Multilevel Issues in the Application of Propensity Score Matching

Simulations and Results for Evaluation of the Advanced Placement Program

Brian F. Patterson, The College Board

In “Complexities Encountered When Modeling Multilevel Data”

Annual Meeting of the American Educational Research Association

April 17th, 2012

CollegeBoard

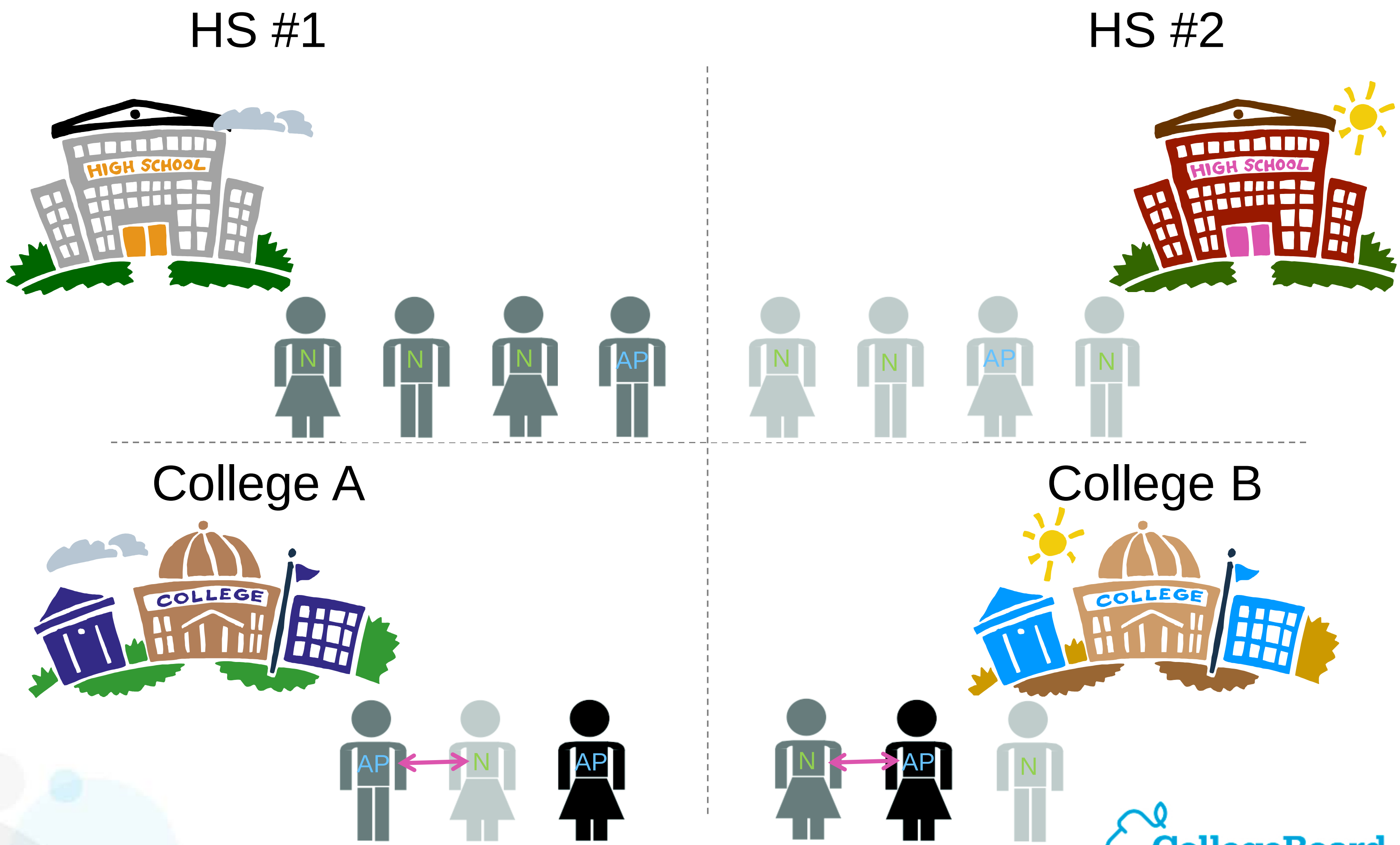
Outline

1. Context for the Empirical Study
2. Simulation to Inform the Empirical Study
3. Some Preliminary Results from the Empirical Study

1.1 Current Empirical Study

- Goal: Estimate unbiased effect of Advanced Placement (AP) on related college grades
  - Propensity score methods may reduce bias
- Problem: Propensity for taking AP varies across high schools, even after conditioning on student characteristics
  - We are unsure of the consequences on our conclusions of ignoring such dependence within high schools
- Solution: Estimate multilevel propensity score model with random high school effects

1.2 Picturing the Empirical Study



CollegeBoard

1.3 If we could design the perfect experiment…

- We might:
  - take a cluster sampling approach to selecting a representative set of high schools;
  - randomly assign students of a variety of ability levels to take the Advanced Placement (AP) course;
  - follow all students to their college of choice and:
    - assign non-AP to take intro and subsequent course; and
    - assign AP to skip the intro and take the sequent.
- We hope to find that the AP group tended to perform at least as well as the non-AP

CollegeBoard

1.4 Choosing to Participate in AP

- Construct a model of propensity for AP participation
- Potentially important predictors of AP participation
  - Academic achievement
  - Subject area interest
  - Achievement motivation
  - Opportunities for participation
  - High school atmosphere (e.g., college-focused; pro-AP)

2.1 Existing Research

- Griswold, Localio, and Mulrow (2010)
  - Compared: ignoring clusters; within-cluster; and multilevel match
- Arpino and Mealli (2011)
  - Fixed cluster effects superior to either random or no effects
  - No normality assumption for cluster effects
- Vanderweele (2008)
  - Ignorability & stable unit assumptions for cluster-level treatment
- Outside the multilevel context, see:
  - Rosenbaum & Rubin — foundational propensity score theory
  - Peter Austin — recent simulations & best practice

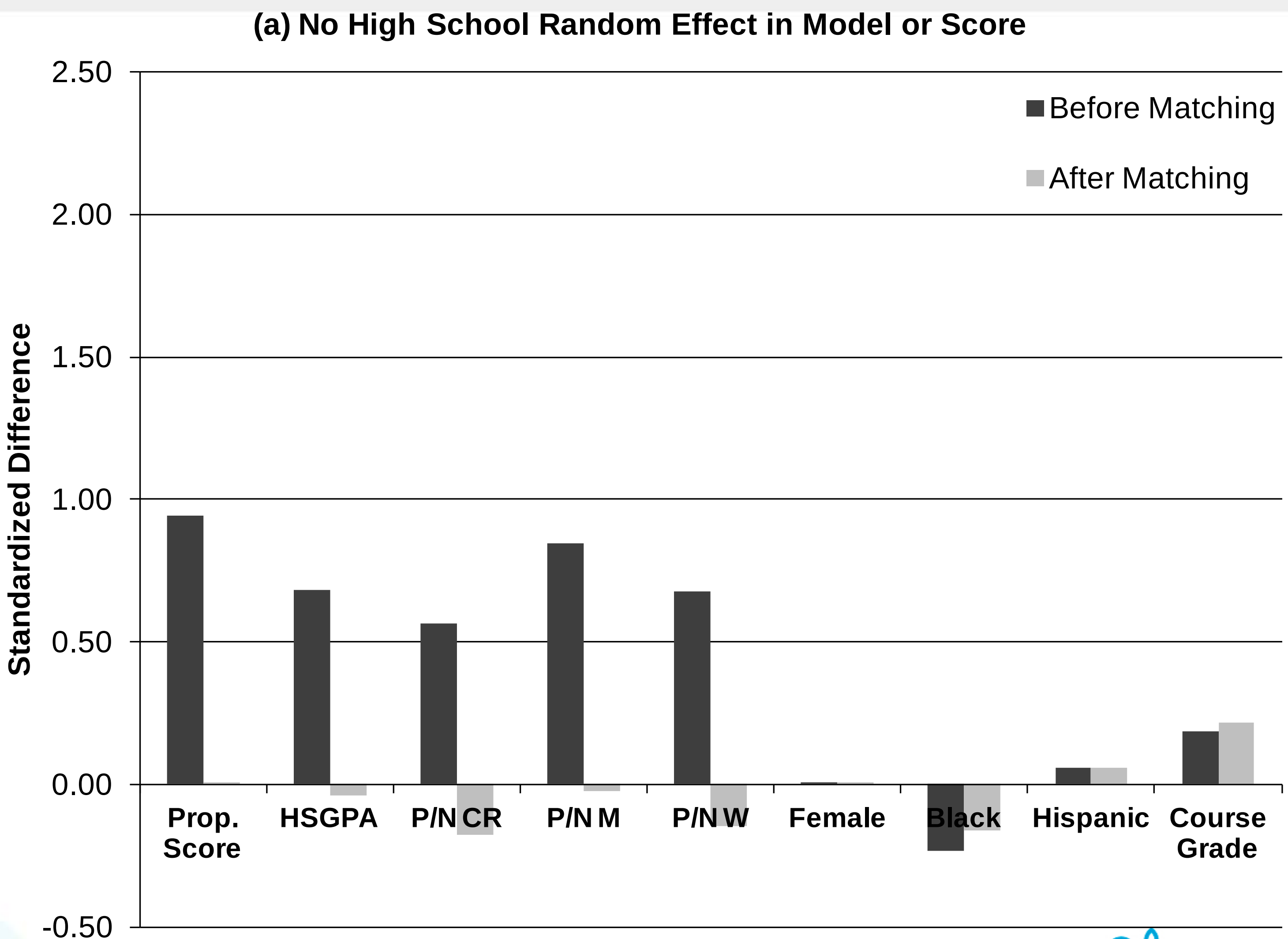
2.2 What about College Effects?

- Ignore college effects in estimating prop. scores
  - Data not cross-classified until students enter college
- Since the outcome is at the college level…
  - Only match AP- and non-AP-examinees:
    - at the same college; and
    - who took the same subsequent course.
  - Referred to as exact matching on these variables
  - Do not require that students attended the same high school

2.3 Some Notes on Propensity Score Matching Procedure

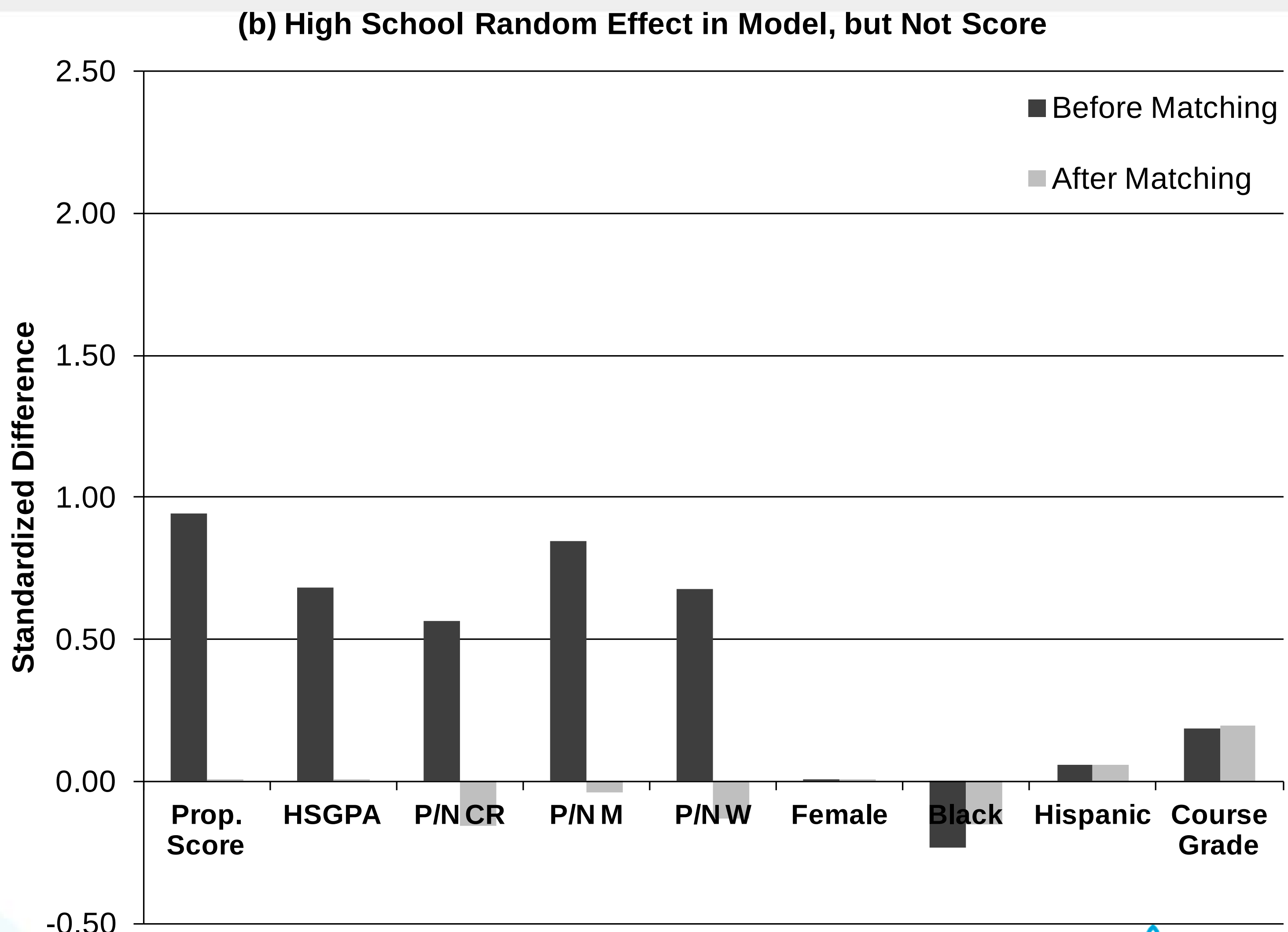
- Greedy matching
  - As opposed to optimal matching
- Within calipers
  - Caliper size = 0.2 * Population SD(propensity score)
- On logistic scale
  - As opposed to probability scale; avoids scale issues
- Use BLUP predicted propensity score?
  - Simulations will examine effects of either including or excluding predicted random intercept effect

2.4 Example w/ $\tau = 6$, No RE



Replicate #59 from condition 4 simulated on 2012-03-13

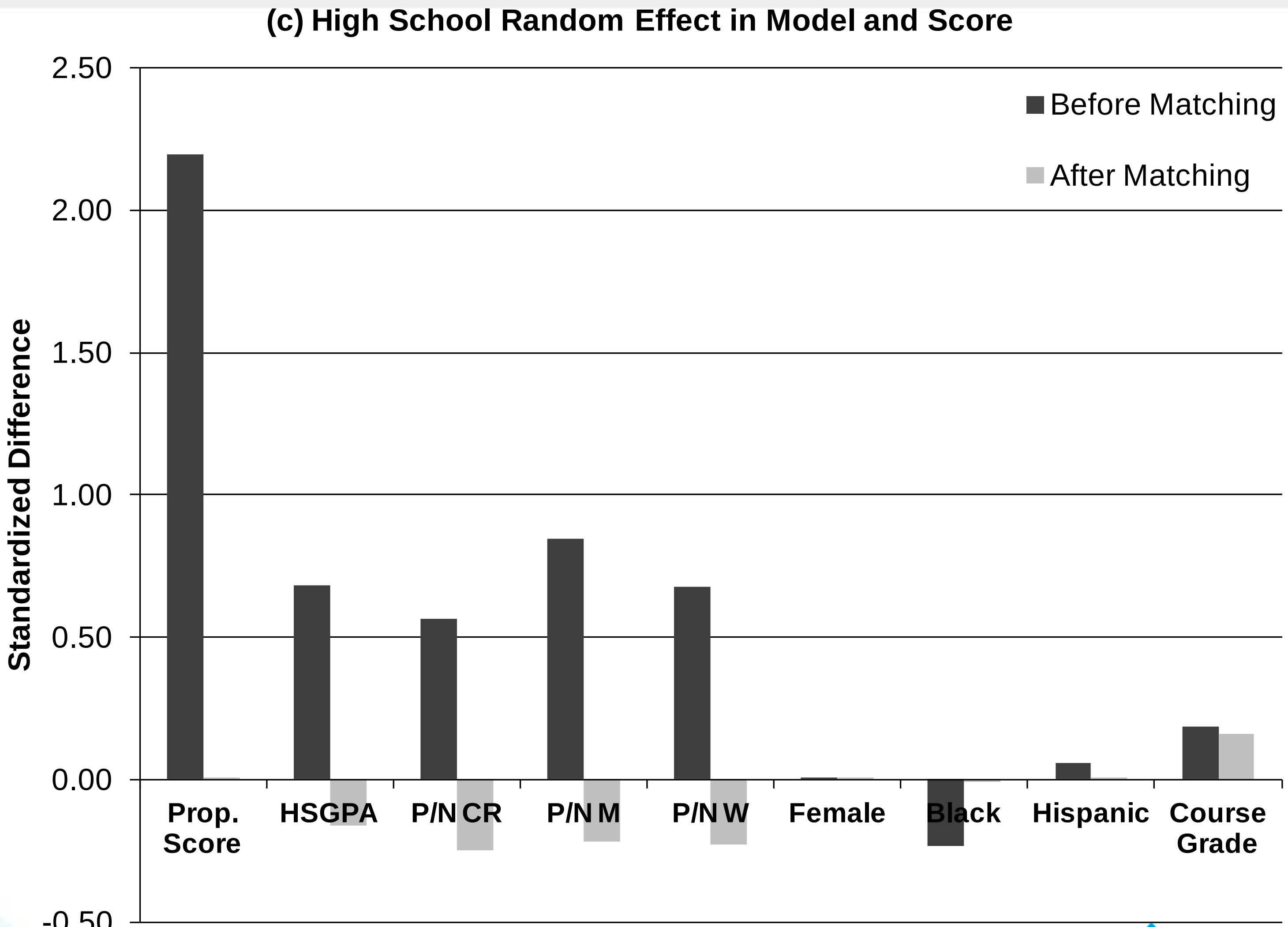
2.5 Example w/ $\tau = 6$, Model RE, Not in PS



Replicate #59 from condition 4 simulated on 2012-03-13

CollegeBoard

2.6 Example w/ τ = 6, Model RE, Inc. in PS

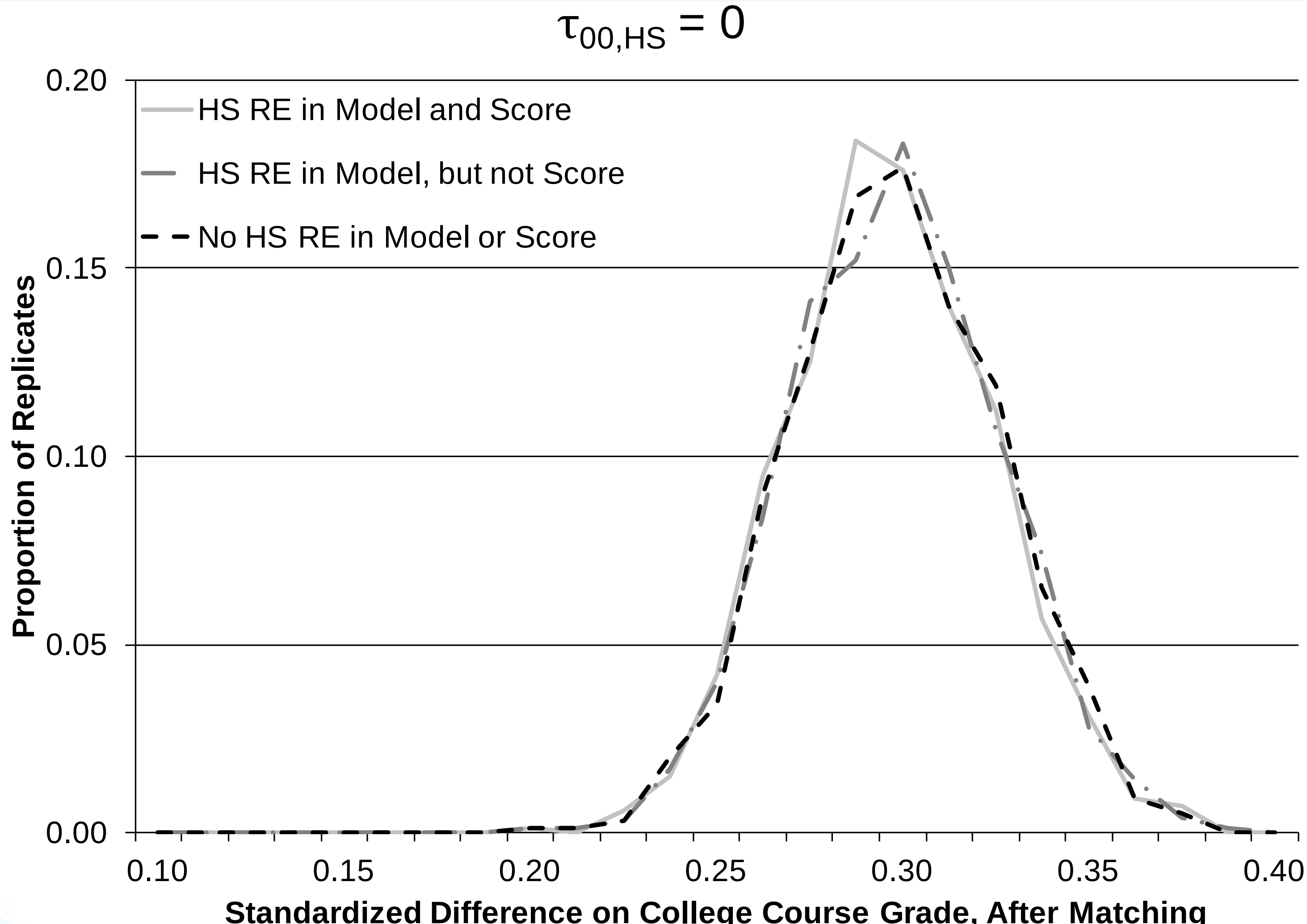


Replicate #59 from condition 4 simulated on 2012-03-13

2.7 Comments on Example Replicate

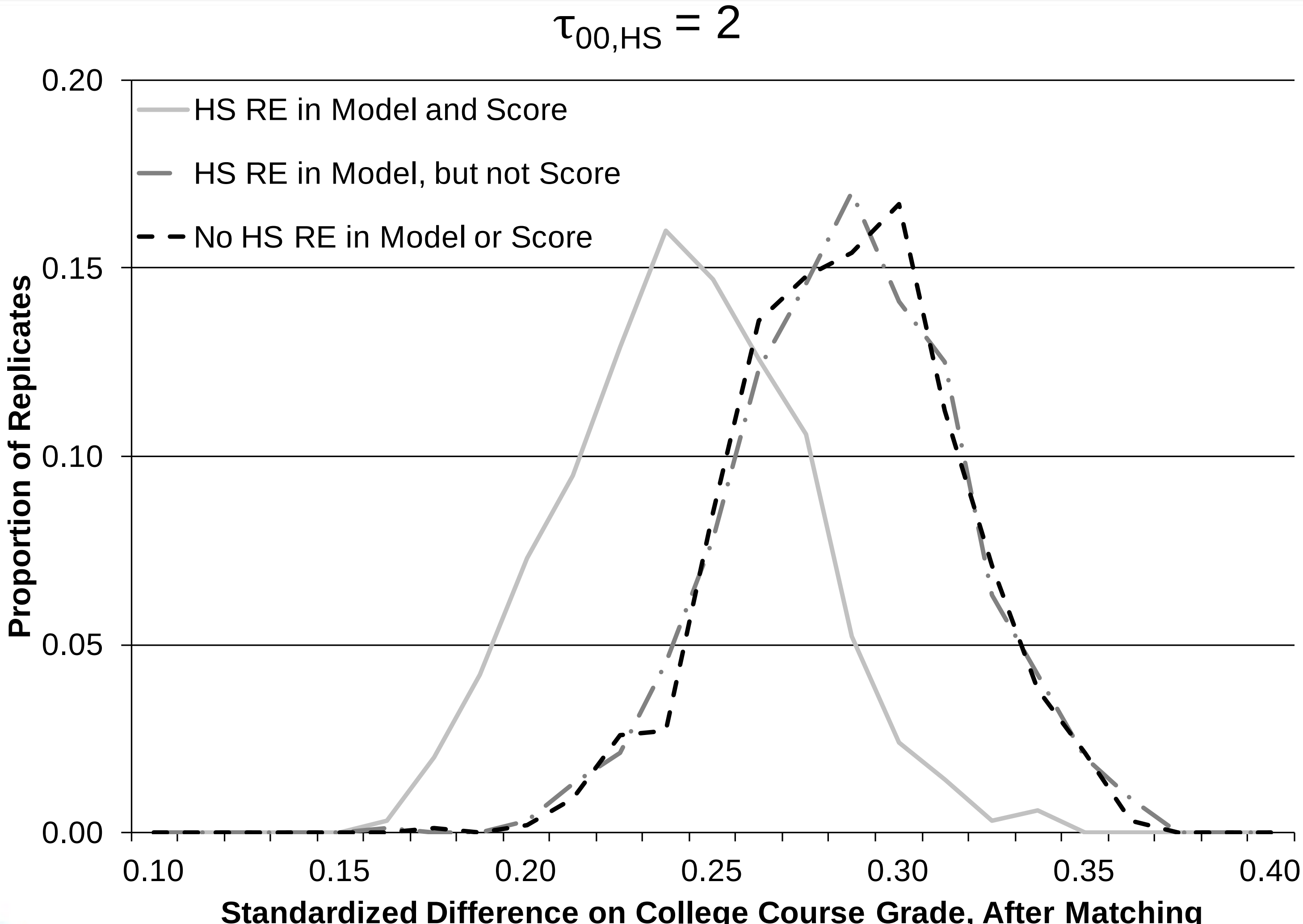
- When including HS random effects:
  - Propensity score *d* much larger, before matching
  - Better balance on gender & race after matching
- Aside from that, either picture looks pretty good:
  - Approximate balance after matching.
  - Non-negative course grade *d*.
- The problem with ignoring random effects is a violation of ignorability
  - Without HS, AP Participation is not MAR.

2.9 Course Grade *d*'s after Matching



1,000 replicates from condition 1 simulated on 2012-03-13

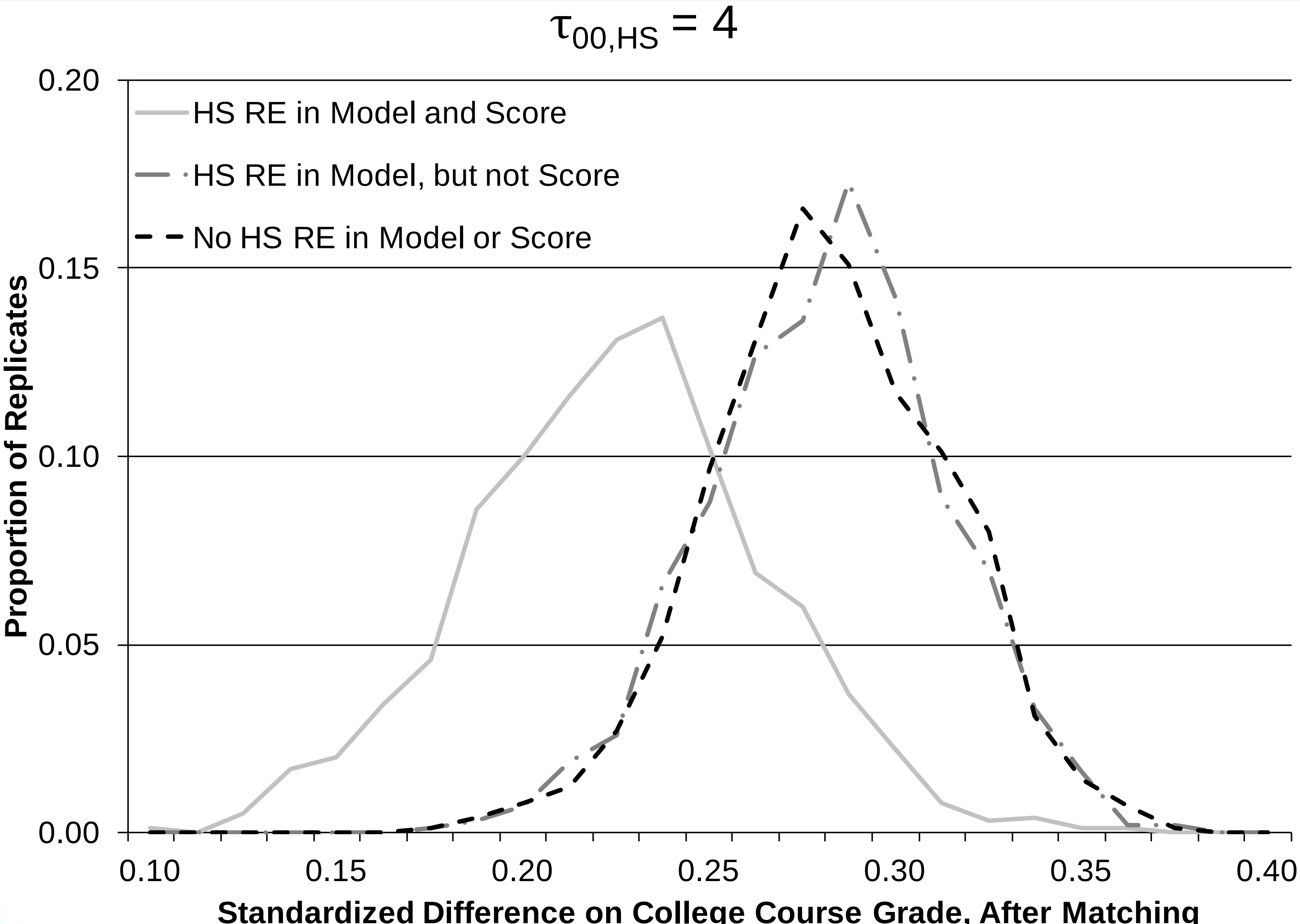
2.10 Course Grade *d*'s after Matching



1,000 replicates from condition 2 simulated on 2012-03-13

CollegeBoard

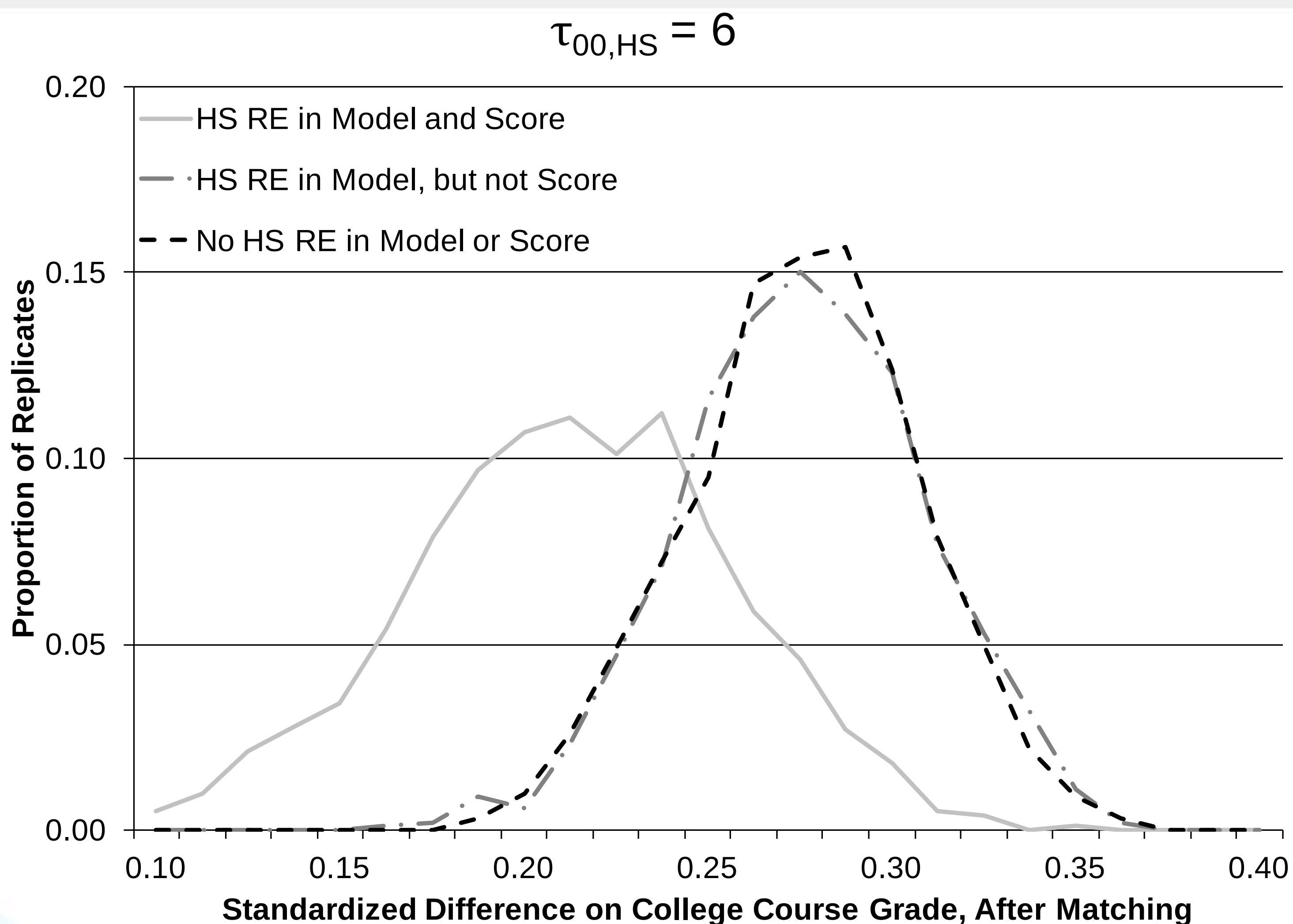
2.11 Course Grade *d*'s after Matching



1,000 replicates from condition 3 simulated on 2012-03-13

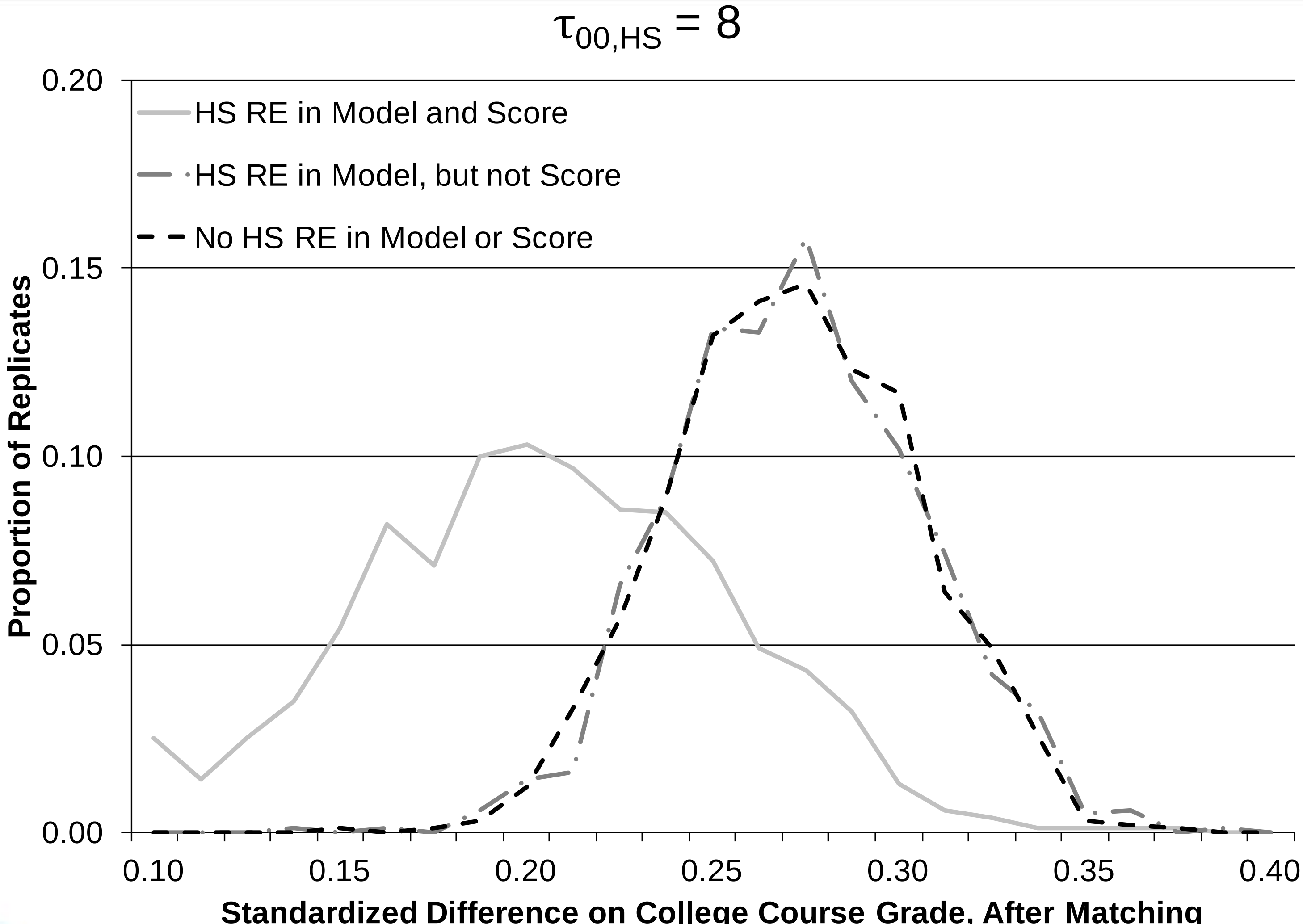
CollegeBoard

2.12 Course Grade *d*'s after Matching



1,000 replicates from condition 4 simulated on 2012-03-13

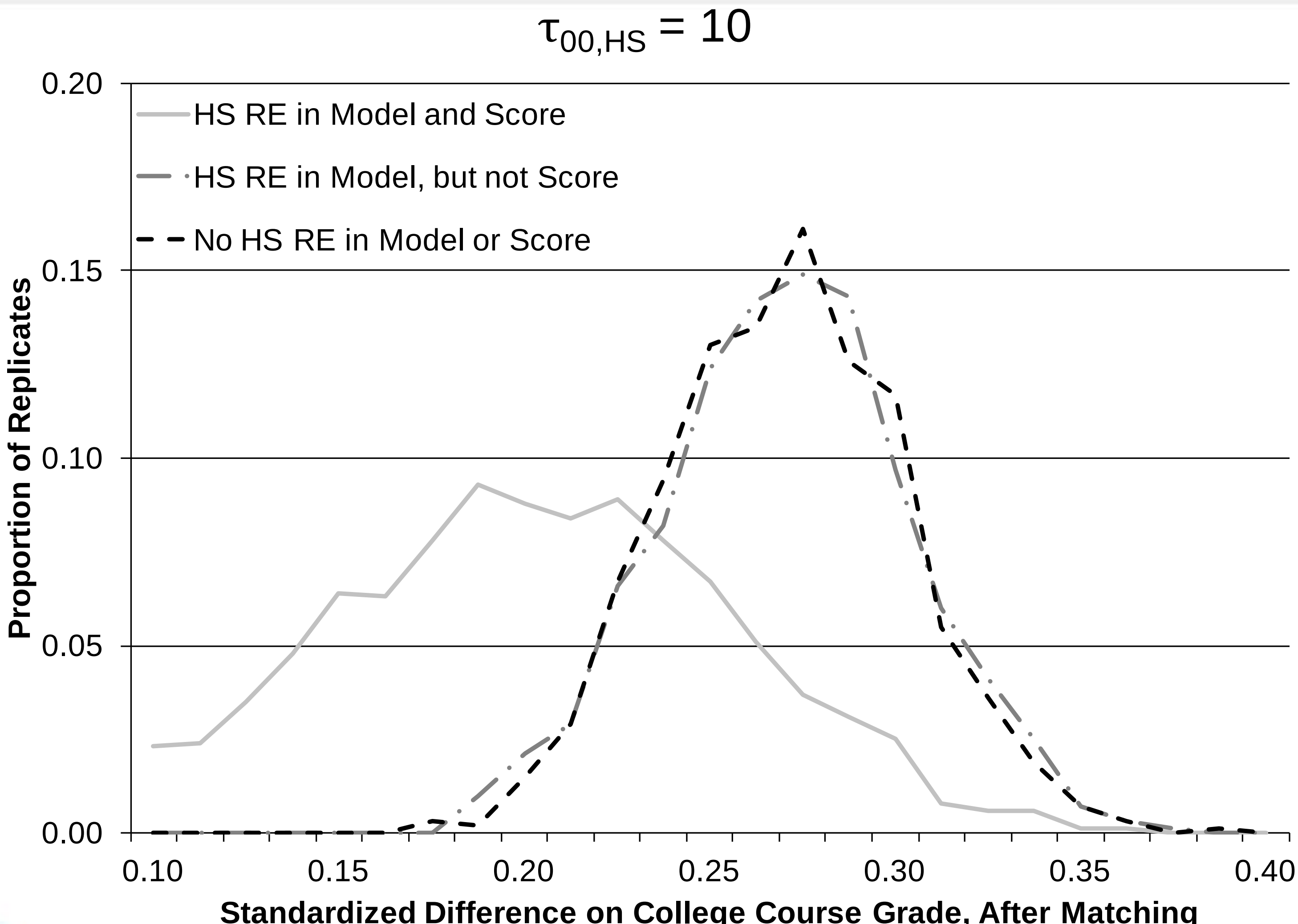
2.13 Course Grade *d*'s after Matching



1,000 replicates from condition 5 simulated on 2012-03-13

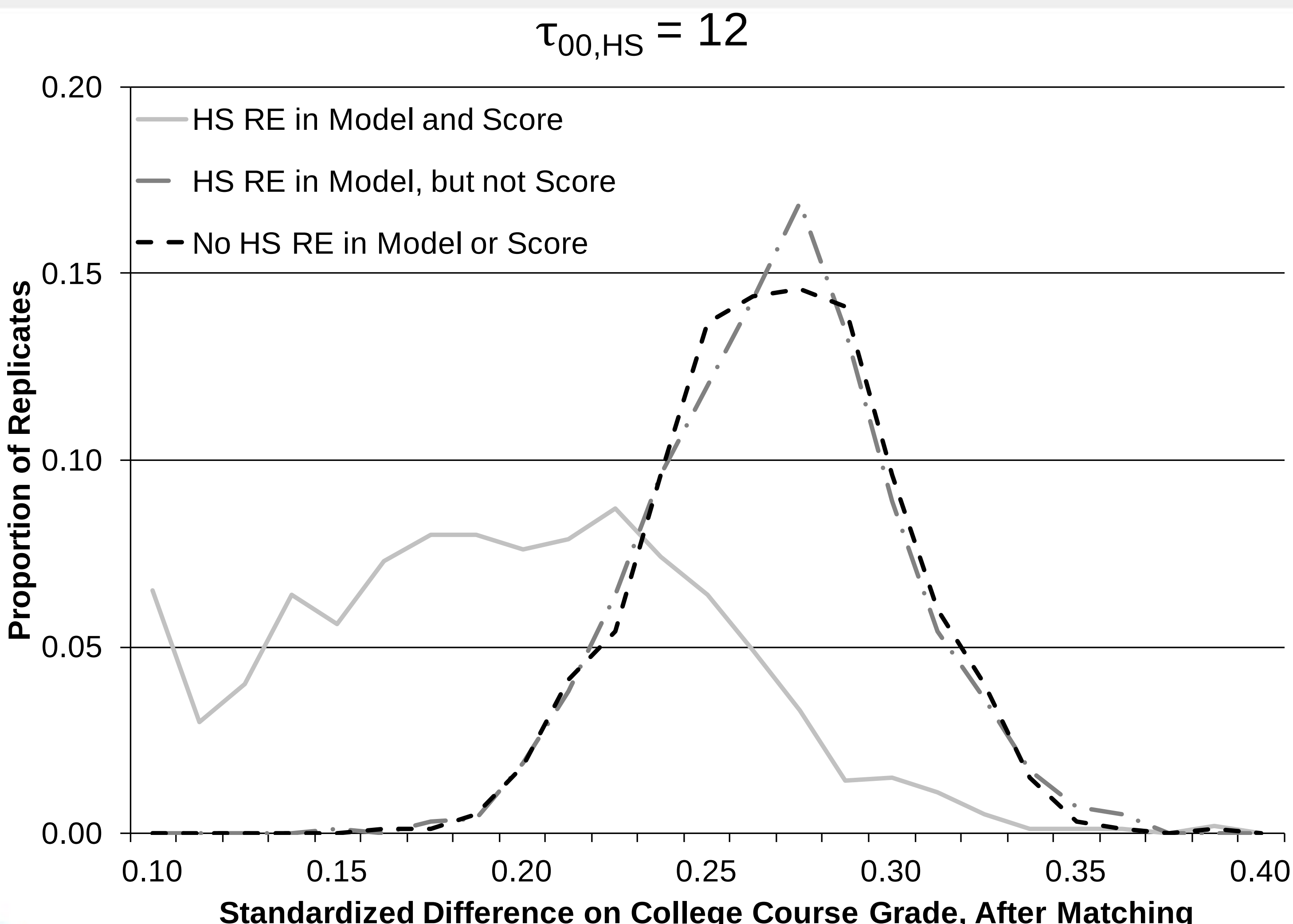
CollegeBoard

2.14 Course Grade *d*'s after Matching



1,000 replicates from condition 6 simulated on 2012-03-13

2.15 Course Grade *d*'s after Matching



1,000 replicates from condition 7 simulated on 2012-03-13

2.16 Average Course Grade Stats After Matching by Condition & PS Model

	HS RE in…			AP		Non-AP		*d*	
τ_{00}, HS	Model?	Prop Score?	Matched Pairs	M	SD	M	SD	SD bef Match	SD aft Match
0	No	n/a	4,392	3.11	0.67	2.89	0.80	0.29	0.21
	Yes	No	4,392	3.11	0.67	2.89	0.80	0.29	0.21
	Yes	Yes	4,392	3.11	0.67	2.90	0.80	0.29	0.21
4	No	n/a	6,487	3.07	0.67	2.87	0.80	0.28	0.19
	Yes	No	6,487	3.07	0.67	2.87	0.80	0.28	0.19
	Yes	Yes	6,270	3.07	0.67	2.91	0.79	0.22	0.15
8	No	n/a	7,817	3.06	0.67	2.87	0.80	0.27	0.19
	Yes	No	7,817	3.06	0.67	2.87	0.80	0.27	0.19
	Yes	Yes	7,201	3.06	0.67	2.91	0.79	0.20	0.14
12	No	n/a	8,743	3.06	0.67	2.87	0.80	0.26	0.19
	Yes	No	8,743	3.06	0.67	2.87	0.80	0.26	0.19
	Yes	Yes	7,667	3.05	0.67	2.91	0.79	0.19	0.13

CollegeBoard

2.17 Simulation Results w/ respect to τ

- As random HS intercept variance (τ) increases…
  - number of within-caliper matches made increases;
  - ignoring HS RE → mean recovered *d* is stable;
  - modeling HS random effects:
    - PS excludes HS RE → similar to ignoring HS; and
    - PS includes HS RE →
      - mean recovered *d* decreases; and
      - variance in recovered course grade *d* increases.

CollegeBoard

2.18 Other Simulation Results

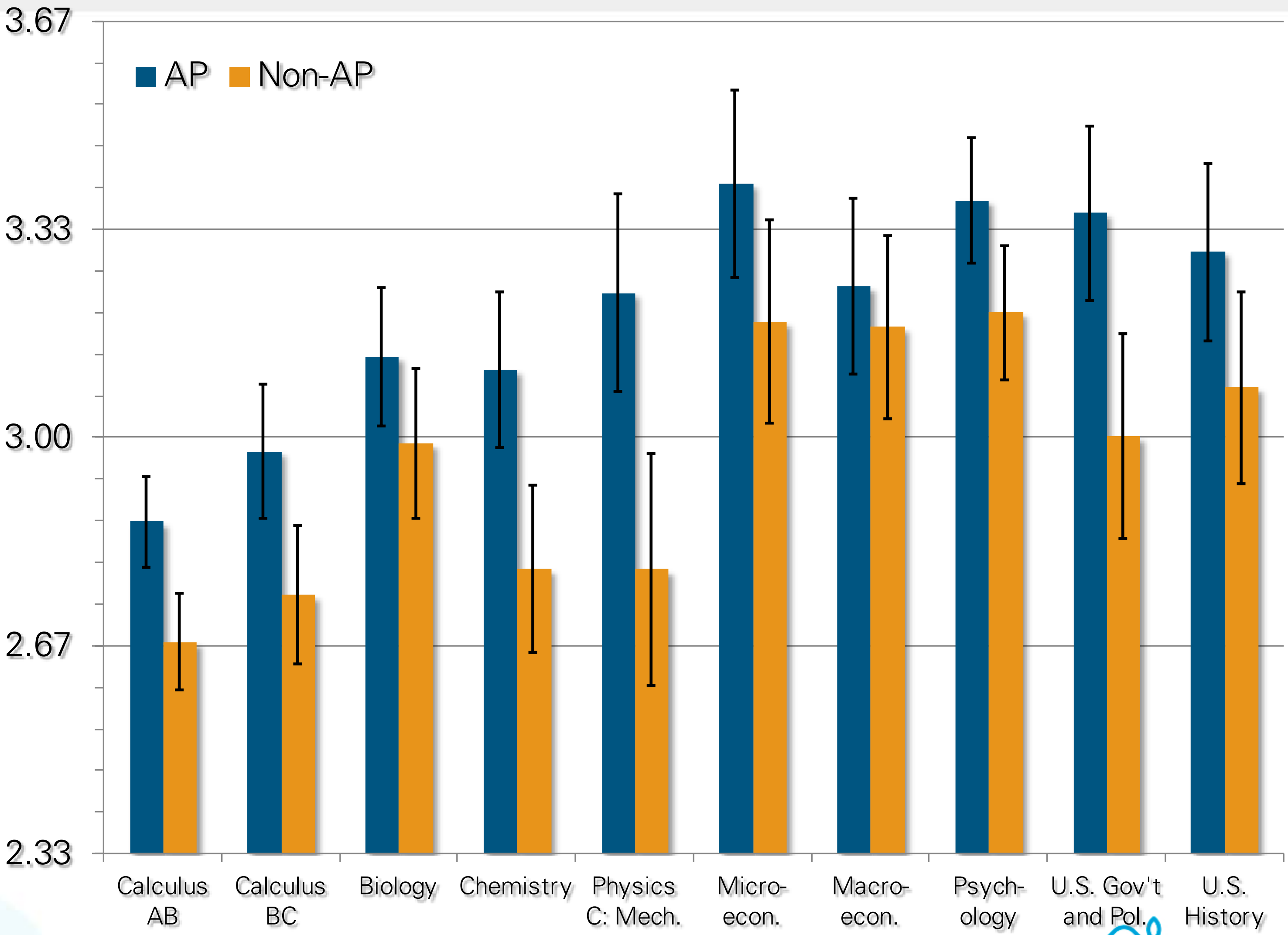
- When ignoring HS or excluding from prop. score:
  - Prop. score SD and therefore caliper size is smaller
  - More matches result
  - Are these better matches, than when modeling and including in the prop. score the HS random effect?

- How to optimize both the quality of matches and sample size

3.1 Tying Simulations back to Application

- A Placement Validity Study for Advanced Placement® Exam Scores
  - Forthcoming study with my colleague Maureen Ewing
  - 2006 cohort of first-time, first-year college students
  - Used official AP credit / placement granting policies
  - Needed sufficient number of AP examinees taking subsequent courses
  - Needed a good propensity score model and to achieve balance
  - Final sample: 10 exams; ≤ 53 colleges

3.2 Mean Course Grades, after Matching



CollegeBoard

3.3 Summary of Results

- AP participation differs across high schools
- AP examinees significantly outperformed matched non-AP counterparts in five AP exams
  - Calculus AB, Calculus BC, Chemistry, Physics C: Mechanics, and United States Government and Politics
- In the remaining five exams, no significant differences existed for course grades
  - Biology, Microeconomics, Macroeconomics, Psychology, and U.S. History
  - Criterion differences? Differential selection?

CollegeBoard

3.4 Questions, Comments, Suggestions?

- Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board presentations do not necessarily represent official College Board position or policy.

- Please forward any questions, comments, and suggestions to:
  - bpatterson@collegeboard.org
- And check out Research & Development's site:
  - http://www.collegeboard.com/research

CollegeBoard

3.5 References

References

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*(4), 1770-1780. doi:10.1016/j.csda.2010.11.008.

College Board. (2006). Annual survey of colleges. New York: The College Board.

Commenges, D., & Jacqmin, H. (1994). The Intraclass Correlation Coefficient: Distribution-Free Definition and Test. *Biometrics, 50*(2), 517. doi:10.2307/2533395.

Griswold, M. E., Localio, A. R., & Mulrow, C. (2010). Propensity score adjustment with multilevel data: setting your sites on decreasing selection bias. *Annals of Internal Medicine, 152*(6), 393.

Patterson, B. F., Packman, S., & Kobrin, J. L. (2010). Advanced Placement® exam-taking and performance: Relationships with first-year subject area college grades. (College Board Research Report No. 2011-4). http://professionals.collegeboard.com/profdownload/pdf/RR2011-4.pdf.

Shaw, E. J. & Patterson, B. F. (2010). What Should Students Be Ready For in College? A Look at First-Year Course Work in Four-Year Postsecondary Institutions in the U.S. (College Board Research Report No. 2010-1). http://professionals.collegeboard.com/profdownload/pdf/RR2010-1.pdf.

Vanderweele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine, 27*(11), 1934-43. doi:10.1002/sim.3139.

CollegeBoard