ACT Research Report

2021-10

Are They Trying?

Motivation in State Census Testing With a College Admissions Exam

JEFFREY T. STEEDLE, PHD

ACT.org

Conclusions

Students participating in statewide testing using the ACT® test exhibited motivation consistent with expectations based on patterns observed in national ACT testing.

So What?

Evidence of testing motivation supports the valid interpretation of ACT results and the claim that students should be more motivated when taking the ACT compared to low-stakes achievement tests.

Now What?

States, districts, and schools should consider carefully how scores, score trends, and indicators of college readiness differ between the ACT and low-stakes achievement tests on which students may not have exerted their best efforts.

About the Authors

Jeffrey T. Steedle

Jeffrey Steedle is a lead psychometrician directing the team responsible for statistical analyses for the ACT test and guiding research studies related to maintaining measurement quality while making changes to the assessment program. Jeff holds advanced degrees in education, statistics, and educational psychology, and his research interests include assessment validation, psychometrics, and motivation on achievement tests.

Acknowledgements

The author acknowledges Krista Mattern for her careful review of this manuscript and recommendations for improvement.

Introduction

Statewide testing with college admissions exams dates to 2001, when Colorado and Illinois introduced the practice. Such testing expanded in subsequent years, and beginning in 2017–2018, the Every Student Succeeds Act (ESSA) allowed states to use a college admissions exam to meet accountability requirements for high schools. Some see this option as beneficial for students, especially those students who might not have taken the ACT or SAT tests otherwise, yet this policy remains controversial among educational assessment researchers (e.g., Camara, Mattern, Croft, Vispoel, & Nichols, 2019 and responses). Indeed, an invited-speaker session at the 2019 Annual Meeting of the National Council on Measurement in Education was dedicated to this topic (*Using the ACT and SAT for Accountability Under the Every Student Succeeds Act: Appropriate or Inappropriate Use*). Critics may question, for example, the assertion that students should exhibit greater motivation on a college admissions test compared to a typical low-stakes achievement test because performance on an admissions test has stakes attached (e.g., admissions, scholarships, and course placement). Generally, low testing motivation negatively impacts the validity of score interpretations and uses. As suggested by some critics, for example, low motivation could manifest in a dramatic increase in the proportion of students scoring near chance level.

With the introduction of statewide testing—or *census testing*—with a college admissions exam, the characteristics of the tested sample are expected to change, as are the distributions of test scores and possibly indicators of testing motivation. The study reported here was designed to estimate such changes using data from five states that recently adopted census testing of 11th graders with the ACT test. Testing motivation was inferred from analyses of item response and score patterns. The results indicated the extent of the changes and, importantly, whether census testing was associated with an unexpectedly large decrease in apparent testing motivation. Mean ACT score differences between racial/ethnic groups from census testing were compared with those from a low-stakes testing context (the National Assessment of Educational Progress). If the two tests revealed different patterns of mean score differences, this could suggest something about the perceived stakes of the ACT in the census testing context.

Study results have direct bearing on the interpretation and use of college admissions test scores from statewide administrations. With the introduction of census testing, the tested sample in each state included more male students, minority students, and students with lower high school grade point averages. Average ACT performance in each state was slightly lower in the census testing sample compared to the self-selected sample prior to census testing, and mean ACT scores decreased by different amounts for different racial/ethnic groups. Decreases were driven by the increase in the number of students tested and their lower average academic achievement levels.

Two out of three motivation indices suggested lower motivation among census testers compared to pre-census testers, but some decline was expected due to the correlation between test scores and the motivation indices. Subsequent regression analyses indicated that observed decreases in apparent motivation were consistent with expectations given decreases in ACT scores. On average, mean score differences between racial/ethnic groups were slightly smaller on the ACT than on the low-stakes National Assessment of Educational Progress. This finding possibly indicated higher motivation among lower-achieving student groups on the ACT than on a low-stakes assessment. Taken together, the results are consistent with the notion that use of a college admissions exam in the census testing context supports testing motivation even for students who might not have taken the exam otherwise.

Background

At present, more than 25 states test all public high school students in 11th grade with a college admissions test (the ACT or SAT), though not all use test results to meet federal accountability requirements. Census testing programs are viewed as beneficial because they help assess college readiness, increase college awareness and recruitment, require less testing time than other assessments, and provide students with a no-cost opportunity to take a college admissions test during the school day. However, the use of college admissions tests under ESSA has hurdles to overcome regarding issues such as alignment to state content standards and performance levels. For more details, refer to the special section of *Educational Measurement: Issues and Practice* (2019, Vol. 38, No. 4). To provide context for the current study, the following literature review focuses on testing motivation and how the introduction of census testing could impact aggregate test scores.

Testing Motivation

The *Standards for Educational and Psychological Testing* acknowledge that testing motivation should be considered when interpreting test scores (Standard 3.18; AERA, APA, & NCME, 2014). This recommendation arises from understanding that motivation—though it may be construct-irrelevant—is significantly associated with test scores, particularly under low-stakes testing conditions when motivation is more variable (e.g., Sundre, 1999). Test takers generally exhibit higher motivation and achieve higher test scores when they perceive stakes attached to test performance, including when tests are graded (Napoli & Raymond, 2004; Wolf & Smith, 1995) or used for admissions (Cole & Osterlind, 2008). In one prior study, average test scores were estimated to increase by 0.41 to 0.50 standard deviations when a state achievement test changed from having low stakes to requiring students to achieve certain scores to graduate high school (Steedle & Grochowalski, 2017). Considering this result, aggregate performance on typical state achievement tests, where students have little incentive to perform well, could grossly underestimate student

achievement. If motivation on college admissions tests is higher—as some proponents claim (e.g., Camara et al., 2019)—then scores from such tests could more accurately reflect students' college readiness levels than scores from low-stakes achievement tests aligned to state college and career readiness standards.

Test Taker Characteristics

Historically, students have chosen to take college admissions tests because they need scores to support college applications. That sample—presumably college-bound—tended to have higher average achievement and socioeconomic status than the general high school population. Thus, whenever a state introduces census testing with a college admissions test, test taker demographics and score distributions are expected to change. One prior study examined data from 12 states that adopted census testing with the ACT, and demographic percentages for male, Black, Hispanic, low-income, and low parental education levels all increased (Allen, 2015). Larger score changes were observed in states with lower percentages of students taking the ACT before census testing (i.e., states where the sample changed the most). Specifically, mean ACT Composite scores (average of English, math, reading, and science) were estimated to decrease by 1.22 (on a 1–36 scale) per additional 25% of students tested.

When students choose to take a college admissions test, one might assume generally high motivation to perform well, in part because those students recognize stakes attached to test performance (i.e., college admissions, scholarships, and course placement). According to expectancy-value theory (Wigfield & Eccles, 2000), such students would have a greater expectation of performing well and place higher value on the test compared to relatively low-ability students required to take the test. Prior research provides no direct comparison between low-stakes and high-stakes conditions for college admissions tests, but one study compared 10th graders to 11th graders taking the ACT (Allen & Mattern, 2019). Considering that 10th graders might perceive the test to be too difficult or see less value in the test, they could be less motivated than 11th graders. An analysis of 53 schools that administered the ACT to nearly all 10th and 11th graders revealed that 11th graders were more likely to respond to every item (89% vs. 84%) and less likely to exhibit "guessing patterns" (e.g., ABCDCBA, BCDBCDBCD, AAAAABBBBB; 18% vs. 23%), but only on the 75-item English test. Differences in completion and guessing rates, which were assumed to indicate motivation, were negligible on the math, reading, and science sections.

Detecting Motivation

Allen and Mattern (2019) used two simple indices to measure apparent testing motivation, but other methods are common in motivation research. As testing programs transition to online administration, it is becoming more common to analyze response latency data (e.g., response-time effort; Wise & Kong, 2005), but nearly all ACT testing in the United States still occurs on paper, which makes it

impossible to gather latency data. Thus, motivation must be inferred from response patterns and item scores. Item omit rate was the first of three approaches used in the present study. In prior research, skipping questions on state standardized tests predicted future educational outcomes even when the researchers controlled for prior test scores (Hernández & Hershaff, 2015), which supports the use of item omit rate as an indicator of student motivation.

The C_z index was proposed to detect unusually long repeating patterns in item responses, which can indicate unmotivated responding (Cui, 2020). This approach automates the search for guessing patterns, which otherwise must be manually specified. Cui (2020) provided a pseudo-code algorithm for computing the C_z index, which is simply the length (number of items) of an examinee's longest repeating response pattern. For example, $C_z = 5$ if the longest repeating pattern is BBBBB, $C_z = 8$ if the longest repeating pattern is ABCDABCD, $C_z = 6$ if the longest repeating pattern is ABABAB, and so forth. Of course, some amount of repeating responses is expected, even for high-ability, highly motivated examinees. For that reason, Cui (2020) proposed setting a cutoff for flagging high C_z values using a scree-like procedure that involves inspecting a C_z frequency plot to see where the "elbow" occurs (i.e., where the distribution becomes more uniform).

In general, person-fit statistics indicate the extent to which item score patterns are consistent with expectations based on typical patterns or a measurement model. Low testing motivation can lead to aberrant responding, which can then be detected using person-fit statistics. The H^T index (Sijtsma & Meijer, 2016) is a non-parametric person-fit statistic that has been shown to perform relatively well for detecting simulated aberrant responding (Karabatsos, 2003). In basic terms, H^T is a correlation indicating similarity between an examinee's item scores and those of all other examinees. H^T for examinee i is calculated as

$$H_i^T = \sum_{j \neq i} \sigma_{ij} \Big/ \sum_{j \neq i} \max\left(\sigma_{ij}\right) \tag{1}$$

where σ_{ij} is the covariance between item scores for persons i and j. H^T is greater when an examinee's item scores are more like those of other examinees, so low H^T values indicate poor person fit. Low H^T is often observed when high-achieving students answer easy items incorrectly or low-achieving students answer difficult items correctly.

Applications of H^T often use the same cutoff to identify low H^T for all examinees, but the variance of H^T differs across raw scores, so a single cutoff is inappropriate. For example, this approach can flag many students with nearly perfect scores who responded incorrectly to a very small number of relatively easy items. A better

approach is to set a unique critical value for each raw score in a way that mimics null-hypothesis significance testing. This involves simulating the H^T null distribution using item scores from simulated examinees with non-aberrant responding. A similar approach has been used to establish cutoffs for methods of detecting unmotivated responding on self-report inventories (Steedle, Hong, & Cheng, 2019).

Research Questions

The current study was designed to address the following research questions:

- To what extent do ACT score distributions change with the introduction of census testing?
- Does the introduction of census testing lead to unexpected decreases in apparent testing motivation?
- How do mean score differences between racial/ethnic groups differ between ACT census testing and a low-stakes achievement test?

Findings from this study indicate whether there is empirical backing for concerns about motivation on a college admissions exam used in the census testing context. Therefore, the results have bearing on the validity of score interpretations and uses for such tests.

Method

Measure

The data analyzed in this study came from the ACT test, which is widely recognized by colleges and universities to support admissions decisions (ACT, 2020). When students take the ACT test battery, they complete four multiple-choice sections: English (75 items, 45 minutes), math (60 items, 60 minutes), reading (40 items, 35 minutes), and science (40 items, 35 minutes). Scores on each section are reported on a 1–36 scale, and a 1–36 Composite is calculated as the average of the four subject test scores. Some students also take a writing test, but it was not analyzed in this study because several statistical methods involved analyzing item scores, which were not available from the one-prompt writing test. When students register for the ACT, they provide demographic information (e.g., gender, race/ethnicity, and income) and report their high school grades.

Sample

For this study, ACT data were gathered for five states that first administered the ACT to all 11th graders in public schools in the spring of 2015 (i.e., students in the high school graduating class of 2016). Prior to the introduction of census testing in spring

2015, most high school students who took the ACT in those five states would have chosen to do so by registering for one of the national administrations occurring on Saturdays throughout the year. Students from the high school graduating class of 2015 who took the ACT served as the comparison group for the class of 2016, which was required to take the ACT during the school day (Table 1). To examine the effect of introducing census testing, analyses were conducted on test scores for the class of 2015 (since some of these students tested multiple times, scores earned nearest in time to the spring of 11th grade were used) and on spring 2015 census testing scores for the class of 2016. Different test forms were administered during different testing windows and for special examinee groups (e.g., requiring certain accommodations). Test forms taken by fewer than 500 students were excluded from analyses because some statistical methods required large samples to estimate form-specific critical values for flagging possible unmotivated responding.

Table 1. Study Groups

Comparison group	Census testing group
In the high school graduating class of 2015	In the high school graduating class of 2016
All high school students who chose to take the ACT	All public high school students
Registered to take the ACT on a Saturday	Took the ACT during the school day
ACT scores closest in time to spring of 11th grade	ACT scores from census testing in spring of 11th grade

Table 2 shows sample demographics for the five states combined; in order to mask identities, demographics from individual states are not reported. After the introduction of census testing, gender balance in the sample improved to nearly 50/50. Historically, White students have taken college admissions tests at disproportionately high rates, so the percentage of White students in the overall sample was expected to decrease. Indeed, the percentage of White students declined from 72.4% to 67.0%, and the percentage of Hispanic students increased from 5.9% to 10.9%. Unexpectedly, however, the percentage of Black students decreased from 11.3% to 10.8% in the overall sample even though the percentage increased in each of the five states (by 0.4% to 0.8%). This result was caused by a certain state where few students took the ACT before census testing (about 40%, compared to 70–80% in the other states). In that state, most of the added students were White or Hispanic, and this had the effect of decreasing the percentage of Black students in the overall sample. The introduction of census testing also caused the distribution of self-reported high school grade point average to shift toward lower grades, since the census testing sample included a greater proportion of lower-achieving students who would not have taken the ACT if not for census testing.

Table 2. Sample Demographics From Complete Records

Category	Group	Pre-census testing (class of 2015)	Census testing (class of 2016)
Gender	Female	54.8%	49.9%
	Male	45.2%	50.1%
Race/ethnicity	Black	11.3%	10.8%
	Amer. Indian/Alaska Native	0.5%	0.8%
	White	72.4%	67.0%
	Hispanic	5.9%	10.9%
	Asian	3.2%	2.9%
	Native Hawaiian/Pac. Islander	0.2%	0.4%
	Two or more races	3.6%	4.4%
	Prefer not to respond	3.0%	2.9%
HSGPA range	F to D (0.0–0.9)	0.0%	0.9%
	D to C– (1.0–1.4)	0.4%	2.9%
	C– to C (1.5–1.9)	2.1%	6.1%
	C to B– (2.0–2.4)	7.9%	14.1%
	B– to B (2.5–2.9)	15.8%	17.0%
	B to B+ (3.0–3.4)	29.3%	25.8%
	A– to A (3.5–4.0)	44.5%	33.2%
Sample size		127,858	201,239

Descriptive Analyses

The descriptive analyses conducted for this study were carried out on all available data from the five states and separately for the four largest racial/ethnic groups (Asian, Black, Hispanic, and White). ACT score distributions were generated for the tests administered before and after the introduction of census testing. Cumulative score distributions were plotted, and mean scores were calculated. Those means were used to calculate changes in mean score differences between racial/ethnic groups. For each of the five states, mean score differences on a low-stakes test (the National Assessment of Educational Progress) were calculated from the most recent publicly available aggregate score data. This included the State Snapshots at 8th grade for the 2015 science test, the 2019 math test, and the 2019 reading test (National Center for Education Statistics, 2016, 2019a, 2019b). To facilitate comparisons, all differences were expressed as effect sizes in standard deviation units.

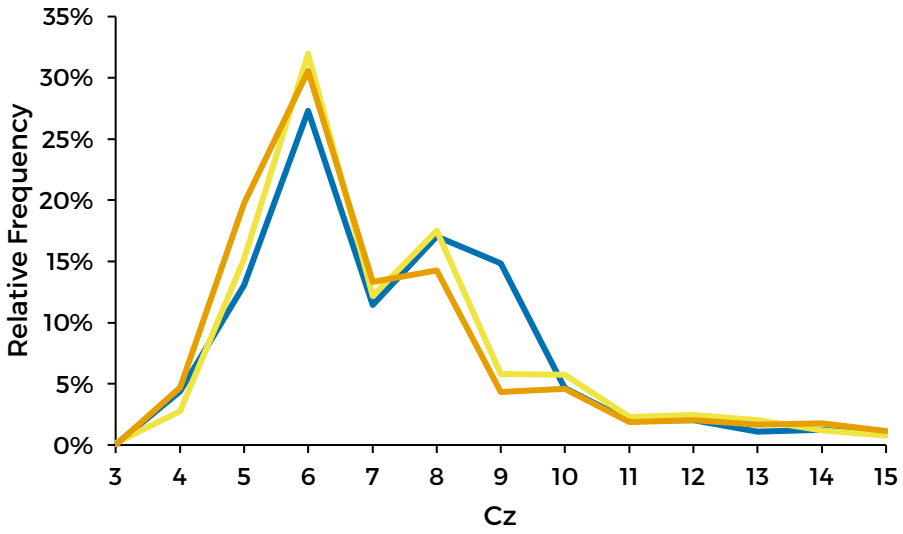
Motivation Indicators

Three methods were applied to identify low motivation. For each method, cutoff values were chosen to flag students for apparent unmotivated responding. Those choices were somewhat arbitrary, and other choices would have changed the percentages of students flagged. However, for this study, the relative percentages of

flagged students before and after the introduction of census testing—not the absolute percentages—were most important for addressing the research questions.

First, the number of items omitted was calculated for each student on each test, and students were flagged when they omitted more than 20% of the items on a test. Omitting items was generally uncommon, especially on the shorter reading and science tests, and this percentage was chosen to avoid false positive flags for students who may have omitted items due to running out of time. Next, the length of the longest repeating response pattern (C_z) was calculated for each student on each test following the algorithm provided by Cui (2020). For each test, a frequency plot of C_z was generated to identify a cutoff that would indicate unusually long repeating response patterns. The plots were quite consistent across test forms and subjects, and a cutoff of 11 was selected for all four subject areas (Figure 1). Finally, critical values for the person-fit statistic H^T were estimated for each test form using the following procedure: fit a three-parameter logistic IRT model to the data for a test form, simulate data for 10,000 examinees from a uniform distribution of ability, calculate their H^T statistics, and determine the 5th percentile H^T value for the simulated examinees at each raw score point. Students were flagged when their H^T values fell below the cutoff for their respective raw scores in the simulated null distributions.

Figure 1. Example C_z Relative Frequency Distributions



Results

Score Distributions

Figure 2 shows the cumulative score distributions before and after the introduction of census testing, and Table 3 shows corresponding means and standard deviations. Compared to the means and standard deviations of the pre-census sample, the

census testing means were 1.5 to 1.9 points lower on the 1–36 score scale, and the standard deviations were 0.1 to 0.5 points higher. In standard deviation units, the score differences were –0.32, –0.30, –0.33, and –0.33 for English, math, reading, and science, respectively. Lower mean scores were expected due to the influx of lower-achieving students, but some of the differences reflected learning that occurred between the spring of 11th grade and the time of testing for the pre-census testing sample, which could have occurred after spring of 11th grade. On average, pre-census testing occurred two months later in a student's high school career than census testing (assuming census testing took place in March of 11th grade); 34% of pre-census testing occurred in September of 12th grade or later.

Figure 2 shows vertical reference lines indicating the approximate scale scores associated with chance-level test performance (the score varies slightly across forms). The percentage of students scoring at or below chance level was higher in the census testing sample by approximately 6, 7, 8, and 7 percentage points for English, math, reading, and science, respectively. This result could reflect lower motivation to some extent, but some increase was expected due to the greater proportion of students in the census testing sample with very low ability.

Figure 2. ACT Score Distributions by Testing Context

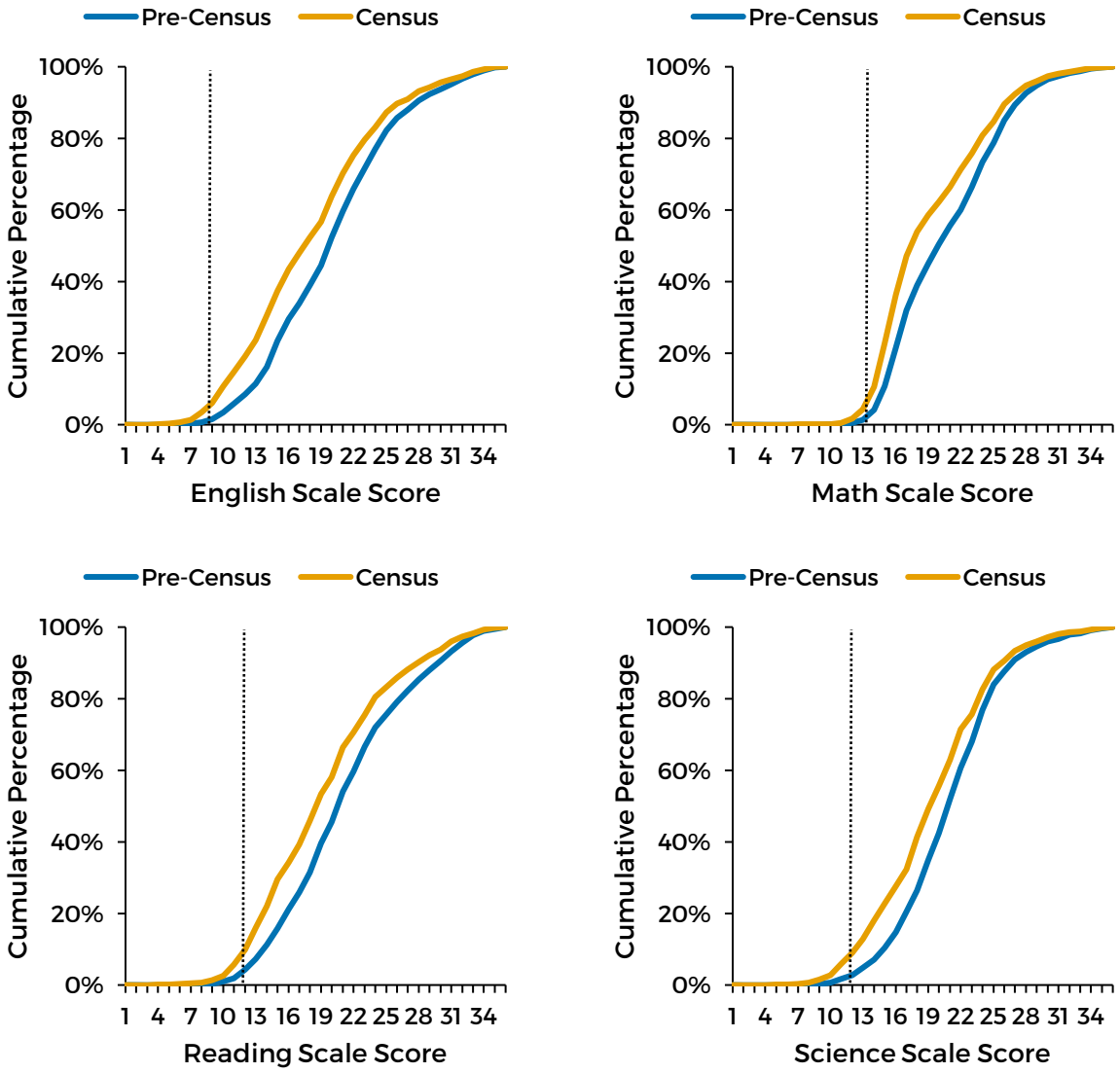


Table 3. ACT Score Means and Standard Deviations by Testing Context and Race/Ethnicity

			English			Math			Reading			Science		
Race/ethnicity	Context	*N*	Mean	SD	*d*	Mean	SD	*d*	Mean	SD	*d*	Mean	SD	*d*
All	Pre-census	127,858	20.3	5.9	−0.32	21.1	4.9	−0.30	21.6	5.8	−0.33	21.4	4.8	−0.33
	Census	201,239	18.4	6.3		19.6	5.0		19.6	6.0		19.7	5.3	
Asian	Pre-census	4,066	20.0	6.6	−0.16	22.2	5.6	−0.15	21.1	6.2	−0.16	21.6	5.2	−0.14
	Census	5,617	18.9	6.8		21.3	5.7		20.1	6.2		20.8	5.5	
Black	Pre-census	14,352	15.1	4.7	−0.20	16.7	3.1	−0.15	16.6	4.4	−0.17	17.0	3.8	−0.24
	Census	20,724	14.2	4.8		16.2	3.0		15.9	4.4		16.0	4.1	
Hispanic	Pre-census	7,560	18.3	5.4	−0.56	19.5	4.3	−0.52	19.8	5.5	−0.49	19.8	4.4	−0.51
	Census	20,818	15.2	5.5		17.4	4.0		17.2	5.3		17.4	4.6	
White	Pre-census	92,065	21.3	5.5	−0.27	21.9	4.8	−0.26	22.5	5.6	−0.29	22.2	4.5	−0.28
	Census	128,490	19.8	6.1		20.6	5.0		20.8	5.9		20.8	5.1	

When racial/ethnic groups were examined individually, the largest score changes occurred for the Hispanic and White student groups (Table 3). Score changes mainly reflected two factors: (1) differences in test scores between students added by census testing and those who would have tested anyway and (2) the percentage increase of students in a racial/ethnic group. Larger score differences and larger sample increases led to greater changes in ACT score distributions. Regarding the first point, the distribution of scores for added students was roughly approximated by subtracting the census testing score frequencies from the pre-census frequencies. From those distributions, it was estimated that the average ACT Composite score was 3.3 points lower for added Asian students (range of –2.5 to –6.0 across states), 2.5 points lower for added Black students (range of –2.5 to –4.3), 4.0 points lower for added Hispanic students (range of –1.6 to –4.6), and 5.2 points lower for added White students (range of –2.6 to –6.2).

Figure 3 illustrates the relationship between percentage increase in sample size and observed score changes (with the five states represented by different colors). The patterns were similar for English, math, reading, and science, so the results are shown in terms of ACT Composite scores. Sample size increases ranged from 3% to 107% in four states, but increases were much greater (177% to 422%) in the state with few ACT examinees before census testing. As expected, score changes were greater in magnitude when more students were added. For example, the score change was approximately –1 point for a 50% sample increase and –2 points for a 100% sample increase.

Figure 3. Scatterplot of Mean ACT Composite Change Versus Percentage Sample Size Increase

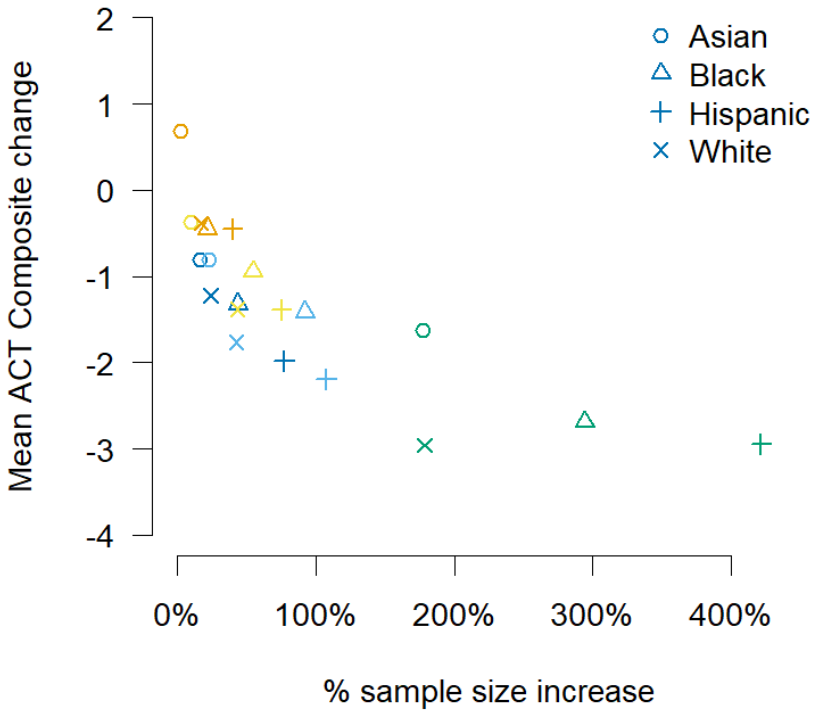


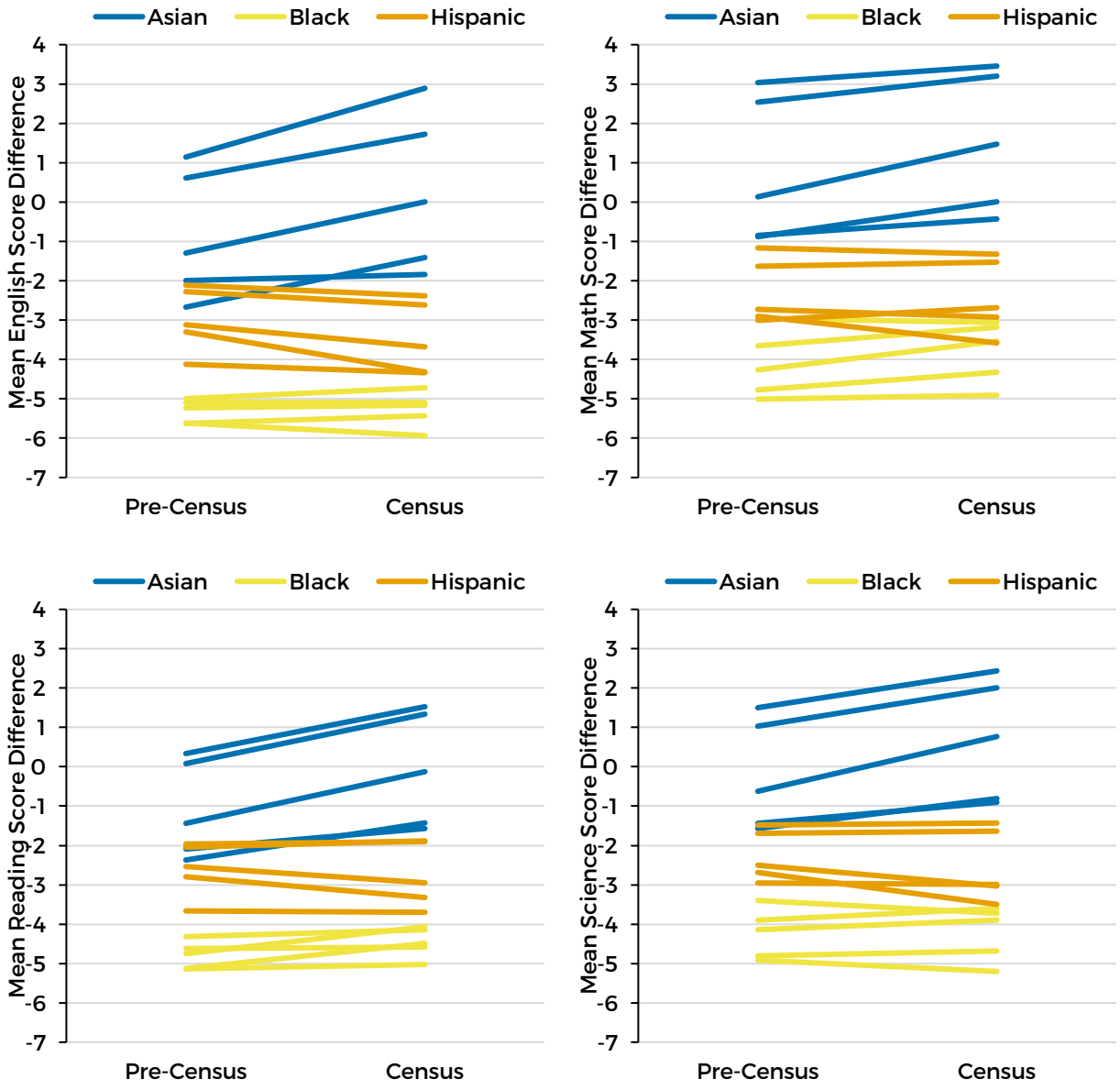
Table 4 shows mean ACT score differences between minority and White students for pre-census and census testing for all five states combined. Note that changes in means and mean differences are not necessarily problematic; they likely reflect true changes due to sample differences (self-selected, college-bound students vs. all students in public high schools). Asian-White mean score differences were smaller in magnitude for the census testing sample except on the math test, where the difference increased from 0.3 to 0.8. Mean score differences between Hispanic and White students were greater in magnitude for census testing. That is, Hispanic student means declined more than White student means after the introduction of census testing. In contrast, mean score differences between Black and White students were smaller in magnitude for census testing, but this result was explained in part by the reduced proportion of Black students in the overall sample. As described below, trends differed somewhat when the five states were examined individually.

Table 4. Racial/Ethnic Mean Score Differences and Standardized Differences (*d*) by Testing Context

		English		Math		Reading		Science	
Comparison	Context	Diff.	*d*	Diff.	*d*	Diff.	*d*	Diff.	*d*
Asian-White	Pre-census	−1.4	−0.25	0.3	0.07	−1.4	−0.25	−0.6	−0.14
	Census	−0.8	−0.14	0.8	0.15	−0.7	−0.12	0.0	0.00
Black-White	Pre-census	−6.2	−1.15	−5.2	−1.14	−5.8	−1.08	−5.2	−1.18
	Census	−5.6	−0.94	−4.4	−0.92	−4.9	−0.86	−4.8	−0.97
Hispanic-White	Pre-census	−3.1	−0.56	−2.4	−0.51	−2.6	−0.47	−2.4	−0.54
	Census	−4.6	−0.76	−3.2	−0.66	−3.6	−0.62	−3.4	−0.68

Figure 4 shows mean score differences between non-White and White students before and after the introduction of census testing disaggregated by state. Note that each racial/ethnic group has five lines per plot (to represent the five states). On the *y*-axis, positive differences indicate higher mean scores for non-White students; negative differences indicate higher mean scores for White students. Across states, the Black-White and Hispanic-White mean score differences were consistently negative regardless of the testing context. That is, White students had higher average scores than Black and Hispanic students before and after the introduction of census testing. However, as indicated by the positive and negative slopes in Figure 4, the changes did not always occur in the same direction. The mean score difference between Hispanic and White students most often increased in magnitude (i.e., became more negative), but there were some negligible changes (e.g., 0.06 in reading and 0.05 in science) and some small decreases in magnitude (0.32 in math and 0.16 in reading). The changes in score differences between Black and White students had greater variability. Those changes ranged from −0.32 to 0.28 in English, −0.10 to 0.73 in math, 0.04 to 0.69 in reading, and −0.33 to 0.30 in science. In contrast, the direction of Asian-White mean score difference changes was consistently positive (i.e., in the direction of Asian students performing relatively well compared to White students), and those changes caused some Asian-White mean score differences to switch from favoring White students to favoring Asian students. As explained above, the changes in each state must be interpreted with respect to the proportional increases in each group and the achievement levels of the added students.

Figure 4. Mean ACT Score Differences (Non-White Minus White) by State and Testing Context



Motivation Indicators

Students' item response and item score patterns were analyzed to identify possibly unmotivated responding using omit rate, C_z, and H^T. It is first important to acknowledge that the motivation indices correlated with test scores. As shown in Table 5, the person-fit statistic H^T correlated the most with test scores (.35 for English, .57 for math, .41 for reading, and .38 for science), followed by percentage of omitted items (–.31, –.17, –.23, and –.19) and C_z (–.17, –.14, –.17, and –.12). That is, students with lower test scores tended to have lower H^T (i.e., worse person fit), higher percentages of omitted items, and higher C_z. To some extent, this result could have reflected lower motivation among lower-achieving students, but it also reflected the methods used

to detect low motivation. That is, higher-achieving students would necessarily have good person fit, omit few items, and have short repeating response patterns. Regardless of motivation, lower-achieving students would be more likely to leave items blank, enter long strings of repeating responses, and guess randomly.

Table 5. Correlation Between ACT Scale Scores and Motivation Indices by Testing Context

Test	Context	Omit rate	C_z	H^T
English	All	−.31	−.17	.35
	Pre-census	−.27	−.17	.27
	Census	−.32	−.16	.37
Math	All	−.17	−.14	.57
	Pre-census	−.15	−.14	.54
	Census	−.18	−.13	.59
Reading	All	−.23	−.17	.41
	Pre-census	−.21	−.14	.29
	Census	−.24	−.19	.49
Science	All	−.19	−.12	.38
	Pre-census	−.17	−.15	.29
	Census	−.20	−.09	.41

In Table 6, a comparison of ACT testers from before and after the introduction of census testing reveals that the percentages of students who omitted more than 20% of items increased in all subject areas, with a range of 0.8 (science) to 2.5 (English) percentage points. Likewise, the percentages of students flagged for unusually long repeating response patterns based on C_z increased in all subject areas, with a range of 0.8 (math) to 2.6 (English) percentage points. The Type-I error rate for the person-fit statistic H^T was approximately 5%, and most H^T flagging rates were close to that. The percentage of students flagged for low H^T increased by 0.9 percentage points in English, decreased by 0.9 percentage points in math, and changed little in reading and science. Across motivation indices, the largest increases were observed on the English test. There was generally low agreement among the three indices, which indicated that they tended to flag different students. For example, only 6.2% of students flagged for H^T were also flagged for C_z, and this was the highest agreement rate.

Table 6. Percentages of Examinees Flagged for Apparent Low Motivation by Testing Context and Racial/Ethnic Group

		English			Math			Reading			Science		
Race/ethnicity	Context	Omit	C_z	H^T	Omit	C_z	H^T	Omit	C_z	H^T	Omit	C_z	H^T
All	Pre-census	3.1%	5.2%	5.8%	2.8%	4.1%	5.5%	2.8%	3.8%	5.4%	1.9%	4.0%	4.8%
	Census	5.6%	7.8%	6.7%	3.7%	4.9%	4.6%	4.4%	5.0%	5.7%	2.7%	5.7%	4.8%
Asian	Pre-census	4.6%	4.1%	8.5%	4.2%	2.7%	5.7%	4.4%	3.2%	4.7%	3.0%	3.3%	5.0%
	Census	6.3%	6.7%	7.5%	4.4%	3.4%	5.3%	5.4%	4.1%	4.7%	3.7%	4.7%	5.0%
Black	Pre-census	10.8%	10.0%	7.9%	7.0%	7.1%	5.8%	7.3%	7.4%	6.1%	4.6%	7.0%	5.5%
	Census	11.7%	14.0%	8.7%	6.1%	8.0%	4.6%	7.6%	8.6%	5.5%	4.4%	9.2%	5.6%
Hispanic	Pre-census	8.2%	5.9%	5.3%	7.7%	4.7%	5.1%	7.7%	4.0%	4.8%	5.5%	4.1%	4.9%
	Census	16.2%	7.6%	6.3%	11.1%	4.5%	4.2%	13.0%	4.7%	5.2%	8.1%	5.4%	5.2%
White	Pre-census	1.4%	4.5%	5.4%	1.5%	3.6%	5.4%	1.5%	3.3%	5.3%	1.0%	3.6%	4.6%
	Census	2.4%	6.9%	6.3%	1.6%	4.5%	4.6%	1.6%	4.5%	4.6%	1.2%	5.1%	4.5%

Table 6 also shows percentages of flagged students disaggregated by racial/ethnic groups. Considering the correlation between test scores and the motivation indices, some differences in motivation flagging rates were expected due to differences in achievement between racial/ethnic groups. Across subjects, White students were least likely to be flagged for omitting items; Black and Hispanic students were most likely. When pre-census and census testing were compared, the percentage point increase in omit rate flagging was greatest for Hispanic students, and this was likely related to the addition of many Hispanic students in one state. For long strings of repetitive responses (C_z) across subjects, White and Asian students were similarly likely to be flagged, and Black students were most likely to be flagged. Most H^T flagging rates were near the expected 5% rate; only on the English test were the flagging rates notably higher for Asian and Black students.

Expected Changes in Motivation Indices

There were observed differences between student groups in test score distributions (Table 3) and in percentages flagged for apparent low motivation (Table 6). Unfortunately, considering the correlation between test scores and the motivation indices (Table 5), differences in apparent motivation were confounded with differences in achievement. For example, more census testers than pre-census testers were flagged for apparently low motivation, but part of that difference reflected the lower average achievement of census testers, not true differences in motivation. Disentangling test scores and motivation indices would require an independent measure of achievement, ideally administered under the same motivational conditions for all students. That is, it would be ideal to compare groups on the motivation indices while controlling for achievement. This would help address the question of whether observed differences in apparent motivation were related to the introduction of census testing or simply expected due to achievement differences.

Figure 5 illustrates this challenge to interpreting the motivation indices. In Figure 5, the plotted points represent the percentages of students flagged for omitting items versus mean ACT scores for racial/ethnic groups in the five states. Linear regression lines were fit to the data for pre-census and census testers separately (with weighting for sample size). Note that lower mean ACT scores were associated with higher flagging rates for pre-census testers (plotted in blue). That is, even for students presumably motivated to perform well on the test, some students still left at least 20% of the items blank, and the expected percentage increased as average test scores decreased. The same trend is apparent in Figure 6 for C_z, but not in Figure 7 for H^T due to the method of setting flagging cutoffs (i.e., different cutoffs for different raw scores).

In this analysis, the relationship between mean test scores and flagging rates for pre-census testers (presumably college-bound, motivated students) was treated as the

criterion. If the relationship was similar for census testers, this would suggest that observed decreases in the motivation indices were consistent with expectations and not, therefore, indicative of unusual decreases in motivation caused by the introduction of census testing. For item omitting (Figure 5), the regression lines for census testers had steeper (negative) slopes than for pre-census testers, and the regression lines crossed near a mean ACT score of 18. This result could suggest unusual increases in omitting items among lower-achieving groups of students, but there were few data points in low mean score ranges (i.e., below 16) except on the English test, which had two such data points. In the range of most observed ACT means (approximately 16–23), the regression lines were either very close or the census testing line was slightly below the pre-census testing line. These results were driven largely by decreases in mean ACT scores that were *not* accompanied by notable increases in the percentages of students flagged for omitting items. Consider, for the example, the circled points in Figure 5. From pre-census (blue) to census (orange) testing, there was a clear decrease in mean ACT scores (points shifting to the left), but there was not a noticeable increase in the percentages of students flagged for omitting items (no shift upward). Such results suggest greater than expected motivation for census testers conditional on average test scores, but this was not replicated for C_z and H^T. Pre-census and census regression results were similar for C_z (Figure 6), and they were practically identical for H^T (Figure 7). Thus, general similarity between pre-census and census testing is the strongest conclusion that should be drawn.

Despite similar trends, there were outlying data points. For example, the Black and Hispanic student groups (from a certain state) were very likely to be flagged for omitting items (Figure 5), though the increases in omit rate flagging between pre-census and census testing appeared commensurate with decreases in mean ACT scores (based on the slope of the blue, pre-census regression line). The C_z plot for English also includes several points with relatively high flagging rates for several racial/ethnic groups (Figure 6). Note that these results may also indicate that observed differences in the motivation indices between racial/ethnic groups (Table 6) were largely reflections of achievement differences, not true differences in motivation.

Figure 5. Relationship Between Omit Rate Flagging Percentage and Mean ACT Score by Testing Context

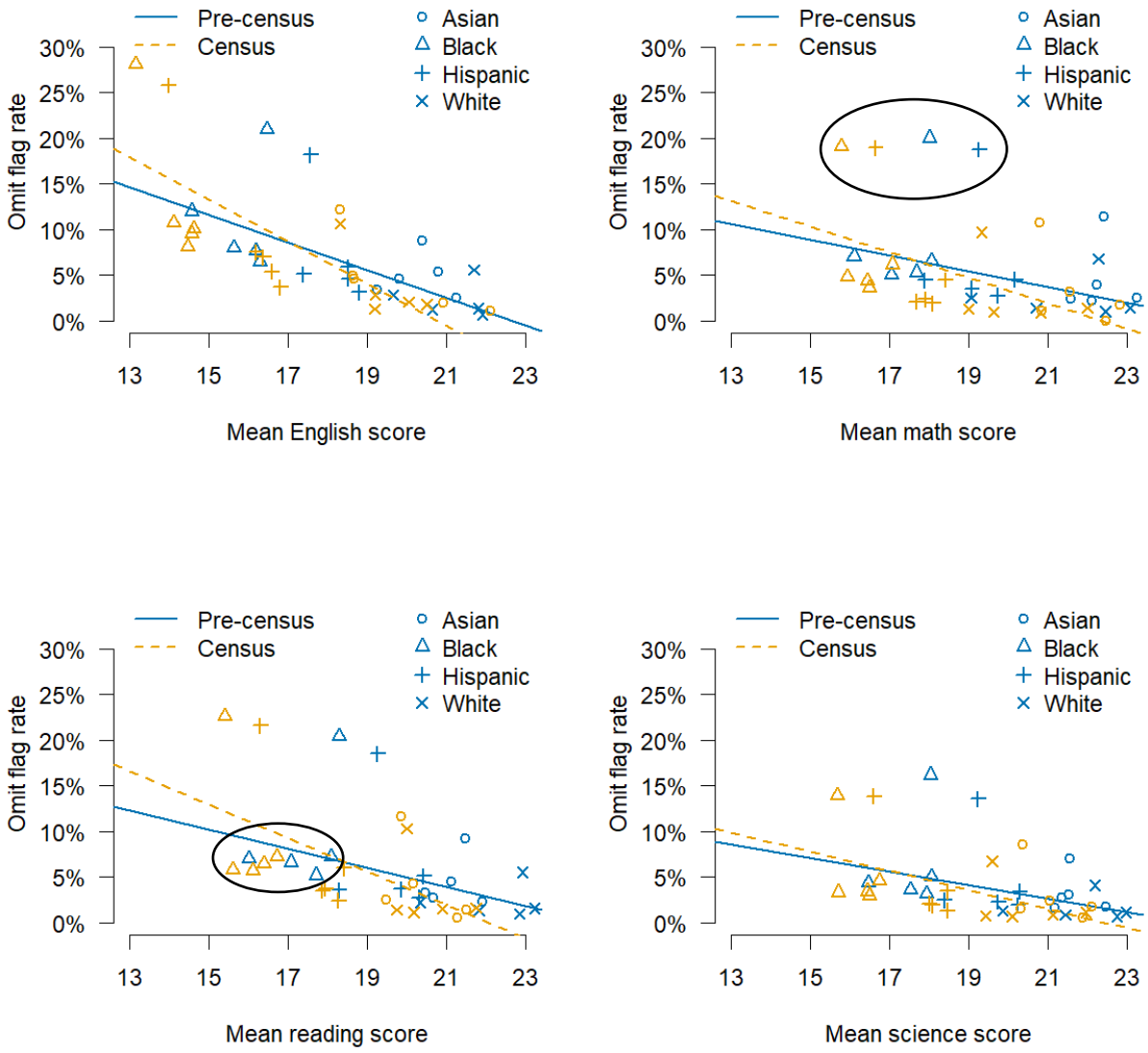


Figure 6. Relationship Between C_z Flagging Percentage and Mean ACT Score by Testing Context

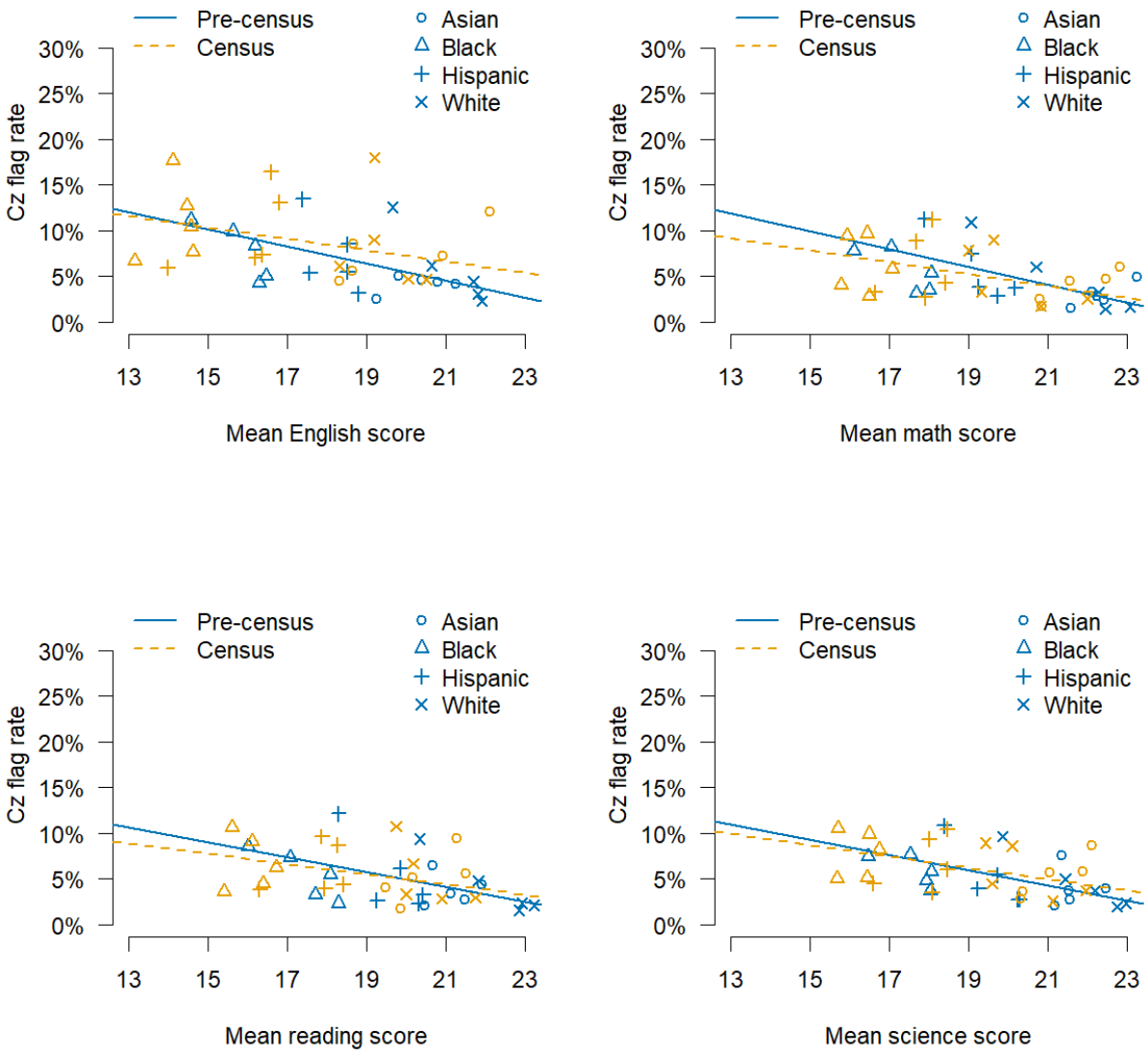
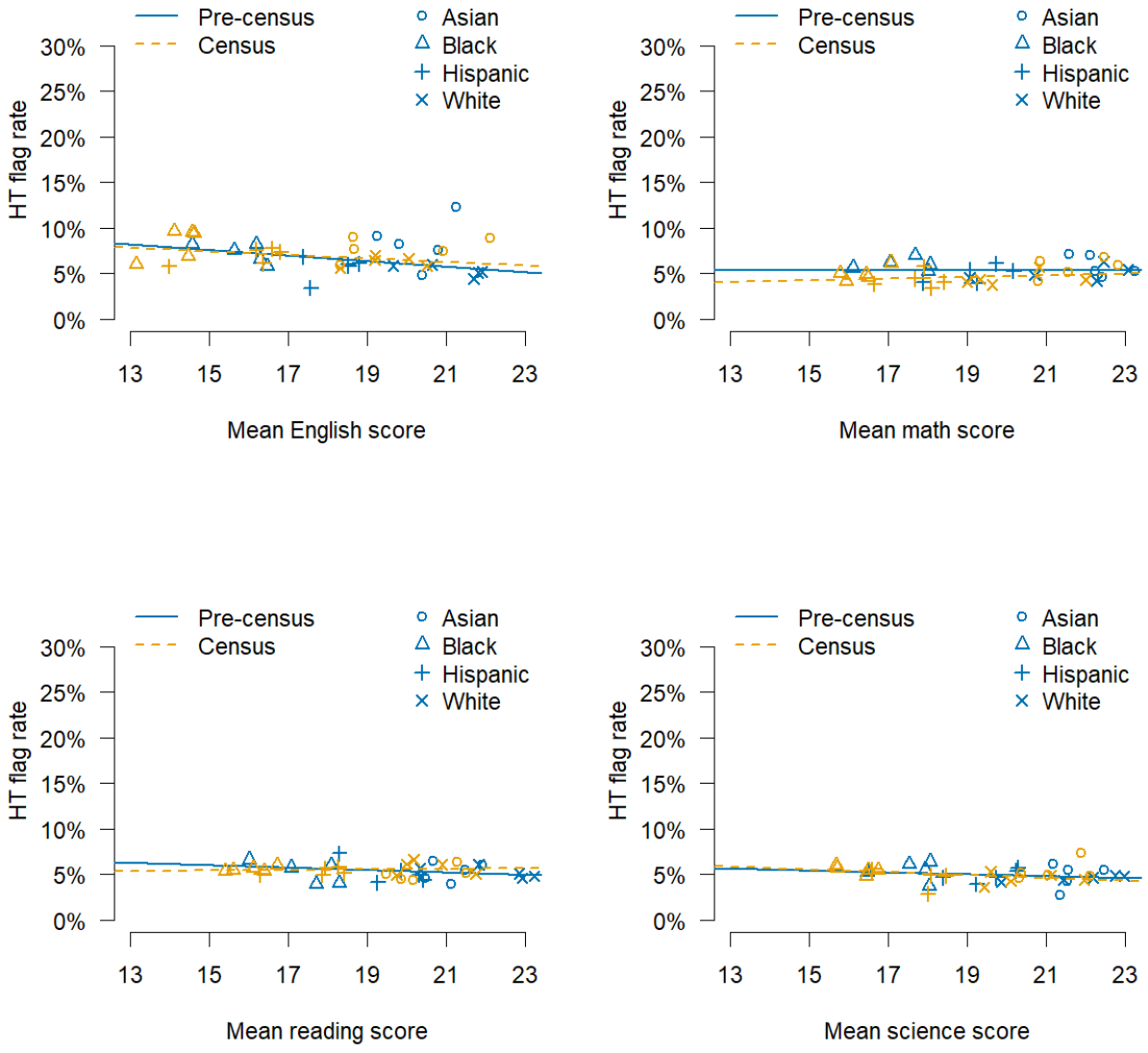


Figure 7. Relationship Between H^T Flagging Percentage and Mean ACT Score by Testing Context

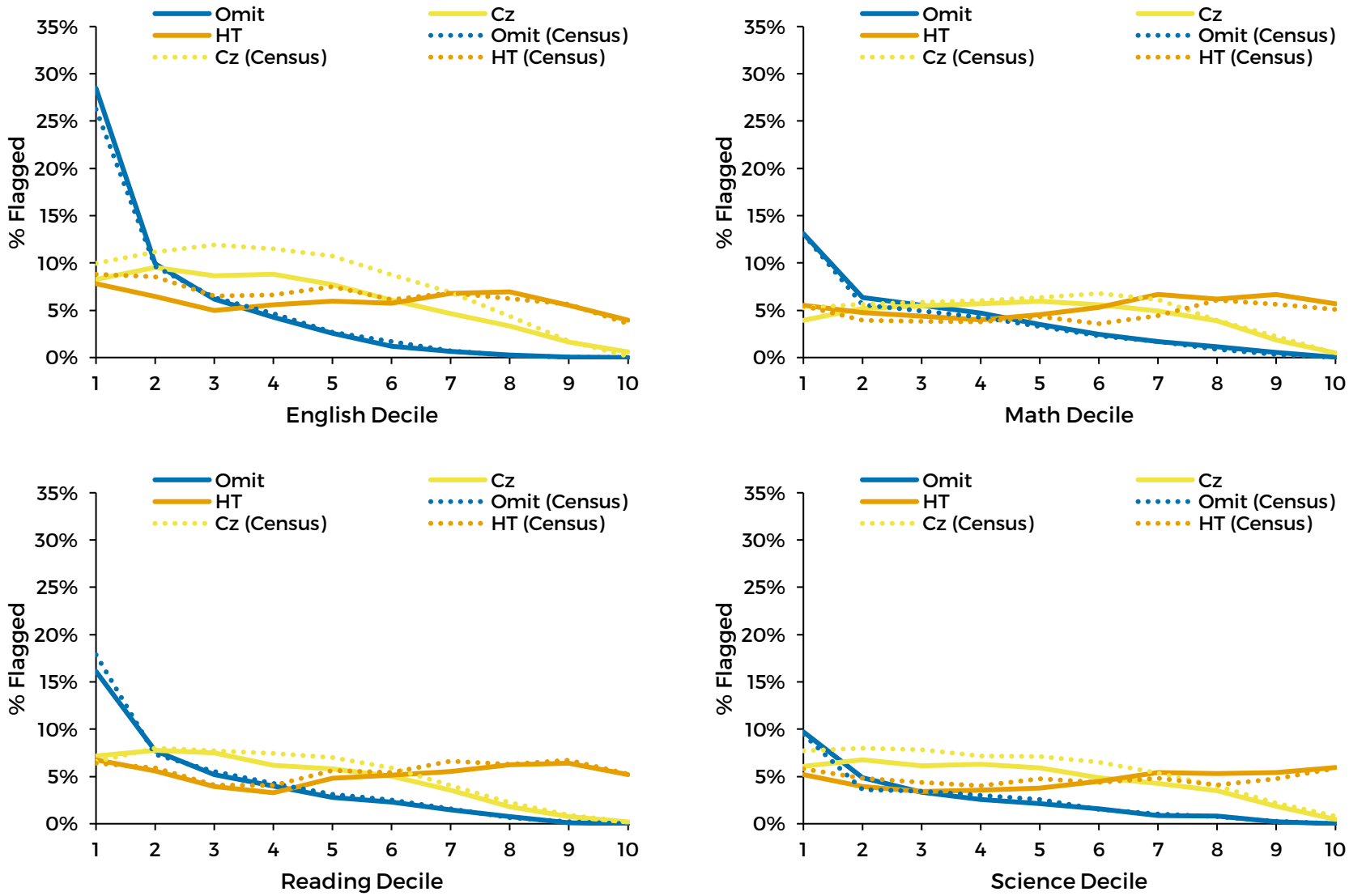


To provide additional evidence concerning motivation among pre-census and census testers, Figure 8 shows plots of motivation flagging rate by ACT score decile (with the same deciles applied to pre-census and census data).[1] As shown in Figure 8, the omit rate and C_z flagging percentages tended toward 0% as ACT score increased. Despite the relatively high correlation between H^T and ACT scores (Table 5), however, the H^T flagging rate remained around 5% throughout the ACT score range due to the flagging method. Between the pre-census and census samples, the flagging rates were sometimes higher for the census testers, though differences were generally

[1] A similar, parametric analysis might have been conducted using logistic regression, but the relationship between motivation flagging probability and ACT score did not appear to be monotonically decreasing for C_z or H^T.

small across the ACT score deciles. Only the C_z flagging rates on the English test ever differed by more than two percentage points. Across subjects and deciles, the average difference in omit rate flagging was 0.0 percentage points, the average difference in C_z flagging was 1.0 percentage points, and the average difference in H^T flagging was 0.1 percentage points. Overall, once differences in achievement were roughly accounted for, the results were consistent with the notion that pre-census and census testers were similarly motivated.

Figure 8. Relationship Between Omit Rate Flagging Percentage and Mean ACT Score by Testing Context



NAEP Comparison

Compared with examinee motivation in high-stakes testing contexts, motivation tends to have greater variance on low-stakes tests (Sundre, 1999). Several studies suggest that higher-ability examinees are more likely than lower-ability examinees to maintain motivation on low-stakes tests, even when test content is cognitively challenging (Barry, Horst, Finney, Brown, & Kopp, 2010; Wise, Pastor, & Kong, 2009). That is, there could be *differential motivation* when lower- and higher-ability examinee groups are compared. If this is true, the average difference in test scores between lower- and higher-ability examinees would be inflated. That is, lower-ability examinee groups would perform worse than if they were motivated, and higher-ability examinee groups would perform better than if they were unmotivated. This applies to the comparison of any groups with different ability distributions, including racial/ethnic groups. Therefore, mean test score differences between racial/ethnic groups ("achievement gaps") could be inflated in low-stakes testing contexts. If this is true, mean score differences should be smaller (and possibly more accurate) in higher-stakes testing contexts.

Proponents of census testing claim that motivation should be higher on college admissions exams compared with typical state achievement tests because there are stakes attached to performance (Camara et al., 2019). This assertion applies mainly to lower-achieving students who probably would not have taken the ACT if not for census testing. Such students are apparently affected by census testing, considering that statewide adoption of college admissions exams is associated with higher rates of college enrollment (ACT, 2015; Hyman, 2017; Klasik, 2013). Thus, lower-achieving students might perceive the importance of doing well and exhibit greater motivation than they would on a low-stakes test.

These ideas were tested by comparing mean differences between racial/ethnic groups on the ACT to mean differences on the National Assessment of Educational Progress (NAEP)—a low-stakes assessment administered to a representative sample of schools across the United States. Table 7 compares standardized mean score differences (effect sizes) between racial/ethnic groups by state for ACT census testing and NAEP. The comparison was not ideal because of differences in test content, testing populations, and education levels, but it still provided a sense of how the ACT mean score differences in census testing might compare to those of a low-stakes assessment in the same content area in the same state with a representative sample. Note that NAEP results for Asian students were not reported in some states due to small sample sizes.

Table 7. ACT and NAEP Racial/Ethnic Standardized Mean Score Differences for Five States

		Standardized mean score difference (*d*)					Difference in differences				
Comparison	Test	State 1	State 2	State 3	State 4	State 5	State 1	State 2	State 3	State 4	State 5
Asian-White	Math (ACT census 2015)	−0.08	0.68	0.82	0.30	0.00					
	Math (NAEP grade 8 2019)	−0.05			0.36	−0.08	0.03			0.06	−0.08
	Reading (ACT census 2015)	−0.27	0.23	0.28	−0.02	−0.24					
	Reading (NAEP grade 8 2019)	−0.08			0.08	0.08	0.19			0.10	0.32
	Science (ACT census 2015)	−0.18	0.41	0.51	0.15	−0.16					
	Science (NAEP grade 8 2015)	−0.73			−0.03	−0.13	−0.55			−0.18	0.03
Black-White	Math (ACT census 2015)	−0.97	−0.70	−0.91	−0.78	−0.87					
	Math (NAEP grade 8 2019)	−1.13	−0.76	−0.83	−0.82	−1.21	−0.16	−0.06	0.08	−0.04	−0.34
	Reading (ACT census 2015)	−0.87	−0.72	−0.88	−0.78	−0.77					
	Reading (NAEP grade 8 2019)	−0.95	−0.75	−0.67	−0.68	−1.05	−0.08	−0.03	0.21	0.09	−0.29
	Science (ACT census 2015)	−1.05	−0.75	−0.86	−0.78	−0.91					
	Science (NAEP grade 8 2015)	−1.30	−1.03	−1.03	−1.03	−1.44	−0.25	−0.28	−0.17	−0.25	−0.53
Hispanic-White	Math (ACT census 2015)	−0.70	−0.33	−0.32	−0.64	−0.59					
	Math (NAEP grade 8 2019)	−0.80	−0.49	−0.31	−0.51	−0.62	−0.10	−0.16	0.02	0.13	−0.03
	Reading (ACT census 2015)	−0.57	−0.33	−0.34	−0.66	−0.51					
	Reading (NAEP grade 8 2019)	−0.65	−0.25	−0.25	−0.47	−0.62	−0.08	0.08	0.09	0.19	−0.12
	Science (ACT census 2015)	−0.71	−0.34	−0.30	−0.63	−0.59					
	Science (NAEP grade 8 2015)	−0.97	−0.72	−0.49	−0.78	−0.84	−0.26	−0.38	−0.19	−0.15	−0.25

The mean score differences were generally similar in magnitude, though the average difference in differences was –0.14 standard deviations for the Black-White comparisons and –0.08 standard deviations for the Hispanic-White comparisons. The negative sign indicates that achievement differences were, on average, slightly smaller for the ACT than for NAEP. When achievement differences were averaged across states, there were several notable negative differences: Black-White in math (–0.10), Black-White in science (–0.30), and Hispanic-White in science (–0.25). Others were close to zero on average. Thus, the overall results for the Black-White and Hispanic-White comparisons were consistent with the hypothesis that mean score differences on the ACT would be smaller than on a low-stakes test, though the results did not always follow that trend (e.g., States 3 and 4). The results for the Asian-White comparisons in three states were inconsistent. In math, the magnitudes of Asian-White differences were generally small on both tests. In reading, the ACT differences were sometimes greater in magnitude than NAEP differences and indicated relatively high scores for White students. In science, the results were inconsistent, with one difference in differences being notably smaller for the ACT than for NAEP (State 1).

Conclusions

This study first addressed the question of how college admissions test score distributions change when states adopt census testing of all 11th graders in public high schools. That transition causes many students—often lower achieving—to take a college admissions test who would not have taken one otherwise. Thus, scores were expected to decline, and they did by about one-third of a standard deviation on the ACT English, math, reading, and science tests. The declines reflected a combination of the number of students added by census testing and the lower average achievement of the added students. Declines occurred to different degrees for different racial/ethnic groups; they were greatest for Hispanic and White students, then for Asian and Black students. Different declines caused changes in the mean score differences between racial/ethnic groups on the ACT. Specifically, mean differences between Hispanic and White students tended to increase, but there were increases and decreases in the mean differences between Black and White students, and Asian students made gains on White students. Changes in mean differences, whether positive or negative, were not necessarily problematic. In large part, changes simply reflected differences in the samples tested. For example, the mean ACT score difference between Black and White students may truly be different for self-selected, college-bound students than for all students in public high schools.

In the next set of analyses, indices of testing motivation were compared for tests administered before and after the introduction of census testing. The percentages of students flagged for high item omit rates, long repeating response strings, and poor person fit all increased with the introduction of census testing, though the effects were small. The percentage point increases were greatest on the 75-item, 45-minute

English test (2.5 for omit rate, 2.6 for C_z, 0.9 for H^T). Differences were smaller on the 60-item math test, even though it was 15 minutes longer, so test duration may not have been the driving factor (perhaps speededness was). Though not a formal indicator of testing motivation, the percentage of students scoring at or below chance level increased with census testing, but not to the extreme degree suggested by critics. This increase likely reflected the combined effects of unmotivated testing behavior and a greater proportion of students whose true scores fell around chance level.

Decreases in apparent motivation were expected after the transition to census testing due to the correlation between test scores and motivation, which was at least partly caused by the methods used to identify low motivation. At issue was whether the introduction of census testing caused an unexpectedly large decrease in the motivation indices. When regression analyses controlled for mean ACT score, results indicated that motivation in the census testing sample was generally aligned with expectations (or even greater than expected). A subsequent descriptive analysis revealed that flagging rates were similar for pre-census and census testers when roughly controlling for achievement. Overall, there appeared to be no greater-than-expected decreases in motivation associated with the introduction of census testing.

The final set of analyses compared mean score differences between racial/ethnic groups on the ACT and NAEP to see whether differences were smaller on the ACT, possibly because of greater motivation (and subsequent effort) among lower-achieving students. Consistent with the hypothesis, achievement gaps on the ACT were, on average, slightly smaller than those observed on NAEP in the same states and content areas with representative samples. Thus, if differential motivation caused inflated mean differences on NAEP, the results possibly indicate that lower-achieving student groups were more motivated on the ACT than they would have been on a low-stakes test.

The motivation indices used in this study had several notable limitations. Like many indicators of testing motivation, omit rate, C_z, and H^T were all correlated with achievement, and this was partly due to calculation methods, which made it nearly impossible for high-achieving students to be flagged for low motivation. This interfered with the interpretation of declines in average test scores and apparent motivation, which precipitated the subsequent regression analyses to attempt controlling for achievement. Perhaps the H^T results were most dependable due to the flagging method, which resulted in apparent statistical independence between test scores and flagging. Another methodological limitation was that students could have been flagged for reasons other than unmotivated responding. For example, omitting items and repetitive responding can occur due to low ability or speededness. Poor person fit can also arise due to factors other than low motivation (Meijer, 1996).

In conclusion, achievement and apparent motivation both decreased with the introduction of census testing. This was expected because of the influx of testers with lower academic achievement and the correlation between test scores and the motivation indices. Differences between pre-census and census testers on the motivation indices were consistent with expectations based on decreases in achievement. Overall, the census testing sample appeared to behave like a motivated, college-bound sample but with lower average achievement. Moreover, mean ACT score differences between racial/ethnic groups were slightly smaller on average than those on a low-stakes assessment, which could indicate that lower-achieving student groups exhibited greater motivation on the ACT than they would have on a low-stakes assessment. In other words, use of a college admissions exam for census testing appears to encourage motivated testing behavior, even among students who would not have taken the exam otherwise. When examinees are motivated, their test scores are more likely to accurately reflect their knowledge and skills, and this is an improvement over typical low-stakes achievement tests, where student motivation may be low.

References

ACT. (2015). *Expanding opportunities: A college choice report for the graduating class of 2014. Part 2: Enrollment patterns*. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/CollegeChoiceReport-2014-Part2.pdf

ACT. (2020). *ACT® technical manual*. Retrieved from http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Retrieved from American Educational Research Association website: https://www.apa.org/science/programs/testing/standards

Allen, J. M. (2015). *Anticipated changes in ACT scores and participation rates with ACT statewide adoption* (Issue Brief No. R3619). Retrieved from ACT website: https://www.act.org/content/dam/act/unsecured/documents/Statewide-Adoption.pdf

Allen, J. M., & Mattern, K. (2019). *Validity considerations for 10th-grade ACT state and district testing* (Insights Report No. R1758). Retrieved from ACT website:

https://www.act.org/content/dam/act/unsecured/documents/R1758-act-grade10-validity-2019-06.pdf

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342–363. doi:10.1080/15305058.2010.508569

Camara, W. J., Mattern, K., Croft, M., Vispoel, S., & Nichols, P. (2019). A validity argument in support of the use of college admissions test scores for federal accountability. *Educational Measurement: Issues and Practice, 38*(4), 12–26. doi:10.1111/emip.12293

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education, 57*(2), 119–130. doi:10.1353/jge.0.0018

Cui, Z. (2020). On a new algorithm for removing repeating patterns in similarity analysis. *Educational and Psychological Measurement, 80*(3), 446–460. doi:10.1177/0013164419882980

Hernández, M., & Hershaff, J. (2015). *Skipping questions in school exams: The role of non-cognitive skills on educational outcomes.* Retrieved from University of Michigan Education Policy Initiative website: https://edpolicy.umich.edu/sites/epi/files/uploads/wp-hernandez-hershaff-skipping-questions-dec-2015.pdf

Hyman, J. (2017). ACT for all: The effect of mandatory college entrance exams on postsecondary attainment and choice. *Education Finance and Policy, 12*(3), 281–311. doi:10.1162/EDFP_a_00206

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298. doi:10.1207/S15324818AME1604_2

Klasik, D. (2013). The ACT of enrollment: The college enrollment effects of state-required college entrance exam testing. *Educational Researcher, 42*(3), 151–160. doi:10.3102/0013189X12474065

Meijer, R. R. (1996). The influence of the presence of deviant item score patterns on the power of a person-fit statistic. *Applied Psychological Measurement, 20*(2), 141–154. doi:10.1177/014662169602000204

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data?: A comparison of the reliability of data produced in graded and un-graded conditions. *Research in Higher Education, 45*(8), 921–929. doi:10.1007/s11162-004-5954-y

National Center for Education Statistics. (2016). *NAEP science 2015 state snapshot reports*. Retrieved from National Center for Education Statistics website: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016157

National Center for Education Statistics. (2019a). *The nation's report card: Mathematics 2019 state snapshot reports*. Retrieved from National Center for Education Statistics website: https://nces.ed.gov/nationsreportcard/pubs/stt2019/2020013.aspx

National Center for Education Statistics. (2019b). *The nation's report card: Reading 2019 state snapshot reports.* Retrieved from National Center for Education Statistics website: https://nces.ed.gov/nationsreportcard/pubs/stt2019/2020014.aspx

Sijtsma, K., & Meijer, R. R. (2016). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*(2), 149–157. doi:10.1177/014662169201600204

Steedle, J. T., & Grochowalski, J. (2017). The effect of stakes on accountability test scores and pass rates. *Educational Assessment, 22*(2), 111–123. doi:10.1080/10627197.2017.1309276

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social–emotional learning competencies. *Educational Measurement: Issues and Practice, 38*(2), 101–111. doi:10.1111/emip.12256

Sundre, D. L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the

Annual Meeting of the American Educational Research Association, Montréal, Canada.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. doi:10.1006/ceps.1999.1015

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A., & Kong, X. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185–205. doi:10.1080/08957340902754650

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227–242. doi:10.1207/s15324818ame0803_3