.005	-0.335
91	0.218	<.0001	-0.512	86	-0.105	0.006	0.248
90	0.209	<.0001	-0.492	53	0.148	0.006	-0.349
149	0.196	<.0001	-0.461	37	0.172	0.007	-0.404
64	0.261	<.0001	-0.613	123	-0.134	0.010	0.314
27	-0.194	<.0001	0.455	100	0.104	0.011	-0.244
105	-0.179	<.0001	0.420	138	0.133	0.011	-0.313
102	-0.193	<.0001	0.453	87	-0.107	0.011	0.252
160	-0.160	<.0001	0.377	63	0.145	0.012	-0.339
82	0.172	0.000	-0.403	83	0.103	0.013	-0.243
45	-0.260	0.000	0.611	10	0.131	0.014	-0.307
85	-0.144	0.000	0.339	39	0.150	0.015	-0.353
150	-0.146	0.000	0.342	119	0.127	0.016	-0.298
68	0.206	0.000	-0.484	57	0.124	0.016	-0.292
25	0.225	0.000	-0.528	5	0.107	0.019	-0.251
147	0.185	0.000	-0.434	36	-0.164	0.019	0.386
157	-0.130	0.001	0.306	44	-0.157	0.024	0.368
143	0.185	0.002	-0.435	70	0.097	0.025	-0.227
50	0.235	0.003	-0.552	2	-0.091	0.037	0.214
120	0.159	0.003	-0.374	101	-0.085	0.040	0.199
121	0.154	0.003	-0.362	34	0.148	0.045	-0.347
8	0.127	0.004	-0.299	136	-0.113	0.048	0.266

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Finally, in Table 14, the results of the logistic regression analysis are presented for 12th grade students. According to the results, after controlling for sub-content ability levels separately, there are 7 items that have a coefficient on the Superior state standard dummy significant at 5% level. When the delta scale is examined for these items, 3 items exhibit Category B DIF and other 4 items exhibit negligible DIF (Category A). These three items are the same as the ones documented in research question 1. This finding suggests that controlling for separate ability levels does not alter any DIF findings based on differences in state science standards as documented in the first research question analysis.

Table 14. Summarized Results of Logistic Regression for 12th Grades controlling for Separate Content Ability

Item #	β_4	P_Val	Delta
68	0.514	0.001	-1.208
101	0.519	0.001	-1.220
166	-0.383	0.010	0.900
36	0.398	0.012	-0.936
31	0.471	0.018	-1.106
96	0.324	0.031	-0.760
66	-0.282	0.048	0.662

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Chapter 6: Discussion

DIF items in 12th grade exams

The results in the previous suggest that differences in state science standards lead to DIF only in 12 grade students and do not lead to DIF in 4th and 8th grade students. In this section, I examine the nature and type of the questions that exhibit DIF in NAEP science exams. The summary of these properties is presented in Table 15. According to Table 15, 2 of the items, 68 and 36, are from Physical Science area and one item, item 101, is from Life Science area. Item 68 is about Source of Carbon in Plant Tissue, item 36 is about Photons of Microwave Radiation, and item 101 is about Classifying Observations about Molecular Motion. It is important to note that, these content areas have more weight in 12th grade science exam (a total of 76% of the exam) and therefore, more detailed questions from variety of topics are more likely to be asked in the exam.

Table 15. Subject and Content Areas of DIF Items

Item #	Subject	Content Area
68	Source of Carbon in Plant Tissue	Physical Science
36	Photons of Microwave Radiation	Physical Science
101	Classify Observations about Molecular Motion	Life Science

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Next, the response patterns for the DIF items is examined in individual superior vs inferior states. Note that these results only show the general success of students in these states and cannot be comparable to DIF results as the ability is not controlled in these tables. According to the results in Table 16,[13] the success rate in DIF items in inferior states are 53.1%, 67.9%, and 65.7% for items 101, 36, and 68, respectively. For superior states, these numbers become 60.1%, 73.1%, and 71.2% for items 101, 36, and 68, respectively. These results show that item 101 is a more difficult question for the examines since correct response percentages are lower in both superior and inferior states. In item 101, North Dakota, Oklahoma, Oregon, Wisconsin exhibit a relatively higher success rate compared to other inferior states. For this question, among superior states, while the correct response rate is above 60% in majority of the superior states, Arkansas, Kansas, Maryland, South Carolina perform relatively weaker. For item 36, New Jersey and Wisconsin have weaker performance among inferior states and almost all states perform well among superior states. Finally, in item 68, Nevada, North Dakota, and Wyoming perform relatively poorer among inferior states and Indiana, Kansas, South Carolina are the worst performers among superior states.

[13] Due to the confidentially requirements of NCES, the number of examinees in the Table are rounded to the next decimal points.

Table 16. States' Performances in DIF Items

Inferior States

Item #		All	Alaska	Colorado	Idaho	Iowa	Kentucky	Nebraska	Nevada	New Jersey	North Dakota	Oklahoma	Oregon	Pennsylvania	Wisconsin	Wyoming
101	**# Examinees**	340	10	10	10	20	30	10	30	50	0	20	20	60	40	20
	%Success	53.10%	37.5%	58.3%	40.0%	44.4%	53.1%	50.0%	56.7%	48.1%	100.0%	65.2%	62.5%	48.7%	68.6%	40.9%
36	**# Examinees**	360	10	10	10	20	20	20	40	40	20	20	90	20	20	20
	%Success	67.90%	88.9%	64.3%	81.8%	66.7%	81.0%	66.7%	60.5%	58.3%	70.8%	73.9%	67.8%	70.0%	58.8%	70.6%
68	**# Examinees**	340	10	10	10	10	40	0	30	50	0	30	20	70	40	20
	%Success	65.70%	75.0%	61.5%	85.7%	64.3%	57.1%	100%	58.6%	64.7%	33.3%	68.0%	66.7%	63.5%	81.6%	58.8%

Superior States

		All	Arkansas	California	Indiana	Kansas	Louisiana	Maryland	Massachusetts	New York	Ohio	South Carolina	Virginia	Utah
101	**# Examinees**	750	180	170	20	20	30	0	40	80	80	50	50	20
	%Success	60.20%	52.0%	60.6%	66.7%	52.6%	70.6%	50.0%	59.5%	64.3%	75.3%	50.0%	58.5%	87.0%
36	**# Examinees**	720	160	170	20	20	30	0	40	80	80	40	60	20
	%Success	73.1%	74.1%	69.0%	80.0%	73.3%	69.7%	80.0%	81.4%	74.7%	74.7%	68.3%	71.2%	78.9%
68	**# Examinees**	740	170	160	30	20	40	20	60	80	60	50	30	10
	%Success	71.2%	69.6%	75.3%	57.7%	55.0%	73.8%	76.5%	74.6%	79.7%	78.6%	58.3%	64.7%	80.0%

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2015 Science Grades 4, 8, and 12 Assessments Restricted-Use Data Files Data

Differences in DIF results in 4th, 8th, and 12th grades

So far, the analysis suggests that there are the differences in DIF findings between 4th and 8th grades versus 12th grade students as no standards based DIF is present for 4th and 8th grade students. According to Gross et al. 2013, while differences in state science standards result in differences in curricular coverage, high school students are affected more from lack of higher standards. For example, on page 10 of the report, they mention that “the physical science standards fail to lay the foundation for advanced study in high school and beyond, and there is so little advanced content that it would be impossible to derive a high school physics or chemistry course from the content included in the NGSS and inferior science standards.

Therefore, the finding that standards based DIF only exists among 12th grade students is somehow not that surprising. Indeed, as discussed before, the reasons that are mentioned to lead to DIF are more likely to be related to high school science education. For example, an important source of DIF might be omitted subjects in inferior standard states. Since high school students are taught more detailed and variety of subjects compared to 8th and 4th grade students, which is also evident from higher number of questions asked in the NAEP science exam, it is more likely to have more omitted subjects in inferior standard state curriculums. Also, the fact that the sequence and detailed teaching in certain subject affects the learning in the following years, a weak education standard in the early years would have a bigger impact on students learning in the following years. As an example, 4th grade questions are not as detailed and information focused as 12th grade questions and 4th grade students might be successful in answering a question in a certain topic with their knowledge in other topics even that topic is not covered in detailed in lower grades. However, questions become more detailed and information oriented in higher grades which creates a higher potential for omitted and neglected topics to lead to DIF.

In order to examine whether this pattern of DIF is only unique to science standards differences based DIF, this study also examines race and gender based DIF in 2015 NAEP Science exam. While NAEP also conducts gender and race base DIF analysis, they do not have the results available for 2015 exam yet. According to the analysis (not tabulated), 3 items exhibit strong or weak DIF based on gender (2 strong and 1 weak) among 4th grades, 7 items exhibit strong or weak gender based DIF among 8th grades, and 14 items exhibit strong or weak gender based DIF among 12th grades. I also checked NCES report of DIF in 2009 science exam and observed a similar pattern number of gender-based DIF across different grade levels. Hence, it is clear from these results that 12 grade exams are more prone to exhibit DIF in its items compared 4th or 8th grade exams.

Chapter 7: Conclusion

This study examines whether differences in K-12 state science standards across US states create differential item functioning in 2015 NAEP science exam among 4th, 8th, and 12th grade students. Furthermore, whether controlling for two well established source of DIF, gender and race, alters any of the findings in the first question is examined in the second research question. Finally, the analysis is repeated using dissected ability levels as measured sub content area scores rather than the total exam score. Differences in state science standards have a significant potential to create DIF because it creates differential curriculum coverage, both in content, depth, the teaching techniques, and the order of the material covered, across different states. The analysis is conducted using restricted examinee level data from NCES. The science standards ranking of Gross et. al 2013 is used to categorize states into inferior and superior standard states. Mantel Haenszel and Logistic Regression analysis are used to conduct the statistical analysis for the first question and Logistic Regression analysis is used to conduct the analysis for the second and third research questions.

According to the results, while none of the items exhibit DIF in 4th and 8th grade students, three questions in 12th grade exam exhibit DIF. Among 12th grade students, two of the DIF items are from the physical science content area, the third one is from the life science content. For these three items, the probability of students successfully answering these questions is 1.537 to 1.664 times higher in superior states than that of students in inferior standard states. Further analysis show that one of the three DIF questions is significantly harder than the others and the results not confined to some of the inferior or superior states. The study also shows that gender and race differences cannot account for the findings and controlling for individual content area ability measures does not alter any of the findings.

Implications and Limitations

The analysis provides important insights to the fairness and measurement invariance concerns related to the NAEP assessments and other high-stake tests particularly for 12th grade students. Findings from this study provides meaningful evaluation criteria and standards for DIF detection and help researchers to create more homogenous groups which is vital for a thorough and thus, accurate DIF detection. In addition, students from all states take the NAEP assessments, and it is vital to establish measurement equivalence of these measures, since inaccurate assessments might lead to incorrect measurement of true ability, and to incorrect decisions at the student, educator, and school/district levels. This is true across all NAEP measures (e.g., reading, mathematics, science, writing, etc.). Therefore, since comparability of measurement at the item and test levels are an essential part of validity and fairness investigations (Ercikan et al., 2004), any significant DIF detection would raise concerns about the validity of the NAEP tests for all states. Moreover, the findings might help the states to revise and potentially improve their science standards and curriculum focus if desired. And, findings in science provide guidance to such evaluations in the other NAEP tested content areas.

One limitation of this study is that, because the items in NAEP 2015 exam are still under NAEP's own review, one cannot examine the questions in detail. In other words, while the analysis suggests that DIF exists in 3 items in 12th grade exams, the exact source of DIF (vocabulary or omitted content) cannot be detected. Therefore, while the results of this study can be used to flag DIF items, the decision of whether these questions should be excluded from the exams should be done after a thorough investigation of these items.

References

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of DIF on verbal items. Journal of Educational Measurement, 36, 185.198.

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two metaanalyses. Journal of Research in Science Teaching, 26, 141–169.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27, 165–174.

Braden L, Finn CE, Lerner LS, Munroe S, Petrilli MJ, Raimi RA, Saxe DW, Smith T, Stotsky S. The State of State Standards. Washington: Thomas Fordham Foundation; 2000.

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L, & Wells, C. S. (2009). Evaluation of the National Assessment of Educational Progress: Final report. Washington, D.C.: U.S. Department of Education.

Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. American Educational Research Journal, 34, 297–331.

Cohen, A. S., & Kim, S.-H. (1993). A Comparison of Lord's Chi Square and Raju's Area Measures in Detection of DIF. Applied Measurement in Education, 17(1), 39-52

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35–66). Hillsdale, NJ: Erlbaum.

Emenogu, B. C., & Childs, R. A. (2005). Curriculum, translation, and differential functioning of geometry items. Canadian Journal of Education, 28, 123–142.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada’s national achievement tests. Applied Measurement in Education, 17, 301–321.

Ercikan, K. (2006). Developments in assessment of student learning and achievement. In P. A. Alexander & P. H. Winne (Eds.), American Psychological Association, Division 15, Handbook of educational psychology (2nd ed.) (pp. 929–953). New York, NY: Routledge.

Ercikan, K., Arim, R. G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Application of think-aloud protocols in examining sources of differential item functioning. Educational Measurement: Issues and Practice, 29(2), 24–35.

Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), Validity, fairness and testing of individuals in high stakes decision-making context (pp. 69–86). Bingley, UK: Emerald Publishing.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), APA handbook testing and assessment in psychology (Vol. 3; pp. 545–569). Washington, DC: American Psychological Association.

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? Applied Measurement in Education, 24, 1–18.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. Journal of Educational Measurement, 23, 185.196.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. Journal of Educational Measurement, 38, 164.187.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). Using statistical and judgmental reviews to identify and interpret translation DIF. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC.

Gross, P. (2005). The state of state science standards. Washington, DC: Thomas B. Fordham Institute.

Gross, P.R. (2013). Thomas B. Fordham Institute Final Evaluation of the Next Generation Science Standards. Retrieved from: http://edexcellence.net/publications/final-evaluation-of-NGSS.html

Halpern, D. (1992). Sex differences in cognitive abilities. Hillside, NJ: Lawrence Erlbaum.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed response science test. Applied Measurement in Education, 12, 211–235.

Harnish, D., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133.146.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P., & Wainer, H. (Eds.). (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. International Journal of Selection and Assessment, 9, 152-194.

Huang, X. (2010). Differential item functioning: The consequence of language, curriculum, or culture? (Unpublished doctoral dissertation). University of California, Berkeley, Berkeley, CA.

Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? Education and Urban Society, 26, 352–366.

Kind, P. M. Conceptualizing the Science Curriculum: 40 Years of Developing Assessment Frameworks in Three Large-Scale Assessments. Science Education 97(5): 671-694, 2013

Landis, J. R., Heyman, E. R., and Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. Internat. Statist. Rev. 46, 237-254.

Lawson, A., Bordignon, C., & Nagy, P. (2002). Matching the Grade 8 TIMSS item pool to the Ontario curriculum. Studies in Educational Evaluation, 28, 87.102.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? Journal of Educational Measurement, 18, 85.95.

Lerner, S. (2000a). Good science, bad science: Teaching evolution in the states. Washington, DC: Thomas B. Fordham Institute.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.

Mead, L. S., & Mates, A. (2009). Why science standards are important to a strong science curriculum and how states measure up. Evolution: Education & Outreach, 2, 359-371.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. Journal of Educational Measurement, 23, 185.196.

Meulders, M., & Xie, Y. (2004). Person-by-item Predictions. In P. Boeck & M. Wilson (Eds.), Explanatory Item Response Models (pp. 213-240). New York: Springer.

Millsap, R., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. Applied Psychological Measurement, 17(4), 297- 334

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. Journal of Educational Measurement, 25, 205.219.

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Applications of a new IRT-based detection technique to mathematics achievement test items. Journal of Educational Measurement, 28, 1.22.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. Journal of Educational Measurement, 25, 205.219.

NCES.ed.gov. (2014). Overview. Retrieved from http://nces.ed.gov/nationsreportcard/about/

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? Applied Measurement in Education, 24, 1–18.

Petersen, N. (1988). DIF Procedures for Use in Statistical Analysis. Internal memorandum.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika 53 (4), 495–502.

Rogers, H.J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105-116.

Sipahi, Rabia & Poggio, John (2020). Dissecting Ability in NAEP Science Exams and DIF. Working paper, University of Kansas.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361–370.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & W. Howard (Eds.), Differential item functioning. Hillsdale: NJ: Lawrence Erlbaum

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. Journal of Educational and Behavioral Statistics, 27(1), 53. 66.

Wenning, R., Herdman, P.A., Smith, N., McMahon, N., & Washington, K. (2003). No child left behind: Testing, reporting, and accountability. ERIC Digest. New York, NY: Eric Clearinghouseon Urban Education.

Whitmore, M.L. & Schumacker, R.E. (1999). A comparison of logistic regression and analysis of vari-ance differential item functioning detection methods. Educational and Psychological Measurement, 59, 910-927.

Young, D. J., & Fraser, B. J. (1994). Gender differences in science achievement: Do school effects make a difference? Journal of Research in Science Teaching, 31, 857–871.

Zenisky, A., Hambleton, R., & Robin, F. (2003). Detection of differential item functioning in large scale state tests: A study evaluating a two-stage approach. Educational and Psychological Measurement, 63, 51–64.

Zhang, Y., Dorans, N. J., & Matthews-López, J. L. (2005). Using DIF dissection method to assess effects of item deletion (College Board Research Report No. 2005-10). New York, NY: The College Board.

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337–347). Hillsdale, NJ: Erlbaum

Zwick, R., & Ercikan, K. E. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55–66.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. Journal of Educational and Behavioral Statistics, 25(2), 225.247.

Zwick, R., Ye, L., & Isham, S. (2012, April). Investigation of the efficacy of DIF refinement procedures. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.