Educational Evaluation and Policy Analysis
September 2014, Vol. 36, No. 3, pp. 371–393
DOI: 10.3102/0162373713517935

Improving the Targeting of Treatment: Evidence From College Remediation

Judith Scott-Clayton
Peter M. Crosta
Teachers College, Columbia University

Clive R. Belfield
The City University of New York

Remediation is one of the largest single interventions intended to improve outcomes for underprepared college students, yet little is known about the remedial screening process. Using administrative data and a rich predictive model, we find that severe mis-assignments are common using current test-score-cutoff-based policies, with "underplacement" in remediation much more common than "overplacement" college courses. Incorporating high school transcripts into the process could significantly reduce placement errors, but adding test scores to already available high school data often provides little marginal benefit. Moreover, the choice of screening policy has significant implications for the racial and gender composition of college-level courses. Finally, the use of more accurate screening tools would enable institutions to remediate substantially fewer students without compromising college success.

Keywords: *remediation, higher education, standardized testing*

ONLY about half of degree-seeking college entrants will complete any type of degree or certificate within 6 years.[1] One of the primary explanations for college non-completion is that many entrants, despite having graduated from high school, nonetheless lack the basic academic skills required for success in college coursework (Bailey, Jeong, & Cho, 2010; Greene & Forster, 2003). As a result, most 2-year colleges and many 4-year colleges require incoming students to be screened for possible remediation, which provides basic skills instruction but does not bear college credit, before they may enroll in college-level courses.

Besides financial aid, remedial education is perhaps the most widespread and costly single intervention aimed at improving college completion rates. Half of all undergraduates will take one or more remedial courses while enrolled; among those who take any, the average is 2.6 remedial courses.[2] With over 3 million new students entering college each year, this implies a national cost of nearly US$7 billion dollars annually.[3] This figure accounts only for the direct cost of remediation: It does not include the opportunity cost of time for students enrolled in these courses, nor does it account for any impact, positive or negative, that remediation may have on students' future outcomes.

The impacts of remediation are likely heterogeneous across individuals, depending upon prior academic preparation as well as non-academic characteristics such as motivation and

grit.[4] Thus, like a costly medical intervention with non-negligible side effects, the net value of remediation in practice depends not just on the average effectiveness of the treatment, but also on whether or not the individuals most likely to benefit can be identified in advance. Of the public 2-year institutions where remediation is particularly concentrated, virtually all use brief, standardized tests administered to new students just prior to registration to determine who needs remediation (Parsad, Lewis, & Greene, 2003). Often, assignment is determined solely on the basis of whether a score is above or below a certain cutoff. While several studies have leveraged the somewhat arbitrary nature of these cutoffs to identify the causal effect of remediation, very little attention has been paid to the diagnostic value of the tests themselves.

This is surprising given the potentially serious adverse consequences of incorrectly assigning a truly prepared student to remediation. Prepared students who are assigned to remediation may garner little or no educational benefit, but incur additional tuition and time costs and may be discouraged from or delayed in their degree plans. Indeed, several studies using regression-discontinuity (RD) analysis to compare students just above and just below remedial test-score cutoffs have generally found null to negative impacts of remediation for these "marginal" students. For example, Martorell and McFarlin (2011) examine administrative records for over 250,000 students in Texas public 2- and 4-year colleges: Those just below the test-score threshold had significantly lower rates of persistence and college credit accumulation, with no impact on degree attainment and future labor market earnings. Studies in the state of Florida and an anonymous large northeastern urban community college system using similar data and methods found similarly null to negative effects on academic outcomes (Calcagno & Long, 2008; Scott-Clayton & Rodriguez, 2012).

A typical caveat in RD studies is that they identify average treatment effects that are local to students scoring near the cutoff—that is, the highest scoring remediated students—and thus one interpretation of the RD evidence may be that the existing remedial cutoffs are set too high. The available evidence regarding heterogeneity by ability does in fact suggest that the negative effects of remediation may be largest for higher ability or lower academic–risk students (Martorell & McFarlin, 2011; Scott-Clayton & Rodriguez, 2012).[5]

Moreover, assigning truly unprepared students directly to college-level coursework implies a different, but no less important set of potential costs. First, there is strong evidence of peer effects in higher education, meaning that truly unprepared students who are incorrectly assigned to college-level coursework might not only do worse academically than they would have otherwise, they might depress the achievement of their better-prepared peers (Carrell, Fullerton, & West, 2009; Sacerdote, 2001; Winston & Zimmerman, 2004; Zimmerman, 2003). Second, there is evidence that at least some students fare better in college (in terms of persistence and degree outcomes) if they enter remediation, and those wrongly assigned directly to college level would forgo these potential benefits. Taking advantage of arbitrary variation in test cutoffs across 4-year campuses in Ohio, Bettinger and Long (2009) use distance to college as an instrument for the stringency of the cutoff policy an applicant was likely to face. They find that students who were more likely to be remediated (by virtue of the cutoff policy at the nearest school) were also more likely to complete a bachelor's degree in 4 years. Similarly, some RD studies examining very low-scoring students at the margin between higher and lower levels of remediation have found less negative and some positive effects of being assigned to the more intensive remedial treatment (Boatman & Long, 2010; Dadgar, 2012; Hodara, 2012, however, finds large negative effects of assignment to a lower level).

Improving the accuracy of the assignment process is thus of particular importance given the evidence for heterogeneous impacts across individuals and given that the dominant pattern of null to negative effects suggests remediation may be overprescribed as a treatment. Indeed, many institutions and several states, including Connecticut and Florida, are currently moving away from mandatory test-based remedial placement systems out of concern that too

many students are being assigned to too many remedial courses (Fain, 2013).

The contribution of our study is to use a rich predictive model of college grades to simulate the prevalence of mis-assignment using common cut-off rules with the two most commonly used remedial screening tests, to explore whether high school transcript information might be a more valuable screening device, and to examine empirically the trade-offs institutions face in assigning either too many or too few students to remediation. We also test whether the choice of remedial screening device has disparate impacts by race or gender. Our analysis uses administrative data, including high school transcripts, remedial test scores, and college grades for tens of thousands of students, in two large but otherwise distinct community college systems. One is a large urban community college system (LUCCS) with six affiliated campuses; the other is a state-wide community college system (SWCCS) of over 50 community colleges.[6]

To preview our results, we find that roughly one in four test takers in math and one in three test takers in English are severely mis-assigned, with severe underplacements in remediation much more common than severe overplacements in college-level coursework. Holding the remediation rate fixed, we find that using high school transcript information for remedial assignment—either instead of or in addition to test scores—could significantly reduce the prevalence of these assignment errors. In contrast, incorporating test scores when high school information is already available often provides virtually no additional benefit. To address concerns that our predictive model relies too heavily upon extrapolation, we conduct a sensitivity analysis in which we exclude students scoring substantially below the existing test-score cutoffs and find that our conclusions are highly robust. Furthermore, the choice of screening device has significant implications for the racial and gender composition of both remedial and college-level courses. Finally, we find that if institutions took account of students' high school performance, they could remediate substantially fewer students without lowering success rates in college-level courses.

The article proceeds as follows: we first provide background on remedial testing and summarize the relevant research on test validity. Next, we describe the methodology, including our institutional context and data. We then present our results, and conclude with a discussion of policy implications.

Background on Remedial Testing and Test Validity

At non-selective, "open-access" 2- and 4-year institutions, many students' first stop on campus will be to a testing center to be screened for remediation in reading/writing and math. In practice, institutional decisions about which screening tools to use and where to establish cutoffs for college-level coursework appear to be somewhat ad hoc (Bettinger & Long, 2009).[7] The affordability and efficiency of the screening tool itself are clearly important, particularly for large institutions that may need to process thousands of entrants within a matter of weeks.

Currently, two remedial placement exams dominate the market: COMPASS®, developed by ACT, is used by at least 61% of community colleges, and ACCUPLACER®, developed by the College Board, is used by at least 39% of community colleges (Fields & Parsad, 2012).[8] Both testing suites offer a written essay exam as well as computer-adaptive tests in reading comprehension, writing/sentence skills, and several modules of math (of which pre-algebra and algebra are most common). The tests are not timed, but on average each test component takes less than 30 minutes to complete, such that an entire battery of placement exams may be completed in less than 2 hours (ACT, 2006; College Board, 2007).[9] Typically, colleges waive the placement test for students with high ACT or SAT scores. Those who fail the test(s)—meaning they score below their institution's designated cutoff score on one or more modules of the given test battery—are assigned to remedial coursework, which may stretch from one to several courses depending upon the student's score. Unlike the SAT and ACT exams used for college admissions, no significant test preparation market has sprung up around placement exams, perhaps because many students are not even aware of these exams and their consequences until after admission. One recent qualitative study found that students were generally

uninformed about remedial assessments, with some students even believing it would be "cheating" to prepare (Venezia, Bracco, & Nodine, 2010).

Related Literature on Test Validity

Perhaps the simplest approach to evaluating the validity of a screening test is to identify the key outcome of interest and regress it on the predictor(s) of interest, either alone or in conjunction with other available predictors.[10] The researcher then examines goodness-of-fit statistics (R^2 or correlation coefficients) as well as the size and significance of the resulting regression coefficients. This method has been used, for example, to examine the predictive validity of the SAT and ACT (Bettinger, Evans, & Pope, 2011; Bowen & Bok, 1998).

With respect to remedial placement exams, the College Board has published correlation coefficients relating each of the ACCUPLACER® modules to measures of success in the relevant college credit-bearing course, with correlations ranging from .23 to .29 for the math exams and from .10 to .19 in reading/writing (Mattern & Packman, 2009). In two working papers related to this study, Scott-Clayton (2012) finds comparable correlation coefficients for the COMPASS® in a LUCCS (ranging from .19 to .35 in math and .06 to .15 in English), while Belfield and Crosta (2012) find much lower correlations for both COMPASS® and ACCUPLACER® at a statewide system of community colleges.

Goodness-of-fit analyses, however, necessitate several caveats. Linearity and distributional assumptions may be violated in the case of dichotomous or ordinal outcomes. Moreover, while in theory one could examine the relationship between test scores and college grades for any student who ever makes it to college coursework, for students initially assigned to remediation the treatment may confound the relationship between initial scores and future performance. This necessitates a restriction of the sample to only those who are placed directly into college-level courses, and this restricted range of variation can bias goodness-of-fit statistics downward (ACT, 2006).[11] More fundamentally, these measures provide no tangible estimates of how many students are correctly or incorrectly assigned under different screening devices, nor any practical guidance for policymakers wondering whether test cutoffs are set in the right place.

A second approach is to examine success rates in the college-level course for students selected on the basis of different screening devices and assignment thresholds. Bettinger et al. (2011) perform this type of analysis with respect to the ACT, simulating the college dropout rates that would result depending upon how ACT subtest scores are weighted in a college admissions process with a fixed number of spots. Examining test validity in a different context, Autor and Scarborough (2008) observe how the productivity of job hires (as measured by length of employment) changes when employment tests are introduced into the applicant screening process. These types of analyses are useful but focus on only one side of the assignment process. In the case of remediation, policymakers may worry not only about unprepared students being assigned to college-level work but also about adequately prepared students being assigned to remediation. As discussed above, both types of mistakes have potentially significant costs.

A third approach, which we develop for our primary analyses, is to analyze measures of diagnostic accuracy, or "the ability to correctly classify subjects into clinically relevant subgroups" (Zweig & Campbell, 1993). This approach has a long history in the medical screening literature and a more recent history in educational measurement, but has not been widely applied in economics or education policy research. This could be due to a longstanding focus on identifying average treatment effects: As long as such effects are constant, then the matter of identifying whom to treat is less important. But given an increasing interest in the potential heterogeneity of treatment effects, it will become increasingly important to develop assignment tools to more accurately target interventions. Analyses of diagnostic accuracy may utilize a variety of metrics, but all aim to quantify the frequencies of accurate diagnoses, false-positive diagnoses, and false-negative diagnoses using a given test and classification threshold.[12] If decision makers also have information on the costs and benefits of each

Treatment assignment	Predicted to Succeed in College-Level Course?	
	No	Yes
Assigned to remediation	(1) Accurately placed (true positive)	(2) Under-placed (false positive)
Assigned to college-level	(3) Over-placed (false negative)	(4) Accurately placed (true negative)

FIGURE 1. *Classifications based on predicted outcomes and treatment assignment.*

type of event (as well as the cost of testing itself), the event frequencies can be weighted accordingly and combined into a welfare function (or loss function) that can guide the selection of the optimal screening tool and cutoff.

Sawyer (1996) is the first to apply this type of decision theory framework to the choice of remedial screening tests. He notes that no assignment rule can avoid making errors—some students who could have succeeded in the college-level course will be assigned to remediation (an underplacement error), while some students who cannot succeed at the college level will be placed there anyway (an overplacement error). Figure 1 summarizes the four potential events that result from an assignment decision by cross-tabulating potential outcomes in the college-level course against actual treatment assignments.

The assignment accuracy rate, which adds the proportions of students in cells (1) and (4) of Figure 1, derives from an implied welfare function in which the decision maker gives equal weight to students placed accurately into remediation or college-level coursework, and zero weight to under- and overplacement errors. Publishers of the two most commonly used remedial placement exams now provide estimated placement accuracy rates, ranging from 60% to 80%, to help support their validity (ACT, 2006; Mattern & Packman, 2009). In related working papers using the same data utilized here, Scott-Clayton (2012) and Belfield and Crosta (2012) also find accuracy rates in this range, at least when "success" in college coursework is defined as earning a B or better.

But accuracy rates may vary depending upon how success is defined: This can be seen in Figure 2, which provides a schematic plot of college math success rates against placement test scores. Among students scoring at the hypothetical cutoff, 45% earn a B or better in college-level math (bottom line), 62% earn a C or better (middle line), and 74% can at least pass (top line). Thus, if placed in remediation 45% of these students at the cutoff (as well as the proportion indicated by the B-or-better line for students with scores below the cutoff) will be underplaced by any criterion; if placed in college level, then 26% of those at the cutoff (as well as the proportion indicated by one minus the passing percentage for student with scores above the cutoff) will be overplaced by any criterion. The remaining proportion who would earn a C or D are ambiguously classified; placing them into the college-level course is correct under a passing criterion for success but is a mistake under the B-or-better success criterion. Prior research consistently finds that remedial tests are more accurate at classifying students based on the B-or-better criterion than on lower success criteria (ACT, 2006; Belfield & Crosta, 2012; Mattern & Packman, 2009; Scott-Clayton, 2012). Scott-Clayton (2012) and Belfield and Crosta (2012) find that when the goal is simply identifying who will pass versus fail, accuracy rates range between just 36% and 50%.

Our analysis (described in detail below) will focus on error rates rather than accuracy rates, for two reasons. First, Sawyer's (1996) study demonstrates how policy conclusions based on accuracy rates can shift dramatically depending upon the definition of success. He compares accuracy rates

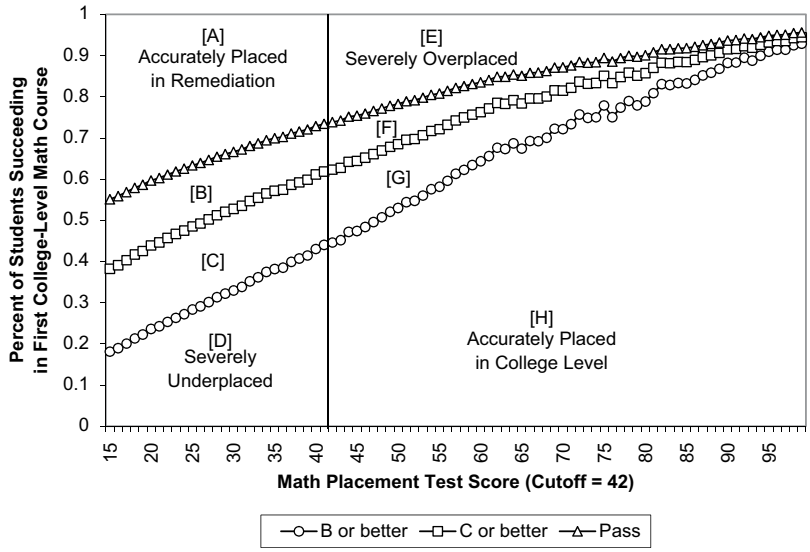

FIGURE 2. *Percent succeeding in college-level math, by math test score (schematic).*
Note. This schematic diagram illustrates the concept of accuracy and error rates using alternative definitions of success in the college-level course. The vertical line indicates a hypothetical cutoff for remedial assignment. Students scoring at this hypothetical cutoff have a 45% chance of earning a B or better in college-level math, 62% chance of earning a C or better, and 74% chance of passing. Thus, if placed in remediation, 45% of these students will be severely underplaced; if placed in college level, then 26% (100% – 74%) of students with this score will be severely overplaced. The region of the chart that is unlabeled, lying between the "B or better" line and the "Passed" line, represents ambiguous classifications (i.e., the proportion likely to earn only a C or D at college level, meaning their classification will depend upon the standard of success chosen).

using ACT math subtest scores versus using a locally developed test for math placement at a large public institution in the Midwest. He finds that if success is defined as earning a B or better, using the ACT math subscore with a relatively high cutoff generates the best accuracy rates, while if success is defined as earning only a C or better, using the locally developed test with a relatively low cutoff generates the best accuracy rates. Second, his results indicate that a wide range of potential cutoffs can generate similar accuracy rates, even as the mix of overplacement and underplacement errors changes substantially. As these errors may have different costs (and will fall on different students), it is useful to consider them separately.

The Potential Value of High School Transcript Data

Even the test publishers themselves emphasize that test scores should not be used as the sole factor in placement decisions (see, for example, *Accuplacer Coordinator's Guide*; College Board, 2007). One potentially rich source of additional information is a student's high school transcript, used either in conjunction with or as an alternative to placement tests for deciding on remedial assignment. Transcripts are readily accessible, as most students submit their high school transcripts as part of the admissions process, and may yield a wealth of information on cognitive skills, subject-specific knowledge, as well as student effort and motivation. Moreover, because they are accumulated over time across a range of courses and instructors, high school grade point averages (GPAs) and courses completed may simply be less noisy than brief, "one-off" exams. Yet to the best of our knowledge, high school grades and coursework have not been widely utilized or even studied as potential screening tools for assignment into remediation.

This is surprising given their demonstrated explanatory power for college outcomes and beyond. Studies have found strong associations between high school GPA and freshman GPA (Rothstein, 2004), as well as between high school efforts and college enrollment (on high

school algebra, see Gamoran & Hannigan, 2000; on high school coursework, see Long, Conger, & Iatarola, 2012; and on curricular intensity in high school, see Attewell & Domina, 2008). A related study by Long, Iatarola, and Conger (2009) looks at the influence of high school transcripts on the need for math remediation in Florida. However, remediation is identified as failing the Florida Common Placement Test, which presupposes the validity of the placement test. Nevertheless, the results from Long et al. suggest a strong influence of high school curriculum: Remediation need varies inversely with eighth-grade math scores and with the level of math taken in high school. Plausibly, information from high school appears to be predictive of performance in college.

The optimal decision rule may be a combination of placement tests and transcripts (Noble & Sawyer, 2004). A major contribution of our study is to compare the usefulness of high school transcript information either instead of or in addition to remedial test scores, and to explore whether the choice of screening device has disparate impacts by race or gender.

Method

We use a rich predictive model of college grades to examine several validity metrics under alternative policy simulations, focusing on three questions. First, how well do remedial screening tests identify students who are likely or unlikely to succeed in college-level coursework? Second, what is the incremental value of such tests above and beyond the information provided by high school transcripts generally and high school GPA in particular? We examine these questions for the full sample and for subgroups by race/ethnicity and gender. Finally, what are the trade-offs involved in establishing higher versus lower screening thresholds for remedial "treatment," and what does the chosen threshold reveal about institutional preferences?

Validity Metrics and Alternative Screening Policies

To address the potential oversimplification of examining a single placement accuracy rate, the simple two-by-two chart in Figure 1 could be expanded to include multiple gradations of success, and policymakers could assign separate weights to every possible outcome. But it would be presumptuous for researchers to attempt to completely specify the weights in a highly intricate welfare function. Instead, we propose a simple alternative to the accuracy rate: a loss function that we call the severe error rate (SER). Specifically, the SER combines the proportion of students predicted to earn a B or better in college level but instead placed into remediation (the severe underplacement rate, or Region D in Figure 2) with the proportion of students placed into college level but predicted to fail there (the severe overplacement rate, or Region E in Figure 2).

We see at least two advantages of the SER relative to placement accuracy rates. First, it focuses attention on the most severe assignment errors, which may be associated with the highest costs. While there may be disagreement about the "correct" placement for a student predicted to earn only a C or D in a college-level course, it seems uncontroversial that a student likely to earn an A or B should be placed directly into college level and a student likely to fail should not. Second, by breaking the SER into its two components, we allow for severe overplacements and severe underplacements to have different weights in a welfare analysis.

Finally, to acknowledge that policymakers may care about factors beyond mis-assignment rates, we show two additional metrics for each policy simulation: the predicted success rate among those placed directly into the college-level course (using the C-or-better criterion) and the remediation rate. For example, given two different assignment systems with the same overall error rates, policymakers may prefer the system that has a higher success rate in the college-level course. And even when we hold the remediation rate fixed overall, alternative screening devices may differentially affect remediation rates within race or gender subgroups, something that we examine below.

We examine these metrics under the current test-score cutoff-based policies in place in each system (using pre-algebra and algebra test scores to screen for remedial math, and reading/writing test scores to screen for remedial English). We then compare the results with

those obtained with two alternative screening devices, holding the proportion assigned to remediation fixed: (a) using an index of high school achievement alone, using information from high school transcripts; and (b) using an index that combines both test scores and high school achievement. Later, we examine how these metrics vary as we vary the proportion assigned to remediation, holding the choice of screening device fixed.

Estimating Severe Under- and Overplacement Rates

The SER combines the proportion of students predicted to earn a B or better in college level but instead placed into remediation (the severe underplacement rate) with the proportion of students placed into college level but predicted to fail there (the severe overplacement rate). The first step in calculating SERs is thus to estimate rich predictive models of students' probability of failing the college-level course as well as the probability of earning a B or better.[13] To do this, we restrict the sample to those who ever enrolled in a college-level course in the relevant subject (math or English) without taking a remedial course in that subject first.[14] We refer to this as the math or English estimation sample. Separately for college-level math and English courses, we run the following two probit regressions:

$$\Pr(\text{Fail}=1) = \alpha + (\mathbf{TestScores})\beta_1 + (\mathbf{HSAch})\beta_2 + \mathbf{X}\beta_3 + \varepsilon, \quad (1a)$$

$$\Pr(\text{BorBetter}=1) = \alpha + (\mathbf{TestScores})\beta_1 + (\mathbf{HSAch})\beta_2 + \mathbf{X}\beta_3 + \varepsilon, \quad (1b)$$

where **TestScores** is a vector of pre-algebra and algebra test scores for college math outcomes, and reading/writing test scores for college English outcomes; **HSAch** is a vector of high school achievement measures, including cumulative GPA and credits accumulated (the precise measures, described in the data section below, vary somewhat across our two systems); and **X** is a vector of other demographic variables that have predictive value. For the LUCCS analysis, **X** includes race/ethnicity, gender, age, English as second language (ESL) status, years since high school graduation, and an indicator of whether or not the individual previously attended a local high school. For the SWCCS analysis, the model includes race/ethnicity and gender. For both systems, we also include interactions of test scores and high school achievement with race/gender.[15] Even though these demographic variables cannot be used in the assignment process, they help improve the predictions that underlie our estimated error rates.[16]

After running these two regressions for the estimation sample, we then compute predicted probabilities of failing or earning a B-or-better for all students with available data, including those scoring below the cutoff (we call this larger group the prediction sample). The following equations describe how these predicted probabilities are used to compute the probability of severe underplacement or overplacement for each individual under a given assignment rule:

$$\Pr(\text{SeverelyUnderplaced}=1) = \Pr(\text{BorBetter}=1) \text{ if remediated, 0 otherwise,} \quad (2)$$

$$\Pr(\text{SeverelyOverplaced}=1) = \Pr(\text{Fail}=1) \text{ if NOT remediated, 0 otherwise.} \quad (3)$$

An individual's probability of being severely misplaced is simply the sum of overplacement and underplacement probabilities from Equations 2 and 3. The SER for the sample as a whole, or for a given subgroup, is simply the average of these individual probabilities.

When we simulate SERs using alternative screening devices, the underlying probabilities of success from Equations 1a and 1b remain fixed and we simply vary the assignment rule. When comparing across screening devices, we initially choose cutoffs that ensure the proportions assigned to remediation remain roughly constant. If the alternative device were a single measure, such as cumulative high school GPA, we could simply set the cutoff at the percentile corresponding to the current test-score-based cutoff. But as we are simulating alternative *sets* of predictors, we first combine these multiple measures into a single regression-based index.[17]

Addressing Extrapolation Concerns

A limitation of this type of analysis is that it requires extrapolation of relationships that are observed only for those placing directly into

college level to those who score below the current test cutoff. While this restriction-of-range in our initial predictive model does not *necessarily* lead to biased accuracy and error rates (in contrast to goodness-of-fit statistics, which can be biased by range restrictions even when regression coefficients are unbiased), the analysis rests on the underlying assumption that the observed relationship between scores and outcomes for high scorers is equally applicable to very low scorers.

For several reasons, however, for our analysis this concern is less worrisome in practice than in theory. First, the test scores themselves are extremely noisy: For example, the COMPASS algebra module has a standard error of measurement of 8 points, meaning a score of 30 (LUCCS cutoff for the most recent cohorts) is not distinguishable with 95% confidence even from the lowest possible score of 15 (ACT, 2006). Second, the earlier cohorts in LUCCS were subject to lower cutoffs (27 for the two math modules, 65 for the reading module) meaning that we *do* have some observations below the current cutoffs that do not rely upon extrapolation, and there is no indication of a different relationship between scores and outcomes for these additional observations. Moreover, while each sample (LUCCS and SWCCS, COMPASS and Accuplacer, English and math) we analyze involves extrapolation, the extrapolation is not the same in every case because the cutoffs occur at different points in the distribution and the tests themselves are different. Finally, it is worth emphasizing that our underlying predictive model includes more than just test scores; it also includes our measures of high school achievement as well as basic demographic information that are not explicitly range-restricted in the estimation sample.

Nonetheless, to explicitly address extrapolation concerns, we perform a sensitivity analysis in which we exclude at the outset all students with test scores substantially below the current cutoffs. While the characteristics of this more limited sample are very different than our primary analysis sample (with much lower rates of remediation, for example), we will show that the conclusions from our analysis remain unchanged.

Institutional Context and Data

We analyze two very large, but distinct community college systems to improve the generalizability of our results. The data sets for this analysis were provided under restricted-use agreements with a LUCCS including six individual institutions, and a SWCCS comprising 50 separate institutions. The LUCCS data come from four cohorts of nearly 70,000 first-time degree-seekers who entered one of the system's colleges in the fall of 2004 through 2007. The SWCCS data are from two cohorts of 49,000 students who enrolled in the academic years 2008 to 2010, almost all of whom are in degree programs. For additional detail on institutional context, see Scott-Clayton (2012) for LUCCS and Belfield and Crosta (2012) for SWCCS.

During our study period, LUCCS was using the COMPASS® test, with modules for numerical skills/pre-algebra, algebra, and reading, as well as a writing exam adapted slightly from the standard COMPASS® writing module (each writing exam is graded in a double-blind system by two LUCCS readers at a central location). The SWCCS permits a range of placement tests, although the majority of students took either ACCUPLACER® or COMPASS® tests (we analyze the ACCUPLACER® and COMPASS® samples separately at SWCCS). In both systems, test cutoffs are established centrally, and students' compliance with course assignment decisions is high: While some students may not enroll in the required remedial course immediately, relatively few circumvent remediation to enroll directly in a college-level course. Re-testing is not allowed at LUCCS until after remedial coursework has been completed; at SWCCS approximately 10% to 15% of students retest prior to initial enrollment. In both cases, we use the maximum test score (prior to enrollment) for our simulations because this is what is actually used for placement in practice.

Table 1 provides demographic information on the full sample and main subsamples for the predictive validity analysis, and also shows the percentages assigned to remedial coursework in each subject as a result of their placement exam scores. The first column describes the overall populations. Subsequent columns are limited to students who took a placement exam in the

TABLE 1
Selected Demographics

	Math sample			English sample	
	(1) All degree-seeking entrants	(2) Math test takers	(3) Math test takers with HS achievement data	(4) Reading/ writing test takers	(5) Reading/ writing test takers with HS achievement data
A. LUCCS sample					
% female	56.8	57.3	58.2	56.7	57.2
% minority	85.4	85.5	84.5	86.8	86.0
Age (years)	21.0	21.1	20.8	21.5	21.2
Years since HS graduation	2.6	2.7	2.2	3.0	2.4
% entering <1 year after HS	55.0	53.4	62.8	50.1	59.3
Average cumulative HS grades[a]	70.3	69.7	72.6	69.5	72.5
Average COMPASS algebra score	27.0	26.9	27.5	26.5	27.1
Average COMPASS reading score	70.8	71.2	71.1	70.9	70.9
% assigned to remediation					
In math	63.0	78.9	77.8	70.1	68.5
In English (reading or writing)	59.4	63.8	61.8	76.1	75.4
In either subject	81.5	91.4	90.8	92.2	91.7
Sample size	68,220	54,412	37,860	50,576	34,808
B. SWCCS sample					
% female	53.7	51.9	49.9	53.8	51.0
% minority	33.1	29.9	27.0	33.3	28.8
Age (years)	22.5	22.0	18.7	22.5	18.7
Average cumulative HS GPA	2.5	2.5	2.6	2.5	2.5
Average COMPASS algebra score	34.8	34.5	39.7	34.4	38.9
Average COMPASS reading score	79.9	82.6	81.9	79.8	78.9
% assigned to remediation					
In math	70.2	70.2	60.7	70.9	61.5
In English (reading or writing)	58.4	55.1	59.5	58.4	62.3
In either subject	74.8	80.6	76.7	75.5	75.5
Sample size	48,735	31,587	10,897	47,230	14,789

Source. Administrative data from LUCCS (2004–2007 entrants) and SWCCS (2008–2009 entrants).
Note. At LUCCS, Approximately 30% of test takers do not have HS achievement data available because they enrolled directly at an institution instead of via a centralized application system. For SWCCS, full transcript data from the public school system within the state was matched to college enrollees. Thus, data are available only for those matriculating from the public schools. LUCCS = large urban community college system; HS = high school; SWCCS = state-wide community college system; GPA = grade point average.
[a]At LUCCS, the HS grade average is converted to a 0 to 100 grading scale.

respective subjects, and then further restricted to those with high school information available. The samples described in columns 3 and 5—math and reading/writing test takers, respectively, who have high school transcript data available—will be the focus of the remainder of our analyses. For LUCCS, we have high school data for any student that applied to the system centrally (about 70% of all test takers) while for SWCCS we only have data for recent graduates of the state's public school system (about 30%–35% of all test takers). Note that the students in these primary analysis samples tend to be younger and are more likely to have entered college directly from high school.

For LUCCS, as at higher education institutions generally, nearly 6 out of 10 entrants are female. While more than half of LUCCS entrants are age 19 or below and come directly from high school, nearly one quarter are 22 or

older, and on average entrants are 2.6 years out of high school. Finally, LUCCS is highly diverse (over one third of students are Hispanic, over one quarter are Black, and over 10% are of Asian descent). Across these four cohorts of LUCCS entrants, more than three quarters were assigned to remediation in at least one subject: 63% in math, 59% in writing or reading. The proportions among those who actually take the placement exams are necessarily higher, with 78% of math test takers assigned to math remediation, and 76% of reading/writing test takers assigned to writing remediation.

For SWCCS, a slight majority of students are female and the typical entrant is a couple of years out of high school. In contrast to LUCCS, only one third of the students are minorities. But SWCCS shows similarly high rates of remedial assignment: 70% in math, 58% in English, and three quarters overall. These rates are slightly higher for our math and English testing samples.

Our measures of high school achievement differ somewhat between LUCCS and SWCCS.[18] For LUCCS, the high school data comes from transcripts that are submitted as part of a system-wide college application process.[19] Staff at the system's central office identify "college-preparatory" courses in key subjects from the transcripts and record the total number of college-preparatory units and average grades earned within each subject and overall. Thus, our high school measures for LUCCS include cumulative GPAs both overall and in the relevant subject; cumulative numbers of college-preparatory units completed, both overall and in the relevant subject; and indicators of whether any college-preparatory units were completed, both overall and in the relevant subject.

For SWCCS, our high school data come from an administrative data match to state-wide K–12 public school records (and thus are only available for students who attended a public school).[20] The high school measures we use for SWCCS are unweighted high school GPA, and from 11th- and 12th-grade transcripts, the total number of courses taken, the number of honors/advanced courses, the number of math courses, the number of English courses, the number of F grades received, and the total number of credits taken.

Results

SERs and Other Validity Metrics

Table 2 reports SERs and other validity metrics using alternative screening devices for remedial placement. Focusing first on the "test-scores" column, which simulates current policy at LUCCS and SWCCS, we see that one quarter to one third of tested students are severely misplaced depending upon the sample and subject. Recall that this does not imply that the remainder are all accurately placed, just that they are not *severely* misplaced. With the exception of the ACCUPLACER® math sample at SWCCS, severe underplacements are two to six times more prevalent than severe overplacements. In LUCCS, for example, nearly one in five students who take a math test, and more than one in four students who take the English tests, are placed into remediation even though they could have earned a B or better in the college-level course. This implies that nearly a quarter of remediated students in math (18.5/76.1), and one third of remediated students in English (28.9/80.5), are students who probably do not need to be there.

In all of our samples for both subjects, holding the remediation rate fixed but using measures of high school achievement instead of test scores to assign students results in both lower SERs and higher success rates among those assigned to college level. The reduction in SERs comes from reductions in both underplacements and overplacements, so unlike debates about where cutoffs should be optimally set, there is no trade-off here between these two types of errors. With the exception of math placement in LUCCS, the reductions are substantial, suggesting that out of 100 students tested, 4 to 8 fewer students would be severely misplaced, representing up to a 30% reduction in severe errors compared with test-based placements. Also with the exception of math placement in LUCCS, for which improvements are more modest, using high school achievement instead of test score results improves the success rate among those placed in college level by roughly 10 percentage points. For example, among students assigned directly to college level, the percentage earning at

TABLE 2

Predicted Severe Error Rates and Other Validity Metrics Using Alternative Measures for Remedial Assignment

	Measures used for remedial assignment					
	Test scores	HS GPA/ units[a]	Test + HS combined	Test scores	HS GPA/ units[a]	Test + HS combined
A. LUCCS sample	COMPASS® sample					
Math	n = 37,813					
Severe error rate	23.9	22.9	21.4	—	—	—
Severe overplacement rate	5.3	5.0	4.7	—	—	—
Severe underplacement rate	18.5	17.9	16.7	—	—	—
CL success rate (≥C), if assigned to CL	67.5	69.8	72.4	—	—	—
Remediation rate	76.1	74.7	74.7	—	—	—
English	n = 34,697					
Severe error rate	33.4	29.4	29.3	—	—	—
Severe overplacement rate	4.5	2.2	2.7	—	—	—
Severe underplacement rate	28.9	27.2	26.6	—	—	—
CL success rate (≥C), if assigned to CL	71.6	81.8	81.4	—	—	—
Remediation rate	80.5	79.8	79.8	—	—	—
B. SWCCS sample	COMPASS® sample			ACCUPLACER® sample		
Math	n = 4,881			n = 6,061		
Severe error rate	34.2	26.9	27.2	26.6	18.9	18.9
Severe overplacement rate	5.8	2.5	2.7	12.3	8.2	8.2
Severe underplacement rate	28.4	24.4	24.5	14.3	10.7	10.7
CL success rate (≥C), if assigned to CL	76.4	88.5	88.1	65.1	74.5	74.4
Remediation rate	68.5	70.0	70.0	54.0	55.0	55.0
English	n = 8,307			n = 6,573		
Severe error rate	26.2	19.6	19.6	33.5	26.9	26.8
Severe overplacement rate	8.8	4.9	5.0	5.6	2.7	2.6
Severe underplacement rate	17.3	14.7	14.6	27.8	24.3	24.2
CL success rate (≥C), if assigned to CL	72.6	82.4	82.4	76.0	86.4	86.5
Remediation rate	57.6	60.0	60.0	70.2	70.0	70.0

Source. Administrative data from LUCCS (2004–2007 entrants) and SWCCS (2008–2009 entrants).

Note. All figures in the table are percentages. The severe error rate is the sum of the percentage of students (a) placed into CL and predicted to fail there and (b) placed into remediation although they were predicted to earn a B in the CL. The CL success rate is the proportion of students assigned directly to college-level coursework in the relevant subject who are predicted to earn at least a C grade or better. The remediation rate is the percentage of all students assigned to remediation. HS = high school GPA = grade point average; LUCCS = large urban community college system; CL = college level; SWCCS = state-wide community college system.

[a]The measures included for HS GPA/Units varies for LUCCS and SWCCS due to data availability. For LUCCS, it includes cumulative GPA both overall and in the relevant subject; cumulative numbers of college-preparatory units completed, both overall and in the relevant subject; and indicators of whether any college-preparatory units were completed, both overall and in the relevant subject. For SWCCS, it includes unweighted cumulative GPA, and from 11th- and 12th-grade transcripts: the total number of courses taken, the number of honors/advanced courses, the number of math courses, the number of English courses, the number of F grades received, and the total number of credits taken.

least a C or better increases from 76% to 89% in the SWCCS COMPASS® sample, even though the same number of students are admitted.

Utilizing both test scores and high school transcript data for assignment generates the best placement outcomes at LUCCS, although the incremental improvement beyond using high

school data alone is small. At SWCCS, the combination yields no additional improvement beyond using high school information alone.[21]

Holding remediation rates fixed as we compare alternative screening tools is a useful benchmark, but it also limits the potential for major improvements particularly with respect to the severe underplacement rate. With remediation rates of 60% to 80%, it is possible that many students might be underplaced regardless of what screening device is used to select them. (Note that as the remediation rate approaches either 0% or 100%, the choice of screening device is irrelevant.) In an extension below, we examine our validity metrics across the full range of possible diagnostic thresholds for remediation.

Sensitivity Analysis: Excluding Low-Scoring Students

As noted above, one concern is that our underlying predictive models (expressed in Equations 1a and 1b) may not extrapolate to students far below the current test-score cutoffs. To address this concern, we re-run the entire analysis with very low-scoring students excluded from the sample.[22] These restrictions exclude approximately 25% to 50% of test takers depending upon the sample and subject.

The results are presented in Table 3. We first note that there are some level shifts in these validity metrics between Tables 2 and 3. For example, because we have explicitly excluded very low scorers, the remediation rates under current policy for these restricted samples are uniformly lower than those in Table 2. For the same reason, overplacement rates are higher and underplacement rates generally lower after low scorers are excluded, although the overall SERs remain very similar.

Overall, throwing out these low scorers does little to alter the pattern of findings from Table 2. Using high school achievement measures instead of test scores still improves both overall error rates and college-level success rates. And it is still the case that combining these two types of measures generates the best results in math at LUCCS, but for all other samples and subjects the combination provides little added value above using high school achievement alone.

Do Alternative Screening Tools Have Disparate Impacts by Gender or Race?

Even if high school transcript–based assignments are more accurate than test-based assignments on average, the use of high school transcripts might systematically disadvantage some students relative to others. In the spirit of Autor and Scarborough (2008), who examined the trade-offs between test accuracy and equity in the context of employment screening, we examine our validity metrics by gender and racial/ethnic identity for evidence of disparate impacts under alternative assignment rules. As with job screening tests, there is potentially an equity-efficiency trade-off in the choice of remedial screening tools if one tool more accurately identifies those likely to succeed, but as a result more minorities and/or females are placed in remediation.[23] Note that while we include gender and race/ethnicity in the underlying model predicting college-level outcomes (described in Equations 1a and 1b), we assume that these demographic factors cannot be used in any assignment rule. Thus, while we establish our cutoffs for the high school index and test-plus-high-school index at levels that keep the overall remediation rate fixed, the rate among any particular subgroup may change.

We present the results by gender in Table 4.[24] The first thing to note is that the pattern we found in Tables 2 and 3 holds within each gender subgroup as well: Using high school transcript data instead of test scores for placement would reduce the SER and increase college-level success rates for all subjects and samples; combining test scores and high school information would lead to additional incremental improvements in LUCCS math placement.

Nonetheless, there is some evidence of disparate impacts, in the direction that one might anticipate. Using high school information instead of test scores has the effect of decreasing the remediation rate for women but increasing it for men, for both SWCCS and LUCCS samples. This reinforces findings from prior research that men tend to do better on standardized tests while women tend to earn higher grades (see Hedges & Nowell, 1995).

Thus, even while high school transcript information may be more accurate for students

TABLE 3
Predicted Severe Error Rates and Other Validity Metrics, Restricting Analysis to Exclude Low-Scoring Students

	Measures used for remedial assignment					
	Test scores	HS GPA/ units	Test + HS combined	Test scores	HS GPA/ units	Test + HS combined
A. LUCCS sample	COMPASS® sample					
Math		*n* = 21,894				
Severe error rate	25.6	23.9	21.9	—	—	—
Severe overplacement rate	10.0	9.1	8.8	—	—	—
Severe underplacement rate	15.6	14.7	13.0	—	—	—
CL success rate (≥C), if assigned to CL	66.9	69.0	71.0	—	—	—
Remediation rate	56.4	54.7	54.5	—	—	—
English		*n* = 26,246				
Severe error rate	33.5	29.4	29.2	—	—	—
Severe overplacement rate	5.9	3.5	3.8	—	—	—
Severe underplacement rate	27.5	25.8	25.4	—	—	—
CL success rate (≥C), if assigned to CL	71.6	79.9	80.1	—	—	—
Remediation rate	74.2	74.7	74.7	—	—	—
B. SWCCS sample	COMPASS® sample			ACCUPLACER® sample		
Math		*n* = 2,431			*n* = 3,461	
Severe error rate	28.7	17.6	17.8	27.6	20.5	20.5
Severe overplacement rate	11.7	7.5	7.6	21.6	17.6	17.6
Severe underplacement rate	17.0	10.1	10.1	6.0	2.9	2.9
CL success rate (≥C), if assigned to CL	76.4	84.6	84.3	65.1	70.1	70.1
Remediation rate	36.7	35.0	35.0	19.4	20.0	20.0
English		*n* = 4,780			*n* = 3,333	
Severe error rate	25.2	17.3	17.4	29.8	20.6	20.6
Severe overplacement rate	15.3	12.0	12.1	11.7	6.7	6.7
Severe underplacement rate	9.9	5.3	5.3	18.1	13.9	13.9
CL success rate (≥C), if assigned to CL	72.6	78.2	78.2	76.1	84.5	84.6
Remediation rate	26.4	25.0	25.0	38.1	40.0	40.0

Source. Administrative data from LUCCS (2004–2007 entrants) and SWCCS (2008–2009 entrants).
Note. LUCCS: Math analysis excludes students scoring more than 10 points below the current test-score cutoff on either of the two math test modules. English analysis excludes students scoring more than 3 points below the current writing test-score cutoff or 10 points below the current reading test-score cutoff. SWCCS: Math and English analysis excludes students scoring more than 10 points below the current test-score cutoff on either of the math or English test modules, respectively. See Table 2 for additional notes. HS = high school, GPA = grade point average; LUCCS = large urban community college system; CL = college level; SWCCS = state-wide community college system.

of both genders, some may object to a policy change that impacts men and women differentially. At least at LUCCS, using the combined test-plus-high-school index for remedial assignment appears to be a win-win situation for both genders relative to the current test-score-based policy. Using the combined index for assignment would not raise the remediation rate for either subgroup relative to current policy, but would lower both over- and underplacements for both genders in both subjects, and would noticeably increase success rates for those placed directly into college-level work.[25] At SWCCS, using the combined measure moderates, but does not eliminate, the disparate impact on remediation rates by gender.

An online appendix provides the same analysis by race/ethnicity, focusing on LUCCS which has sufficiently large sample sizes within each subgroup. Again, we find that using high school

TABLE 4
Predicted Severe Error Rates Using Alternative Measures for Remedial Assignment, by Gender

	Men			Women		
	Test scores	HS GPA/ units	Test + HS combined	Test scores	HS GPA/ units	Test + HS combined
A. LUCCS (COMPASS®) sample						
Math		*n* = 15,814			*n* = 22,046	
Severe error rate	22.6	21.7	20.0	24.8	23.8	22.5
Severe overplacement rate	7.0	5.2	6.0	4.2	4.9	3.7
Severe underplacement rate	15.6	16.5	14.0	20.6	18.9	18.6
CL success rate (≥C), if assigned to CL	62.2	66.6	67.8	72.2	72.0	76.3
Remediation rate	73.4	76.2	72.7	78.1	73.7	76.2
English		*n* = 14,884			*n* = 19,924	
Severe error rate	29.5	26.3	25.8	36.2	31.8	31.9
Severe overplacement rate	4.5	2.2	2.7	4.4	2.2	2.7
Severe underplacement rate	25.0	24.1	23.0	31.8	29.5	29.2
CL success rate (≥C), if assigned to CL	67.1	79.7	78.5	74.3	83.0	83.2
Remediation rate	82.7	82.5	82.3	78.8	77.8	77.9
B. SWCCS (ACCUPLACER®) sample						
Math		*n* = 2,975			*n* = 3,086	
Severe error rate	27.0	19.3	19.3	26.2	18.4	18.5
Severe overplacement rate	14.7	7.8	8.0	10.0	8.6	8.5
Severe underplacement rate	12.3	11.5	11.3	16.2	9.8	10.0
CL success rate (≥C), if assigned to CL	59.9	71.3	71.1	70.7	76.7	76.9
Remediation rate	51.7	61.8	61.2	56.2	48.4	49.1
English		*n* = 3,220			*n* = 3,353	
Severe error rate	32.7	27.6	27.8	34.2	26.3	25.9
Severe overplacement rate	7.4	2.9	2.9	3.9	2.4	2.4
Severe underplacement rate	25.3	24.7	24.9	30.3	23.9	23.5
CL success rate (≥C), if assigned to CL	69.0	82.9	82.8	83.1	88.7	88.8
Remediation rate	69.4	75.7	76.1	71.1	64.5	64.1

Source. Administrative data from LUCCS (2004–2007 entrants) and SWCCS (2008–2009 entrants).
Note. All figures in the table are percentages. See Table 2 for additional notes. HS = high school; GPA = grade point average; LUCCS = large urban community college system; CL = college level; SWCCS = state-wide community college system.

information in combination with test scores maintains or reduces SERs and increases college-level success rates for all racial groups across all subjects. Even using high school information alone would reduce SERs for all groups and subjects except for Black students in English and Asian students in math. Again, however, we find that these improvements in accuracy must be weighed against disparate impacts on remediation rates, though the pattern of these disparate impacts is not always in the direction one might expect.[26] In math, using high school information instead of test scores lowers the remediation rate for Hispanic students by 7 percentage points and increases it for Asian students by 10 percentage points, though these changes are moderated by using the combined measure for placement. In English, using high school information would increase the remediation rate by 11 percentage points for Black students and reduce it for Asian students by nearly 25 percentage points relative to the current test-score-based policy.

Table 5 summarizes the consequences of these disparate impacts by simulating class compositions at LUCCS under our alternative

TABLE 5
Simulated Composition of College-Level Courses, Using Alternative Measures for Remedial Assignment (LUCCS Only)

	All tested students	Students placed in college-level (simulation)		
		Test scores	HS GPA/units	Test + HS combined
Math				
% female	58.2	53.4	60.6	54.8
% White	14.8	18.9	19.3	19.4
% Black	28.8	23.7	20.6	21.2
% Hispanic	34.2	22.3	30.8	26.0
% Asian	10.4	22.7	17.3	21.4
% Other/unknown race/ethnicity	11.8	12.5	12.0	12.1
Sample size	37,860	9,041	9,465	9,465
English				
% female	57.2	62.1	63.0	62.6
% White	13.4	17.9	18.4	20.9
% Black	28.1	31.2	14.6	19.5
% Hispanic	35.0	30.0	33.9	31.7
% Asian	12.0	8.2	22.8	15.9
% Other/unknown race/ethnicity	11.5	12.7	10.4	12.0
Sample size	34,808	6,787	6,962	6,962

Source. Administrative data from LUCCS (2004–2007 entrants).
Note. For comparison to subsequent columns, the first column provides the demographic breakdown of all students in our analysis sample (corresponding to the sample in columns 3 and 5 of Table 1). Subsequent columns indicate the simulated composition of college level classrooms that would result from the three alternative placement strategies. See Table 2 for additional notes. LUCCS = large urban community college system; GPA = grade point average; HS = high school.

screening devices. If high school information were used for screening instead of test scores, college-level math classes would have substantially higher proportions of female and Hispanic students; however, representation of Black and Asian students would fall. In college-level English, switching to high school achievement would not change the gender composition, but representation of Black students would fall by half (from 31% to 15%) and Asian students' representation would more than double (from 8% to 23%). These compositional changes are moderated, but not eliminated, when a combined measure of test scores and high school achievement is used for placement.

Optimal Cutoffs: Trading Off Underplacement and Overplacement

So far, we have presented results that compare alternative screening devices while holding the overall percentage of students remediated fixed at current levels. But in considering the optimal screening policy, the diagnostic threshold can be allowed to vary along with the instrument used, allowing for greater potential improvements in accuracy. For a given instrument, if policymakers weight overplacement and underplacement errors equally, then the optimal instrument and cutoff can be chosen to minimize the overall SER.

Figure 3 shows the overall SERs, underplacement and overplacement rates for math using alternative screening instruments in both LUCCS and SWCCS. As the percentile cutoff increases, increasing proportions of students are assigned to remediation and so underplacement rates grow sharply and overplacement rates fall. Error rates for alternative instruments must converge at both the high and low end of the potential cutoff range when either no students or all students are assigned to remediation.[27]

In math (Panels A and B), the SERs using test scores alone are higher than under the alternative instruments we simulate, except for very low cutoffs. Using high school achievement alone or in addition to test scores reduces SERs most noticeably for cutoffs between the 50th and 80th percentiles. If policymakers cared only about the SER, the optimal policy would be to

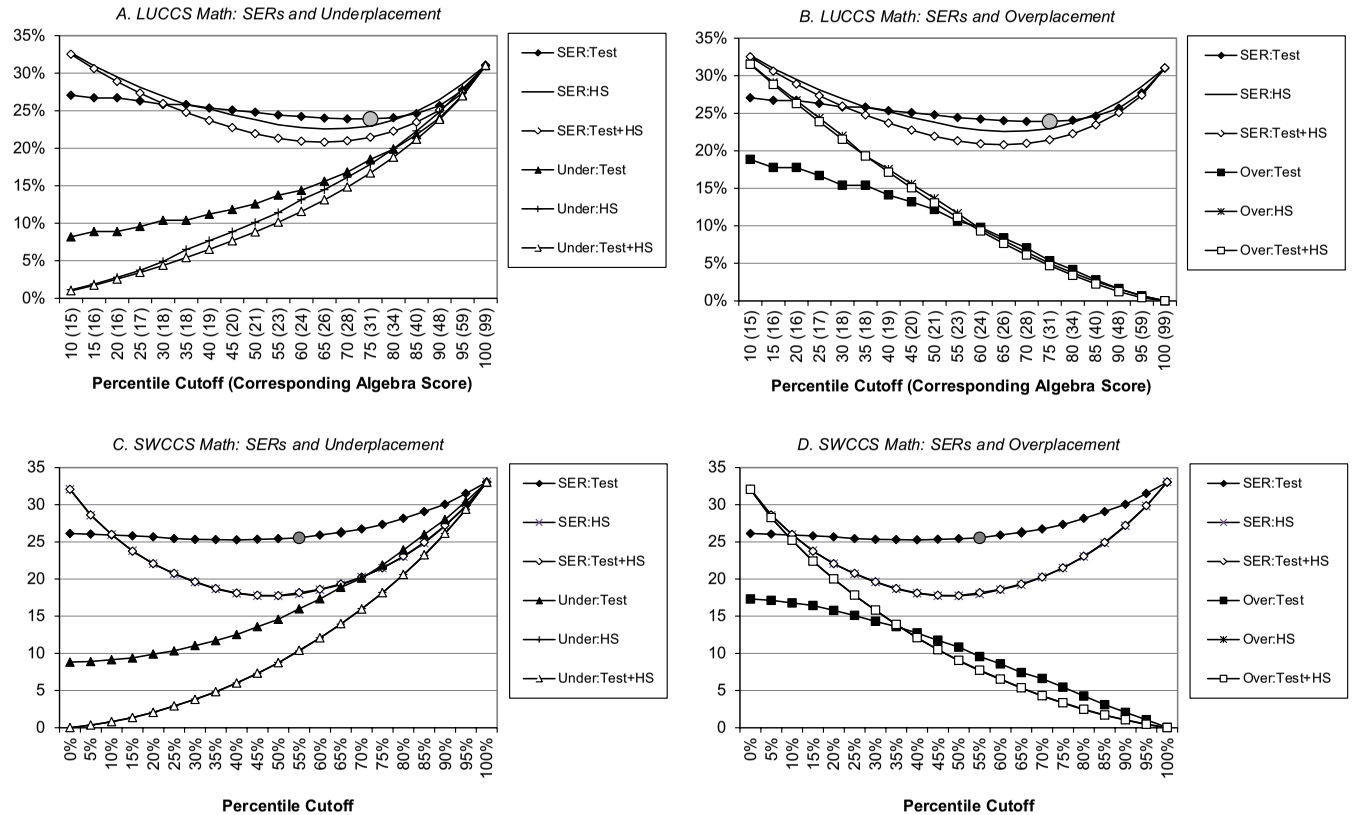

FIGURE 3. *Assignment outcomes by SERs and overplacement rates by simulated cutoff.*
Source. Administrative data from LUCCS (2004–2007 entrants) and SWCCS (2008–2009 entrants).
Note. Gray dot indicates simulated current policy. For LUCCS, test-only results are based on varying the more binding algebra test cutoff while the pre-algebra cutoff is fixed at the current cutoff of 30. The fixed pre-algebra cutoff explains why the test-only results begin to diverge sharply from the HS-only and Test + HS results for lower simulated algebra cutoffs: Even when the algebra cutoff is very low, the fixed pre-algebra cutoff will continue to assign students to remediation, increasing underplacements but decreasing overplacements relative to HS-only and Test + HS models with similarly low cutoffs. For SWCCS, test-only results are based on varying the algebra test cutoff while the arithmetic cutoff is fixed at the current cutoff of 55. Note that for SWCCS the results for HS-only and Test + HS are virtually indistinguishable, making it appear that there are fewer lines in the figure than indicated by the legend. SERs = severe error rates; LUCCS = large urban community college system; SWCCS = state-wide community college system; HS = high school.

assign students based on the combined test-plus-high-school-achievement index with a cutoff at the 65th percentile. This policy would reduce the SER by 3.1 percentage points (13%) while actually slightly *improving* the success rate among those placed into college level (not shown), but perhaps the most notable difference is that it would achieve these outcomes with a remediation rate 10 percentage points lower than the rate under current policy.

Interestingly, current policy at both LUCCS and SWCCS—indicated by the large gray circular marker on the test-only line—appears to be near the SER-minimizing level for the test-only instrument; however, the test-only SER line is relatively flat around the current cutoffs. As these percentile cutoffs roughly correspond to remediation rates (this correspondence is exact for our two alternative instruments, which are one-dimensional indices), this implies that a very wide range of remediation rates can generate similar SERs.

In an online appendix figure, we show findings that are even more striking for English. For example, at both LUCCS and SWCCS, utilizing the high school achievement index with a cutoff at just the 35th percentile could reduce remediation rates by 35 to 45 percentage points while also reducing the SER by 10 to 17 percentage points and holding the college-level success rate essentially flat. Moreover, the figures indicates that using the current test-score-based instrument, the SER-minimizing policy in English would be to admit virtually everyone to college level (though the SER is flat between the 5th and 35th percentiles).

Institutions' choice of cutoff can reveal information about how they perceive the costs of different types of assignment errors. In math, the test-score-only SER is flat across a wide range of cutoffs, but both systems choose a cutoff near the top of this range; in English, both systems choose a cutoff higher than the SER-minimizing level. This suggests that institutions perceive the costs of overplacement to be significantly higher than the costs of underplacement.

Discussion

Our results underscore the reality that it is difficult to predict who will succeed in college by any means: Regardless of the screening tool we examine, one fifth to one third of students are likely to be severely misplaced. Yet among a set of feasible, if imperfect screening devices, high school transcript information is at least as useful as and often superior to placement test scores. In both math and English, using high school GPA/units alone as a placement screen results in fewer severe placement mistakes than using test scores alone (with error reductions of 12% to 30% relative to test scores, in all samples/subjects except LUCCS math). There is no assignment trade-off: Both underplacement and overplacement errors can be reduced, and the success rate in college-level courses increased, without changing the proportion of students assigned to remediation. At LUCCS, these errors are further reduced when placement tests and high school information are used in combination, while at SWCCS we find that placement tests have little incremental value if high school information is already available. Our results are not driven by the predicted outcomes for very low-scoring students (for whom our model relies more heavily on extrapolation); the pattern of findings holds even when these students are excluded.

One potential explanation for the limited utility of placement exams is that they are simply quite short (taking just 20–30 minutes per module) and thus very noisy, as noted above. Another possible factor may be a disconnect between the limited range of material tested on the exam and the material required to succeed in the typical first college-level course (Jaggars & Hodara, 2011). For example, ACT's own (2006) analysis suggests that the COMPASS algebra exam is more accurate for predictions of success in "college algebra" versus "intermediate algebra," but many students meet their college-level math requirement by taking courses that are not primarily algebra-based, such as introductory statistics. In comparison, high school transcript information may be both less noisy (because it is accumulated over years instead of minutes), and may capture broader dimensions of college readiness, such as student effort and motivation.

Compared with current test-score-based policies, using high school information for remedial assignment not only reduces severe placement errors overall but also within each racial/ethnic

and gender subgroup we examine. Despite these universal improvements in accuracy, some subgroups in some subjects do better on the tests while others do better on a high school achievement index—meaning that the choice of screening device has implications for the gender and racial/ethnic composition of college-level courses. For example, if the remediation rate is held fixed, then switching to assignment based on high school information only would increase the representation of women and Hispanics in college-level math at the expense of men, Black, and Asian students; while in college-level math, the switch would dramatically increase the representation of Asian students while lowering representation of Black students. Using a combined measure for placement could moderate the disparate impacts of this potential policy shift. An alternative approach to addressing these disparate impacts would be to use high school information but lower the cutoffs such that no subgroup would face a higher remediation rate.

Our findings provide new insights regarding how institutions weigh overplacement errors versus underplacement errors. Faculty and institutions may be aware of the obvious connection between course failure and student dropout but may have failed to consider that the discouragement of underplaced students may similarly increase the risk of dropout. While the overplacement problem—students admitted to college-level courses even though they end up failing there—is well known and much discussed, we find that severe underplacements are actually far more common. Our estimates suggest that one quarter to one third of students assigned to remediation could have earned a B or better had they been admitted directly to college-level work. Moreover, we find evidence that institutions could substantially lower their remediation rates without increasing the SER. That they have not done so—in fact LUCCS has increased its cutoffs recently—suggests that institutions are more concerned about minimizing overplacements than underplacements.

This may be because the costs of overplacement fall not just on the overplaced student (who may be discouraged and/or risk losing financial aid eligibility) but also on faculty members who dislike having to fail students, as well as on other students in the college-level course who may experience negative peer effects. The costs of underplacement, in contrast, fall primarily on the institution and the underplaced student. Moreover, overplacements may simply be easier to observe: It is straightforward to document how many students are placed into a college-level course fails there, while underplacements can only be estimated statistically.

The apparently greater weight given to overplacements also appears consistent with the financial incentives of colleges. These incentives depend on the cross-subsidy (revenues minus costs) between remedial and college-level courses. In most states, revenues through state aid formulas are equal across remedial and college-level courses, although for six states the funding formula is more generous for remedial courses (in only three states, it is less generous). Very few states provide data on the costs of remedial courses specifically, although these courses are more often taught by lower paid faculty and use limited technology. However, data for Ohio's 2-year colleges show that remedial courses cost 9% less than college-level courses. It thus seems quite possible that remedial courses subsidize college-level courses, giving colleges an implicit incentive to underplace students.[28] If so, colleges may face a financial constraint if remediation rates are reduced without any additional resources provided.

Finally, our findings have implications for the interpretation of prior estimates of the impact of remedial assignment, which are largely based upon RD designs. First, the relatively low predictive validity of placement exam scores (the running variable in RD studies) suggests that RD estimates may generalize beyond just students scoring near the cutoffs. This is an important conclusion, because a common critique of prior null-to-negative impact estimates has been that these estimates are local to students scoring near the cutoff, and that students well below the cutoff may experience more positive effects. However, even if test scores were as good as random—meaning that the existing null-to-negative RD estimates could be interpreted as global average treatment effects—this would not rule out the possibility

of heterogeneity in treatment effects. It may simply be that treatment effects vary along some dimension other than test scores. Indeed, Scott-Clayton and Rodriguez (2012) provide evidence using LUCCS data that RD estimates of the impact of remediation are more negative for subgroups identified as low risk on the basis of high school transcript data. It is possible that there are positive impacts of remediation for some subset of students who are underprepared, but that current policies simply catch too many prepared students in a widely cast remedial net.

Acknowledgments

The authors gratefully acknowledge the community college personnel for access to the data; research support from Olga Rodriguez, Michelle Hodara, and Emma Garcia; comments from Davis Jenkins and Tom Bailey; and editorial assistance from Betsy Yoon and Doug Slater.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding was provided by the Bill and Melinda Gates Foundation and the National Center for Postsecondary Research, Teachers College, Columbia University.

Notes

1. Authors' calculations based on beginning postsecondary student (BPS) 2009 data (National Center for Education Statistics [NCES], 2012). Bachelor's degree attainment rates are 59% for those entering with a 4-year-degree goal, and bachelor's/associate's degree attainment rates are 30% for those entering with a 2-year-degree goal.

2. Estimate based on BPS:2009 transcript data for 2003–2004 entrants (NCES, 2012). Estimates based on student self-reports are substantially lower, potentially because students do not realize the courses are remedial.

3. This estimate is based on first-time degree-seeking fall enrollees (Snyder & Dillow, 2012, Table 207). We estimate a cost of roughly US$1,620 per student per remedial course, making the assumption that each course is equivalent to a three-credit course or roughly 1/8th of a full-time year of college, and assuming the costs are comparable with the costs at public 2-year colleges which have total expenditures of US$12,957 per full-time equivalent (FTE) per year (Desrochers & Wellman, 2011). With an average of 1.3 remedial courses per entrant, this implies costs of 1.3 course × US$1,620 per course × 3.1 million students = US$6.7 billion annually.

4. Effects may also vary by group characteristics such as age and gender; however, in this article, we focus primarily on ability as predicted by test scores and prior high school achievement.

5. Both of these studies find some evidence that regression-discontinuity (RD) estimates are more negative when cutoffs fall lower in the ability distribution; Scott-Clayton and Rodriguez (2012) also use preexisting characteristics to examine impacts for high- and low-academic risk students who all score around the same test-score cutoff.

6. Both systems requested confidentiality in exchange for permission to freely analyze and report on the data.

7. However these decisions are made, they are increasingly made at a system- or even state-wide level (Hodara, 2012).

8. Some colleges may "mix and match" so these numbers do not add to 100%; moreover, ACT and SAT scores are also for placement at some colleges.

9. Scores on the COMPASS® algebra exam may be determined by as few as eight questions (ACT, 2006).

10. We recognize that a test per se cannot be validated: It is its use in a given context that is validated (Brennan, 2006). We focus here on screening devices for course placement in math and English, under the hypothesis that if the tests are not valid for placement in their own subject, they are unlikely to be valid for placement in other subjects less directly related to the material on the exams.

11. Statistical corrections that are sometimes employed in an effort to address this type of bias may themselves rely on implausible assumptions (Rothstein, 2004). This range restriction also introduces an extrapolation problem: It is not obvious that the relationships between predictors and outcomes will be the same for students above and below the cutoff for remediation. We discuss this extrapolation problem further at the end of our "Method" section.

12. For example, researchers studying the accuracy of an automated Pap smear test in the 1950s analyzed rates of false-positive and false-negative classifications for a range of possible diagnostic thresholds, then used this information to determine the optimal threshold (Lusted, 1984). The automated Pap smear test was analyzed using something similar to receiver-operating

characteristic (ROC) plots, which for any given diagnostic threshold, plot what proportion of the healthy are falsely identified as sick against what proportion of the sick are correctly identified as such.

13. We group withdrawals and incompletes as failures given evidence that these outcomes are grade related (Ang & Noble, 1993).

14. Analyzing the relationship between pretreatment predictors and grades for those who took remediation could confound the estimates for two reasons: (a) The remedial treatment may effectively eliminate skill deficiencies or (b) the only remediated students who make it to college-level courses may have high levels of unobserved motivation.

15. We do not use reading/writing test scores in our predictive model for college math grades or vice versa because this would require limiting the sample to students who took tests in both subjects, and the incremental predictive power of the cross-subject test scores was comparatively small.

16. Because we are ultimately interested in estimating overall error rates and not in predicting individual outcomes per se, the inclusion of these demographic variables turns out to make virtually no difference to our full-sample estimates of our validity metrics. Full regression results are available upon request.

17. So, for example, to select the cutoff in math using high school information, we regress college-level math grades (among only those assigned directly to college level) on the set of high school achievement variables described above and establish the cutoff as the 75th percentile on this index of predicted grades. Note that while additional transcript information improves predictive power, the overall cumulative high school grade measure is by far the most powerful single component.

18. Despite the differences in transcript measures available for each system, it is worth noting that in both cases the overall high school grade measure is the driving component of the high school achievement index.

19. Students who simply show up on a given campus are known as "direct admits" and typically have much more limited background information available in the system-wide database.

20. Although most of these students had both grade point average (GPA) and detailed transcript data, for some we only had GPA information. Differences between our sample and students without high school GPAs were not large.

21. In some cases, the combination actually appears to do marginally worse than using high school data alone, which can result if test scores are extremely noisy.

22. For large urban community college system (LUCCS), the math analysis excludes students scoring more than 10 points below the current test-score cutoff on either of the two math test modules. English analysis excludes students scoring more than 3 points below the current writing test-score cutoff or 10 points below the current reading test-score cutoff. For state-wide community college system (SWCCS), the math and English analysis excludes students scoring more than 10 points below the current test-score cutoff on either of the math or English test modules, respectively.

23. There are two differences with our context, however: First, in our setting, the test-score-based policy is the default already in place, and we examine replacing or augmenting this with additional quantitative, externally verifiable measures (as opposed to a version of managerial discretion). Second, as 85% of LUCCS testers are minorities (with roughly 30% Black, 34% Hispanic, and 10% Asian), any disparate impacts are likely to be between minority groups rather than between minorities and non-Hispanic Whites.

24. For brevity, we show only the LUCCS COMPASS® and SWCCS ACCUPLACER® results to demonstrate the consistency across samples/exams. The patterns for the SWCCS COMPASS® sample are very similar.

25. This slight decline in the remediation rate when using alternative assignment rules is also reflected in the full-sample results in Table 3; it reflects the fact that we cannot set the cutoff at a point that will precisely preserve the original 76.1% remediation rate in math and 80.5% rate in English.

26. Our results reinforce Autor and Scarborough's (2008) observation that just because a group has lower test scores in general does not necessarily mean that they are disadvantaged when tests are used as a screening device; it depends where group members would fall in the distribution of the alternative measure that would be used instead.

27. In our figures, this is complicated by the fact that the current test-only placement rule is actually based upon two subscores, only one of which we allow to vary here—we hold the easier pre-algebra test cutoff fixed at its current level, which matters only at the very low range of algebra cutoff scores. Because we hold the pre-algebra cutoff fixed, even with a very low algebra cutoff high proportions of students will be assigned to remediation, which tends to increase underplacements but limits overplacements, as reflected in Figure 3.

28. For funding formulae, see http://faccc.org/research/FTEspending_bystate.pdf. For costs of remediation in Ohio, see http://regents.ohio.gov/perfrpt/special_reports/Remediation_Consequences_2006.pdf.

References

ACT. (2006). *COMPASS/ESL reference manual*. Iowa City, IA: Author.

Ang, C. H., & Noble, J. P. (1993). *The effects of alternative interpretations of incomplete and withdrawal grades on course placement validity indices* (Research Report No. 93-3). Iowa City, IA: American College Testing.

Attewell, P., & Domina, T. (2008). Raising the bar: Curricular intensity and academic performance. *Educational Evaluation and Policy Analysis, 30*, 51–71.

Autor, D. H., & Scarborough, D. (2008). Does job testing harm minority workers? Evidence from retail establishments. *Quarterly Journal of Economics, 123*, 219–277.

Bailey, T., Jeong, D. W., & Cho, S.-W. (2010). Referral, enrollment and completion in developmental education sequences in community colleges. *Economics of Education Review, 29*, 255–270.

Belfield, C., & Crosta, P. (2012). *Predicting success in college: The importance of placement tests and high school transcripts* (CCRC Working Paper No. 42). New York, NY: Community College Research Center.

Bettinger, E. P., Evans, B. J., & Pope, D. G. (2011). *Improving college performance and retention the easy way: Unpacking the ACT exam* (NBER Working Paper No. 17119). Cambridge, MA: National Bureau of Economic Research.

Bettinger, E. P., & Long, B. T. (2009). Addressing the needs of underprepared students in higher education: Does college remediation work? *Journal of Human Resources, 44*, 736–771.

Boatman, A., & Long, B. T. (2010). *Does remediation work for all students? How the effects of postsecondary remedial and developmental courses vary by level of academic preparation* (NCPR Working Paper). New York, NY: National Center for Postsecondary Research.

Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton, NJ: Princeton University Press.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Westport, CT: ACE/Praeger Publishers.

Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance* (NBER Working Paper No. 14194). Cambridge, MA: National Bureau of Economic Research.

Carrell, S., Fullerton, R., & West, J. E. (2009). Does your cohort matter? Measuring peer effects in college achievement. *Journal of Labor Economics, 27*, 439–464.

College Board. (2007). *ACCUPLACER coordinator's guide*. New York, NY: College Board.

Dadgar, M. (2012). Essays on the economics of community college students' academic and labor market success (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. [3506175])

Desrochers, D. M., & Wellman, J. V. (2011). *Trends in college spending, 1999-2009. A report of the Delta Cost Project*. Washington, DC: Delta Project on Postsecondary Education Costs, Productivity, and Accountability.

Fain, P. (2013, June 5). Remediation if you want it. *Inside Higher Ed*. Retrieved from http://www.insidehighered.com/news/2013/06/05/florida-law-gives-students-and-colleges-flexibility-remediation

Fields, R., & Parsad, B. (2012). *Tests and cut scores used for student placement in postsecondary education: Fall 2011*. Washington, DC: National Assessment Governing Board.

Gamoran, A., & Hannigan, E. C. (2000). Algebra for everyone? Benefits of college-preparatory mathematics for students with diverse abilities in early secondary school. *Educational Evaluation and Policy Analysis, 22*, 241–254.

Greene, J. P., & Forster, G. (2003). *Public high school graduation and college readiness rates in the United States* (Manhattan Institute Education Working Paper No. 3). New York, NY: Center for Civic Innovation, Manhattan Institute for Policy Research.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41–45.

Hodara, M. (2012). Language minority students at community college: How do developmental education and English as a second language affect their educational outcomes? (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. [3505981])

Jaggars, S. S., & Hodara, M. (2011). *The opposing forces that shape developmental education: Assessment, placement, and progression at CUNY community colleges* (CCRC Working Paper No. 36). New York, NY: Community College Research Center.

Long, M. C., Conger, D., & Iatarola, P. (2012). Effects of high school course-taking on secondary and postsecondary success. *American Education Research Journal, 49*, 285–322.

Long, M. C., Iatarola, P., & Conger, D. (2009). Explaining gaps in readiness for college-level math: The role of high school courses. *Education Finance and Policy, 4*, 1–33.

Lusted, L. B. (1984). ROC recollected [Editorial]. *Medical Decision Making, 4*, 131–135.

Martorell, P., & McFarlin, I. J. (2011). Help or hindrance? The effects of college remediation on academic and labor market outcomes. *Review of Economics and Statistics, 93*, 436–454.

Mattern, K. D., & Packman, S. (2009). *Predictive validity of ACCUPLACER scores for course placement: A meta-analysis* (Research Report No. 2009-2). New York, NY: College Board.

National Center for Education Statistics. (2012). *2004/09 Beginning Postsecondary Students Longitudinal Study (BPS:04/09) restricted use data and codebook*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Noble, J. P., & Sawyer, R. L. (2004). Is high school GPA better than admissions test scores for predicting academic success in college? *College & University Journal, 79*, 17–23.

Parsad, B., Lewis, L., & Greene, B. (2003). *Remedial education at degree-granting postsecondary institutions in fall 2000: Statistical analysis report* (NCES 2004-101). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics, 121*, 297–317.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics, 116*, 681–704.

Sawyer, R. (1996). Decision theory models for validating course placement tests. *Journal of Educational Measurement, 33*, 271–290.

Scott-Clayton, J. (2012). *Do high stakes placement exams predict college success?* (CCRC Working Paper No. 41). New York, NY: Community College Research Center.

Scott-Clayton, J., & Rodriguez, O. (2012). *Development, discouragement, or diversion? New evidence on the effects of college remediation* (National Bureau of Economic Research Working Paper No. 18328). Cambridge, MA: National Bureau of Economic Research.

Snyder, T. D., & Dillow, S. A. (2012). *Digest of education statistics 2011* (NCES 2012-001). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Venezia, A., Bracco, K. R., & Nodine, T. (2010). *One shot deal? Students' perceptions of assessment and course placement in California's community colleges*. San Francisco, CA: WestEd.

Winston, G. C., & Zimmerman, D. J. (2004). Peer effects in higher education. In C. Hoxby (Ed.), *College choices: The economics of where to go, when to go, and how to pay for it* (pp. 395–423). Chicago, IL: National Bureau of Economic Research and University of Chicago Press.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics, 85*, 9–23.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*, 561–577.

Authors

JUDITH SCOTT-CLAYTON, PhD, is an assistant professor of economics and education at Teachers College, Columbia University. She is also a faculty research fellow of the National Bureau of Economic Research and senior research associate at the Community College Research Center. Her interests are in labor economics and higher education policy.

PETER M. CROSTA, PhD, is the director of research at 2U, Inc.He previously served as data scientist at the Community College Research Center, Teachers College, Columbia University. His research interests within the economics of education include quantitative analysis of educational policy change and the visualization of educational data.

CLIVE R. BELFIELD, PhD, is an associate professor of economics at Queens College, City University of New York. He is also co-director of the Center for Benefit-Cost Analysis at Teachers College, Columbia University, and research fellow at the Community College Research Center. His interests are in economic evaluation and cost–benefit analysis of education.

Manuscript received May 21, 2013
Revision received September 10, 2013
Accepted November 1, 2013