Uncertainty in Accountability: How High-Stakes Test Scores Compare to Grades When Predicting College Achievement

Jason Giersch

UNC Charlotte

Political Science and Public Administration

jgiersch@uncc.edu

Abstract

One of the criticisms of accountability policies is that high-stakes tests distort instruction, causing those tests to lose validity, especially for students at risk of missing benchmarks. Using a longitudinal dataset from North Carolina, I compare the usefulness of high-stakes test scores in predicting college grades in honors track classes to standard track classes to test whether test scores lose predictive validity in the lower-tracks, where teaching to the test is more likely to occur. According to the results, while test scores for standard track students are indeed less useful predictors of later academic achievement, high school GPA remains just as strong a predictor for standard as for honors classes. Interviews with teachers suggest that this pattern results from the limitations of the exams in contrast to the decisions by teachers to shape instruction and grading to the needs and characteristics of their students.

Keywords

College achievement, high-stakes testing, accountability, social justice, school organization.

Acknowledgement

I am grateful to the Roots of STEM research team for allowing me access to their data for this project, specifically Elizabeth Stearns, Stephanie Moller, Melissa Dancy, and Roz Mickelson.
Uncertainty in Accountability: How High-Stakes Test Scores Compare to Grades When Predicting College Achievement

Few policies receive such consistent criticism from researchers yet enjoy nearly universal application as academic tracking. Decades of research have shown that separating students into homogenous groups has more negative than positive effects on academic outcomes of students (Oakes, 2005 & 1985) yet the practice remains firmly engrained in schools of all types (Rubin, 2006; Welner, 2001). Although recent school reforms have attempted to “leave no child behind,” it has been exceedingly rare for schools to dismantle their tracking policies. Instead, recent reforms have taken an accountability approach, whereby states assess student achievement on a regular basis using scores on standardized tests, announce the results, and reward and punish schools accordingly (Hanushek & Raymond, 2004).

Implicit in this strategy of school reform is the issue of consistency across classrooms, instructors, and schools. Using state-wide academic standards and uniform exams, states label stronger and weaker performers at the student, teacher, and school levels and provide incentives to raise test scores, especially for the lowest-achieving students, according to the rationale for such policies (Reback, 2008). But the push for consistency in achievement has not carried over to consistency in instruction. Schools, overall, continue to segregate students by academic track, even though all students in the same course have to meet the same standards and take the same test within each state.

While testing blossomed and academic tracking remained a fixture in high schools, higher education has continued to be a popular goal for students. Since the early 1980s, a
majority of the nation’s graduates have enrolled in college soon after high school; for more than a decade that rate has hovered between 65 and 70% (Norris, 2014). Preparation for higher education, once the domain of the academic and economic elite, has been a goal of public education more broadly for generations. Yet through academic tracking, high schools continue to segregate their students and deliver different sets of instructional opportunities as though substantial portions of their student bodies will never pursue any further schooling. At the end of their high school experiences, the gulf between students who pursued upper-track classes and those in the lower-track classes is wider than when they entered (Klugman, 2012; Mickelson, Giersch, Stearns, & Moller, 2013).

This study utilizes student scores on standardized tests to investigate the difference in college achievement that is associated with high school academic track. Because accountability policies hold students in both upper and lower tracks to the same academic standards, these policies provide an opportunity to assess the academic consequences of taking a slate of lower-track classes before entering college. This study also serves to assess the success with which high-stakes tests reveal students’ readiness for college by track.

**Research on Academic Tracking**

An important issue in educational equity that has gone unaddressed by both No Child Left Behind and Race to the Top is academic tracking, even though scholars have been criticizing the practice for decades for its exacerbation of educational inequality (Werblow, Urick, & Duesbery, 2013; Lucas & Berends, 2002; Lucas, 1999; Oakes, 1985; Page, 1991; Page, 1989). Harris and Anderson (2012) argue that in high-track classes, teachers are more likely to engage students in intellectual discourse, thus providing those students with more opportunities
to practice critical thinking. Teachers in low-track classes, however, tend to reduce academic rigor and limit instructional content and higher-order thinking.

These differences in opportunities to learn influence educational outcomes, as was found in Klugman’s (2012) national study of tracking’s effects on college destinations. In it, he finds that a student’s academic track will magnify the effects of family background on students’ pursuit of higher education. The result of this pattern is the reproduction of social inequality, as a new generation of an advantaged family accumulates more advantages than the generation before it, and the next generation from a disadvantaged family falls further behind.

Of course, family background influences a student’s academic track which in turn influences achievement, as was found in Kornhaber’s (1997) study of a large metropolitan school district in the 1990s, where track placement, is set for most students while in middle school. Administrators who served as subjects in the study were frank in their descriptions of the Academically Gifted (AG) tracks in middle school as the domain of middle and upper class children whose parents knew the value of the AG track and used their resources to get their sons and daughters enrolled in it. The differences in course content and teaching methods observed by Kornhaber were stark, and by the time the children got to high school the lower-track students could not keep up with the upper-track students.

Even when controlling for family background, tracking still has remarkable effect on academic outcomes, as discovered by Werblow, Urick, and Duesbery (2013). Their study found that students were 60% more likely to drop out of school if they took lower-track courses, even when controlling for race, gender, individualized education plans (IEPs), improvement in math, and socio-economic status. Even in the absence of any evidence that student outcome improve as
a result of tracking, parents cling to the belief that academic tracks are critical to achievement and demand their offering (Rubin, 2006; Loveless, 2009).

Lower track classes harm achievement through several mechanisms, according to scholarship addressing the topic. Hallinan (2003) argues that not only do schools do a poor job of sorting students into academically homogenous groups, but they also fail to deliver quality instruction to lower-track classes. The curriculum in lower track classes tends to be lower quality (Loveless, 1999; Oaks, 2005) and the teachers tend to have less experience, fewer resources, and lower expectations for their students’ achievement (Blanchett, 2006). Students in lower tracks also tend to develop poorer academic self-concepts (Chiu et al., 2008). Finally, the experience of learning in lower track classes often brings with it exposure to students with similarly low self-concepts and less interest in schooling (Gamoran & Berends, 1987; Gilbert & Yerrick, 2001).

**Research on High-Stakes Testing**

Standardized tests are useful tools in education, but, as with any tool, their usefulness has many limitations, especially in area of summative evaluation (Hopkins, 1998). Compared to teacher grades, which typically reflect both cognitive and non-cognitive aspects of student performance, standardized tests emphasize cognitive domains. If standardized tests are by their nature limited in their ability to assess student progress, attaching high stakes to them can be educationally dangerous. Linn (2000) cautions that tests that would otherwise be dependable and credible are not so once policymakers apply high stakes to their outcomes. Two years before NCLB was passed, Heubert and Hauser (1999) argued that attaching high stakes to tests causes educators to narrow the curriculum, reduce instructional quality, dilute learning, increase dropouts, reduce graduation rates, encourage or even engage in cheating, and favor some
students over others. Jones, Jones, and Hargrove (2003) found that high stakes testing policies are associated with reduced curricula, poor teaching practices, reduced motivation, increased retention, degraded reputations of schools, more difficult teacher recruitment, and disproportionate harm to special populations. In short, for many students standardized tests interfered with their learning experiences rather than reflect them.

Berliner (2011) argues that these consequences reflect rational decisions by teachers, administrators, and students in an era of test-based accountability and are the foundation of what he calls the Uncertainty Principle. An ambitious educator acting in his or her own best interest would be reasonable to consider engaging in excessive test preparation to boost students’ scores higher than what students of similar levels of actual academic progress might earn. Of course, doing so damages the morale of both teachers and students and causes teachers to engage in educational triage in which they devote resources toward getting students on the bubble to reach proficiency benchmarks rather than to overall educational progress for all students (Nichols & Berliner, 2007). Amrein and Berliner (2002) studied 18 states to find out if the introduction of high-stakes testing had a measurable effect on ACT, SAT, NAEP, and AP test scores and found that scores declined or stayed flat more often than they increased. Even more alarming, they found that high-stakes tests prompted an increase in dropouts, most likely because the high stakes created an environment unwelcoming to struggling students.

Similarly, Haney (2000) and McNeil (2005) showed that the “Texas miracle” of rapidly increasing scores on standardized tests was made possible by the dropping out, “pushing out”, transferring, incarceration, or deportation of low-achieving students, who happened to be disproportionately Black and Hispanic. In Haney’s study, the high-stakes test scores showed remarkably weak correlations with one another and with related grades in school, suggesting that
the test scores were not accurately measuring student learning. McNeil also found correlations among various tests to be weak and that achievement on the high-stakes exam was a poor predictor of college success. Braun, Chapman, and Vezzu (2010) examined 8th grade NAEP scores in math from 1992 to 2007 across 10 different states to see if the introduction of NCLB improved scores and closed achievement gaps. They detected only very modest gains for Blacks and persistent achievement gaps.

One consequence of current accountability policies, therefore, seems to be that high stakes cause teachers and students to focus on the content of the exam more than skills and knowledge needed for a well-rounded education. Case studies by Perna and Thomas (2009) showed that state-mandated testing creates an atmosphere in high schools that reduces academic preparation, knowledge, and information and lowers graduation rates. The authors contend that such tests are not sufficiently aligned with the academic skills necessary in college and distract teachers and students from focusing on more productive skills and knowledge, particularly in settings where students are at the greatest risk of failing to meet standards of “proficiency” or “progress”. Overall, these studies suggest that an emphasis on test scores only contributes to leaving more students behind.

**Tracking and the Uncertainty Principle**

The characteristics of many lower-track classes resemble the characteristics critics argue high-stakes tests encourage in schools: when stakes are high, learning is reduced to the memorization of facts rather than elevated to higher-order thinking experiences. In the instance of high-stakes testing, Berliner and colleagues claim that the Uncertainty Principle accounts for the shift in educational emphasis. With roots in quantum mechanics (Heisenberg, 1927) and
social science (Campbell, 1979), the Uncertainty Principle as applied by Berliner maintains that our ability to measure academic achievement is diminished when consequences are attached to the results of the tests intended to measure that achievement. For this reason some researchers have used NAEP scores, which have no sanctions or rewards attached to them, to check the construct validity of high-stakes tests, yielding mixed results (Amrein and Berliner, 2002; Braun, Chapman, and Vezzu, 2010).

If the loss of validity associated with high-stakes tests is a result of teachers changing their teaching strategies in order to prepare students for exams, and those changes sacrifice comprehensive learning, we should see the distortion of high-stakes test scores to be greatest among students least likely to meet proficiency requirements. Academic tracking concentrates such students in lower track classrooms. The distortion of test score measurements should be greatest, therefore, in lower-track classes.

The intersection of academic tracking and high-stakes testing has received limited attention by scholars. Watanabe’s ethnographic case study of North Carolina schools subject to high-stakes testing policies reveals in detail how track influenced the way educators and their students dealt with standardized testing. Watanabe completed 68 hours of observation, examined student work, and conducted multiple interviews with two focal teachers who taught middle school classes at both honors and standard levels. Watanabe observed that students in upper track classes received more reading and writing assignments, more feedback on completed work, more discussion questions requiring more analysis and synthesis, and more creative and independent projects. Students in lower-track classes received more test preparation, even when they were taught by the same teacher as the upper-track classes. The difference in time spent on test preparation was drastic: while upper-track classes spent 15% of their instructional time on
activities directly related to test preparation, the lower-track classes devoted triple the instructional time on such exercises, and in the process skipped higher-order thinking activities (Watanabe, 2008).

There was nothing unusual about the schools, teachers, or classrooms in Watanabe’s study and nothing to suggest that behaviors observed in them are unique. Even though test-based accountability policies are intended to lift the expectations for and performance of even the lowest-achieving students, evidence suggests that learning experiences continue to differ drastically by academic track.

Research Question and Hypothesis

Previous research has demonstrated that upper and lower track classes provide starkly different learning experiences and these differences seem to persist in the era of standards and standardized testing. With college continuing to be an important and common step for high school graduates pursuing both academic tracks, it would be useful to know more about the intersection of tracking and testing in terms of preparing students for further education. More specifically, are high-stakes tests better at predicting college success for honors students than for lower-track students? If so, the results of this study will add to the skepticism of the ability of current policies of high-stakes tests and practices of academic tracking to adequately address unequal academic outcomes. I hypothesize that due to the Uncertainty Principle, student scores on standardized tests will be a better predictor of college achievement for upper track students than for lower track students. Additionally, I hypothesize that compared to high-stakes test scores, grade point averages are better predictors of college achievement for both upper- and lower-track students.
The rest of this paper details the study I conducted to test these hypotheses. It proceeds with a description of the data, methods, and results of my quantitative analysis. That section is followed by the qualitative component, intended to dig deeper into the mechanisms at work in the quantitative analysis. The paper closes with a discussion of the results and policy implications.

**Quantitative Investigation**

I test the hypotheses using a dataset comprised of the population of North Carolina public school students who graduated high school in 2004 and then matriculated into the UNC system. North Carolina is especially useful for this study due to its role as a pioneer in using high-stakes tests for accountability purposes (Grissmer & Flanagan, 1998). Between 1995 and 1998 the state rolled out the ABCs (Accountability and high standards, the Basics, and local Control) program for every public school from kindergarten to 12th grade. The program continues to evolve, but the use of its End-of-Grade (EOG, for elementary and middle school students) and End-of-Course (EOC, for high school students) test results to evaluate students and schools continues.

Measuring college performance is a key element of this study that sets it apart from other research on tracking and high-stakes testing. Besides the fact that this approach allows for using individual students as the unit of analysis, the state of North Carolina has an affordable, diverse, and geographically dispersed public system of higher education which attracts a large portion of its high school graduates. The sixteen campuses of the University of North Carolina (UNC) system are managed by a single General Administration and include historically Black institutions, both urban and rural campuses, and highly competitive liberal arts and engineering-focused schools.
Data

All models in this study use the same longitudinal dataset, the North Carolina Roots of STEM dataset originally conceived by Drs. Stearns, Mickelson, Dancy, and Moller at UNC Charlotte. The dataset follows one complete statewide North Carolina public school graduating class from middle school through high school and then into the state’s UNC system. More than 18,000 students graduated from NC public schools in 2004 and then matriculated into the UNC system; all of them appear in these data. The North Carolina Education Research Data Center (NCERDC) assembled the dataset with cooperation from North Carolina’s Department of Public Instruction (DPI) and the UNC General Administration and prepared the data for use in the project by developing crosswalk files among the sources relying on anonymous, encrypted student identifiers.

Variables

Data from high school years include students’ test scores, grade point averages (GPA), academic tracking, high school academic climate, and demographics such as race and gender. Data from college experiences include campus attended, grade point averages, and declared majors. The analyses in this study focus on student-level variables, but the models control for the effects of high schools and college campuses to account for variance that may be associated with attendance at a particular high school or UNC campus (see Laird & Ware, 1982). Additionally, I include a measure of the high school performance composite, which is the percentage of students enrolled at the high school achieving at or above grade level, according to the NC Department of

---

1 Their research is supported by a grant from the National Science Foundation (NSF REC-0635004) and seeks to understand influences on women and minorities entering STEM disciplines.
Public Instruction (DPI), to account for the effects of the academic climate. Note that while differences in achievement associated with sex are not a focus of this study, I include sex as a control variable because an achievement gap between males and females, favoring the latter, has been well documented (Corbett, et al., 2008; Orr, 2011).

Most of the variables used in this study came through NCERDC from North Carolina’s DPI. Variables for sex and limited English ability needed no recoding. Race variables were created by breaking apart the categorical variable provided in the dataset and creating a dummy variable for each race but leaving out “white” as the reference category. A dummy variable indicates whether a student is from his/her family’s first generation to attend college.

I used DPI data to create measures of mean EOC scores. At the time of this cohort’s high school education students in the Core curriculum were required to take multiple choice high-stakes tests in Algebra I, Geometry, Algebra II, English I, Biology, U.S. History, and ELP (Economics, Law, and Politics). DPI converts students’ raw scores into scale scores for accurate comparisons. To create the individual-level mean EOC variable, I averaged each student’s required EOC test scale scores.

Data on EOC courses include an indicator of course level. From these variables I created a proportion honors classes variable by dividing the number of EOC tested courses a student took at the honors, AP, or IB level by the total number of EOC courses that student had taken. On average, the students in this dataset took 45% of their EOC courses in an upper track. About 7% took all of the EOC courses in upper track classes and about 19% took none of their EOC courses in upper track classes. The majority took some mix of upper and lower tracks.
Table 1. Descriptive summaries for variables used.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Observations</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>College GPA</td>
<td>19251</td>
<td>2.629</td>
<td>0.881</td>
<td>0</td>
<td>4.52</td>
</tr>
<tr>
<td><strong>Independent Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean EOC score</td>
<td>19251</td>
<td>64.189</td>
<td>6.156</td>
<td>41.7</td>
<td>88</td>
</tr>
<tr>
<td>Male</td>
<td>19251</td>
<td>0.439</td>
<td>0.496</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>White</td>
<td>19251</td>
<td>0.686</td>
<td>0.464</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Black</td>
<td>19251</td>
<td>0.251</td>
<td>0.433</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hispanic</td>
<td>19251</td>
<td>0.013</td>
<td>0.112</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Asian</td>
<td>19251</td>
<td>0.030</td>
<td>0.171</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>American Indian</td>
<td>19251</td>
<td>0.011</td>
<td>0.104</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>19251</td>
<td>0.003</td>
<td>0.057</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Limited English</td>
<td>19251</td>
<td>0.008</td>
<td>0.088</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1st generation college student</td>
<td>19251</td>
<td>0.423</td>
<td>0.363</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Proportion honors classes</td>
<td>19251</td>
<td>0.449</td>
<td>0.334</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>High school GPA</td>
<td>19251</td>
<td>3.596</td>
<td>.660</td>
<td>0</td>
<td>5.53</td>
</tr>
<tr>
<td>H.S. Performance Composite</td>
<td>19251</td>
<td>72.371</td>
<td>11.543</td>
<td>25.5</td>
<td>94.9</td>
</tr>
<tr>
<td>Science major</td>
<td>19251</td>
<td>0.179</td>
<td>0.383</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Professional major</td>
<td>19251</td>
<td>0.031</td>
<td>0.173</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Health major</td>
<td>19251</td>
<td>0.056</td>
<td>0.230</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Business major</td>
<td>19251</td>
<td>0.133</td>
<td>0.340</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Humanities major</td>
<td>19251</td>
<td>0.128</td>
<td>0.334</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Education major</td>
<td>19251</td>
<td>0.122</td>
<td>0.327</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Social science major</td>
<td>19251</td>
<td>0.183</td>
<td>0.386</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Data from UNC’s General Administration included records of each student’s declared majors and grades. I recoded the variables for declared major according to the Classification of Instructional Programs (CIP) codes to create seven different variables, each representing a different category of disciplines. A student who declared a major in the humanities, for example, received a 1 for the humanities major variable while others received a 0. This method allowed students to have majors in more than one subject area and allowed models to account for differences in difficulty. I calculated college GPA in the typical fashion by dividing the quality points earned by the credits attempted.
Analytic Steps

This study tracks thousands of individual North Carolina students in a single cohort through high school and college. I developed several statistical models to measure the success with which high-stakes tests and other factors predict college GPA. In each model I control for the effects of clustering by schools through the use of multi-level modeling. A standard multiple regression equation includes an intercept, a coefficient for each independent variable, and an error term. Multi-level regression models add another term that represents another set of coefficients and intercept for each additional level of clustered observations. The models in this study that include student-, high school-, and college-level data are not only multi-level, but also cross-classified because groupings are not nested. I used the software package Stata and its `xtmixed` procedure to run multiple regression models that account for the effects of high school and college clusters (Rabe-Hesketh & Skrondal, 2012; Stata Corp, 2013). To account for collinearity problems that may occur in multilevel regression analysis, I centered the variables by the grand means.

Quantitative Results

Because the hypotheses assert that the relationship between high school achievement and college achievement is conditional upon academic track, the analyses separate the observations by track and add high school GPA as an additional independent variable. Most students took a mix of upper-level and lower-level track classes; these students appear in Model 2 of Table 2. Students who took all honors classes appear in Model 1 and students who took no honors classes appear in model 3.
Table 2. Predicted college GPA for high-, mixed-, and low-track students.

<table>
<thead>
<tr>
<th></th>
<th>Model 1: Students with all honors courses</th>
<th>Model 2: Students with some honors courses</th>
<th>Model 3: Students with no honors courses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>S.E.</td>
<td>Estimate</td>
</tr>
<tr>
<td>Mean EOC score</td>
<td>.021***</td>
<td>.004</td>
<td>.004***</td>
</tr>
<tr>
<td>High school GPA</td>
<td>.006***</td>
<td>.000</td>
<td>.007***</td>
</tr>
<tr>
<td>1st generation</td>
<td>-.139**</td>
<td>.040</td>
<td>-.089***</td>
</tr>
<tr>
<td>Black</td>
<td>-.109</td>
<td>.067</td>
<td>-.162***</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-.158</td>
<td>.140</td>
<td>-.140***</td>
</tr>
<tr>
<td>Asian</td>
<td>-.140</td>
<td>.085</td>
<td>.023</td>
</tr>
<tr>
<td>American Indian</td>
<td>-.075</td>
<td>.189</td>
<td>-.024</td>
</tr>
<tr>
<td>Other race</td>
<td>.089</td>
<td>.278</td>
<td>.046</td>
</tr>
<tr>
<td>Male</td>
<td>-.319***</td>
<td>.037</td>
<td>-.190***</td>
</tr>
<tr>
<td>Limited English</td>
<td>.125</td>
<td>.225</td>
<td>.146*</td>
</tr>
<tr>
<td>School Composite</td>
<td>-.000</td>
<td>.002</td>
<td>.004***</td>
</tr>
<tr>
<td>Constant</td>
<td>-.940***</td>
<td>.295</td>
<td>-.620***</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-1320.196</td>
<td></td>
<td>-13740.206</td>
</tr>
<tr>
<td>N</td>
<td>1399</td>
<td></td>
<td>14265</td>
</tr>
</tbody>
</table>

* significant at .05  ** significant at .01  *** significant at .001

NOTE: To save space, variables for declared major do not appear in the table.

Before addressing the main independent variables, some interesting patterns appear among the coefficients for several of the control variables. School composite had no effect on the students who were taking all upper-track classes, whereas higher school-wide achievement in high school raised the college performance of students in mixed- and lower-track classes. These results suggest that honors classes take on an academic culture that is separate from the rest of the school. The effects of gender are apparent in all three models, but strongest among the all-honors classes. Females do better than males in college, net the effects of high school achievement, but that gap is even larger among highest achieving students. In contrast, race had no significant effect among the all-honors-track students, but for the other students Blacks did
worse than Whites in college. Hispanics were only different from Whites in the mixed-track group. Finally, being a first-generation college student had a negative effect on college GPA when the student took more honors classes. Having parents who attended college had the greatest benefit among students taking upper-level classes.

As for the main independent variables of interest, results in Table 2 indicate that while test scores are good predictors of college GPA for all-honors students (with a statistically significant coefficient of .021), they were not nearly as good for mixed-track students (with a statistically significant coefficient of .004), and even worse for non-honors students (with a coefficient that is not significantly different from zero). The coefficient for the all-honors track students was significantly different from the coefficients for the other two groups, but the coefficients for the mixed and all-standard track students were not statistically different from each other.

High school GPA tells a different story. Results show that while the correlation between test scores and college GPA varies by academic track, the correlation between high school GPA and college GPA remains consistent across the three tracks. Both high school GPA and high-stakes test scores measure academic performance in high school, but high school GPA is substantially better at predicting college success for lower-track students than test scores.

Why do the three data points align so well for honors classes but not for standard track classes? To investigate this question I chose a sequential explanatory mixed methods approach (Cresswell and Clark, 2011) by following the quantitative analysis with a qualitative component that sheds light on what occurs in the typical honors and standard classrooms for state-tested high school courses.
Qualitative Investigation

Based on the Uncertainty Principle and the expected loss of predictive validity when stakes are high, I hypothesized that because lower track classes have more students at risk of failing high-stakes tests, those students’ test scores would have a weaker relationship with college achievement than students’ scores in upper track classes. While the results of my quantitative analysis supported that hypothesis, they also showed that high school grades do not follow the same pattern. Because my data did not include measures regarding curriculum, climate, and instructional method within classrooms, I took a grounded theory approach (Strauss and Corbin, 1990) for the next portion of the study by collecting relevant data from North Carolina classroom teachers and using it to develop an understanding of why test scores would lose predictive validity in the lower track but grades would not.

Methods

I began by developing a list of concepts relevant to the research question, such as tracking, test preparation, high stakes, and curriculum. I then developed questions that could be asked of educators in interviews regarding those concepts. As a former high school teacher who taught both honors and standard courses for state-tested and non-state-tested subjects, I thought it important to consult with another education researcher who, while also a former educator, had never taught under high-stakes testing programs. She suggested interview questions very similar to my own.

I then conducted separate interviews with ten different high school teachers currently employed in North Carolina public schools. Each of them teaches a state-tested course at both honors and standard tracks. The subjects represent five different high schools and two different
school districts, each selected based on geographic convenience and setting; one is a large metropolitan district while the other is a smaller and more rural district. The set of schools included both middle class suburban populations and populations with high rates of racial minorities, immigrants, and poverty. The teachers varied in subjects taught and years of experience.

During the interviews I asked teachers to comment on differences between their honors and standard tracked sections of the same course across five topics, including curriculum, materials, assignments, grading, and classroom climate. I also asked teachers to reflect on the effects state exams have on their instruction. To verify the accuracy of my coding and reporting of interview results, I shared my analysis with the interviewees and asked for their feedback. I also asked the same education researcher who had suggested interview questions at the start of this stage of the study to serve as auditor and review the analysis.

Results

Subjects were remarkably consistent in their responses. All teachers agreed that the curricula, standards, and state assessments were identical across the academic tracks. Several noted that they thought the curricula were quite good and no one complained that the content required by the state was insufficient or unrealistic. Several teachers, however, said that the tests were not designed, administered, or utilized in ways that were fair or accurate. These teachers remarked that the exams assumed a level of literacy, particularly in terms of vocabulary, that put some students, and therefore their teachers and schools, at a severe disadvantage.

Teachers reported that while the tests are always a factor that “looms in the background and shapes our classes,” none of them felt that they engaged in excessive test preparation for
either of their academic tracks. The reasons they gave varied. Some said that their love of teaching would end if they were forced to “teach to the test.” Others said that they could adequately prepare their students without engaging in “drill and kill.” Yet another said that her lower track students would benefit more from learning real life skills than from test preparation.

Teachers described several substantial differences between their upper and lower-track classes. English teachers said that they used different novels to teach the same concepts in standard and honors classes. While science classes all conducted the same lab experiments, honors students would do more with them, such as extended analysis, oral presentations, or video production. One English teacher said she regularly “flips” the honors classes but prefers to lead her standard level students through the content herself. These differences were expressed in one form or another by nearly every teacher, underlining the message that learning experiences differ by track, even if the course standards and exams do not.

Another difference between tracks that nearly all the teachers described relates to grading. Teachers describe the grading polices in their honors classes as “very strict”. One teacher said that her department has a rule that late work is never to be accepted in honors classes. Grading standards are not always as formal in the honors classes, but every teacher described them as “rigorous,” “academic,” or “tough.” Even when assigned the same work as the other track, honors students were expected to “do more with it” or “demonstrate higher-order thinking, such as with analysis, synthesis, or evaluation.” In their standard classes, teachers looked for mastery of the same content, but often calculated grades in a different fashion. Several said that effort would earn more points in their standard classes than in their honors classes, in which grades were a reflection of the quality of the finished product. Teachers also said they are willing to give standard-level students “the benefit of the doubt” whenever students’ work is
unclear; their honors students are required to communicate their ideas with precision. Finally, teachers said that they reward standard-level students for following directions whereas honors students are simply expected to do so. For each of these aspects of grading – effort, communication, and following directions – teachers say that they are trying to prepare students for their futures. Honors students, they believe, need more academic challenges while standard students need encouragement, direction, and achievable goals.

Finally, teachers consistently described stark differences in learning atmosphere between honors and standard classes. Honors students, one teacher remarked, can run their own discussions with minimal guidance from the teacher, while standard classes need teachers to act as a “traffic cop.” Several teachers noted that it was the attitude students had toward learning and school, not so much differences in academic ability that distinguished honors from standard classes. “It’s not that standard kids can’t do the work,” said one teacher, “it’s that they choose not to.”

To summarize, the teachers described several substantive ways in which their honors classes differ from their standard classes, including their materials, assignment depth and breadth, grading, and classroom climate. Such differences are typical, according to prior research. Furthermore, according to the interviews the presence of high-stakes testing policies had little influence over how they run their classes. Indeed, the different experiences in upper and lower track classes would likely exist even without the demands of high-stakes tests.

Discussion

This study began with the intention of assessing the success with which high-stakes tests can be used to assess college readiness. The quantitative portion of the study confirmed what
earlier scholarship would predict: even though they take the same tests as the upper track students, the test scores of students in lower track classes do a much poorer job predicting college performance than do the scores of upper track students. High school GPAs, however, are good predictors of college performance across both tracks.

The Uncertainty Principle, as well as the interviews with teachers in this study, explains why the predictive power of test scores might differ by track. Upper-track classes are academically rigorous and emphasize course content, content which appears on the exam. Success in an upper-track class depends upon students performing tasks that will payoff for them both on the standardized tests and in their college courses. Lower track classes, in contrast, do not place as much emphasis on academic rigor. Even though the state’s standards and exams remain the same for lower track students, their teachers modify the classroom expectations and instructional methods in ways that de-emphasize the content of the exams. Grades in upper-track classes reflect mastery of content, mastery that comes from academic rigor. Grades in lower track classes reflect less course content and more student behaviors, such as effort, participation, and deadlines, habits that will pay off for the lower-track students who go on to college.

In other words, this study uses three data points to assess student achievement: standardized test scores, high school GPA, and college GPA. For upper track students, these three points are well aligned with one another because of the emphasis on curriculum content that these students face in the exams, their high school classrooms, and in their college courses. Lower track students have a different experience however, one that de-emphasizes the rigorous pursuit of mastering course content that is so important on the exams and instead rewards behaviors that are associated with teacher-assigned grades, both in high school and in college. Teachers explained that they made this adjustment in order to encourage students to
develop persistence and train them for future education or work. As one teacher put it, “so much of success in college is just following directions.” If that premise is correct, then the type of performance that brings better grades in lower-track classes resemble those that bring better grades in college, but do not necessarily produce better scores on high-stakes tests.

Such a conclusion, however, does not justify the practice of academic tracking, let alone the stark differences in academic rigor. Even though teachers are thoughtful and even successful at choosing materials and planning assignments in ways that prepare students for college, it is also clear from this study that honors students are still getting more opportunities to learn than their standard-track counterparts. This difference can be seen in the interaction model where the same increase in test scores yields bigger gains in college GPA for honors students than for standard students. Secondly, teachers are very clear about honors classrooms being spaces where students have opportunities to lead discussions, present work, pursue additional scientific inquiry, read more challenging novels, and engage in higher-order thinking activities. Teachers attribute these pedagogical differences to the behavior and attitudes of students. Teachers feel they can turn their honors students loose on the course material but must run a tight ship in their standard classes. As a result, lower-track students fall further behind.

Limitations

Perhaps the greatest limitation to this study is the fact that it only uses data for North Carolina public school students who entered the state’s public university system. Students who did not attend college are not included in this study, so the interpretation of results only applies to a limited (yet important) segment of high school students. The intersection of tracking and
standardized testing may have a different dynamic for the non-college bound, but that question goes beyond the scope of this study.

Students enrolling in private institutions, out-of-state public schools, or community colleges likewise do not appear in the data. These missing students went on to elite schools, somewhat selective schools, and non-selective schools and are thus distributed across the full range of student achievement and backgrounds, much like the students found in this study. Roughly 8% of the students included in this study took all of their tested courses in upper-track classes, about 20% took none of their tested classes in an upper track, and the other 72% took a mix of upper- and lower-track classes. Within each of the three categories, the UNC-bound students in the dataset on average scored roughly 5 points higher on the EOCs than the non-UNC bound students, less than one standard deviation. Thus, while the dataset fails to capture all the students, it succeeds at capturing a segment that is not drastically different from the full population.

Conclusion

This study contributes to the steadily growing vein of research that suggests high-stakes tests are not as useful or as reliable as the public may hope. Furthermore, the study renews questions about the practice of academic tracking and its deleterious effects on equity of academic outcomes. However, despite the apparent disadvantages to policies that assess with tests and divide by tracks, it is likely that both high-stakes tests and honors and standard level courses will be around for a long time. There is too much momentum behind test-based accountability policies and too few opponents of tracking to envision alternative methods of
evaluating and organizing schools. Given that reality, policymakers should focus on how to make the practice less harmful to students, especially those on the lower tracks.

Perhaps the most urgently-needed strategy is to make lower-track classrooms more like upper-track classrooms. Teachers want all of their students to succeed, but many of them have found it difficult to adapt the best instructional strategies from their honors courses to their non-honors courses. New research and professional development might help in that regard. Higher expectations, higher-order thinking activities, and more varied opportunities to learn may be more difficult to employ in lower-track classes, but they may also be the key to improved outcomes.

Finally, this study serves as a reminder that education reforms such as test-based accountability programs do not occur in a vacuum. They come to schools that already have their established organizations and cultures. In this example, it is the presence of academic tracking that presents an obstacle to a policy that attempts to treat all students the same by enforcing universal standards among students receiving very different educational experiences. Achievement and equity are worthwhile goals for our public school systems; the outcomes for each goal depend greatly on how educators teach within each school.

References


Stata Corporation (2013). *Stata multilevel mixed-effects reference manual*. College Station, TX: Stata Corp LP.


