Measuring the Power of Learning.®

Research Report
ETS RR–17-36

Gender and Minority Achievement Gaps in Science in Eighth Grade: Item Analyses of Nationally Representative Data

Xiaoyu Qian

Ratna Nandakumar

Joseph Glutting

Danielle Ford

Steve Fifield

July 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Gender and Minority Achievement Gaps in Science in Eighth Grade: Item Analyses of Nationally Representative Data

Xiaoyu Qian,[1] Ratna Nandakumar,[2] Joseph Glutting,[2] Danielle Ford,[2] & Steve Fifield[2]

1 Educational Testing Service, Princeton, NJ
2 University of Delaware, Newark, DE

In this study, we investigated gender and minority achievement gaps on 8th-grade science items employing a multilevel item response methodology. Both gaps were wider on physics and earth science items than on biology and chemistry items. Larger gender gaps were found on items with specific topics favoring male students than other items, for example, an earth science item requiring visual – spatial ability. Minority students were more likely than White students to score lower on harder constructed-response (CR) items. Some teachers were more likely to reduce minority achievement gaps on easier CR items than other teachers. Implications for instruction in terms of improving visual – spatial awareness, efficacy of female students, and modeling scientific literacy for minority students were discussed.

Keywords Achievement gaps; middle-school science education; science assessment; multilevel item response models

doi:10.1002/ets2.12164

According to the Trends in International Mathematics and Science Study (TIMSS), U.S. eighth graders ranked 11th among the 48 participating countries in science in 2007 (National Center for Educational Statistics [NCES], 2009, p. 32, Table 11). Among these top 11 countries, females performed significantly less well than male students in the United States and also in the Republic of Korea, Hungary, and the Czech Republic (NCES, 2009, p. 45, Figure 20). However, in the top three performing countries — Singapore, Chinese Taipei, and Japan — there was no such significant gender gap. The minority achievement gap in the United States was even larger than the gender gap in science. For example, Black students performed about 1 *SD* below White students in Grades 4, 8, and 12 in the National Assessment of Educational Progress (NAEP) science assessment in 2009 (NCES, 2011).

Women's disproportionate representation in science is well documented, especially in academia. According to the National Science Foundation (NSF; 2013, Tables 22 & 24), in 2010, the two categories in science-related areas with relatively lower doctoral degree attainments in percentage for women as compared to men were mathematics and computer sciences (3.6% vs. 9.5%) and physical sciences (5.6% vs. 11.5%). This gap was much wider in engineering (7.7% vs. 22.7%).

The minority achievement gap in science starts as early as kindergarten (Chapin, 2006). It is more prominent than the gender gap, as indicated by NAEP (NCES, 2011) and TIMSS (Gonzales et al., 2009). The gap becomes increasingly wider in higher education and academia. According to NSF (2011), in 2008 Black and Hispanic individuals' science and engineering degree attainments at the bachelor's (8.3% and 8.2%, respectively), the master's (9.4% and 6.6%, respectively), and the doctoral (4.9% and 5.8%, respectively) levels were much lower than their population percentages (12.2% and 15.5%, respectively).

The National Academies (2011) have described the expansion of the demand in the science and engineering workforce in the United States: As projected by the U.S. Bureau of Labor Statistics, more than five million people in the science and engineering workforce are needed to meet this demand. This growth is much faster than that of any other sectors. However, the growth in the science and engineering talent pool is unbalanced. Non – U.S. citizens, especially those from China and India, accounted for almost all the growth in the doctorates awarded in science, technology, engineering, and mathematics (STEM). It is uncertain whether the United States can keep relying on non – U.S. citizens to meet the needs of the U.S. STEM workforce, especially in national security and defense.

Corresponding author: X. Qian, E-mail: xqian@ets.org

Sources of Gaps

Gaps in Verbal, Quantitative, and Visual–Spatial Abilities

As early as the mid-20th century, using factorial analysis techniques, Thurstone and Thurstone (1941) found three distinct factors—verbal, quantitative, and visual–spatial—underlying 60 tests they administered to eighth-grade students. More recently, Halpern (2011) and Halpern et al. (2007) commented that the majority of the literature on the gender differences in mathematics and science covered the same three core cognitive abilities essential for learning. Research (e.g., Chiu & McBride-Chang, 2006) is plentiful in recording females' advantages in various verbal abilities such as literacy and passage comprehension, at all ages across different cultures and countries, as indicated by the Progress in International Reading Literacy Study (PIRLS; Mullis, Martin, Kennedy, & Foy, 2007) in 43 of the 45 participating countries. However, males' advantage is significant in quantitative ability from high school years, especially in complex problem solving (Hyde, 2005; Hyde & Linn, 2006).

The largest gender difference was found in visual–spatial ability, favoring males. Two meta-analyses (Coluccia & Louse, 2004; Voyer, Voyer, & Bryden, 1995) suggested that males score significantly higher than females on spatial perception (the ability to determine spatial relations in spite of distractions) and mental rotation (the ability to rotate a two- or three- dimensional object using mental images). Some researchers (e.g., Halpern, 2011; Halpern et al., 2007) believe that visual–spatial ability is a decisive factor related to, if not causal of, gender differences in mathematics and science that demand a high level of mental rotation and imaging skills, which are indispensable in occupations such as physics, engineering, surgery, architecture, and chemistry.

Although males have been reported to have higher quantitative and visual–spatial abilities, results of international comparison studies (e.g., TIMSS) indicated that, in some countries (e.g., Singapore, Chinese Taipei, and Japan; Martin, Mullis, & Foy, 2008), there was no significant gender difference in science achievement at the middle-school level. Researchers (Hyde, 2007; Hyde & Linn, 2006) have suggested that favorable social and educational factors may be the reasons for the disappearance of the gaps in these countries.

Gaps Due to Social Factors

In addition to the three decisive abilities for learning science (verbal, quantitative, and visual–spatial), science learning efficacy is a commonly studied variable due to its importance in the prediction of science engagement, achievement, and aspiration for a science-related career. Several studies (Beghetto, 2007; Caprara et al., 2008; Mantzicopoulos, Patrick, & Samarapungavan, 2008) indicated that self-efficacy correlated with or predicted science achievement significantly.

Another important social factor, socioeconomic status (SES), is regarded as a major factor correlated to, if not causal of, the lower achievement of minority students (e.g., Chiu, 2007; Chudgar & Luschei, 2009). Analyzing the 1990 High School Effectiveness Study (HSES) in a model including only SES, minority status, and gender, Von Secker and Lissitz (1999) found that average science achievement increased by 0.44 *SD* for every standard deviation increase in SES. Minority students' average science achievement was 0.58 *SD* lower than the White students' average.

Gaps Due to Instructional Factors

Among the instructional methods researched, science inquiry instruction has been extensively investigated for its effectiveness in science achievement between the genders and among the racial groups. Research results are mixed regarding whether science inquiry instruction is effective in building female students' science learning efficacy and narrowing the science achievement gap. Female students reported less teacher support (Wolf & Fraser, 2008) in an inquiry class as compared to a regular class. However, Patrick, Mantzicopoulos, and Samarapungavan (2009) found that kindergarten girls in inquiry science classes gained more learning motivation than girls in a regular science class.

Several studies have investigated whether effective instruction could narrow the achievement gap of minority students (Johnson, 2009; Johnson, Kahle, & Fargo, 2007; Von Secker & Lissitz, 1999; Wilson, Taylor, Kowalski, & Carlson, 2010), especially in urban schools (Geier et al., 2008). Johnson (2009) and Johnson et al. (2007) found that effective science instruction is fruitful in narrowing the achievement gap. Effective instruction was defined as purposeful, engaging, and flexible teaching methods that met the needs of all students.

However, after reviewing 101 studies, Minner, Levy, and Century (2010) suggested that there was no statistically significant association between the amount of inquiry and increased student science conceptual learning. Using national assessment data (National Education Longitudinal Study and HSES), Von Secker and Lissitz (1999) did not support the effectiveness of inquiry instruction in narrowing the achievement gap between White and low SES minority students. The authors stressed that instruction that accommodates individual differences, learning styles, and backgrounds is important for educational equity.

Current Study

In the current study, we investigated gender and minority achievement gaps on eighth-grade science items. We were able to quantify the magnitude of achievement gaps at the item level employing a multilevel differential item functioning (DIF) methodology. At the same time, we dug deeper to investigate whether achievement gaps at the item level could be accounted for by both student- and teacher-level variables. We also discussed how teacher instruction can potentially reduce both achievement gaps.

Research Questions

More specifically, we asked five research questions:

1. What are the item-level achievement gaps—first, between female and male students, and second, between minority (Black and Hispanic students combined) and White students?
2. Can some student variables (geometry score, science learning efficacy, and books at home) explain the item-level achievement gaps; if so, how?
3. Are certain items more likely than others to have achievement gaps?
4. Are some teachers more capable than others in reducing the gender and minority achievement gaps at the item level?
5. Can inquiry instruction explain why some teachers are more successful in reducing the gaps?

Research questions are summarized in Appendix A, along with their corresponding statistical models of investigation and the findings.

To answer these questions, the DIF methodology will are applied in the context of multilevel data to account for the student-level and teacher-level effects.

Significance

Studies using statistics such as means and effect sizes (e.g., Martin et al., 2008; NAEP, 2011) have shown that achievement gaps exist on total test scores. However, these studies could not identify items in the assessment that contributed to the achievement gaps. Thus, little effort can be made to reduce such gaps. Other studies used traditional DIF methodology to identify achievement gaps at the item level (Hamilton, 1999; Le, 2009). However, such traditional DIF methodology cannot investigate the relationship between the achievement gaps and student variables (e.g., science learning efficacy) and teacher variables (e.g., inquiry instruction). In summary, no studies have investigated the relationship of achievement gaps with the student and the teacher variables at the item level for a large-scale science assessment.

In some earlier studies (e.g., Martin et al., 2008; NCES, 2011), total test scores were compared between gender groups and between one of the ethnic groups to the White group to investigate whether there was an achievement gap and how large the gap between the two groups of students within a subject area was. However, because the comparisons were made on the total scores, neither content-specific comparisons nor abilities and skills needed for understanding the content could be evaluated. On the other hand, if comparisons are made at the item level, people who are interested in identifying abilities and skills can conduct a review of items identified as having achievement gaps. In the review, they can analyze characteristics of those items that are likely to contribute to the achievement gaps, identify abilities and skills required to address these gaps, and investigate whether the achievement gaps are caused by the content coverage in a class. Finally, based on the item review, potential improvement for instruction can be discussed and implemented to address the causes of the achievement gaps.

The multilevel item analysis methodology (Binici, 2007; Kamata, 2001) is a comprehensive statistical tool. Although means, effect sizes, and *t*-tests have been used for studying achievement gaps at the total score level, the multilevel item analysis methodology compares the odds of getting an item correct between two student groups conditional on a matched ability measure. Using this methodology, two groups can be matched after controlling for student covariant variables (e.g., geometry score, science learning efficacy, and number of books at home). In this study, these covariant variables served as control variables in the analyses so that individual student differences in these three variables were accounted for by the model. With this statistical control, an achievement gap comparison can be viewed as a purer achievement comparison, teasing out some of the social factors at the personal level within the scope of the available data. In addition, the multilevel item analysis model included an instructional variable—inquiry instruction—at the third level of the multilevel item analysis model to study how the instruction is contributing to gender and minority achievement gaps at the item level.

Doing DIF analysis in the context of a multilevel model, we can investigate how item difficulties, variance of student ability, and social and instructional factors serve as correlates in the study of gender and minority achievement gaps.

Method

Data and Sample

The study data was a representative sample from the U.S. TIMSS science assessment in 2007. TIMSS used a two-stage stratified cluster sampling design. First, schools were sampled randomly to be representative with respect to school sizes in the United States. Second, classes of the target grade were sampled within these schools (Olson, Martin, & Mullis, 2008). This methodology aimed to ensure that the sampled students were representative of the U.S. students and that their performance can be compared from one cycle to the next of the trend study. The trend study started in 1995. Assessments were conducted once every 4 years internationally.

To investigate the gender and minority achievement gaps, the study sample was a subsample from the TIMSS U.S. data. It included White (62.7%), Black (12.9%), and Hispanic (24.4%) students. When investigating the minority achievement gap, Black and Hispanic students were combined as one group, and White students were the other comparison group. The Black and Hispanic groups were combined because around two thirds of the two ethnic groups were in the same science classes (either with or without the White students) and shared the same instructional characteristics; the other one third was either the only ethnic group sharing the same science classes with the White students or in a class made up of only their ethnicities. Asians were excluded because they were not the focus of the current study. The study sample included 4,281 students, who were instructed by 189 teachers. Each teacher had at least 15 students in the class. The subsample was comparable to the original data set in student gender and ethnicity compositions, and also in school sectors (private or public) and school sizes.

The TIMSS uses a scale of $M=500$ and $SD=100$. The mean of science achievement for the current study sample was $M=520$ ($SD=79$). The male students obtained a higher mean on the science achievement than the female students for all three ethnic groups—White, Hispanic, and Black; see Table 1.

The largest gap was observed between White and Black students on science achievement. For both female and male students, Black students were 1 *SD* below White students in science achievement.

Dependent Variable

The dependent variable of the study was the student responses to the 202 dichotomously scored items—104 multiple-choice (MC) and 98 constructed-response (CR) items. It is a set of outcome variables analyzed together as a vector with 202 elements. Items with partial credits were excluded from the current study. We coded 1 for the correct responses and 0 for the incorrect responses. Responses that were not reached or omitted were coded as incorrect responses in the study, based on the assumption that TIMSS is a basic proficiency test and students are expected to reach all items in the assessment. Items covered four content domains: 70 items in biology, 37 in chemistry, 39 in earth science, and 56 in physics. The TIMSS 2007 assessment used a matrix-sampling design with a spiraling booklet administration so that students randomly received a booklet containing a subset of the total assessment items. The assessment consisted of 14 item blocks (S1, S2, . . ., and S14) distributed into 14 booklets. Each booklet had two item blocks, and each block appeared in two different booklets. Items not appearing in the administered booklet were assumed to be missing completely at random (Little & Rubin, 2002) statistically for each student. The design is shown in Table 2.

Table 1 Descriptive Statistics of Science and Geometry Scores

	Female			Male		
	Science *M* (*SD*)	Geometry *M* (*SD*)	*N*	Science *M* (*SD*)	Geometry *M* (*SD*)	*N*
White	546 (62)	496 (57)	1,331	556 (68)	501 (59)	1,354
Black	446 (69)	429 (56)	299	456 (69)	436 (58)	252
Hispanic	471 (71)	447 (60)	534	483 (75)	454 (63)	511

Table 2 Science Assessment Matrix-Sampling Design

Item block/booklet	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
1	R	R												
2		R	R											
3			R	R										
4				R	R									
5					R	U								
6						U	R							
7							R	U						
8								U	U					
9									U	U				
10										U	U			
11											U	U		
12												U	U	
13													U	U
14	R													U
No. of CR	8	8	4	6	6	7	5	10	9	8	6	9	3	9
No. of MC	4	8	8	8	9	7	7	7	6	9	7	9	9	6

Note. U = unreleased science item blocks; R = released science item blocks; MC = multiple-choice items; CR = constructed-response items.

Independent Variables

We selected three student variables — geometry score, science learning efficacy, and number of books at home — and one teacher variable, inquiry instruction, as the independent variables. The three student variables also served as control at the student level for matching two groups of students on their overall ability. Selecting these three student variables is theory driven. Earlier literature review has indicated that these variables are significantly correlated to students' science achievement. After the control, gender and minority achievement gaps were reduced at the item level. Inquiry instruction was selected as a predictor for the variations in the achievement gaps from teacher to teacher at the item level.

Geometry score was selected as an independent variable. We believe that, as compared to the other quantitative measures such as number, algebra, data, and chance, it reflects students' visual–spatial ability in addition to their quantitative ability.

The mean of the geometry score for the current study sample was $M = 478$ ($SD = 65$). The male students obtained a higher mean on the geometry score than the female students for all three ethnic groups — White, Hispanic, and Black; see Table 1.

The TIMSS study group constructed a science learning efficacy variable from students' responses to six survey questions on a 3-point Likert scale, with 3 denoting the highest self-efficacy and 1 the lowest self-efficacy. This variable with a 3-point scale was used in this study as a measure of learning efficacy. These six survey questions are "I usually do well in science," "I would like to take more science," "Science is more difficult for me," "I enjoy learning science," "Science is not my strength," and "I learn things quickly in science."

A higher percentage of male students from all three ethnic groups reported a high level of science learning efficacy than their female counterparts; see Table 3.

Number of books at home was selected as an indicator of SES. Other researchers have used number of books as a proxy for the SES (Chudgar & Luschei, 2009). Student parental income was not surveyed by the TIMSS research team. Although

Table 3 Descriptive Statistics of Science Learning Efficacy

		Female		Male	
	Science learning efficacy	*N*	Percentage	*N*	Percentage
White	Low	192	14%	144	11%
	Medium	376	28%	338	25%
	High	763	57%	872	64%
	Total	1,331	100%	1,354	100%
Black	Low	61	20%	34	13%
	Medium	93	31%	74	29%
	High	145	48%	144	57%
	Total	299	100%	252	100%
Hispanic	Low	116	22%	76	15%
	Medium	200	37%	181	35%
	High	218	41%	254	50%
	Total	534	100%	511	100%

Table 4 Number of Books at Home

		Female		Male		Combined	
	Number of books at home	*N*	Percentage	*N*	Percentage	*N*	Percentage
White	1	87	7%	180	13%	267	10%
	2	174	13%	235	17%	409	15%
	3	428	32%	406	30%	834	31%
	4	303	23%	260	19%	563	21%
	5	339	25%	273	20%	612	23%
	Total	1,331	100%	1,354	100%	2,685	100%
Black	1	59	20%	83	33%	142	26%
	2	106	35%	65	26%	171	31%
	3	83	28%	54	21%	137	25%
	4	29	10%	22	9%	51	9%
	5	22	7%	28	11%	50	9%
	Total	299	100%	252	100%	551	100%
Hispanic	1	149	28%	175	34%	324	31%
	2	156	29%	140	27%	296	28%
	3	135	25%	119	23%	254	24%
	4	57	11%	40	8%	97	9%
	5	37	7%	37	7%	74	7%
	Total	534	100%	511	100%	1,045	100%

Note. 1 = 0 to 10 books; 2 = one shelf of books; 3 = one bookcase of books; 4 = two bookcases of books; 5 = three or more bookcases of books.

parental educational level was surveyed, which is usually a good indicator of SES, 20% missing data were found on that variable.

Number of books at home was measured as a categorical variable on a scale ranging from 1 to 5, with 1 indicating *zero to ten books*, 2 for *one shelf of books*, 3 for *one bookcase of books*, 4 for *two bookcases of books*, and 5 for *three or more bookcases of books* at home. The descriptive statistics of this variable are shown in Table 4, disaggregated by ethnicity and gender. A much higher percentage of White students (more than 40%) reported that they had more than two bookcases of books at home than Black (less than 20%) and Hispanic students (less than 20%), as can be seen from the last column of Table 4 of the combined ethnic groups.

The TIMSS study team surveyed teachers on 10 questions about how often they conducted science inquiry instruction versus traditional instruction with fact memorization. They responded on a 4-point Likert scale, with 1 denoting *never*, 2 *some lessons*, 3 *about half the lessons*, and 4 *every or almost every lesson*. We used an exploratory factor analysis (principal axis factoring and direct oblimin rotation) to investigate the construct structure of the survey questions. The results suggested that these 10 items measure two factors. Five items reflected science inquiry instruction — how often students were asked to conduct experiments, work in small groups, design or plan, observe natural phenomena, and relate to daily life. The other five items reflected traditional instruction—how often students were asked to memorize facts, give explanations, read textbooks, use scientific formulas, and watch the teacher demonstrate. The factor loadings ranged from .32 to .85 on items measuring the science inquiry instruction construct, and from .36 to .60 on items measuring the traditional instruction construct. Responses to the five items measuring science inquiry instruction were summed up and rescaled to a *Z* score to form the teacher-level predictor.

Analysis Methods

In the current study, the achievement gaps were investigated using a more advanced multilevel item analysis methodology (Binici, 2007; Kamata, 2001), mathematically equivalent to a Rasch model. As compared to the traditional research on achievement gaps using effect size or *t*-test on the total score, there are several advantages to using the current methodology: First, the multilevel models estimate the item difficulties, the student variables, and the teacher variable independently at Level 1, Level 2, and Level 3 of the models, respectively, while these parameters are calibrated simultaneously. Second, the student variables (e.g., geometry score and science learning efficacy) can be used as controls at Level 2 while estimating the achievement gaps. Thus, after controlling for student-level variations, achievement gaps can be regarded and further analyzed as what is left due to the instructional variations. Third, achievement gaps can be modeled randomly from teacher to teacher (or in another way, class to class) so that an instructional variable can be modeled as a predictor for the gaps.

Five multilevel models were specified to answer the five research questions. These multilevel models are the log odds that a student will answer an item correctly as a linear function of the difficulties of the items and three various characteristics of the students. These models were mathematically equivalent to an item response theory (IRT) Rasch model (Hambleton, Swaminathan, & Rogers, 1991). For a summary of the research questions, their statistical models of investigation, and findings, please refer to Appendix A.

Model 1: Item Difficulty

Model 1 was used to estimate both item difficulty (Kamata, 2001) and student ability. The log odds of student *j* answering each item *i* correctly is a linear function of student ability (α_j) and item difficulty (β_i), as shown below. The probability of answering each item correctly by students in the model follows a Bernoulli distribution. Model 1 estimated each item's difficulty for all groups of students. Model 1 helped understand whether item difficulty is related to achievement gaps.

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j - \beta_i.$$

Model 2: Items with Achievement Gaps

Model 2 answered Research Question 1. Gender and minority achievement gaps were studied in the context of DIF by employing the methodology of multilevel DIF (Binici, 2007). Model 2 added a DIF parameter (γ_j) to Model 1, as below. This parameter indicates the odds of getting an item correct between two groups of comparison when their ability is matched. The group variable was a dichotomously coded variable. When identifying items with gender achievement gaps, the group variable was coded as 1 for the male students and 0 for the female students. When identifying items with minority achievement gaps, it was coded 1 for the White students and 0 for the minority students.

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j - \beta_i + \gamma_i * \text{group variable}_j.$$

Items with significant γ_j values were then classified into Class A, B, and C achievement gaps. These items' identification followed the DIF item classification of Dorans and Holland (1993). The coefficient γ_i was first transformed into $\Delta = -2.35\gamma_i$, and then they were classified as follows:

- Class A denoted negligible magnitude of DIF, where $|\Delta| < 1.00$
- Class B denoted moderate magnitude of DIF, where $1.00 \leq |\Delta| < 1.50$
- Class C denoted large magnitude of DIF, where $|\Delta| \geq 1.50$.

In the current study, items with Class B and Class C DIF were identified as items with achievement gaps. Items with a Δ value of 1.00 and above have an odds ratio of 1.5 and above between the two comparison groups. For example, with a Δ value of 1.00, the odds ratio of White students to get an item correct is 1.5 times that of minority students.

Model 3: Items with Achievement Gaps Controlling for Student Variables

Model 3 answered Research Question 2: Are students' geometry score, science learning efficacy, and number of books at home related to the item-level achievement gaps? Model 3 added these three student variables to Model 2, as denoted below. In addition, item-level achievement gaps estimated by Model 3 were compared to the results of Model 2.

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_{\mathrm{j}} - \beta_{\mathrm{i}} + \gamma_i * \text{group variable}_{\ j} + \pi_{1i} * \text{geometry}_j + \pi_{2i} * \text{efficacy}_j + \pi_{3i} * \text{books}_j.$$

Model 4: Gender Achievement Gaps—Random Effects across Teachers

Model 4 seeks to answer Research Questions 4 and 5: Are some teachers more capable than others in reducing the gender achievement gaps at the item level, and if so, is inquiry instruction associated with the gender achievement gaps at the item level? In Model 4, the parameter for achievement gaps at the item level (γ_{it}) was modeled as varying from teacher to teacher, as indicated by t, for items showing gender achievement gaps identified in Model 3. This approach was adopted by Prowker and Camilli (2007) for estimating item difficulty variations across jurisdictions analyzing the NAEP data. In the current study, a variance of item-level gender achievement gap was estimated for each item across 189 teachers. A significant variance indicated that teachers' instruction was associated significantly with the gender achievement gap of that particular item—some teachers may do better in instruction for narrowing the gaps than other teachers. It is helpful to know, after controlling for individual variances in geometry score, learning efficacy, and number of books at home, whether gender achievement gaps were wider in some classes than the other classes, and on what types of items. Finally, science inquiry instruction was then used to predict the variance of the achievement gap parameters to analyze whether inquiry instruction was associated with the significant achievement gap at the item level. Model 4 is shown below.

$$ln\left(\frac{p_{ijt}}{1-p_{ijt}}\right) = \alpha_j - \beta_{\mathrm{i}} + \gamma_{it} * \text{gender}_{\ j} + \pi_{1i} * \text{geometry}_j + \pi_{2i} * \text{efficacy}_j + \pi_{3i} * \text{books}_j,$$

$$\gamma_{it} = \tau_{0i} + \tau^*_{1i}\text{inquiry}_t.$$

Model 5: Minority Achievement Gaps—Random Effects across Teachers

Model 5 seeks to answer Research Questions 4 and 5: Are some teachers more capable than others in reducing the minority achievement gaps at the item level, and is inquiry instruction associated with the minority achievement gaps at the item level? Unlike gender groups, where there was almost an equal number of male and female students in each class, the students instructed by each teacher in a class were either mostly White or mostly minority students. This lack of equal numbers of White and minority students in each class made the estimation of minority achievement gaps randomly from teacher to teacher impossible. Thus, Model 5 was revised from Model 4; the γ_{it} parameter was dropped and the item difficulty (β_{it}) was modeled randomly from class to class for items identified by Model 3 as having minority achievement gaps. Inquiry instruction was then used to predict the random item difficulty. A significant variance indicated that some teachers could reduce the minority achievement gaps on certain items, whereas other teachers could not. An insignificant variance indicated that the item was of similar difficulty level for students in different classes and no achievement gap existed among classes. It is of interest to researchers, educators, and teachers regarding what type of items showed more item difficulty variations from class to class than other items so that efforts can be made to reduce the item-level gaps. Model 5 is shown below.

$$ln\left(\frac{p_{ijt}}{1-p_{ijt}}\right) = \alpha_j - \beta_{it} + \pi_{1i} * \text{geometry}_j + \pi_{2i} * \text{efficacy}_j + \pi_{3i} * \text{books}_j,$$

$$\beta_{it} = \tau_{2i} + \tau^{*}_{3i}\text{inquiry}_t.$$

For more accurate estimation, especially in the analysis of Model 4 and Model 5, teachers with fewer than 15 surveyed students were deleted, following Hox's (1998) suggestions on multilevel random parameter estimation. In this study, the random parameters were parameters γ_{it} in Model 4 and β_{it} in Model 5. After the deletion, the study sample was representative of the original data in the student ethnicity and the gender compositions. Thus, the teachers for the variance parameter estimation in the study were assumed to be a random sample from the population of U.S. teachers. This is an improvement from the earlier studies that had either small student numbers per class or small school size (e.g., Von Secker & Lissitz, 1999; the smallest size was four students per school) or did not report the size at all. The number of students instructed by each teacher ranged from 15 to 44.

Fully Bayesian Estimation

Fully Bayesian and Markov Chain Monte Carlo (MCMC) estimation was employed for model estimation. Fully Bayesian estimation was used for two reasons. First, students responded to only one booklet administered to them; thus, their responses to the other booklets in the assessment were missing. The software commonly used for the multilevel modeling, HLM, as implemented by Kamata (2001), cannot handle missing data on the dependent variable. An open software program, WinBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009), was used. Using WinBUGS, the 202 items in the science assessment can be modeled at the same time, assuming missing responses to be missing completely at random. Second, the estimation of random variables is more accurate using the fully Bayesian estimation, especially when sample sizes are small (Brown & Draper, 2006; Gelman, 2006). Priors were specified for parameters to be estimated, and they were estimated to posterior distributions. In this study, all parameters were specified as noninformative priors, following Gelman (2006) and Gelman and Hill (2007).

Results

Items With Gender Achievement Gaps

We identified 34 items with gender achievement gaps favoring male students (items with Class B and C, or medium and large achievement gaps) using Model 2. Further, controlling for the three student variables, the geometry score, the science learning efficacy, and the number of books at home in Model 3, we identified 21 items with Class B or medium achievement gap items and five Class C or large achievement gap items; the other eight items' gaps became Class A, negligible. These 26 Class B and C achievement gap items had the odds ratios of getting an item correct versus incorrect between the males and the females larger than 1.5, meaning males had 1.5 and more chance of getting these items correct after controlling for individual differences.

Ten were physics (18% of total physics items), seven biology (7% of total biology items), five earth science (13% of total earth science items), and four chemistry (11% of total chemistry items). Higher percentages of these items were in physics and earth science than in biology and chemistry. All seven items in biology were relatively lower in the odds ratio of achievement gaps as compared to items in physics and earth science between male and female students. Average item difficulty of these 26 items was only 0.17 as estimated by Model 1. This suggested that gender achievement gap items favoring male students were not highly difficult items. In fact, among the five items with large gender achievement gaps (Class C), the difficulties of three items (Items 47, 55, and 202) were below average difficulty of all 202 items. The item analyses suggested that the item content areas were more related to gender achievement gaps than the item difficulty.

Among the five items with Class C, large gender achievement gaps, two items (Items 29 and 47) were on the topic of chemical change. Both items tested the function of oxygen in chemical change. Item 29 questioned the cause of rust on a metal can. The four choices were hydrogen, oxygen, nitrogen, and helium. Item 47 asked students to choose the substance necessary for burning to take place among four choices — ozone, oxygen, hydrogen, and carbon dioxide. The other three

Table 5 Items With Large Gender Achievement Gaps

Item	Item difficulty	γ_i	*SD*	Δ	Odds ratio γ_i	Odds ratio science learning efficacy
47	−0.50	0.90***	0.16	−2.11	2.46	1.10
202	−0.46	0.84***	0.16	−1.98	2.32	1.27*
29	0.91	0.73***	0.16	−1.73	2.08	1.41
144	2.13	0.71**	0.22	−1.67	2.04	1.60***
55	−0.63	0.64***	0.16	−1.50	1.90	1.28*

*$p \le .05$, **$p \le .01$, ***$p \le .001$.

items were two earth science items (55 and 202, on the topic of Earth in the solar system and universe) and one physics item (144, on the topic of light). Difficulties of Items 55 and 202 were below zero.

Estimation results of five Class C, large achievement gap items, are shown in Table 5, along with the item difficulties, sorted in descending order by the magnitude of achievement gap parameter (γ_i).

As to item type, only six, or 6%, of the CR items had medium to large gender achievement gaps, a much lower percentage (19%, or 20) than the MC items.

In the current study, it was found that the gender achievement gaps were more related to items' content areas than to item difficulty. The content areas related to gender achievement gaps favoring males were (a) the function of oxygen in chemical change, (b) physics items about light requiring abstract understanding, and (c) earth science items about the solar system requiring visual–spatial ability.

Items With Minority Achievement Gaps

We decided to use Model 3 to identify items with minority achievement gaps, with the three student variables controlled. The reason is that if we used Model 2, simply too many items were statistically significant. Using Model 3, 71 items were identified as having medium to large minority achievement gaps. Among these 71 items, 44 items were CR items, or 45% of the total 98 CR items. The percentage of CR items showing minority achievement gaps was higher than that showing gender achievement gaps (6%). On the other hand, 26% MC items had minority achievement gaps (27 of the 104 MC items). Thirty of the total 71 items with minority achievement gaps showed Class C, large achievement gaps. The odds of getting these 30 items correct for White students were almost twice that of minority students. Relatively low average item difficulty of these 30 items ($M = 0.17$, $SD = 0.98$) suggested that difficulty is not related to minority achievement gaps. Among these 30 items, nine were biology items (70 in total); one chemistry item (37 in total); eight earth science items (39 in total); and 12 physics items (56 in total). A higher percentage of earth science (20.51%) and physics (21.43%) items showed large achievement gaps as compared to biology (12.86%) and chemistry items (2.7%). The majority of items with large, Class C, achievement gaps in biology, eight out of nine, were CR items, while in total there were 19 Class C CR minority achievement gap items. Among the 10 largest items with achievement gaps, seven were CR items. Mean difficulty of these items was 0.16 ($SD = 1.07$). Results of parameter estimation of these 30 items are shown in Table 6, sorted in descending order by the magnitude of achievement gap parameter (γ_i).

Science Learning Efficacy

Among items showing gender achievement gaps, science learning efficacy was a significant predictor for the odds ratio of getting 15 items (Items 1, 11, 12, 53, 55, 63, 69, 71, 75, 129, 144, 152, 168, 175 and 202) correct between males and females even after controlling for geometry score. For example, the odds ratio between the two genders of getting Item 144 correct increased by 1.60 times, with a 1-point increase in science learning efficacy, on a 3-point scale. Among these 15 items, nine items were earth science and physics items, suggesting, for females, that efficacy is a more important predictor of learning in these two subjects than biology and chemistry.

Among the 71 items with large minority achievement gaps, science learning efficacy was a significant predictor for 37 items in the odds ratio of getting these items correct. Among these 37 items, a higher percentage (76% or 28 items) were CR items than MC items (9 items) suggesting that science learning efficacy had a stronger association to minority students' correct response to CR items than their response to MC items.

Table 6 Items With Large Minority Achievement Gaps

Item	Item difficulty	γ_i	*SD*	Δ	Odds ratio γ_i	Odds ratio science learning efficacy
197	0.37	1.17***	0.17	−2.75	3.23	1.34**
186	−1.09	0.99***	0.18	−2.32	2.68	0.86
198	−1.37	0.98***	0.19	−2.29	2.65	0.97
151	2.56	0.97**	0.32	−2.28	2.64	1.33
182	1.22	0.96***	0.20	−2.26	2.62	1.12
1	0.36	0.93***	0.17	−2.18	2.53	1.37**
52	0.21	0.92***	0.17	−2.15	2.50	1.39**
53	0.76	0.91***	0.19	−2.13	2.48	1.50***
110	−0.91	0.89***	0.17	−2.09	2.43	1.39**
172	−0.69	0.89***	0.17	−2.09	2.43	1.43**
95	1.26	0.84***	0.21	−1.96	2.31	1.60***
174	0.16	0.83***	0.17	−1.95	2.30	1.04
15	−0.59	0.82***	0.17	−1.92	2.26	1.14
196	−0.57	0.80***	0.17	−1.87	2.22	1.38**
6	1.61	0.79***	0.22	−1.86	2.21	1.22
57	−0.12	0.79***	0.16	−1.86	2.20	1.11
45	2.00	0.75**	0.25	−1.77	2.13	1.47*
34	−0.30	0.71***	0.17	−1.67	2.03	0.82
14	−1.74	0.71***	0.21	−1.66	2.02	1.41**
72	0.06	0.70***	0.17	−1.64	2.01	1.34**
5	0.69	0.69***	0.18	−1.62	1.99	1.34**
73	1.39	0.69**	0.21	−1.62	1.99	1.51**
91	0.06	0.69***	0.17	−1.62	2.00	0.99
11	−0.55	0.68***	0.17	−1.59	1.97	1.47***
38	−0.38	0.68***	0.17	−1.59	1.97	1.08
193	0.43	0.68***	0.18	−1.59	1.97	1.22
202	−0.46	0.65***	0.17	−1.54	1.92	1.29*
12	−0.25	0.64***	0.17	−1.52	1.91	1.26*
78	0.95	0.64***	0.19	−1.51	1.90	1.25*
117	0.08	0.64***	0.17	−1.50	1.90	1.21

$^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.

Significant Difference Across Teachers

In order to understand whether teachers' inquiry instruction was a significant source of gender and minority achievement gaps, Models 4 and 5 were applied. In Model 4, the achievement gap parameters (γ_{it}) for the items showing medium or large gender achievement gaps were modeled randomly across 189 teachers, and then the teacher's inquiry instruction was used to predict the random achievement gaps. Nine items showed significant achievement gap variations across teachers. Significant gender achievement gap variations across classes suggested that the gaps of these nine items could be larger between male and female students instructed by some teachers, whereas for other teachers the gaps could be smaller. Items 29 and 144 had the highest mean on the odds ratio of getting the items correct between male and female students among the nine items identified. These two items required abstract understanding of the concepts of oxygen's function in chemical change and light traveling in different media. Significant achievement gap variations from class to class were possibly due to teachers' coverage of the concepts in the class and teachers' skills in delivering the concepts to students. Except for two items (Item 11, a biology item, and Item 29, a chemistry item), the remaining seven items were either earth science or physics items, suggesting that the achievement gaps between the two genders varied a lot from class to class as a result of teachers' instruction in these two subjects. The estimated results are shown in Table 7. For example, the mean log odds ratio of the achievement gap from class to class for Item 29 is 1.20 (the mean odds ratio of the gap is 3.32 between males and females), and the standard deviation of the random parameter is 0.36. However, our further analysis did not support the hypothesis that science inquiry instruction is a significant predictor for the variations in the gender achievement gaps from class to class.

Similarly, item difficulties (β_{it}) were estimated randomly across teachers for items with minority achievement gaps in Model 5. A significant item difficulty variation across the classes implies that instruction could have made a difference in minority students' performance. Further, inquiry instruction was used to predict the random item difficulties of these

Table 7 Gender Achievement Gaps: Random Effects

Item	γ_{it}	*SD*	Odds ratio achievement gap
29	1.20***	0.36	3.32
144	1.19**	0.47	3.29
171	0.89**	0.37	2.44
182	0.88*	0.40	2.41
4	0.86**	0.33	2.36
53	0.85*	0.39	2.34
11	0.83*	0.39	2.29
69	0.78*	0.40	2.18
56	0.77*	0.35	2.16

$^*p \le .05$, $^{**}p \le .01$, $^{***}p \le .001$.

Table 8 Minority Achievement Gaps: Random Effects Item Difficulty

Item	β_{it}	*SD*	Odds ratio achievement gap
45	1.71***	0.31	5.53
110	1.23***	0.24	3.42
14	1.20***	0.30	3.32
92	1.08***	0.21	2.94
159	0.89***	0.23	2.44
29	0.86***	0.23	2.36
195	0.83***	0.21	2.29
172	0.81**	0.26	2.25
197	0.81***	0.22	2.25
53	0.77**	0.25	2.16
41	0.76*	0.38	2.14
65	0.76**	0.27	2.14
15	0.75**	0.25	2.12
82	0.73**	0.28	2.08
182	0.72**	0.25	2.05

$^*p \le .05$, $^{**}p \le .01$, $^{***}p \le .001$.

items. Thirty-eight out of these 71 items showed significant item difficulty variations across teachers. Among these 38 items, 19 were CR items (19% of all CR items); the mean difficulty of these 19 CR items was 0.14 ($SD = 0.89$). Among the remaining 33 items with minority achievement gaps without significant item difficulty variations across teachers, 25 were CR items (25% of all CR items). The mean difficulty of these 25 CR items was 0.67 ($SD = 1.20$). The CR items with significant item difficulty variations across teachers had lower average item difficulty. This suggested that it is harder for teachers to narrow the achievement gaps of the minority students on more difficult CR items. On the other hand, it also suggested that instruction of some teachers is more effective than others in narrowing the achievement gaps on CR items of less difficulty. Minority students were more likely to perform poorly when items were harder CR items, and the teachers' instruction was less helpful in narrowing the gaps between minority and White students on these items. The estimations are shown in Table 8. For example, the mean log odds ratio of the achievement gap for Item 45 is 1.71 (the mean odds ratio of the gap is 5.53 between White and minority students); the standard deviation of the random parameter is 0.31. Inquiry instruction is not a significant predictor for item difficulty variances across teachers.

Item Difficulty and Content

Item difficulties were estimated with Model 1, a one-parameter IRT model. The reason for using a one-parameter model is that the estimated item difficulties can be directly compared among all 202 items, whereas for a two- or three-parameter model, the difficulties cannot be directly compared. The TIMSS group has investigated the model fit of all these items through IRT two-parameter and three-parameter models for dichotomous CR items and MC items, respectively (Olson et al., 2008). The accession numbers of the 202 items and their corresponding analysis ID are shown in Appendix B. With the accession numbers, interested readers can look up more detailed item information from the

TIMSS study link http://timss.bc.edu/timss2007/idb_ug.html. The information includes item content, cognitive domains, topics, item types, and their release status. For items released in 2007, their content is available as well. Content of some unreleased items in 2007 was released after the 2011 study at the link http://timssandpirls.bc.edu/timss2011/international-database.html. The item ID is the order these items were analyzed in the current study; they can be used to identify the items' analysis results in the tables starting from Table 5.

Difficulties of the total 202 items ranged from −2.70 to 4.28 ($M = 0.08$, $SD = 1.04$). Difficulty range for biology items was −2.70 to 2.73 ($M = -0.08$, $SD = 0.94$); for chemistry items −2.53 to 4.28 ($M = 0.24$, $SD = 1.23$); for earth science items −1.23 to 1.61 ($M = -0.11$, $SD = 0.78$); for physics items −2.63 to 2.72 ($M = -0.11$, $SD = 1.10$). CR item difficulties ranged from −2.02 to 4.28 ($M = 0.39$, $SD = 1.10$). MC item difficulties ranged from −2.70 to 2.13 ($M = -0.22$, $SD = 0.87$). Due to limited space, many detailed item contents and characteristics cannot be included. Interested readers may contact the first author of this paper.

Discussion

This study has several important results. First, items having gender achievement gaps favoring male students were more noticeable in the content areas of physics and earth science as compared to biology and chemistry. Female students also performed less well on the items requiring visual–spatial ability. The items identified were not all difficult items. In fact, many easy items were identified, suggesting that item difficulty may not be the only source of the gender gaps; rather, there are various reasons for the gaps.

We analyzed two items' contents with large (Class C) gender achievement gaps. Item 55 was on the topic of "Earth in the solar system and the universe" in earth science. Item 55 asked students to choose a correct answer from four choices to complete the statement "An Earth year is the length of time it takes for ...." The four choices are as follows: (A) Earth to rotate once on its axis; (B) The moon to revolve once around earth; (C) The sun to revolve once around Earth; and (D) Earth to revolve once around the sun. The correct answer was the last choice. Although this item did not require students to draw the relative position of the sun, the Earth, and the moon, the item tested the concept of the Earth's relative position to the sun at different seasons and how an Earth's year occurs. The best way to understand this concept is through visual–spatial instruction. Teachers can use computer simulation to improve students' understanding of the sun, the Earth, and the moon's motion orbits in a dynamic animation. Mathewson (1999) suggested that visual–spatial ability is central in both teaching and understanding of many science concepts, such as planetary orbits in the solar system; motion in heat energy, ocean currents and winds; and cycles in earthquake, light waves, and seasons (p. 40); however, this ability, as suggested by the author, is often overlooked by educators. The gender gap on Item 55 suggested that female students perform less well on items requiring visual–spatial ability. The finding is consistent with an earlier study (Hamilton, 1999) in which 12th-grade male students scored an average of nearly one half standard deviation higher than female students on spatial–mechanical reasoning MC items, the largest among all three different types of science items, with quantitative science and basic knowledge and reasoning as the other two types of science items. Hamilton's study also found that a CR item on the topic of eclipse showed gender DIF favoring male students. Findings of the current study and Hamilton's study suggest that visual–spatial ability is essential in teaching and understanding of the solar system and the universe.

Item 144, on the topic of light in physics, had a difficulty rating of 2.13. Its cognitive domain is *knowing*, basic knowledge required for students. It asked, "Light travels fastest through which of the following?" The four choices were air, glass, water, and a vacuum. The correct choice is the last option. This item requires students' understanding of an abstract concept: Compared to traveling in a vacuum, speed reduction of light occurs due to traveling in a medium. We believe that instruction of the concept of light is better to follow force and motion. The friction of an object can be used as an analogy for the understanding of light traveling in different media. It is easier for teachers to instruct, demonstrate, and prove (with more easily available lab equipment) the concept of an object traveling and its friction than a more abstract concept of light traveling. Another option for teachers is the use of computer simulation with assisted visual animation to explain light traveling.

Unfortunately, as Halpern et al. (2007) commented, unlike verbal and quantitative skills, visual–spatial ability is not routinely assessed in school settings. Mathewson (1999) suggested that although visual–spatial ability is central in both teaching and understanding of many science concepts, this ability is overlooked by educators. A lack of coherence in the science curriculum was criticized and suggestions were made (Mathewson, 1999) for developing a thematic approach to fostering visually assisted learning in the science curriculum. Teaching strategies such as analogy and computer-assisted learning, which employ visual–spatial thinking, were discussed in his study. It can be helpful for educators and teachers to classify science contents that require different skills, especially visual–spatial skills, and to deploy different instructional tactics such as visual modeling through computer animation, especially when visual modeling is essential in understanding and also enhances students' learning interest, engagement, and ultimately, performance.

Second, science learning efficacy is more likely a significant predictor for the gender achievement gap when an item is a physics or an earth science item. Improving female students' science learning efficacy can potentially enhance their performance on physics and earth science items. Early on in the learning experience of female students, it will be more helpful if families, schools, and communities foster and encourage these students' pursuit of, and interest and engagement in, science by meeting their different and necessary needs and connecting them to various opportunities and resources at school, community, and national levels. Also, we should inspire and empower female students by telling them stories of the many difficulties female scientists have had in the pursuit of scientific knowledge and how they persevered and eventually made huge contributions to society.

Third, a much higher percentage of CR items were Class C minority achievement gap items than MC items, especially in biology, indicating that curriculum, literacy, and cognitive thinking may be the cause. These items were not difficult items. The average difficulty of items showing Class C gaps was only slightly higher than the average difficulty of the total testing items. We reviewed two CR items' contents with the largest gaps: Items 52 and 53, two earth science CR items. Item 52 asked students, "Describe one way groundwater can become polluted." Item 53 asked students, "Explain why soil erosion can be reduced by planting trees." We believe that performance gaps can be caused by two factors. First, the item contents may not have been covered by the school curriculum or discussed at home, and students did not have any knowledge of it. To improve performance, teachers and parents can think about what the basic concepts are in different science areas in our daily life, such as pollution and environment protection. Then these concepts can be covered in a prioritized manner in school: Teachers can introduce concepts that are easier to understand and closer to life and then abstract concepts that are based on prior knowledge and less related to daily life. Second, the students may have understood the concept vaguely; however, they were not able to express it clearly and logically in writing. The second factor is related to students' cognitive understanding and their verbal ability to express a concept. Boys and girls learn things through playing and engaging. A field trip to a park can help kids understand the concept of erosion. A video clip contrasting barren sandy land to tree-covered rich soil can help in visualizing the concept. Teachers can also discuss how eroded land can do harm to human beings, such as weather change and agriculture production reduction. They can also talk about how soil has become eroded by such things as overplanting and pasturage.

Fourth, minority students performed significantly worse on physics and earth science items as compared to biology and chemistry items. Over 20% of earth science and physics items showed large achievement gaps favoring White students. As discussed earlier, many physics and earth science contents require students' abstract thinking skills to visualize and develop a logical notion of complex phenomena. The way teachers can help student understanding is instructing through simple life analogies to these phenomena through lab demonstration, visually assisted computer animation, and extensive discussion of how these phenomena occurred, what the impacts are on human beings, and what we can do to avoid, preserve, and improve the current situations.

Fifth, our results did indicate that both achievement gaps could be narrowed by teachers in some classes, especially in physics and earth science. We feel that many teachers did well in helping female and minority students' understanding of many abstract physics and earth science concepts. These teachers with better knowledge about how to engage female and minority students can introduce their instructional ideas and successful examples to those who know less at different levels—school, state, and national.

Although teachers' instruction was helpful in narrowing minority achievement gaps on less difficult CR items, instruction did not make a difference when CR items became harder. It is a challenge for teachers who teach in urban school districts with a higher concentration of minority students. These students have less resources at home, financially and also in social capital. These students need extra care in their lives through social volunteering to engage and encourage them in academics. There are many successful stories. For example, JPMorgan Chase & Co.'s Fellowship Initiative Program (Chiang & Trowbridge, 2014) helped many high school students in minority low-income neighborhoods graduate and get admitted into colleges through after-school programs and Saturday classes. The program is expanding from New York City to two other large cities, Chicago and Los Angeles, where there are many minority low-income students. Doing well in science requires all-around abilities—verbal, quantitative, and visual–spatial (Halpern et al., 2007).

Therefore, science instruction does not mean teaching science itself, but rather enhancing students' basic literacy skills as well as other reasoning and cognitive thinking skills. Brockton High School became an exemplary public school in narrowing the achievement gaps of its minority and free lunch student body (Ferguson, Hackman, Hanna, & Ballantine, 2010). Leaders and teachers of Brockton High School identified four skills—reading, writing (including writing skills to open response question), reasoning, and speaking—as basics for improving achievement not only in English but also in mathematics, science, social science, and other electives. Curriculum built around these four skills evidenced narrowing of achievement gaps of all students from Grades 8 to 10.

In our study, we did not find that inquiry instruction was effective in reducing either the gender or the minority achievement gaps. Doing science needs multiple skills: Cognitive understanding, reasoning, and speaking are basic skills to scaffold inquiry instruction. Merely using a single instructional method may not be effective in narrowing the achievement gaps for females and minority students. There may be no one-formula-fits-all curriculum (Von Secker & Lissitz, 1999) to reduce gender and minority achievement gaps because students each have their own learning needs—cognitive, efficacy, and instructional. It is essential to create a more, diversified, individualized, engaging, caring, and equitable, yet competitive environment with fair rules to play and study (Johnson, 2009; Johnson et al., 2007). Reducing gender and minority achievement gaps in science requires efforts from school science teachers, science educators, assessment experts, and other social workers and philanthropists to work together. School teachers work most closely with students; they have the best understanding of students' learning tracks and diverse needs. By working closely with school teachers, science educators can develop teacher professional development programs by identifying instruction methods that are effective in narrowing the achievement gaps through different styles of instruction and for different subject content areas. Assessment experts should use the most equitable and comprehensive measures to evaluate students' learning and achievement and to select best-fit students for various social needs.

Narrowing achievement gaps of both females and ethnic minorities at the middle-school level is indispensable to supporting the leading position of the United States in science at the high school and higher education levels, especially in the science and engineering workforce and research fields. Given the increasing demands of the science and engineering workforce and the growing reliance on foreign talents (National Academies, 2011), it is imperative to learn where the achievement gaps are and to consider ways to create a full participatory environment for all citizens to learn science, including both genders and all ethnic minorities, from a young age and especially when the gender and minority achievement gaps start to emerge. Gender and minority achievement gaps in science do not exist at the middle-school level in many countries that rank at the top of science achievement (e.g., Singapore, Chinese Taipei and Japan; NCES, 2009). However, the achievement gaps are significant in the United States. We do feel, with females accounting for half of the population and ethnic minorities comprising around one third of the population, that there is an urgent need to narrow the gaps at the middle-school level. And eventually, we may see more females and minorities holding more science and engineering positions in a more important way.

With a national representative sample, findings of the current study can be generalized to the population. One limitation of this study is that although we found that teacher instruction was related to gender and minority achievement gaps on various items, we could not identify with available data the particular instructional method that was related to the gaps. The instructional variable we used, inquiry instruction, was not significantly related to the achievement gaps. There are several reasons that inquiry instruction was not a significant predictor. This variable was a teacher self-reported variable. Teachers' understanding of inquiry instruction may vary and thus made this variable hard to measure and use. Another reason can be that, as implied by Von Secker & Lissitz (1999) and our own study, instruction should fit the needs of each individual student. Inquiry instruction may be effective for some students; however, it may not be useful for all students. In the future, we may look at how instruction that addresses each individual's needs can be more useful than inquiry instruction in predicting achievement gaps.

References

Beghetto, R. A. (2007). Factors associated with middle and secondary students' perceived science competence. *Journal of Research in Science Teaching*, *44*(6), 800–814.

Binici, S. (2007). *Random-effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods* (Unpublished doctoral dissertation). Florida State University.

Brown, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vicchio, G. M., Barbaranelli, C., & Bandura, A. (2008). Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology 100*(3), 535–534.

Chapin, J. R. (2006). The achievement gap in social studies and science starts early: Evidence from the early childhood longitudinal study. *The Social Studies*, *97*(6), 231–238.

Chiang, L., & Trowbridge, A. (2014, June 24). *Jamie Dimon touts program aimed at young men of color* [CBS online news]. Retrieved from http://www.cbsnews.com/news/jamie-dimon-touts-program-aimed-at-young-men-of-color

Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries: Country-, school-, and student-level analyses. *Journal of Family Psychology, 21*(3), 510–519.

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading, 10*(4), 331–362.

Chudgar, A., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal*, *46*(3), 626–658.

Coluccia, E., & Louse, G. (2004). Gender differences in spatial orientation: A review. *Journal of Environmental Psychology, 24*(3), 329–340.

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: Lawrence Erlbaum.

Ferguson, R. F., Hackman, S., Hanna, R., & Ballantine, A. (2010). *How high schools become exemplary: Ways that leadership raises achievement and narrows gaps by improving instruction in 15 public high schools.* Report on the 2009 Annual Conference of the Achievement Gap Initiative at Harvard University, Cambridge, MA.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, *45*(8), 922–939.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York, NY: Cambridge University Press.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2009). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context.* Retrieved from http://nces.ed.gov/pubs2009/2009001.pdf

Halpern, D. F. (2011). *Sex differences in cognitive abilities* (4th ed.). New York, NY: Psychology Press.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(3), 211–235.

Hox, J. (1998). Multilevel modeling: When and why. In R. Mathar & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–165). Berlin, Germany: Springer-Verlag.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581–592.

Hyde, J. S. (2007). New directions in the study of gender similarities and differences. *Current Directions in Psychological Science, 16*, 259–263.

Hyde, J. S., & Linn, M. C. (2006). Gender similarities in mathematics and science. *Science, 314*, 599–600.

Johnson, C. C. (2009). An examination of effective practice: Moving toward elimination of achievement gaps in science. *Journal of Science Teacher Education*, *20*(3), 287–306.

Johnson, C. C., Kahle, J. B., & Fargo, J. D. (2007). Effective teaching results in increased science achievement for all students. *Science Education*, *91*(3), 371–383.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

Le, L.T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing, 9*(2), 122–133.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28*(25): 3049–3067.

Mantzicopoulos, P., Patrick, H., & Samarapungavan, A. (2008). Young children's motivational beliefs about learning science. *Early Child Research Quarterly, 23*(3), 378–394.

Martin, M. O., Mullis, I. V. S., & Foy, P. (with Olson, J. F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Mathewson, J. H. (1999). Visual – spatial thinking: An aspect of science overlooked by educators. *Science Education, 83*(1), 33 – 54.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction — What is it and does it matter? Results from a research synthesis years 1984 – 2002. *Journal of Research in Science Teaching 47*(4), 474 – 496.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's Progress in International Reading Literacy Study in primary schools in 40 years*. Retrieved from http://timss.bc.edu/PDF/PIRLS2006_international_report.pdf

National Academies of Science, Engineering, and Medicine. (2011). Expanding underrepresented minority participation: America's science and technology talent at the crossroads. Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/read/12984/chapter/5

National Center for Educational Statistics. (2009). Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth- grade students in an international context. Retrieved from http://nces.ed.gov/pubs2009/2009001.pdf

National Center for Educational Statistics. (2011). *The nation's report card. Science* 2009. National assessment of educational progress at grades 4, 8, and 12. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2009/2011451.pdf

National Science Foundation. (2011). *Women, minorities, and persons with disabilities in science and engineering: 2011*. Retrieved from http://www.nsf.gov/statistics/wmpd/pdf/wmpd2011.pdf

National Science Foundation. (2013). *Science and engineering degrees: 1966 – 2010*. Retrieved from http://www.nsf.gov/statistics/nsf13327/pdf/nsf13327.pdf

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Patrick, H., Mantzicopoulos, P., & Samarapungavan, A. (2009). Motivation for learning science in kindergarten: Is there a gender gap and does integrated inquiry and literacy instruction make a difference? *Journal of Research in Science Teaching, 46*(2), 166 – 191.

Prowker, A., & Camilli, G. (2007). Looking beyond the overall scores of NAEP assessments: Applications of generalized linear mixed modeling for exploring value-added item difficulty effects. *Journal of Educational Measurement*, *44*(1), 69 – 87.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence*. Chicago, IL: University of Chicago Press.

Von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching*, *36*(10), 1110 – 1126.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250 – 270.

Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, *47*(3), 276 – 301.

Wolf, S. J., & Fraser, B. J. (2008). Learning environment, attitudes and achievement among middle-school science students using inquiry-based laboratory activities. *Research in Science Education*, *38*(3), 321 – 341.

Appendix A

Research Questions, Statistical Models, and Results

	Research question	Statistical model	Parameter	Result
RQ1	What are the item-level achievement gaps: first, between female and male students; and second, between minority and White students?	Model 2: gender gaps Model 3: minority gaps	γ_i, the parameter for achievement gaps at the item level	Using Model 3, 34 items were identified with Class B and C (medium and large) gender gaps, favoring male students. Using Model 3, 30 items were identified with Class C (large minority achievement gaps) favoring White students.
RQ2	Can some student variables explain the item-level achievement gaps? If so, how?	Model 3	π_{1i}, π_{2i} and π_{3i}, the parameters associated with the three student variables, at the item level	*Geometry score* was a significant predictor for getting all the items correct; learning efficacy was a significant predictor for the chance of getting many items correct even after the effect of the *geometry score*.
RQ3	Are certain items more likely than others to have achievement gaps analyzing those items identified by RQ1?	Item analyses included item difficulty, item content, item type, and cognitive abilities required to solve the questions	β_{i}, the parameter for item difficulty in Model 1	Higher percentages and wider gaps were found on physics and earth science items than for biology and chemistry items, in terms of both gender and minority gaps. Females performed less well on items requiring visual–spatial ability (Item 55). In terms of item type, higher percentage gender gap items were MC items, versus higher percentage minority gap items that were CR items, especially biology CR items. Item difficulty is not related to achievement gaps using item difficulty (β_i) estimated by a Rasch model.
RQ4	Are some teachers more capable than others reducing the gender and minority achievement gaps at the item level?	Model 4 & Model 5	γ_{it} (Model 4) β_{it} (Model 5)	Nine items had significant gender achievement gap variations (γ_{it}) from teacher to teacher; seven of them were in physics and earth science. The findings suggested that teachers' instruction varied on these 7 items, which caused gender gaps. Thirty-eight out of 71 minority achievement gap items showed significant item difficulty variations across teachers (β_{it}). It was harder for teachers to narrow achievement gaps on more difficult CR items.
RQ5	Can *inquiry instruction* explain why some teachers are more successful in reducing the gaps?	Model 4 & Model 5	τ_1 (Model 4) τ_2 (Model 5)	*Inquiry instruction* was not a significant predictor for gender or minority achievement gaps.

Appendix B
Summary of Science Assessment Items

ID	Accession	ID	Accession	ID	Accession	ID	Accession	ID	Accession	ID	Accession	ID	Accession
1	S032115	33	S022019	65	S032519	97	S042408	129	S042017	161	S042110	193	S042319
2	S032565	34	S022002	66	S032683	98	S042015	130	S042007	162	S042222A	194	S042068
3	S032403	35	S022294	67	S032258	99	S042309	131	S042024	163	S042222B	195	S042216
4	S032273	36	S022244	68	S032120A	100	S042049A	132	S042095	164	S042222C	196	S042249
5	S032019A	37	S022150	69	S032120B	101	S042049B	133	S042022	165	S042065	197	S042094
6	S032019B	38	S022042	70	S042304	102	S042182	134	S042063	166	S042280	198	S042293A
7	S032516	39	S022069	71	S042038	103	S042402	135	S042197	167	S042088	199	S042293B
8	S032620	40	S022268	72	S042298	104	S042228A	136	S042112	168	S042218	200	S042195
9	S032693A	41	S042013	73	S042261	105	S042228B	137	S042173A	169	S042104	201	S042400
10	S032693B	42	S042006	74	S042051A	106	S042228C	138	S042173B	170	S042064	202	S042164
11	S032697A	43	S042054	75	S042051B	107	S042126	139	S042173C	171	S042273		
12	S032697B	44	S042043	76	S042076	108	S042210	140	S042173D	172	S042301		
13	S042009	45	S042196	77	S042306	109	S042176	141	S042173E	173	S042312		
14	S042313	46	S042061	78	S042403	110	S042211	142	S042407	174	S042217		
15	S042059	47	S042109	79	S042272	111	S042135	143	S042278	175	S042406		
16	S042011	48	S042232A	80	S042238A	112	S042257	144	S042274	176	S032611		
17	S042028	49	S042232B	81	S042238B	113	S032542	145	S032465	177	S032614		
18	S042001	50	S042232C	82	S042141	114	S032645	146	S032315	178	S032156		
19	S042276	51	S042294	83	S042215	115	S032530A	147	S032640	179	S032056		
20	S042279	52	S042149	84	S032606	116	S032530B	148	S032579	180	S032087		
21	S042106	53	S042155	85	S032015	117	S032007	149	S032570	181	S032279		
22	S042071	54	S042150	86	S032310A	118	S032502	150	S032024	182	S032238		
23	S042101	55	S022290	87	S032310B	119	S032679	151	S032272	183	S032160		
24	S042307	56	S022292	88	S032672	120	S032184	152	S032141	184	S032654		
25	S042405	57	S022054	89	S032392	121	S032394	153	S032060	185	S032126		
26	S042244A	58	S022181	90	S032425	122	S032151	154	S032463	186	S032510		
27	S042244B	59	S022208	91	S032257	123	S032651A	155	S032650A	187	S032158		
28	S042153	60	S022078	92	S032663	124	S032651B	156	S032650B	188	S042258		
29	S022183	61	S022126	93	S032660	125	S032665A	157	S032514	189	S042016		
30	S022276	62	S022281	94	S032555	126	S032665B	158	S042042	190	S042300A		
31	S022115	63	S032385	95	S032122	127	S032665C	159	S042030	191	S042300B		
32	S022022	64	S032035	96	S042053	128	S042073	160	S042003	192	S042300C		

Suggested citation:

Qian, X., Nandakumar, R., Glutting, J., Ford, D., & Fifield, S. (2016). *Gender and minority achievement gaps in science in eighth grade: Item analyses of nationally representative data* (Research Report No. RR-17-36). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12164

Action Editor: Heather Buzick

Reviewers: Katherine Castellano, Samuel Livingston, Maria Oliveri, and Elizabeth Stone

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/