The Relationship between Test Preparation and State Test Performance: Evidence from the Measure of Effective Teaching (MET) Project

Hongli Li
Georgia State University

&

Yao Xiong
University of Pittsburgh
United States

Citation: Li, H., & Xiong, Y. (2018). The relationship between test preparation and state test performance: Evidence from the Measure of Effective Teaching (MET) project. Education Policy Analysis Archives, 26(64). http://dx.doi.org/10.14507/epaa.26.3530

Abstract: The passage of the NCLB Act enhanced accountability policies in the United States, and standardized testing became prevalent as a policy tool to ensure accountability in K-12 education. Given the high stakes of state administered accountability tests, more school teachers have adopted test-preparation strategies to ensure satisfactory student performance on state tests. However, it remains unclear as to whether and how test preparation relates to students’ state test performance. In this study, by drawing on the Measure of Effective Teaching (MET) longitudinal dataset, we examined the relationship between test preparation and students’ state test performance. We found that students with lower test performance in Year 1 received more test preparation in Year 2; however, the effects of test preparation on students’ state test performance were rather small and...
mixed. In regard to racial differences, we found that Black and Hispanic students received more test preparation than White students. Further, the effect of test preparation measured by the item “practicing for the state test” on state test performance was significantly greater for Black and Hispanic students than for White students. The implications of the study, its limitations, and directions for future research are also discussed.

**Keywords**: test preparation; state test performance; racial differences; Measure of Effective Teaching (MET)

La relación entre la preparación de la prueba y el rendimiento de la prueba estatal:

**Evidencia del proyecto Measure of Effective Teaching (MET)**

**Resumen**: La aprobación de la Ley NCLB mejoró las políticas de rendición de cuentas en los Estados Unidos, y las pruebas estandarizadas prevalecieron como una herramienta de política para garantizar la rendición de cuentas en la educación K-12. Dado el alto nivel de las pruebas de responsabilidad administradas por el estado, más maestros de escuela han adoptado estrategias de preparación de exámenes para garantizar el rendimiento satisfactorio de los estudiantes en los exámenes estatales. Sin embargo, aún no está claro si la preparación para la prueba se relaciona con el rendimiento de la prueba estatal de los estudiantes y cómo lo hace. En este estudio, al utilizar el conjunto de datos longitudinales de la Measure of Effective Teaching (MET), examinamos la relación entre la preparación de la prueba y el rendimiento de la prueba estatal de los estudiantes. Encontramos que los estudiantes con un rendimiento de prueba más bajo en el año 1 recibieron más preparación para la prueba en el año 2; sin embargo, los efectos de la preparación de la prueba en el rendimiento de la prueba estatal de los estudiantes fueron más bien pequeños y mixtos. Con respecto a las diferencias raciales, encontramos que los estudiantes negros e hispanos recibieron más preparación para la prueba que los estudiantes blancos. Además, el efecto de la preparación de la prueba medida por el ítem "practicando para la prueba estatal" en el rendimiento de la prueba estatal fue significativamente mayor para los estudiantes afroamericanos que para los blancos. También se discuten las implicaciones del estudio, sus limitaciones y las direcciones para futuras investigaciones.

**Palabras clave**: preparación de la prueba; rendimiento de la prueba estatal; diferencias raciales; Measure of Effective Teaching (MET)

A relação entre a preparação do teste e o desempenho do teste de estado:

**Evidências do projeto Measure of Effective Teaching (MET)**

**Resumo**: A aprovação da Lei NCLB aprimorou as políticas de responsabilidade nos Estados Unidos, e os testes padronizados tornaram-se predominantes como uma ferramenta política para garantir a responsabilização na educação básica. Dadas as altas apostas dos testes de responsabilização administrados pelo estado, mais professores adotaram estratégias de preparação de testes para garantir um desempenho satisfatório dos alunos nos testes estaduais. No entanto, ainda não está claro se e como a preparação do teste se relaciona com o desempenho do teste de desempenho dos alunos. Neste estudo, com base no conjunto de dados longitudinais Measure of Effective Teaching (MET), examinamos a relação entre a preparação do teste e o desempenho do teste de desempenho dos alunos. Descobrimos que os alunos com menor desempenho no teste no Ano 1 receberam mais preparação para o teste no Ano 2; no entanto, os efeitos da preparação do teste no desempenho do teste do estado dos alunos foram bastante pequenos e mistos. Em relação às diferenças raciais, descobrimos que os estudantes negros
The Relationship between Test Preparation and State Test Performance

Keywords: preparação para teste; desempenho do teste de estado; diferenças raciais; Measure of Effective Teaching (MET)

Introduction

In the past several decades, there has been a growing emphasis on testing for accountability in the United States (Smith, 2014). By the end of the 1970s, many states had established a link between test scores and school accountability (Dorn, 2007). The passage of the No Child Left Behind Act of 2001 (NCLB, 2002), technically a reauthorization of the Elementary and Secondary Education Act, became the first national framework to link school performance with student scores on standardized tests. Schools are rewarded or sanctioned depending on whether their students have achieved adequate yearly progress (AYP) according to state standards. Further, as part of the Race to the Top initiative (U.S. Department of Education, 2009), states are rewarded for implementing value-added systems through which teachers are evaluated based on their students’ test performance. Important decisions, such as tenure appointments, layoffs, and compensation are based on the results of these evaluations. In December 2015, the Every Student Succeeds Act (ESSA) was signed into law, which maintains the expectation that there will be accountability and action to effect positive changes in low-performing schools (U.S. Department of Education, 2015). As a result of the increasing emphasis on accountability and large-scale assessment, it is expected that teachers may be incentivized to adopt more test-preparation activities in their classes (Cuban, 2007). However, it remains unclear as to whether and how test preparation is related to students’ state test performance in K-12 settings.

Due to the controversial nature of test preparation and a lack of high-quality data, no systematic examinations have been conducted on test preparation in K-12 settings at the student level. The present study draws on the Measure of Effective Teaching (MET) dataset (Bill and Melinda Gates Foundation, 2012) in order to investigate the relationship between test preparation and students’ state test performance, especially among different racial groups. Through the MET project, the largest study of classroom teaching ever conducted in the United States, researchers collected teaching effectiveness indicators from 2,741 teachers at 317 schools in six major school districts over a two-year period. This dataset, therefore, offers a good opportunity to address the focal question. Based on the analysis of the MET dataset, the present study provides important empirical evidence on test preparation in the context of K-12 education.

Literature Review

What is Test Preparation?

Test preparation refers to “any intervention procedure specifically undertaken to improve test scores, whether by improving the skills measured by the test or by improving the skills for taking the test, or both” (Messick, 1982, p. 70). It involves a variety of activities, such as reviewing test content, familiarizing students with test questions, teaching test-taking strategies, and mimicking the test-taking atmosphere. There are different kinds of test-preparation activities, such as after-class test preparation offered by private tutors or commercial organizations, and in-class test preparation
offered by classroom teachers. The former is more common for admission tests such as the SAT, the ACT, and the GRE (Buchmann, Condron, & Roscigno, 2010) where the test has high stakes for students, whereas the latter is more common for state administered accountability tests (Popham, 1991) where the test has high stakes for schools and teachers. The latter is the focus of the present study.

In the literature, the phrase “test preparation” is sometimes used interchangeably with “teaching to the test.” According to Popham (2008), “teaching to the test” has two distinct meanings: (1) a teacher closely follows the curriculum such that he/she directs instruction toward the knowledge, skills, content, or affective domains represented by the test, i.e., curriculum-teaching (2) a teacher directs instruction specifically to the actual items on the test and/or practices test-taking strategies, i.e., item-teaching. The first type is regarded as appropriate as long as the test is well constructed such that it aligns with the curriculum and constitutes a good representation of the knowledge and skills students need to master. In this case, students may not be aware of test preparation because it is integrated seamlessly into the instruction. The second type is regarded as inappropriate and even unethical. However, teachers use a mixture of activities such that the distinction between these two types of “teaching to the test” activities is blurred in practice (Popham, 1991). In summary, “test preparation” is a much broader term than “teaching to the test.” Despite its true neutral meaning, “teaching to the test” is generally used pejoratively (Popham 2008); therefore, we use “test preparation” throughout this paper given that it is more generally used in a neutral sense.

Test Preparation in K-12 Settings

Given the NCLB’s “unprecedented” power, its passage is viewed as “an evolution of previous attempts to use high-stakes tests to improve educational outcomes” (William, 2010, p. 110). The NCLB linked school performance with student scores on standardized tests, as a result of which, standardized testing became prevalent as a policy tool for ensuring accountability. The effects of the accountability policy in K-12 settings are the subject of ongoing debate. On the one hand, the goal of implementing accountability is to improve student access to a high-quality and standards-based educational curriculum (Goertz & Duffy, 2001), and it has been reported that school accountability improved student performance (Hanushek & Raymond, 2005; Li, Fortner, & Lei, 2015). On the other hand, many negative impacts have been reported as well: for example, narrower curriculum under the pressure to increase test scores, lack of time for student-centered learning, student and teacher anxiety, and resentment against high-stakes testing (Menken, 2006; Musoleno & White, 2010; UNESCO, 2017; Watanabe, 2007). Despite widespread debate in regard to accountability policies, the literature lacks a systematic examination of the effects of test preparation on student achievement.

Test-preparation studies in the K-12 setting have focused on describing teachers’ perceptions of high-stakes testing and their use of test-preparation practices. Teaching test-taking skills, using data from the previous year’s test to inform instruction, and integrating test content and test format into instruction are the test-preparation practices that most teachers report implementing (Croft, Waltman, Middleton, & Stevenson, 2005; Lai & Waltman, 2008). Teachers have also reported frequent use of test-preparation activities such as practicing items similar to those included in the test immediately before the test (Firestone et al., 2002). Given the accountability policies, it is not uncommon for schools to devote a tremendous amount of time and money to test preparation (Nelson, 2013). For example, based on data collected from two urban school districts, students in grades 3–8 in one district spent at least 16 full school days preparing for state tests each year; in the other district, students in grades 6–11 devoted approximately one full month of the school year to direct test-preparation activities (Nelson, 2013). In particular, test preparation in K-12 classrooms
targets more low-performing students (Firestone et al., 2002), because boosting these students’ test scores is more effective in helping schools meet the accountability requirement.

As discussed in Cuban (2007), despite the many discussions focused on the negative consequences of test preparation in classrooms, direct evidence to support such claims is mixed and insufficient. Similarly, evidence to support the positive effects of test preparation is also lacking. We were able to find two studies that provide implicit evidence to suggest that test preparation has some positive effects on test results in the K-12 context. In an analysis of state test items in Texas, New York, and Massachusetts, Jennings and Bearak (2014) found that students were more likely to answer an item correctly if that item was designed to test frequently assessed standards. The authors attributed this result to the fact that teachers may have focused their instruction on skills that are tested frequently because these can be identified by simply looking at the test items from the previous year. Nevertheless, this evidence is not direct. A qualitative study by Welsh, Eastwood, and D’Agostino (2014), in which 34 teachers were interviewed about their test-preparation practices, found no clear effect of test-preparation practices on student test performance. However, the researchers did find that students taught by teachers who could identify items from both their own state and other states performed better than students taught by teachers who were not able to do so. In summary, given the insufficient empirical evidence collected to date, it is important to ask whether students who receive more test preparation perform better on tests than students who receive less test preparation.

Racial Group Differences in Test Preparation

The persistence of the link between socioeconomic status and educational attainment is well established in the literature (Breen & Jonsson, 2005). In the US, the most salient and persistent achievement gap is ethnicity-related (Lindsey, Graham, Westphal, & Jew, 2008). A consistent observation is that ethnic minority students (especially Black and Hispanic) lag behind their White peers (Bali & Alvarez, 2004; Fleischman, Hopstock, Pelczar, & Shelley, 2010). Researchers have studied test preparation for different racial groups; however, most studies of this kind focus on college admission tests rather than K-12 accountability tests. Drawing on the National Education Longitudinal Study (NELS), Buchmann et al. (2010) found that compared with White students racial minority students were significantly more likely to participate in SAT test preparation. For example, Black students were 3.3 times as likely as White students to have a private tutor, and Hispanic students were 1.9 times as likely as White students to have a private tutor. Asian students were 2.2 times as likely as White students to take a private course. Minority students’ advantages in test preparation existed regardless of a set of family background characteristics. Based on data from the 2007 National Household Education Survey, Devine-Eller (2012) also found that Black non-Hispanic students were more likely to participate in college entrance examination preparation than White students. Similarly, using a sample of 170 college students, Ellis and Ryan (2003) reported that Black students participated in more cognitive ability test preparation than White students. Given the long-standing Black–White achievement gap (Lee, 2002), it appears surprising that racial minority students (especially Black students) were found to be more likely to receive test preparation than White students. According to research (e.g., Alon, 2010; Buchmann et al., 2010), a possible reason is that minority students are likely to receive a larger boost in their college acceptance odds as a result of higher admission test scores than White students given the diversity-sensitive enrollment policies. This preference increases minority students’ motivation to put more effort into test preparation for college admission. Devine-Eller (2012) also pointed out that cultural influences and school context might have shaped racial minority students’ test-preparation practices.

The literature indicates that racial minority students are more likely to participate in test preparation than their White peers. A remaining question is whether test preparation boosts
minority students’ test performance more than that of White students. In a study conducted in a work setting, Chung-Herrera et al. (2009) examined racial difference in test preparation for a promotion test on job-related knowledge. Using Pearson correlation coefficients and a moderated regression analysis, they found that test preparation had a larger effect on test performance for the White participants than for the Black participants. However, the difference was not statistically significant. Ellis and Ryan (2003) studied 170 undergraduate students’ test-preparation practice with a cognitive-ability test. Their regression analysis suggests that test preparation mediated the relationship between race and test performance. However, it is unclear whether test preparation reduced or widened the Black–White achievement gap. Based on a sample of 36 low-performing high school students, Justus (2010) found that participating in the ACT math test preparation course had a greater impact on students’ ACT scores than race did, although the impact of the test-preparation course did not vary across racial groups.

In summary, research has consistently shown that in regard to college admission tests, racial minority students (especially Black students) tend to have a higher participation level in test-preparation activities than is the case for White students. However, it is unclear as to whether test preparation boosts minority students’ test performance more than it boosts that of White students. Given that racial group differences have not been systematically examined in regard to state test preparation in K-12 settings, in this study we examine whether test preparation has different effects on students from different racial groups.

The Present Study

Measure of Test Preparation in the MET

As stated in the literature (Popham 1991; Welsh et al., 2014), test preparation is a multi-dimensional construct, and measures of test preparation vary from study to study. In most of the studies on test preparation in K-12 settings, test preparation was measured by asking teachers questions regarding whether or not they offered test preparation to students (e.g., Firestone et al., 2002; Welsh et al., 2014). Test preparation has been studied mainly from the perspective of teachers such that there is a lack of literature on test preparation from the perspective of students. In the MET student survey, two student survey items were directly related to their test-preparation activities: (1) We spend a lot of time practicing for the state test; (2) Getting ready for the state test takes a lot of time in our class. We used these two items to measure student test preparation in terms of the time and/or effort they put into preparing for the state test, which is regarded as an important aspect of test preparation (Firestone et al., 2002; Nelson, 2013). The measure of test preparation in the MET project is rather simple, and we have discussed the limitations and implications of using such a coarse measure in the final section of this paper.

Hypothesized Theoretical Model

Drawing on the MET dataset, we test the theoretical model as shown in Figure 1. Based on our literature review, all test-preparation activities have a common aim—that of improving students’ test performance (Crocker, 2005). We, therefore, hypothesize that students who experienced more test preparation in Year 2 would have higher state test scores in Year 2. Also, it is reported that lower performing students are more likely to receive test preparation (Buchmann et al., 2010; Firestone et al., 2002). We, therefore, hypothesize that students with lower state test scores in Year 1 would receive more test preparation in Year 2.

Furthermore, prior achievement is probably the most significant predictor of current achievement (Buchmann et al., 2010). Therefore, as shown in Figure 1, we control for students’ Year
The Relationship between Test Preparation and State Test Performance

1 test scores in the model in order to test the effects of test preparation after accounting for prior achievement. We expect to find that students with higher Year 1 test scores would have higher Year 2 test scores. In addition, some student demographic variables show substantial relationships with test performance, such as student social economic status (Sirin, 2005), sex (Logan & Johnston, 2009), and ethnicity (Museus, Harper, & Nichols, 2010). Also, it has been observed that English language learners (ELLs) tend to perform less well on tests than non-ELLs; students in gifted programs tend to perform better on tests than students not in such programs; and students in special education programs tend to perform less well on tests than students not in such programs. Therefore, we added these demographic variables to the model as covariates too. We expect that students would have lower Year 2 test scores if they received free or reduced-price lunch, were not in a gifted program, were in a special education program, were ELLs, and were Black or Hispanic. We also expect male students to perform better on math and female students to perform better on reading (Stoet & Geary, 2013).

Figure 1. Conceptual framework
Methods

Data Sources and Participants

Over two years (AY 2009–2010 and AY 2010–2011), MET researchers collected a variety of indicators of teaching quality in the classrooms of six large school districts in the US: Charlotte-Mecklenburg (NC) Schools, the Dallas (TX) Independent School District, Denver (CO) Public Schools, Hillsborough County (FL) Public Schools, Memphis (TN) City Schools, and the New York City (NY) Department of Education. In Year 2, teachers were randomly assigned to classrooms within the schools, though a small proportion of the randomization assignment was not complied with (Bill and Melinda Gates Foundation, 2012).

The MET student-level core file in Year 2 was used in this study. One district did not collect information on whether students received free or reduced-price lunch and was excluded from the analysis for this reason. We included only the cases that responded to at least one test-preparation item in Year 2, which resulted in a sample of 24,013 students taught by 1,353 teachers from 254 schools. In this sample, 50.2% of the students responded to the survey in regard to their ELA teachers and 49.8% responded to the survey in regard to their math teachers; 49.8% were male; 8.0% were in a gifted program; 8.6% were in a special education program; 12.6% were ELLs; 56.6% had free or reduced-price lunch; and 20.8% were White, 33.4% were Black, 35.3% were Hispanic, and 7.6% were Asian.

Measures and Variables

As part of the MET project, a student perception survey was administered to all consenting students taught by teachers who participated in the MET project. For teachers who taught multiple subjects (ELA and math), their students were randomly assigned to refer to either their experience in an ELA class or in a math class when they responded to the perception survey. Different surveys were given to elementary school students (i.e., grades 4 and 5) and secondary school students (i.e., grades 6, 7, 8, and 9). The two items related to test-preparation practice, however, were identical in the elementary school student survey and the secondary school student survey:

1. We spend a lot of time practicing for the state test.
2. Getting ready for the state test takes a lot of time in our class.

Both items were scored on a 5-point Likert scale. For the elementary school student survey, the response options were 1 = no, never, 2 = mostly not, 3 = maybe/sometimes, 4 = mostly yes, 5 = yes, always. For the secondary school student survey, the response options were 1 = totally untrue, 2 = mostly untrue, 3 = somewhat, 4 = mostly, 5 = totally true. We treated the response categories from the elementary school and the secondary school students as equivalent.

Students’ state test scores in ELA and math in Year 1 and Year 2 were on a z-score scale. Other variables included in the analysis were Sex (coded as 1 if male, 0 if female), ELL status (coded as 1 if an ELL, 0 if not), Free or reduced-price lunch (coded as 1 if receiving free or reduced-priced lunch, 0 if not), Gifted program (coded as 1 if participating in a gifted program, 0 if not), Special education program (coded as 1 if participating in a special education program, 0 if not), Black (coded as 1 if Black, 0 if not), Hispanic (coded as 1 if Hispanic, 0 if not), and Asian (coded as 1 if Asian, 0 if not).
Data Analysis Procedures

Path analysis is an extension of multiple regression, which estimates hypothesized causal relationships between sets of observed variables (Kline, 2010). We used the Mplus 7.0 software (Muthén & Muthén, 1998–2012) for all the path analysis in this study. Because the two test-preparation items showed a slight departure from a normal distribution, we used the Robust Maximum Likelihood (RML) estimation method to account for the multivariate non-normality (Kline, 2010).

To begin with, we performed a path analysis to test the proposed model for ELA and math separately. Model fit was examined using the criteria recommended by Hu and Bentler (1999): the root mean square error of the approximation (RMSEA) value should be equal to or lower than .06; a comparative fit index (CFI) value of .95 or higher indicates a close fit, and values above .90 indicate a reasonable fit; the standardized root mean square residual (SRMR) should be equal to or lower than .08. In addition, a \( \chi^2 \) test is known to be sensitive to sample size, and a significant \( \chi^2 \) may be acceptable when other fit indices indicate good model fit (Markland, 2007). In the present study, we reported the \( \chi^2 \) values but did not refer to them for model fit judgment due to the large sample size in the analysis.

We also performed a multi-group path analysis in order to determine whether the relationships between test preparation and state test performance vary across racial groups (i.e., White, Black, Hispanic, and Asian). Only data from students who were White, Black, Hispanic, or Asian were included in this analysis. The path coefficients between the groups were compared using the Wald test, which has a \( \chi^2 \) distribution and is asymptotically equivalent to the Likelihood ratio test (Molenberghs & Verbeke, 2007). A significant Wald test result indicates that one should reject the null hypothesis that the path coefficients are equal across groups.

The current dataset had a four-level structure, i.e., students were nested within teachers, teachers nested within schools, and schools nested within school districts. We performed our analysis with the student-level data because the test-preparation measure was based on the student perception survey. In addition, we used the Mplus “Type = Complex” function to control for the fact that students were nested within teachers. In this way, the software adjusted the standard errors and chi-square tests of model fit by taking into account the structure whereby students were nested within teachers (Huber, 1967; White, 1982). However, we did not account for either the school level or the district level. Ignoring these two higher levels did not bias the parameter estimates but may have slightly biased the standard errors, although any such impact will have been minimal (Raudenbush & Bryk, 2002).

Results

Descriptive Statistics of Test-Preparation Practice

Table 1 provides the descriptive statistics of the sample used in the present study. As shown in the table, more than half the students responded with “Mostly yes” or “Yes, always” to the two test-preparation items. The average response value was higher than 3.5 on a 5-point scale for both items. This indicates that test-preparation practice was frequent and prevalent.

Table 2 shows the mean, standard deviation, and sample size of the responses for each racial group for ELA and math. The mean differences between the White group and the three minority groups are also presented. For both ELA and math, the mean of both items was significantly higher for the Black and the Hispanic students than for the White students at the .001 level. This result indicates that the Black and Hispanic students received more test preparation than the White students in both ELA and math classes. Asian students also had a significantly higher mean than the
White students except for the item “getting ready for the state test takes a lot of time in our class” for math.

Table 1
Descriptive Statistics of the Total Sample (N=24,013)

<table>
<thead>
<tr>
<th>Percentage of responding to each category</th>
<th>We spend a lot of time practicing for the state test</th>
<th>Getting ready for the state test takes a lot of time in our class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (No, never)</td>
<td>5.3%</td>
<td>6.8%</td>
</tr>
<tr>
<td>2 (Mostly not)</td>
<td>7.5%</td>
<td>12.7%</td>
</tr>
<tr>
<td>3 (Maybe/sometimes)</td>
<td>21.9%</td>
<td>28.2%</td>
</tr>
<tr>
<td>4 (Mostly yes)</td>
<td>27.2%</td>
<td>25.1%</td>
</tr>
<tr>
<td>5 (Yes, always)</td>
<td>36.3%</td>
<td>25.4%</td>
</tr>
<tr>
<td>Missing</td>
<td>1.8%</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

Mean 3.83 3.51

Standard deviation 1.166 1.200

Skewness - .783 -.388

Kurtosis - .211 -.729

Note: Students in grades 4 and 5 took the elementary school student survey, and students in grades 6, 7, 8, and 9 took the secondary school student survey. The two items had equivalent response categories in the two surveys. The statistics reported here combine item responses from the two surveys.

Table 2
Descriptive Statistics across Racial Groups

<table>
<thead>
<tr>
<th>Subject</th>
<th>Race</th>
<th>We spend a lot of time practicing for the state test</th>
<th>Getting ready for the state test takes a lot of time in our class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean difference against the White</td>
<td></td>
</tr>
<tr>
<td>ELA</td>
<td>White</td>
<td>3.60</td>
<td>1.184</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>3.92</td>
<td>1.175</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>3.78</td>
<td>1.160</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>3.84</td>
<td>1.152</td>
</tr>
<tr>
<td>Math</td>
<td>White</td>
<td>3.66</td>
<td>1.177</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>3.98</td>
<td>1.153</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>3.89</td>
<td>1.130</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>3.92</td>
<td>1.133</td>
</tr>
</tbody>
</table>

Note: ** indicates p < .01, *** indicates p < .001

Path Analysis Results

Figure 2 shows the standardized path analysis results for ELA and math. The model fit was adequate for ELA ($\chi^2 = 98.002, df = 16, p < .001, RMSEA = .022, CFI = .990, SRMR = .013$) and math ($\chi^2 = 104.686, df = 16, p < .001, RMSEA = .023, CFI = .988, SRMR = .011$). The correlation between the two test-preparation items, “practicing for the state test” and “getting ready for the state test in class” was .423 for ELA and .371 for math. The effects of Year 1 students’ state test
scores on the two test-preparation items were significantly negative for both ELA and math. For example, students with lower state test scores in Year 1 spent more time “practicing for the state test” in Year 2 for both ELA ($\beta = -.061, p < .001$) and math ($\beta = -.031, p < .05$). Similarly, students with lower state test scores in Year 1 spent more time “getting ready for the state test in class” in Year 2 for both ELA ($\beta = -.106, p < .001$) and math ($\beta = -.098, p < .001$).

The effects of the two test-preparation items on the students’ state test scores were statistically significant in all cases but in opposite directions. For example, students who spent more time “practicing for the state test” had significantly higher state test scores in Year 2 for both ELA ($\beta = .028, p < .01$) and math ($\beta = .034, p < 001$). However, students who spent more time “getting ready for the state test in class” had significantly lower state test scores for both ELA ($\beta = -.019, p < .01$) and math ($\beta = -.024, p < .01$).

Finally, the effects of the students’ Year 1 test scores and their demographic characteristics were generally in the direction we expected. The path coefficient from the students’ state test scores in Year 1 to Year 2 was .691 for ELA and .724 for math. Also, the state test scores in Year 2 were significantly lower for students who were not in a gifted program, had participated in a special education program, received free or reduced-price lunch, were ELLs, and/or were Black. Hispanic students had significantly lower ELA scores than White students, and Asian students had significantly higher math scores than White students. Sex was not a significant factor for either ELA or math.

![Path analysis results for ELA and math](image)

*Figure 2. Path analysis results for ELA and math*

Note: The number above is for ELA, and the number below is for math. All coefficients are standardized. * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$
A multi-group analysis was performed with race as the grouping variable, and all the parameters were allowed to be freely estimated across the four racial groups. The model fit was adequate for ELA ($\chi^2 = 100.447$, df = 40, RMSEA = .024, CFI = .994, SRMR = .016) and math ($\chi^2 = 90.007$, df = 40, RMSEA = .022, CFI = .994, SRMR = .015). Table 3 summarizes the coefficients related to the two test-preparation items only. For both ELA and math, the path from “practicing for the state test” to “state test score in Year 2” was statistically significant for Black and Hispanic students but not significant for White or Asian students. For ELA, the path coefficient for Black students (.032) was significantly higher than that for White students (.007) according to the Wald test ($\chi^2 = 3.943$, df = 1, $p < .05$). This path was also significantly higher for Hispanic students (.044) than for White students according to the Wald test ($\chi^2 = 6.700$, df = 2, $p < .01$). However, in the case of math, the Wald test did not detect any significant differences among the racial groups, although the coefficient appeared to be larger for Black and Hispanic students than for White students. Further, the path coefficient from “getting ready for the state test in class” to “state test score in Year 2” was negative but not statistically significant for any of the racial groups for ELA. For math, this path coefficient was negative and statistically significant for any of the racial groups for ELA.

Table 3

<table>
<thead>
<tr>
<th>Subject</th>
<th>Race</th>
<th>From “practicing for the state test” to “state test score in year 2”</th>
<th>From “getting ready for the state test in class” to “state test score in year 2”</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELA</td>
<td>White</td>
<td>-.007</td>
<td>-.008</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>.032*</td>
<td>-.020</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>.044**</td>
<td>-.025</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>.052</td>
<td>-.013</td>
</tr>
<tr>
<td>Math</td>
<td>White</td>
<td>.036</td>
<td>-.024</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>.046**</td>
<td>-.029*</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>.033**</td>
<td>-.016</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>.012</td>
<td>-.043</td>
</tr>
</tbody>
</table>

Note: ** indicates $p < .01$, *** indicates $p < .001$

Discussion

Relationship between Test Preparation and Students’ State Test Performance

The path analysis indicates that low-performing students consistently received more test preparation for both ELA and math than was the case for high-performing students. The hypothesis that students with lower state test scores in Year 1 would receive more test preparation in Year 2 is, therefore, fully supported. The NCLB enhanced the implication of accountability policies in the US, and standardized testing became prevalent as a policy tool to ensure accountability across states. Teachers and schools are held accountable for student performance. Students who are close to passing the state tests are even referred to as “bubble kids” (Booher-Jennings, 2005). These students are likely to receive more test preparation in an effort to ensure that schools satisfy the accountability requirement and make sure “no one is left behind.” It is, therefore, not surprising to find that students with lower state test scores in Year 1 received more test preparation in Year 2.

The path analysis also shows that students who spent more time “practicing for the state test” had significantly higher state test scores in Year 2 for both ELA and math. However, students who spent more time “getting ready for the state test in class” had significantly lower state test scores.
scores for both ELA and math. The hypothesis that students who received more test preparation in Year 2 would have higher state test scores in Year 2 is thus partially supported. Further, the correlation between the two test-preparation items “practicing for the state test” and “getting ready for the state test in class” was moderate. This result indicates that these two related variables are distinct. In other words, they measure two different aspects of test preparation. This also highlights the importance of examining the two aspects separately. The mixed findings of the two items seem to contradict each other. However, this is understandable within the context of the accountability policies.

Spending time practicing for the state test helps students to become familiar with both the test format and content, and knowing the test is an important test-preparation strategy (Jackson, & McGlinn, 2014). When students are more familiar with the test format and test-taking procedure, they can better demonstrate their knowledge and ability (Burns, Siers, & Christiansen, 2008). Also, practicing with previous test forms or sample test items is likely to boost students’ test performance. A meta-analysis conducted by Hausknecht, Halpert, Di Paolo, and Moriarty Gerrard (2007) shows that the average effect size of practicing sample items on test performance is .26 and the effect depends on the number of sample test items a test-taker practiced and the time devoted to practicing them. It is, therefore, reasonable to find that students who spent more time “practicing for the state test” had slightly higher state test scores in the present study.

However, “getting ready for the state test in class” had a negative effect on students’ test performance. With the current measure of test preparation in the MET database, we do not know the specific activities used by the teachers to prepare students for state tests. It is possible that some of the in-class test-preparation activities focus very narrowly on drills and test-taking strategies and thus do not necessarily improve students’ actual knowledge, skills, or cognitive abilities (Popham, 1991). Opponents of the current accountability and testing system claim that students have sacrificed developmentally appropriate learning time in favor of test preparation (Nelson, 2013). Too much test preparation in class may have taken up time that could have been dedicated to more meaningful instructional activities. Therefore, it is possible that “getting ready for the state test in class” may actually have a negative impact on test performance.

However, it should be noted that despite statistical significance, the effects of test preparation found in this study were very small in a practical sense. The statistically significant results, therefore, may not have practical significance given the small effect sizes. In fact, the effects of test preparation have never been shown to be either substantial or consistent. For example, although it has been reported that test preparation has positive effects on student performance on admission tests such as the SAT, the ACT, and the GRE (Buchmann et al., 2010; Park & Becks, 2015; Powers & Rock, 1999), such effects are usually not large. For foreign-language tests for admission purposes, such as the International English Language Testing System (IELTS) and the Test of English as a Foreign Language Internet-based Test (TOEFL), the effects of test preparation on test performance are mixed, depending on the actual test-preparation and/or coaching strategies used (Green, 2007; Liu, 2014). Given the sparse literature on the effects of test preparation on student achievement in K-12 settings, additional research is needed to gain a better understanding of the effects of test preparation in this particular context.

**Racial Group Differences**

Previous research indicates that racial minorities (especially Black students) tend to have a higher participation level in test-preparation activities than White students. These studies are based either on college admission tests (Alon, 2010; Buchmann et al., 2010; Devine-Eller, 2012) or on tests given to college students (Ellis & Ryan, 2003). The present study shows that a pattern of this nature
also seems to exist in K-12 settings in regard to state tests. Descriptive statistics presented in Table 2 show that both Black and Hispanic students received more test preparation than White students. A possible reason for this difference is that compared with White students, Black and Hispanic students in general had lower academic performance. For this reason, it is likely that in an effort to maximize the schools’ chances of meeting the state accountability requirements, teachers focused more on preparing Black and Hispanic students for the test than on preparing White students. This observation is in accord with the observation that Black and Hispanic students are more likely to participate in tutoring in school-based afterschool programs (U.S. Department of Education, 2007). Asian students also received more test preparation than White students, especially in ELA. In fact, in the MET project, Asian students’ performed as well as White students in ELA, yet still received more test preparation than White students in ELA. It has been found that Asian students are more likely to participate in tutoring in SAT preparation (Buchmann et al., 2010; Byun & Park, 2012). In the present study, however, the test preparation pertains to state tests. A potential reason could be that Asian students were more likely to be ELLs, so that they received more test preparation in ELA than White students.

Previous research on racial differences in regard to the effects of test preparation on students’ test performance is very limited. The few available studies (e.g., Chung-Herrera et al., 2009; Justus, 2010) focus on racial differences in college admission tests. The present study, however, provides preliminary evidence on racial differences related to state test preparation. The effect of “practicing for the state test” on students’ test scores was significantly stronger for Black and Hispanic students than for White students. This indicates that test preparation boosts the state test performance of Black and Hispanic students to a greater extent than is the case for White students. However, we did not detect a statistically significant racial group difference in regard to the effect of “getting ready for the state test in class” on “state test score in Year 2.” A possible reason is that this effect is rather small to begin with, so that it is difficult to observe any cross-group difference. In summary, the racial group difference in this study appears to be rather small in a practical sense. More empirical evidence is needed in order to understand whether this difference is meaningful enough to constitute a way to reduce the achievement gap.

Conclusions, Implications, Limitations, and Future Research

Drawing on a large-scale dataset, in this study we examined the relationship between test preparation and students’ state test performance. We found that students with lower test performance in Year 1 received more test preparation in Year 2. However, the effects of test preparation on students’ state test performance were mixed. Students who spent more time “practicing for the state test” had significantly higher state test scores, but those who spent more time “getting ready for the state test in class” in fact had lower test scores. In addition, we found that the effect of “practicing for the state test” on the “state test score in Year 2” was significantly higher for Black and Hispanic students than for White students when Year 1 test score and other demographic characteristics were controlled for. Although the racial group differences detected were small, this result is intriguing given that the present research joins a very limited number of studies on racial group differences in test preparation in K-12 settings. In future research, it would be worthwhile to focus a more thorough investigation on the issue of whether test preparation has different effects on different racial groups.

Given the paucity of empirical studies on the relationship between test preparation and students’ test performance in K-12 education, the evidence provided in this study is informative and useful. However, a primary limitation of this study is that the measure of test preparation is rather
The effect of test preparation on test performance found in this study, though statistically significant in most cases, is small in a practical sense. This is probably because the effect of test preparation is limited in nature (Messick, 1982). Also, it could be because the measure of test-preparation practice is rather simple in the present study. Test preparation is a complex multi-dimensional construct, and the measure in this study focused only on time and effort, but it is unclear what kinds of test-preparation strategies were used. Because of the moderate correlation between the two test-preparation items, we included them in the path analysis as separate variables instead of combining them into one variable. This approach had the benefit of enabling us to detect the unique effect of each variable. However, although a single-item measure can be as effective as a multi-item measure in certain circumstances (Wanous, Reichers, & Hudy, 1997), a single-item measure is likely to have low reliability, which may have weakened the validity of this study.

In addition, our test-preparation measure was based on the student self-reported survey only. English, Burniske, Meibaum, and Lachlan-Haché (2016) provided a detailed summary of the benefits and limitations of using student reports to measure teacher performance. Many studies have found that students can provide useful and reliable information about teachers’ teaching performance. However, a drawback is that students may lack the knowledge to understand the full range of teaching requirements and responsibilities. For example, in our study, when teachers skillfully integrate test preparation to their daily instruction, it might be difficult for students to tell the difference between test preparation and instruction while responding to the survey. In future research, more specific measures of test preparation from multiple sources (e.g., students, teachers, principals, or other stakeholders) are needed. In particular, classroom observations by trained observers would be very useful to capture the reality in the classroom. Also, it will be important to study the specific test-preparation practices used for accountability tests in K-12 settings and whether these practices are appropriate.

Further, we did not address district-, state-, or grade-level differences in the data analysis. In the MET project, each participating school district was from a different state, and it is likely that state tests and state accountability policies vary across states. We did not statistically control for either the district or the state level because the MET project included only six districts (or states), too small a number for modeling district or state as a higher level (Raudenbush & Bryk, 2002). Also, for reasons of confidentiality, the MET dataset did not provide the district names. It is advisable for future research to connect test-preparation practice to the specific accountability policies at the state level if such information is available. In addition, although there are certainly differences in terms of instructional focus and student development across grades (Mihaly & McCaffrey, 2014), we did not address grade-level differences in terms of test-preparation practice and effects. Also, state test programs may have different requirements for different grade levels. It is, therefore, important to perform a separate analysis at different grade levels in future research.

Finally, it is unclear as to whether higher test performance is a reflection of score inflation. Many studies have examined whether state tests show a trend toward inflating scores (e.g., Ho, 2007; Linn, Graue, & Sanders, 1990; Klein, Hamilton, Koretz, & Barron, 1998; Koretz, 1988). For example, Ho (2007) compared discrepancies between score trends from NAPE and state tests for the period of 2003 to 2005. He found that state test trends were significantly more positive than NAEP trends. However, given the many initial differences between NAEP tests and state tests in terms of the stakes involved, content, and scoring, Ho cautioned against the tendency to consider only one trend to be “true.” In the present study, we do not have sufficient information to draw a conclusion regarding whether high test performance indicates more learning. In future research, in addition to comparing NAPE scores with state test scores, firsthand information, such as data
collected through interviews and classroom observations, would be very helpful in understanding the actual processes and mechanisms whereby test preparation influences students’ learning.

**References**


U.S. Department of Education. (2007). *State and local implementation of the No Child Left Behind Act: Volume I—Title I school choice, supplementary educational services, and student achievement: A report from the National Longitudinal Study of No Child Left Behind (NLS-NCLB).*


About the Authors

Hongli Li
Georgia State University
hli24@gsu.edu
http://orcid.org/0000-0002-1039-7270
Hongli Li, PhD, is an Associate Professor in the Department of Educational Policy Studies at Georgia State University. Her research interests include applied measurement in education and quantitative methods.

Yao Xiong
University of Pittsburgh
yxiong@pitt.edu
http://orcid.org/0000-0002-9057-3464
Yao Xiong, PhD, is a Research Associate at the Learning Research and Development Center of the University of Pittsburgh. Her research interests include web-based peer assessment, applying Bayesian methods in educational measurement, and educational data mining.
education policy analysis archives
editorial board

Lead Editor: Audrey Amrein-Beardsley (Arizona State University)
Editor Consultant: Gustavo E. Fischman (Arizona State University)
Associate Editors: David Carlson, Lauren Harris, Eugene Judson, Mirka Koro-Ljungberg, Scott Marley, Iveta Silova, Maria Teresa Tattó (Arizona State University)

Cristina Alfaro San Diego State University
Gary Anderson New York University
Michael W. Apple University of Wisconsin, Madison
Jeff Bale OISE, University of Toronto, Canada
Aaron Bevanot SUNY Albany

David C. Berliner Arizona State University
Henry Braun Boston College
Casey Cobb University of Connecticut

Arnold Danzig San Jose State University
Linda Darling-Hammond Stanford University
Elizabeth H. DeBray University of Georgia
Chad d'Entremont Rennie Center for Education Research & Policy
John Diamond University of Wisconsin, Madison
Matthew Di Carlo Albert Shanker Institute

Sherman Dorn Arizona State University
Michael J. Dumas University of California, Berkeley
Kathy Escamilla University of Colorado, Boulder
Yariv Feniger Ben-Gurion University of the Negev
Melissa Lynn Freeman Adams State College
Rachael Gabriel University of Connecticut

Amy Garrett Dikkers University of North Carolina, Wilmington
Gene V Glass Arizona State University
Ronald Glass University of California, Santa Cruz
Jacob P. K. Gross University of Louisville
Eric M. Haas WestEd

Julian Vasquez Heilig California State University, Sacramento
Kimberly Kappler Hewitt University of North Carolina Greensboro

Aimee Howley Ohio University

Steve Klee University of Maryland
Jackyung Lee SUNY Buffalo

Amanda E. Lewis University of Illinois, Chicago
Chad R. Lochmiller Indiana University
Christopher Lubienski Indiana University
Sarah Lubienski Indiana University

William J. Mathis University of Colorado, Boulder
Michele S. Moses University of Colorado, Boulder
Julianne Moss Deakin University, Australia
Sharon Nichols University of Texas, San Antonio
Eric Parsons University of Missouri-Columbia
Amanda U. Potterton University of Kentucky

Susan L. Robertson Bristol University
Gloria M. Rodríguez University of California, Davis
R. Anthony Rolle University of Houston
A. G. Rud Washington State University
Patricia Sánchez University of University of Texas, San Antonio

Janelle Scott University of California, Berkeley
Jack Schneider College of the Holy Cross
Noah Sobe Loyola University

Nelly P. Stromquist University of Maryland
Benjamin Superfine University of Illinois, Chicago
Adai Tefera Virginia Commonwealth University
Tina Trujillo University of California, Berkeley
 Federico R. Waitoller University of Illinois, Chicago

Larisa Warhol University of Connecticut

John Weathers University of Colorado, Colorado Springs
Kevin Welner University of Colorado, Boulder
Terrence G. Wiley Center for Applied Linguistics

Kyo Yamashiro Claremont Graduate University

Jennifer R. Wolgemuth University of South Florida
archivos analíticos de políticas educativas
consejo editorial

Editor Consultor: Gustavo E. Fischman (Arizona State University)
Editores Asociados: Armando Alcántara Santuario (Universidad Nacional Autónoma de México), Jason Beech, (Universidad de San Andrés), Angelica Buendia, (Metropolitan Autonomous University), Ezequiel Gomez Caride, (Pontificia Universidad Católica Argentina), Antonio Luzon, (Universidad de Granada), José Luis Ramírez, (Universidad de Sonora), Paula Razquin (Universidad de San Andrés)

Claudio Almonacid
Universidad Metropolitana de Ciencias de la Educación, Chile

Ana María García de Fanelli
Centro de Estudios de Estado y Sociedad (CEDES) CONICET, Argentina

Miriam Rodríguez Vargas
Universidad Autónoma de Tamaulipas, México

Miguel Ángel Arias Ortega
Universidad Autónoma de la Ciudad de México

Juan Carlos González Faraco
Universidad de Huelva, España

José Gregorio Rodríguez
Universidad Nacional de Colombia, Colombia

Xavier Besalú Costa
Universitat de Girona, España

María Clemente Linuesa
Universidad de Salamanca, España

Mario Rueda Beltrán Instituto de Investigaciones sobre la Universidad y la Educación, UNAM, México

Xavier Bonal Sarro Universidad Autónoma de Barcelona, España

Jaume Martínez Bonafé
Universitat de València, España

José Luis San Fabián Maroto
Universidad de Oviedo, España

Antonio Bolívar Boitia
Universidad de Granada, España

Alejandro Márquez Jiménez
Instituto de Investigaciones sobre la Universidad y la Educación, UNAM, México

Jurjo Torres Santomé, Universidad de la Coruña, España

José Joaquín Brunner Universidad Diego Portales, Chile

María Guadalupe Olivier Tellez, Universidad Pedagógica Nacional, México

Yengny Marisol Silva Laya
Universidad Iberoamericana, México

Damián Canales Sánchez Instituto Nacional para la Evaluación de la Educación, México

Miguel Pereyra Universidad de Granada, España

Ernesto Treviño Ronzón
Universidad Veracruzana, México

Gabriela de la Cruz Flores Universidad Nacional Autónoma de México

Mónica Pini Universidad Nacional de San Martín, Argentina

Ernesto Treviño Villarreal Universidad Diego Portales Santiago, Chile

Marco Antonio Delgado Fuentes Universidad Iberoamericana, México

Omar Orlando Pulido Chaves Instituto para la Investigación Educativa y el Desarrollo Pedagógico (IDEP)

Antoni Verger Planells Universidad Autónoma de Barcelona, España

Inés Dussel, DIE-CINVESTAV, México

José Ignacio Rivas Flores Universidad de Málaga, España

Catalina Wainerman Universidad de San Andrés, Argentina

Pedro Flores Crespo Universidad Iberoamericana, México

Juan Carlos Yáñez Velazco Universidad de Colima, México