How Well Do Structural Demand Models Work? Counterfactual Predictions in School Choice*

Parag A. Pathak† Peng Shi‡

October 2018

Abstract

This paper investigates the prediction accuracy of discrete choice models of school demand, using a policy reform in Boston that altered where applicants can apply under school choice. We find that the discrete choice models do not consistently outperform a much simpler heuristic, but their inconsistent performance largely arises from prediction errors in applicant characteristics, which are auxiliary inputs. Once we condition on the correct inputs, the discrete choice models consistently outperform, and their accuracy does not significantly improve upon refitting using post-reform data, suggesting that the choice models capture stable components of the preference distribution across policy regimes.

JEL codes: C53, C25

Keywords: discrete choice models, counterfactual predictions, simulation, out-of-sample validation, school choice

*The first report of this project was released as NBER Working Paper 19859 in January 2014. We thank Boston Mayor Thomas Menino and Boston Public School Superintendent Carol Johnson for authorizing this study. Boston Public Schools staff, including Kamal Chavda, Tim Nicolette, Peter Sloan, Kim Rice, and Jack Yessayan, provided essential help. We are grateful to our discussant Liran Einav for comments inspiring Section 7, Josh Angrist, Dan McFadden, and Ariel Pakes for encouragement, and participants at the McFadden 80th Birthday conference and the NBER Market Design conference for input. We also thank Nikhil Agarwal, Isaiah Andrews, Steve Berry, Glenn Ellison, Drew Fudenberg, Adam Kapor, Patrick Kline, Whitney Newey, and Michael Whinston for feedback. Financial support is from the National Science Foundation under grant SES-1426566 and the W.T. Grant Foundation. Pathak is on the scientific advisory board of the Institute for Innovation in Public School Choice.

†Pathak: Massachusetts Institute of Technology and NBER, Department of Economics, Cambridge, MA 02142. ppathak@mit.edu

‡Shi: USC Marshall School of Business, Department of Data Sciences and Operations, 3670 Trousdale Pkwy, Bridge Hall 308, Los Angeles, CA 90089. 213-821-1005. pengshi@usc.edu. Corresponding author.

1 Introduction

Developing models capable of quantitatively forecasting the effects of policy changes has been an aim of economics since at least Hurwicz (1950) and Marschak (1953). In recent years, design-based research strategies that estimate particular parameters or causal effects have become increasingly popular. The design-based approach, however, does not immediately allow for ex ante policy evaluations of changes that lie far outside historical experience. An alternative approach that allows for ex ante evaluation of new policies is structural modeling, in which we estimate an underlying model of agent decision-making and simulate equilibrium outcomes induced by this model under the new policy.[1] Both Angrist and Pischke (2010) and Heckman (2010) attribute the growth of design-based research to skepticism about structural modeling, particularly its reliance on parametric and behavioral assumptions.

Though opinions vary on the value of structural models, there is one area of consensus: there are relatively few systematic evaluations of their accuracy in predicting counterfactual outcomes. For instance, Angrist and Pischke (2010, "Industrial Disorganization") write:

> "Many new empirical industrial organization studies forecast counterfactual outcomes based on models and simulations, without a clear foundation in experience. [...] At minimum, we'd expect such a judgement to be based on evidence showing that the simulation-based approach delivers reasonably accurate predictions. As it stands, proponents of this work seem to favor it as a matter of principle."[2]

This paper aims to fill this void by evaluating the prediction accuracy of discrete choice models, which are often used in new empirical industrial organization studies to predict demand in counterfactual simulations. We evaluate these models using the same gold standard used to validate

[1] This definition of structural modeling in inspired by current parlance. For example, Reiss and Wolak (2007) state: "today economists refer to models that combine explicit economic theories with statistical models as *structural econometric models*." Two explicit economic theories underlying our two structural models below are random utility maximization and truth-telling by agents.

[2] Misra and Nair (2011) use a structural agency model to design and implement a compensation scheme and report that the new scheme's outcomes match those from the model. For merger analysis, Peters (2006) examines the predictive value of structural simulation methods for airline mergers and finds they do not accurately predict post-merger ticket prices. Ashenfelter and Hosken (2008) argue that design-based estimates of mergers differ markedly from structural estimates. In response to Angrist and Pischke (2010), Nevo and Whinston (2010) describe a few counterfactual validations in the context of merger analysis and Einav and Levin (2010) support more retrospective analyses of past mergers, though they also express skepticism about cross-merger extrapolation. Lumsdaine, Stock, and Wise (1992) estimate models of retirement behavior before and after the introduction of a pension incentive, and using data from before the incentive, they forecast the effect on retirement.

a hypothesis in other scientific disciplines, which is to use the model to predict what would happen beforehand and compare predictions with the actual outcome.[3] In particular, we predict how families in Boston would respond to a large-scale policy change in the school assignment system.

Each year, thousands of Boston families submit rank order lists of public schools to the city's student assignment plan.[4] In 2013, Boston Public Schools' (BPS) officials, the mayor, and members of the school committee sought to modify the plan in order to assign students to schools closer to their homes, in part to reduce transportation costs. BPS publicized a number of plans that redrew neighborhood boundaries and changed applicant choice sets. BPS and the broader community were interested in predicting both the choices families would make under these alternatives and the ensuing final assignments.[5] The mayor of Boston and the superintendent of BPS delayed the timeline for selecting a new plan and asked us to forecast the effects of these alternatives, stating (Menino, 2012b):

> "We have the opportunity to generate an advanced analysis that will allow us to better predict how families would make choices in the real world [...] This is something we have never been able to do before."

Our policy report, Pathak and Shi (2013), uses historical participation and rankings to predict new choices under the different proposals. The methodology is based on discrete choice modeling. BPS administrators and the public referred to the report to compare alternatives and ultimately selected a new plan.

In January 2014, families throughout Boston ranked schools under new choice sets. There are two grades in which most students enter the school system: Kindengarten 1 (K1) or Kindengarten 2 (K2).[6] K2 is when mandatory schooling begins, but about 50% of K2 students in BPS began attending in K1. For a typical applicant in grade K1, the new system added three new school

[3]Some may argue that trying to predict the outcome is too ambitious for an econometric model, as counterfactuals in economics are intended to be "ceteris paribus" (all else remaining the same), while in policy reforms there may be other unforeseen changes. However, we argue that comparing counterfactual predictions with the actual outcome is still a valuable exercise, as it provides the only objective evidence on the performance of counterfactual predictions for policy decisions.

[4]The Boston assignment system has been subject to a number of theoretical studies including Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu, Pathak, Roth, and Sönmez (2005), Pathak and Sönmez (2008), and Dur, Kominers, Pathak, and Sönmez (2016).

[5]For more details, see Goldstein (2012) and Handy (2012). BPS communications reported more than 1,850 residents offered feedback on the plans. For specific reactions to proposed plans, see Vaznis and Andersen (2012) and Burge (2012).

[6]A small proportion of students may also attend pre-Kindergarten, which is before K1 and is sometimes called K0.

choices, removed sixteen choices, and kept nine choices intact. The magnitude of the change is similarly large for grade K2. Figure 1 summarizes the timeline of the reform. Pathak and Shi (2014) published predictions on this policy reform before it happened, using discrete choice models estimated with the latest pre-reform data as well as an ad-hoc alternative. In this paper, we evaluate these predictions using the actual choices of students after the reform.

To describe our approach and questions, we first introduce some notation. Let X encode the characteristics of student and schools, including the set of schools to which each student can apply. Let Y encode student choice outcomes, which is a rank-ordering over eligible school programs for each student. We observe (X, Y) under the existing policy and can compute equilibrium outcomes of interest, such as the chance students from various neighborhoods have of being assigned to higher-performing schools, as well as the distance students travel and the number unassigned.[7] These prediction targets come from Pathak and Shi (2013) and were central to the Boston policy debate. Let $M(X, Y)$ denote these equilibrium outcomes. This is a well-defined function of X and Y, since the assignment mechanism can be exactly recreated before and after the reform.

The forecasting problem is to predict what happens under the new policy. Because the assignment mechanism in Boston is strategy-proof, we assume that the historical choice rankings of students, which is restricted to schools in their choice set, reflect their underlying preferences across the set of all schools, which we do not observe but can estimate using discrete choice models. We use this estimated model to simulate choices under the new choice sets and forecast the equilibrium outcomes of interest. Our paper focuses on two questions:

1. How well do discrete choice demand models predict equilibrium outcomes important for the school choice context?

2. How well do discrete choice demand models predict raw choice patterns?

Let (X^*, Y^*) be the dataset observed under the new policy regime. The first question compares the actual equilibrium outcomes $M(X^*, Y^*)$ with the forecast $M(X, Y)$, which depends both on a forecast of characteristics X and the choice model. The second question compares actual choices Y^* with the predicted choices conditioned on the actual applicant characteristics. Conditioning on

[7] We refer to these as equilibrium outcomes because they depend not only on the behavior of an individual student, but also on the behavior of all applicants.

the actual characteristics isolates the performance of the demand model from auxiliary forecasts of characteristics.

An innovation of our research design, illustrated in Figure 2, is that we published predictions prior to the policy change. Opportunities for the external validation of structural models are rare, especially in high-stakes contexts where they had a direct influence on a policy decision. The benefit of publishing forecasts before the new policy is that it guarantees that our forecasts and hypotheses are in no way biased by the actual outcome. This goes one step further to ensure genuine out-of-sample assessment than papers that use a hold-out sample such as Wise (1985), Todd and Wolpin (2006), and Keane and Wolpin (2007), because it prevents us from possibly estimating multiple structural models and only reporting what matches the outcome. Moreover, even when researchers do not directly look at post-reform data, qualitative information about the outcome could unconsciously influence modeling decisions. We also report forecasts not based on random utility maximization and requiring no estimation.[8] The simple alternative is based on an intuitive heuristic, in which families rank schools based on a hierarchical list of rules, such as preferring schools in a better tier in terms of test performance, and among schools of the same tier, preferring schools closer to home. We adopt this heuristic as a reference point because similar ideas were used in internal analysis by BPS. Moreover, the same heuristic for simulating Boston school assignment was suggested by education researchers (Levinson, 2015). Such heuristics also have theoretical and empirical support from the psychology and marketing literature.[9]

The choice models we fit are the multinomial logit (MNL) model and the mixed MNL model. In a pioneering contribution, McFadden and co-authors used the MNL model to study the impact of BART, San Francisco's rapid transit system (McFadden, Reid, Talvitie, Johnson, and Associates, 1979). Prior to the introduction of BART, they collect data on the travel behavior of a sample of individuals, and estimate MNL models to predict their behavior if BART were to be introduced. After BART began, they compare the predictions with the actual travel behavior of these individuals. McFadden, Talvitie, and Associates (1977) provide a detailed account of the performance of these models. McFadden (2001) argues that the success of BART forecasts provided strong evidence

[8] However, the heuristic-based alternative we propose still models individual student choices and uses simulation to forecast aggregate outcomes, so it is not a design-based approach as described in Angrist and Pischke (2010). We discuss the difficulties of forecasting outcomes of interest without modeling choices in Section 8.

[9] See Section 3.3.3 for a description of the theoretical and empirical support for lexicographic choice rules.

that disaggregate random-utility models could outperform conventional methods and random utility models are now widely employed in travel analysis and other areas of economics. There have also been many developments in choice modeling in the subsequent four decades. Yet, to our knowledge, our paper is the first post-BART study to publish counterfactual predictions of a choice model before the policy reform, and follow up to compare with the actual outcome.

We find that our heuristic-based alternative to choice modeling performs as well as the more sophisticated discrete choice models for one of the two grades we examine. However, the discrete choice models' inconsistent performance is largely due to prediction errors in applicant characteristics, which are auxiliary inputs. Once we condition on the actual applicant characteristics, the discrete choice models outperform the ad hoc alternative in predicting both choice patterns and policy relevant outcomes. Moreover, refitting the models using the new choice data does not significantly improve their prediction accuracy, suggesting that the choice models capture stable components of the preference distribution across policy regimes.

The inconsistent performance of discrete choice models suggests that accurately predicting the outcome of a policy reform using structural modeling remains an elusive goal. In our exercise, the discrete choice models did not systematically outperform the ad hoc alternative, despite our having rich micro-level data[10], the system being strategy-proof[11], and the policy change being relatively simple, as it was primarily only a change in choice sets.

Our second set of findings, that discrete choice models perform relatively well given the actual auxiliary inputs, support the crucial assumption of such models that to a first-order approximation, preferences are stable, can be estimated accurately, and can be used for extrapolation in different environments. Previous work has cast doubt on this assumption by highlighting the importance of considerations such as framing and informational cues. For example, Hastings and Weinstein (2008) show in a field experiment that choice behavior in Charlotte's school choice plan can be swayed by informational cues. In other contexts, interventions simplifying information can significantly alter

[10] Our dataset includes not only the top choice of students but also their entire submitted ranking of schools. This helps us to estimate choice model parameters precisely. Moreover, the data includes a large number of observables, including student characteristics and exact geographic location.

[11] The current strategy-proof system has been in place since 2005, and BPS further advises families to not be strategic about their choices. For instance, the 2012 School Guide states: "List your school choices in your true order of preference. If you list a popular school first, you won't hurt your chances of getting your second choice school if you don't get your first choice." Several authors have argued for strategy-proof mechanisms partly because they are believed to generate reliable demand data: see Abdulkadiroğlu, Pathak, Roth, and Sönmez (2006), Abdulkadiroğlu, Pathak, and Roth (2009) and Sönmez (2013).

choice behavior (Kling, Mullainathan, Shafir, Vermuelen, and Wrobel, 2012). In our context, BPS also presented the new choice set in a different way, which may have induced a change in underlying preferences (Figure 3 shows how choices were presented). However, we show that the underlying preference distributions remain stable, at least to the extent that any changes in preferences have minimal effects on the predicted outcomes of interests. The validation of the assumption of stable preferences is relevant for a large and growing literature using choice modeling to study school demand.[12] It is also relevant for research that uses a random utility choice model as a component of a larger structural model.

Finally, our exercise illustrates that in predicting counterfactuals using structural modeling, having accurate forecasts of auxiliary variables can be more important than obtaining the correct specification of the model itself. While the previous literature in discrete choice modeling has stressed the importance of capturing individual preference heterogeneity using random coefficients, we found that in our setting whether or not we incorporate random coefficients made little difference, possibly due to the availability of rich observable characteristics.[13] In contrast, accurately predicting the applicant demographics turns out to be the first-order issue. This highlights the importance of sensitivity analysis in counterfactual predictions beyond the specification of the discrete choice model: one should experiment with a range of auxiliary variables and report whether the result is consistent across the scenarios involving alternative input variables.

The rest of this paper is structured as follows. Section 2 provides details on the Boston student assignment plan and events leading up to the adoption of a new plan in 2014. Section 3 describes the data, forecast targets, and the prediction generation process. Section 4 discusses how we evaluate predictions, and Section 5 reviews hypotheses motivated by back-testing our framework. Section 6 compares our predictions to the actual outcomes in the first year of the new plan. The section also decomposes sources of prediction error by examining changes to the set of participants and the underlying stability of the demand model. Section 7 reports on how prediction errors affect the

[12] Research estimating school demand from similar datasets includes: Abdulkadiroğlu, Agarwal, and Pathak (2015), Abdulkadiroğlu, Pathak, Schellenberg, and Walters (2017), Agarwal and Somaini (2014), Burgess, Greaves, Vignoles, and Wilson (2015), Calsamiglia, Fu, and Guell (2017), Glazerman and Dotter (2016), Harris and Larsen (2015), Hastings, Kane, and Staiger (2009), He (2012), Hwang (2015), Kapor, Neilson, and Zimmerman (2017), Ruijs and Oosterbeek (2012), and Walters (2014).

[13] In particular, the distance from home to school is a very important to a family's choice of schools, and our data set contains the Google Maps walking distance between every student's residence and every school. Such detailed observable information may not be available in other settings.

ranking of plans other than the one Boston ultimately selected. Section 8 concludes and discusses directions for future work.

2 Background

2.1 School Choice in Boston

Boston Public Schools has one of the nation's most well-known school choice plans. From 1988 to 2013, the city was divided into the North, West, and East Zones, shown in Figure 3, for elementary school admissions. There are roughly 25 to 30 elementary schools in each zone, and each school may have multiple programs. There is often a separate program for regular education students and for English language learners (ELL), and there may be specialized ELL programs for students of a certain language group. Students are allowed to rank programs in any school in the zone in which they reside as well as any school within one mile of their residence and a handful of city-wide schools. Students can rank as many programs as they want, as long as the programs meet certain eligibility criteria.[14]

At each program, students are prioritized as follows: continuing students (who were already assigned to the school in an earlier grade) have the highest priority, followed by students who have an older sibling at the school, followed by other students. Until 2013, for half of the program seats, students residing in the walk zone[15] obtained priority, but this priority did not extend to the other half. A single lottery number serves as the tie-breaker.[16]

Since 2005, after students submitted their choices, they were processed through a version of Gale and Shapley (1962)'s student-proposing deferred acceptance (DA) algorithm (Abdulkadiroğlu, Pathak, Roth, and Sönmez, 2005; Pathak and Sönmez, 2008). This algorithm takes as inputs both students' submitted preference rankings and their priorities to generate an assignment. DA works as follows:

1. Each student applies to his first choice program. Each program ranks applicants by their priority, rejecting the lowest-ranked students in excess of its capacity. The rest of applicants

[14]For example, in order to rank a ELL program, the student must not be a native English speaker and must not exceed a certain score in a BPS administered language test. All students are able to rank regular education programs.

[15]The walk zone of a school is defined to be the region within one mile of the school, according to straight-line distance. Only students outside the walk zone are eligible for school busing.

[16]Dur, Kominers, Pathak, and Sönmez (2016) present additional details on Boston's DA implementation.

are provisionally admitted: they are not rejected at this step but may be rejected in later steps.

2. The rejected students apply to their next most preferred program (if any). Each program considers these new applicants together with applicants that it admitted provisionally in the previous round, ranks them by their priority, rejecting the lowest-ranking students in excess of capacity. This produces a new set of provisionally admitted students at each program.

The algorithm iterates between these two steps and terminates when there are no new applicants in step 1. (Some students may remain unassigned upon termination.) Under DA, it is a weakly dominant strategy for all participants to rank programs truthfully (Dubins and Freedman, 1981; Roth, 1982). Moreover, this algorithm produces a stable matching with respect to student preferences and school priorities (Gale and Shapley, 1962; Abdulkadiroğlu and Sönmez, 2003).

2.2 Policy Reform

The new policy, which began in 2014, affects the set of schools each applicant is allowed to rank. There were two major rationales for the reform. First, there was a desire to assign students to schools closer to home, which many families value and which reduces the school district's busing costs.[17] Second, there were longstanding concerns about inequities in the three zone system.

The reform was informed by the Pathak and Shi (2013) report, which used discrete choice modeling and simulations to predict the effects of the proposed plans. The study was commissioned by a mayoral-appointed city committee, which met for over a year and hosted community meetings to collect feedback and discuss proposals.[18] The report's methodology inspired BPS to later propose a 10 Zone plan and a modified 11 Zone plan while also considering other plans from the community. Shi (2015) provides more details on the role of the report. It's worth noting that there were two prior failed attempts to reform the choice sets of students in 2003 and 2009. Decision-makers did not have access to comparable forecasts during these prior attempts.

[17] This motivation was emphasized by Mayor Menino, who spent the last year of his administration advocating for a "radically different school assignment process—one that puts priority on children attending schools closer to their homes" Menino (2012a). Other districts have similar objectives; see, e.g., the discussion about Seattle in Pathak and Sönmez (2013).

[18] BPS's initial plans either divided the city into 6, 9, 11, or 23 zones or assigned schools based purely on neighborhood. When these plans were publicly unveiled in September 2012, they were met with widespread criticism (Seelye, 2012).

Based on the Pathak and Shi (2013) report and other discussions, the Boston School Committee adopted the Home-Based plan (Seelye, 2013; Shi, 2013). This plan constructs customized choice sets based on applicants' exact residential address. It uses a BPS categorization of schools into quality tiers, which are computed using schools' prior Massachusetts Comprehensive Assessment System (MCAS) test score growth and levels. Tiers were finalized as of January 2013 for 2014 admissions. Under the new plan, every applicant can choose from any school within one mile (straight-line distance), the two closest Tier 1, the four closest Tier 1 or 2, the six closest Tier 1, 2, or 3 schools, and the three closest "option schools" chosen by BPS. The set of choices also includes the closest early learning center (ELC) and the closest school with an advanced work class (AWC) program.[19]

Families access their choice sets via an online portal, which shows a map of all schools in the choice set and a summary of their attributes. Figure 3 illustrates how participants see choice information. The online application platform lists information on transportation, tier category, and why the choice can be ranked. Previous years' school brochures did not include comparable information.

Aside from the changes to choice sets, the new plan also eliminates walk zone priority (Dur, Kominers, Pathak, and Sönmez, 2016). The school priorities are: continuing students, followed by siblings, followed by other students, and a single lottery number serves as tie-breaker. There are no other changes to the implementation of the assignment algorithm.

The new plan involves large changes in applicant choice sets. This fact can be seen in Table 1, which shows that for an average grade K1 student, the reform adds three new options, removes sixteen options, and keeps nine options intact. This implies that the new plan removes 63% of the old choice options, and 21% of choice options under the new plan were not in the old plan. The magnitude of this change is reduced when we focus only on highly ranked choices, but the magnitude is nevertheless substantial given there are thousands of applicants in each grade. Table 1 reports that about 32% of old choice options that are ranked top five were no longer offered in

[19] There are a few exceptions to this formula. First, students residing in parts of Roxbury, Mission Hill, and Dorchester are allowed to rank the Jackson Mann school. Second, because transportation outside of East Boston requires tunnel travel, East Boston students are eligible for any East Boston school. East Boston students have priority over non-East Boston students at East Boston schools. Non-East Boston students have priority over East Boston students for non-East Boston Schools. Finally, students who are English Language Learners or special needs have additional choices. Level 1, 2, and 3 ELL students are allowed to apply to any compatible ELL program within their ELL zone, a six-zone overlay of Boston. Substantially-separate special education students do not apply in round one.

the new plan, while about 10% of top five choices under the new plan were not available options under the old plan. The magnitude of the change in choice sets is similar for grade K2 applicants.

3 Prediction Approach

3.1 Data Sources

Our data come from BPS round one choice and enrollment files covering years 2010 to 2014. We focus on round one assignment, which takes place in January and February, because over 80% of students are assigned then. Forecasts are based on data from 2010 through 2013. We use the data from the first post-reform year (2014) to evaluate these forecasts.

The choice data contain preference rankings and demographic information for every round one participant. The fields include student ID number; English language learner (ELL) status and first language; special education or disability status; geocode (a geographic partition of the city into 868 regions); school program to which the student has guaranteed priority (designation for continuing students); lottery number; first 10 choices and priorities at each; school program to which the student was assigned and the priority used for that assignment. Using the assigned school and program codes, we infer each school program's capacity available for round one assignment. We place students in one of 14 neighborhoods using the geocode.[20]

The enrollment data represent a December snapshot and contain additional student demographics. The fields are enrolled school and program, grade, geocode, address, gender, race, and languages spoken at home. The file covers the vast majority of the students in the choice data and can be linked by student ID number. When there is a conflict between the demographic information in the choice and enrollment files, we use the choice file. We also match geocodes to 2010 census block groups, using which we can obtain the US census data for median household income.

We have school characteristics for each year between 2010 and 2014. The school file has the building code, address, school type, % of students of each race, % of ELL students, % of students who have special education requirements, and % of students who scored Advanced or Proficient in grades 3, 4, and 5 for MCAS math and English in the previous year. To measure distance to school,

[20]For internal reporting, BPS classifies students into 16 neighborhoods. We combine three neighborhoods with few students, Central Boston, Back Bay, and Fenway/Kenmore, into one neighborhood that we call "Downtown."

we use walking distance estimates from Google Maps API.[21]

3.2 Forecast Targets

Each equilibrium outcome we forecast can be represented by a single number for each grade and neighborhood combination.[22] The outcomes are defined as follows:

- **Access to Quality:** The chance an average student from this grade and neighborhood has of being assigned to a top tier school if he wanted to be assigned to such schools, holding fixed the submitted rankings of other students. In particular, we define a "top tier school" to be any Tier 1 or 2 schools within the student's choice menu, and define "wanting to be assigned to such schools" as ranking all eligible programs in such schools, and ranking them above all other programs.[23]

- **Distance:** The average Google Maps walking distance between the residences of assigned students from this neighborhood and their respective school assignments.

- **Unassigned:** The number of students who are unassigned from this neighborhood at the end of round one.

We refer to these as equilibrium outcomes because they depend on the preference submissions of all students since program capacities are limited and assignment depends on each program's level of competition. As previously mentioned, we chose these outcomes because the Boston debate focused on equal access to quality schools and assignments close to home. A forecast of unassigned students at the end of round one is important for facilities' planning, staffing, and other budgeting issues, especially for grade K2 as BPS is required by law to add capacity in later rounds to assign every K2 applicant.

The second set of targets involve student choices. We predict both individual student choices and choice patterns across a group of students.

[21] For students with missing address information, we treat the centroid of the student's geocode as the address.

[22] We separate the gwo grades in reporting outcomes because their are systematic differences: not all schools offer K1 (as mandatory schooling begins only in K2), and availabilities of K1 seats differ by neighborhood. Moreover, assignments for K2 are heavily affected by the presence of students continuing from K1, who take up about half of the total seats, while the same is not true for K1 since very few students attend K0 programs.

[23] One can show that as long as the student ranks the top tier schools first, this metric does not depend on the order the student ranks the top tier schools among themselves, or the order the student ranks the other schools. This observation is based on the strategyproofness of the deferred acceptance algorithm for the proposing side.

For **individual choices**, we compute the best deterministic prediction of the set of k options ranked highest for each student, where k varies from one to three. For a given pair of options in a student's choice set, we also predict whether the student would prefer one option over the other.

For **distribution of choices**, we predict each school's market share of top k choices, broken down by grade and neighborhood. Furthermore, we predict the aggregate distribution of the top two choices for students who rank at least two choices.

3.3 Generating Predictions

We use data from before the reform to fit choice models to forecast outcomes for the reform's first year and compare these forecasts to outcomes induced by the actual choice data submitted in the reform's first year.[24] To protect the integrity of our out-of-sample comparison, we specify choice models and forecasts prior to the reform by posting a pre-analysis plan in Pathak and Shi (2014) before the post-reform choice data were available.[25]

Our counterfactual predictions come from three choice models, two of which are based on random utility models. Each choice model maps an individual student's characteristics and the set of eligible programs in his menu to a ranking over programs. The three choice models we examine are:

- **Multinomial Logit (MNL):** This widely-used and easy-to-estimate model is motivated by random utility maximization. It is developed in McFadden (1974) and used in the BART analysis (McFadden, Talvitie, and Associates, 1977). It is also the basis of the Pathak and Shi (2013) report that informed the Boston reform.

- **Mixed MNL (MMNL):** We consider MMNL models for several reasons. First, MMNL models can capture substitution patterns that violate the Independence of Irrelevant Alternatives property of MNL models. Second, mixture models are a significant development in discrete choice models of demand in the years following McFadden (1974). Indeed, McFadden (2001) states that the MNL methods used in the BART study were inferior to current

[24] The outcome induced by the actual choice data may not be identical to the actual round one assignment outcome, since we use previous year's program capacities in our computation rather than the actual capacities. We abstract away from forecasting capacities as they are at the discretion of the school board and outside the scope of our structural model.

[25] Pathak and Shi (2014) describe the specification of the mixed MNL model, but did not report estimates before posting the report. Estimating the mixed MNL model was too computationally-intensive to complete in time. Pathak and Shi (2015) update the report with the mixed MNL forecasts.

methods. Third, Berry, Levinsohn, and Pakes (2004) argue that 2nd choice data can improve estimates of substitution patterns. Since we observe an applicant's rank order list, we observe several choices for each applicant and the MMNL model better captures correlation between the first and subsequent choices within a student's rank ordering than does the MNL model.

- **Lexicographic:** This model serves as our benchmark for models not motivated by random utility maximization. It assumes that applicants rank programs based on an intuitive heuristic, such as preferring schools of better performance, and among those with similar performance, preferring schools closer to home. The same heuristic has been proposed independently by education researchers (Levinson, 2015), and it has support from the psychology and marketing literature (see Section 3.3.3 for references). There is no estimation involved with this model. While a model of this form is necessarily ad hoc, our back-tests below inform our beliefs about its potential performance.

Before describing each of the choice models in detail, we first describe how they are used to compute forecasts. For choice outcomes, we use the choice model to predict the relative ranking of options by the actual applicants of the first post-reform year. We use the actual set of students to isolate choice prediction from population forecasting. To compute the best deterministic prediction of an individual's choices, we use the *modal* choice among many simulations.[26] To compute predictions of school market shares and of the distribution of top two choices, we use the empirical choice distribution from the simulations.

For forecasting equilibrium outcomes, there are additional simulation layers. Rather than using the actual applicants, as in the choice forecasts, we simulate the pool of applicants and their characteristics. At the time of the typical counterfactual forecast, an analyst does not know future participants. With the simulated applicant pool, we apply each choice model to generate a complete ranking of options within each student's menu, similar to method in the choice forecasts.

Next, we truncate the generated preference rankings to the first ten choices. Truncation is necessary because the choice data we receive from BPS only have the first ten choices, although there is no restriction on the number of choices in the mechanism. In Pathak and Shi (2013), we performed sensitivity analysis on list length and found ten to be reasonable. In Section 6.2, we

[26] Note that the best deterministic prediction of a biased coin that yields heads 60% of the times is that it always yields heads.

further examine this assumption. Another parameter that affects the equilibrium outcome is the number of seats in each school program. For the purpose of the prediction exercise in this paper, we generate predictions based on the assumption that the school board uses the same capacities as in the previous year. In practice, Boston runs DA several times with minor tweaks to capacity but does not report the outcome until the final run of round one assignment.

Finally, we generate uniformly distributed lottery numbers for each student and compute the assignment using the DA algorithm. In computing access to quality, we compute the probability that the student receives a lottery number that is good enough to be assigned one of the Tier 1 or 2 schools in his menu.[27]

3.3.1 Multinomial Logit (MNL) Choice Model

The MNL choice model assumes that students rank program according to a latent utility for each program, which depends on the observed characteristics in a certain way. Let u_{ij} be the unobserved latent utility of student i for program j, and x_{ij} be a K-dimensional vector of observed characteristics corresponding to the student and program, such as the student's distance to the program or whether the student has a sibling at the same school. The k^{th} component of this vector is denoted x^k_{ij}. The utility of student i for program j is assumed to take the form:

$$u_{ij} = \delta_{s(j)} + \sum_{k=1}^{K} \beta^k x^k_{ij} + \epsilon_{ij}, \tag{1}$$

where $s(j)$ denotes the school containing the program[28], $\delta_{s(j)}$ is a school effect, β is a K-dimensional vector of coefficients, and ϵ_{ij} represents an unobserved idiosyncratic taste. We assume that ϵ_{ij} is distributed according to a type-I extreme value distribution, Gumbel(0,1). Since utility has no scale, we normalize the scale parameter to one. The school effect captures unobserved school characteristics such as safety, reputation, facilities, environment, and teacher quality. The estimated

[27] This probability is estimated in a tractable way as follows. If there is at least one Tier 1 or 2 school with a program, with excess capacity, for which the student is eligible, then the student's access to quality is 100%. If all such programs are full, then we compute a lottery cutoff for the student, which is the worst lottery number needed for that student to displace out at least one currently-assigned student from one of these programs, and we report the chance that the student gets a lottery number at least as high. As shown in Azevedo and Leshno (2016), this approach computes access to quality exactly in a continuum large market model, and is a good approximation in a discrete market with many participants.

[28] Each school may have multiple programs such as regular education or a specialized program for English Language Learners. Since students may later transfer between programs within a school, and since Pathak and Shi (2013) did not find significant program fixed effects, we include a school effect rather than a program effect.

parameters are (δ, β).

We selected controls following other papers estimating models of school demand, which typically use measures of distance and school performance as controls. Our pre-analysis plan, Pathak and Shi (2014), reports several MNL specifications and elaborates on why we picked the following list of characteristics in x_{ij}:

- distance: walking distance from the student's residence to the school;
- continuing: indicator for whether the student is already enrolled in the school at the previous grade. This primarily affects grade K2, for which many applicants are already enrolled in BPS in grade K1.[29]
- sibling: indicator for whether the student has one or more sibling(s) already enrolled at the school;
- ell match: indicator for the student being ELL and the program being specialized for ELL;
- ell language match: indicator for the student being ELL and the program having a language-specific ELL program in the student's first language;
- walk zone: indicator for whether the student lives in the school's (one mile) walk zone.

The list of characteristics also includes student-school interaction terms. The interacted student characteristics are race and the median household income of the student's census block group. For interactions with racial characteristics, we've included interaction terms that are statistically significant in Pathak and Shi (2014). White, Asian, and Others are grouped together for all three interactions, and Black and Hispanic are grouped together for two of the interactions. The interacted school characteristics are distance and the following:

- mcas: the proportion of the school's students who score Advanced or Proficient in the previous year's MCAS standardized test for math, averaging the proportions for grades 3, 4, and 5.[30]

[29] Grade K1 is before mandatory school starts, so not everyone has to be enrolled in K1. In fact, BPS only has about half as many K1 seats as it has K2 seats.

[30] The MCAS test begins at grade 3. Grade 5 is the highest grade in many elementary schools. We only choose math because it is highly correlated with English, with a correlation of 0.84 in both 2012 and 2013. MCAS performance levels need not be a measure of school effectiveness. Abdulkadiroğlu, Pathak, Schellenberg, and Walters (2017) show that in New York City, applicant preferences are uncorrelated with effectiveness once we control for peer quality.

- % white/asian: the proportion of the school's students who are White or Asian, averaging across all grades (an elementary school contains at least 6 grades, from K2 to grade 5).

When estimating choices, we use the previous year's published data for "mcas" and "% white/asian," as this is the information available to families when submitting their preferences. Note that because both characteristics are predominantly determined by the students of higher grades who were assigned to the school many years before, they are minimally affected by the assignment reform in the first post-reform year. Moreover, apparent preferences for these attributes may be driven by preferences for unobserved characteristics correlated for these attributes. For example, a student may prefer schools with high percentage of Whites and Asians not because that is what the student wants, but because school demographics is correlated with other attributes such as school reputation, neighborhood safety, and existence of certain school programs.

Hausman and Ruud (1987) extend MNL models to situations with ranking data, and we estimate the parameters (δ, β) by maximum likelihood. Table A1 reports estimated parameters and standard errors. It is important to note that in estimation, we only consider choices among inside goods, and do not explicitly model the outside option, because we have little data on outside choices. Moreover, students often enroll in options they did not rank but could have ranked, undermining the usual assumption that an unranked option is inferior to the student's outside option. In our interactions with parents and BPS staff, it seems that many families are ranking few options not because they have better outside options, but because they feel confident they would get into the ones they picked.

When simulating choices using the MNL model, we first draw parameters (δ, β) using the point estimate and covariance matrix from the maximum likelihood estimation. We then draw idiosyncratic tastes ϵ_{ij} independently for each student i and program j. Hence, the simulated choices capture both uncertainty in the model estimation and the unobserved components of preferences.

3.3.2 Mixed MNL (MMNL) Choice Model

This mixed MNL model adds applicant-specific random coefficients to the MNL model, allowing the model to capture heterogeneous preferences for observed characteristics. Suppose we place random coefficients on the first L components of x_{ij}. The model specifies the latent utility of student i for program j as

$$u_{ij} = \delta_{s(j)} + \sum_{k=1}^{K} \beta^k x_{ij}^k + \sum_{l=1}^{L} \gamma_i^l x_{ij}^l + \epsilon_{ij}, \tag{2}$$

$$\gamma_i \sim \mathcal{N}(0, \Sigma), \tag{3}$$

where δ and β are fixed effects and coefficients in the MNL model and γ_i is a L-dimensional vector of individual coefficients, assumed to be distributed according to a multivariate normal distribution. The mean is zero without loss of generality because it is already captured in β, and the covariance matrix Σ satisfies certain restrictions that we specify below. The idiosyncratic term, ϵ_{ij}, is distributed Gumbel(0,1) as in the MNL model. The estimated parameters are (δ, β, Σ). The set of characteristics x_{ij} is the same as in the MNL model and also include mcas and % white/asian (which are explained in Section 3.3.1).

Random coefficients are for the following characteristics, which we organize into "blocks." We assume independence across blocks but allow arbitrary covariance within each block. The blocks are:

Block	Features
1	ell match
2	walk zone
3	distance, mcas, % white/asian.

The covariance matrix Σ therefore satisfies the restriction

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix}, \tag{4}$$

where Σ_1, Σ_2, and Σ_3 are 1×1, 1×1 and 3×3 symmetric positive definite matrices. This formulation allows students to have heterogeneous preferences for ELL programs (if applicable), for schools in the walk zone, and for distance, school performance, and school demographics.

Given their computationally-intensive nature, we were not able to compute estimates of the mixed MNL model prior to the policy change, but we committed to the specification of variables in our pre-analysis report (Pathak and Shi, 2014). The choice of variables allow for a flexible distribution of tastes for school quality, distance, and demographics, which were shown in other studies to be the most important factors in selecting schools (Hastings, Kane, and Staiger, 2009).

Because the log-likelihood function cannot be expressed in a closed form and is no longer globally concave, we fit the model by Markov Chain Monte Carlo (MCMC) methods. One difficulty with our specification is that there are 75 school effects. As far as we are aware, the state-of-the-art MCMC techniques for including fixed effects in mixed logit models, described in Train (2003), involve adding a layer of Gibbs sampling and simulating the conditional distribution of the fixed effects using the Random Walk Metropolis-Hasting algorithm. However, simulating a 75-dimensional distribution is prohibitively slow using Random Walk Metropolis. We therefore use Hamiltonian Monte Carlo (HMC), which incorporates the gradient of the log likelihood function, to quickly update the 75-dimensional fixed effect (Neal, 2011). We fit the model by using 1,000,000 iterations of MCMC sampling, throwing out the first half as burn-in. To check for convergence, we repeat the MCMC procedure six times with independent draws with random starting values. The results from the seven runs are nearly identical. Tables A1 and A2 report the estimated parameters and standard errors. Additional details on the MCMC procedure are in Appendix A.

When simulating choices using the mixed MNL model, we first draw (δ, β, Σ) from the posterior distribution of the MCMC, then draw individual coefficients $\gamma_i \sim \mathcal{N}(0, \Sigma)$, and finally draw the idiosyncratic taste shocks ϵ_{ij}.

3.3.3 Lexicographic Choice Model

When evaluating choice models, Nevo and Whinston (2010) emphasize the importance of comparing them to an alternative. We therefore consider a model motivated by intuitive heuristics, but one that requires no estimation. A model of this form is ad hoc, but it is motivated in part by a report from an individual consulting company commissioned by BPS when the initial plans were announced, which evaluated plans using tiers first and then distance. We posit that every student ranks the programs in his menu lexicographically based on the following hierarchy, treating the highest hierarchy as most important and using subsequent hierarchies to break ties.

Hierarchy	Criteria
1 (most important)	(for continuing students) current program
2	(for continuing students) another program in current school
3	programs in a school where sibling attends
4	(for ELL students) ELL program
5	(for ELL students) ELL program in home language
6	better tier school
7	closer walking distance

Students only consider the hierarchy that pertains to them. For example, applicants who are not continuing students do not consider hierarchies 1 or 2 and non-ELL students do not consider hierarchies 4 and 5.

This choice model does not require parameter estimation, but it is still motivated by past choice behavior and expectations of how applicants would choose in the new plan. For instance, the vast majority of continuing students (91%) rank their current program first, and we anticipated this pattern to continue under the new choice sets. Similarly, most students who have a sibling at a school rank that school first. Furthermore, from conversation with parents and BPS staff, we learned that many people expect families to simply choose schools in the highest tier first and then break ties within tier using distance. Such a pattern was also suggested by researchers from the educations community. For example, Levinson (2015) proposes exactly the same simulation for evaluating equality of opportunity among non-ELL, non-continuing students: "pretend that every child who entered the 2014 lottery preferred higher-tier schools to lower-tier schools, and within each tier, preferred closer schools to further away schools." For ELL students, BPS staff believed from their interactions with families that the vast majority of ELL students would prefer ELL programs, especially ones targeted to the student's home language.

The Lexicographic model has support from the psychology and marketing literature. The model is related to Tversky (1969)'s lexicographic semi-order choice model in which options are rated with respect to a variety of attributes and there is a lexicographic order across attributes. Between two options, an agent first compares the most important attribute, and if there is significant difference, the agent chooses the better option according to this attribute; if there is little difference, then the agent goes to the next attribute. This encapsulates the Lexicographic model above if we define the academic quality of schools based on tier. Another related choice model is Tversky (1972)'s elimination by aspect. In this model, the agent chooses an option from a set by going through different aspects (discrete attributes) in order of importance and eliminating options that are sub-optimal with respect to that aspect. Although the original paper allowed for probabilistic choice of aspects, subsequent papers use a deterministic order of aspects: see, for example, Thorngate (1980), Johnson, Meyer, and Ghose (1989), and Payne, Bettman, and Johnson (1988). Our lexicographic choice model is a special case of elimination by aspects with a deterministic ordering of aspect given by the hierarchy.[31]

The empirical support behind lexicographic models makes it a useful benchmark for random utility models. Slovic (1975) conducts a laboratory experiment involving a choice between two options evaluated on two dimensions, and show that the majority of subjects chose consistently based on the more important dimension. Tversky, Sattah, and Slovic (1988) conduct other laboratory experiments and show that in cases in which a decision is framed as choosing from a set, then a lexicographic rule is often used. (If the same decision is framed in terms of varying a numerical dimension to make the decision maker indifferent between the two options, then subjects are less biased toward the more important dimension; this suggests that in decisions framed as a choice, the lexicographic rules are good approximations of reality.) In the marketing literature, Drolet and Luce (2004) show that lexicographic rules are adopted more often when consumers have emotional reasons to avoid making trade-offs. Yee, Dahan, Hauser, and Orlin (2007) study consumer cell phone choices and fit a variety of choice models to consumer choice data. A lexicographic rule by aspect predicts 75% of choices, which is comparable to other discrete choice procedures.

3.3.4 Predicting Who Applies

For grades K1 and K2, all students who wish to be assigned to a Boston public school must participate in the choice process. The set of applicants is therefore an important determinant of our equilibrium targets. A large influx of new applicants in a given neighborhood would increase the number of them unassigned and reduce average access to top tier schools. If we had data on all potential applicants and their non-BPS options, we might include the decision to participate as

[31] The lexicographic-type models have also been axiomatized. Fishburn (1974) surveys the older literature. Kohli and Jedidi (2007) study when lexicographic orders can they be represented by a linear utility function. Manzini and Mariotti (2012) generalize the original Tversky (1969) model to choosing from more than two options.

part of the choice model. Since we don't have this data, we need to reflect the uncertainty in the applicant pool and to capture any trends in the neighborhood participation patterns.

To predict who applies, we use demographic trend projections. A similar approach was used in the McFadden, Reid, Talvitie, Johnson, and Associates (1979) BART study; the authors write, in Chapter IV.3,

> "It is in the nature of auxiliary forecasting that one does not have available complete structural or causal models; hence, forecasting must use data analysis and trend projection techniques, combined with available external forecasts."

The BART study uses census demographic data and projections to construct a representative sample of San Francisco households. Since we have the universe of participants in previous years, we directly observe the joint distribution of household characteristics and use it to predict the applicant pool.

All relevant information on the applicant pool can be encoded in a matrix X, where each column vector X_i corresponds to a particular student i and contains all observable characteristics associated with the student. Let the matrix for the actual applicants be X^*. We do not observe X^* at the time of the prediction exercise, but we use the following method to construct a matrix X to act as a proxy. Due to uncertainty in the applicant pool, the process by which we construct X is based on random sampling. There are five sets of inputs based on past data:

- X^{cont}: a matrix of characteristics for all potential applicants who would be continuing students. (These students are already enrolled in BPS at a lower grade at the time of application.) The format of the matrix is the same as described above, with one column for each student.

- X^{new}: a matrix of characteristics of the previous year's new applicants. (New applicant's are those who were not continuing students when they applied, and the characteristics of last year's new applicants act as proxies for this year's new applicants).

- $\mu^p_{gh}, \sigma^p_{gh}$: predicted mean and standard deviation of the proportion of currently enrolled students in grade $(g-1)$ from neighborhood h who would apply as continuing students for grade g in the upcoming application.[32]

[32]For grade K2, $(g-1)$ is K1; for grade K1, $g-1$ is K0. There are 14 possibilities for h, corresponding to the 14 neighborhoods.

- μ^q, σ^q: predicted mean and standard deviation of the total number of new applicants to BPS for grades K0 through K2.
- $\mu^{\alpha}_{gh}, \sigma^{\alpha}_{gh}$: predicted mean and standard deviation of the proportion of new applicants who are applying for the grade-neighborhood combination g and h.

We use the following process to construct a sample matrix X of applicant characteristics. First, initialize X to an empty matrix.

1. Sample the following quantities

$$p_{gh} \sim Normal(\mu^p_{gh}, \sigma^p_g h) \qquad \text{for each } g \text{ and } h, \tag{5}$$

$$q \sim Normal(\mu^q, \sigma^q), \tag{6}$$

$$\alpha_{gh} \sim Normal(\mu^{\alpha}_{gh}, \sigma^{\alpha}_{gh}) \qquad \text{for each } g \text{ and } h. \tag{7}$$

   In each case, if the sampled value is negative, then replace it with zero.[33] Define $n_{gh} = round(\alpha_{gh} q)$, where $round(\cdot)$ denotes rounding to the nearest integer. n_{gh} corresponds to the predicted number of new applicants for grade g from neighborhood h.

2. For each column of X^{cont}, denote the column by X^{cont}_i, where i corresponds to the currently enrolled student who may apply as a continuing student. Let g be the grade that the student is moving up to and h be the neighborhood of the student's residence. Add the column X^{cont}_i to X with probability p_{gh}, independently from everything else.

3. For each grade $g \in \{K1, K2\}$ and neighborhood h, let X^{new}_{gh} be the columns of X^{new} corresponding to new applicants for this grade and neighborhood combination. Sample with replacement n_{gh} columns of X^{new}_{gh} independently and add them to X.

The end product is a matrix X whose columns are drawn either from X^{cont} or X^{new}. Note that the above method of sampling allows for common shocks for all students for a grade and neighborhood combination (g, h) in the continuing probability p_{gh} and the number of new applicants n_{gh}. Moreover, the values of n_{gh} can be correlated across grades and neighborhoods, due to the common term q.

[33] Sampling a negative value from these normal distributions can theoretically happen, but did not actually occur in any simulation because the respective means turns out to be many times larger than the standard deviations.

The parameters $(\mu^p_{gh}, \sigma^p_{gh}), (\mu^q, \sigma^q), (\mu^\alpha_{gh}, \sigma^\alpha_{gh})$ are estimated using past 4 years of data, from years 2010 through 2013. For each tuple of (μ, σ), we estimate a linear regression using the corresponding quantity in the past 4 years, with the year as the independent variable (along with a constant). If the regression slope is significant at a 95% level, then we set (μ, σ) to the predicted mean and standard deviation for 2014 from the regression. This allows us to incorporate time trends, such as an increase in the total number of new applicants to BPS across the years, or an increased trend in a particular neighborhood. If the slope is not significant at a 95% level, then we reject the trend and set (μ, σ) to be the sample mean and sample standard deviation of the corresponding quantity from the past four years. The parameter estimates are in Appendix A of Pathak and Shi (2015).[34]

4 Evaluating Predictions

4.1 Equilibrium Outcomes

Our metrics for evaluating equilibrium outcomes (i.e. access to quality, unassigned, or distance) take several forms of uncertainty into account. Let ω_h be a random variable that corresponds to the simulated outcome for neighborhood h generated by a choice model, and let outcome vector $\omega = (\omega_1, ..., \omega_H)$, where $H = 14$ is the number of neighborhoods. Each of our prediction model implicitly specifies a distribution Ω for the outcome vector, in which the randomness arises from the sampling used to generate the applicant pool, and in sampling the choice model parameters, unobserved error terms, and lottery numbers. The prediction for neighborhood h is $\bar{\omega}_h \equiv \mathbb{E}_{\omega\sim\Omega}[\omega_h]$. This quantity can be estimated by sampling ω_h many times and taking the average.

Let ω^* denote the actual outcome vector, computed using the actual population, actual choices and actual lottery numbers. The root mean squared error (RMSE), which is our main measure of prediction error for equilibrium outcomes, is defined as the Euclidean distance between the predicted and the actual outcome vectors:

$$\textbf{RMSE} \equiv \|\bar{\omega} - \omega^*\| \equiv \sqrt{\sum_{h=1}^{H} (\bar{\omega}_h - \omega^*_h)^2}. \tag{8}$$

[34] In the end, we only kept regression estimates for (μ^q, σ^q) and $(\mu^\alpha_{gh}, \sigma^\alpha_{gh})$ for the grade neighborhood combinations K1 Charlestown and K2 Downtown. For these three sets of estimates, we detected an upward trend. For all other estimates, we set (μ, σ) to be the sample mean and sample standard deviation of the data in the past 4 years.

The RMSE is an overall measure of prediction error across neighborhoods. To measure uncertainty in the prediction, we define

$$\textbf{Expected RMSE} \equiv \mathbb{E}_{\omega\sim\Omega}[\|\bar{\omega} - \omega\|]. \tag{9}$$

The expected RMSE measures how much prediction error we should expect when the choice model is correct, which is the ideal scenario in which the actual outcome is also drawn from the predicted distribution Ω.

For each grade and neighborhood, and each predicted outcome, we estimate a 95% prediction interval by computing the empirical 2.5^{th} and 97.5^{th} percentile of 1000 simulations. Let this interval be denoted as Ω_h. The proportion of neighborhoods for which the actual outcome is within the prediction interval,

$$\textbf{\% in 95\% P.I.} = |h : \omega_h^* \in \Omega_h|/H, \tag{10}$$

is our third measure of prediction accuracy. If the model is correct, then we expect this to be close to 95% on average.

4.2 Choice Forecasts

To measure prediction accuracy for individual choices, we report the percentage of prediction mistakes for top choices and for pairwise comparisons from the rank order list. For a given choice model, let y_i be the vector of simulated preference rankings of student i. y_{i1} is the index of the top choice, y_{i2} is the index of the second choice, and so on. Define the set of top k choices as $Y_{ik} \equiv \{y_{i1}, ..., y_{ik}\}$. Similarly, let y_i^* denote the student's actual choice ranking, r_i denote the actual number of choices ranked, and Y_{ik}^* denote the actual set of top k choices. Denote $I_k = \{i : r_i \geq k\}$ as the set of students that ranked at least k choices.

Given a choice model, define $\hat{Y}_{ik}$ to be the best prediction of the top k choices for student i. We predict by simulating the top k choices $Y_{ik} = \{y_{i1}, \cdots, y_{ik}\}$ many times and taking the k most common choices across simulations. Our first measure computes prediction mistakes among the top k choices:

$$\textbf{\% Mistakes in Top k Choices} = \frac{1}{|I_k|}\sum_{i\in I_k}\frac{|\hat{Y}_{ik}\backslash Y_{ik}^*|}{k}. \tag{11}$$

That is, we report the average proportion of predicted top k choices in the set $\hat{Y}_{ik}$ that are not in the actual set of top k choices, Y^*_{ik}, counting only students who ranked at least k choices. When $k = 1$, for example, this measures the fraction of top choices that turn out to be different from our prediction for the individual.

A second measure of prediction error considers pairwise comparisons. Given the actual ranking y^*_i, define the set of pairwise comparisons implied by this ranking to be the collection of ordered pairs:

$$C^*_i = \{(j, l) : \text{program } j \text{ is ranked before } l \text{ in } y^*_i\}. \tag{12}$$

A pair of programs (j, l) is in this set if both programs j and l are ranked and j is preferred, or if j is ranked and l is unranked. Given a choice model, define $\hat{z}_i(j, l)$ to be the indicator variable for whether the choice model predicts the student to prefer option j over l with greater than 50% probability.[35] Define the percentage of prediction mistakes in pairwise comparisons to be the proportion of comparisons that the choice model predicts incorrectly:

$$\textbf{\% Mistakes in Pairwise Comparisons} = \frac{1}{|I_1|}\sum_{i\in I_1}\frac{1}{|C^*_i|}\sum_{(j,l)\in C^*_i}(1 - \hat{z}_i(j, l)). \tag{13}$$

For predicting the distribution of choices, we measure accuracy using statistical distance, which is a common metric for the distance between two probability distributions. It is also referred to as the total variation distance. We compute this metric for the distribution of top choices and the joint distribution of the top two choices. For neighborhood h, define market share s_{hjk} to be the average proportion of top k choices from this neighborhood that are for school j, counting only students who ranked at least k choices. Let I_{hk} denote the set of students from this neighborhood who ranked at least k choices. The predicted top k market share of school j in neighborhood h is

$$s_{hjk} = \frac{1}{|I_{hk}|}\sum_{i\in I_{hk}}\mathbb{E}[|\text{programs in } Y_{ik} \text{ at school } j|]/k. \tag{14}$$

Similarly, define the actual market share s^*_{hjk} using the actual set of top k choices Y^*_{ik} instead of the predicted set Y_{ik} for each student i. The statistical distance between two vectors is the

[35]Following this 50% rule results in the fewest number of prediction mistakes if the choice model were exactly correct. This is because for any Bernoulli random variable X with parameter $p > 0.5$, the prediction that minimizes ex-post mistakes is that $X = 1$ every time.

minimum mass needed to transform one vector to the other. The average statistical distance across H neighborhoods is defined as:

$$\textbf{Statistical Distance in Top } k \textbf{ Market Share} = \frac{1}{2H}\sum_{h=1}^{H}\sum_{j}|s_{hjk} - s^*_{hjk}|. \tag{15}$$

The final metric we examine is for the joint distribution of the two highest-ranked schools. For the pair of schools (s_j, s_l), define p_{jl} as the proportion of students who ranked at least two choices and ranked a program in school s_j first and a program in school s_l second. Similarly, we define p^*_{jl} for the corresponding actual choice rankings. The following measures the minimum mass needed to transform one distribution to the other:

$$\textbf{Statistical Distance in Joint Distribution of Top 2 Choices} = \frac{1}{2}\sum_{jl}|p_{jl} - p^*_{jl}|. \tag{16}$$

5 Back-testing and Hypotheses Formulation

How well do we expect to predict the outcomes using our approach? To set expectations, we report on a back-testing exercise that applies our methodology to data from two years before the reform (2012) to predict outcomes one year before the reform (2013). Since applicant choice sets did not change between these years, we expect the results from the back-test to provide a best-case scenario for what we might expect following the reform.

To illustrate the underlying calculations for prediction error, we plot in Figure 4 the predicted and actual access to quality by neighborhood in 2013. For each prediction, the figure also plots the 95% prediction interval. These estimates allow us to compute the three metrics for prediction accuracy of equilibrium outcomes described in Section 4.2. For each choice model, the root mean squared error (RMSE) measures the overall difference between the predicted and the actual access to quality. The expected RMSE corresponds to how long the error bars are in Figure 4. The % of predictions within the 95% prediction interval corresponds to the proportion of actual access to quality that falls within the error bars.

For each grade and each prediction target, the MNL and MMNL models exhibit nearly-identical RMSE. This is shown in Table 2. Moreover, for every outcome except access to quality in grade

K1, the MNL-based models exhibit smaller RMSE than the Lexicographic model. However, the absolute performance of the MNL-based models is not as high as one might have hoped: the RMSE is larger than expected, and the actual outcome is within the 95% prediction interval only about 70% of the time, averaging across outcomes. For the Lexicographic model, the performance is worse: the actual outcome is inside the 95% prediction interval less than 40% of the time.

The MNL and MMNL models also outperform the Lexicographic model for individual choice predictions. Table 3 reports on MNL performance for each grade and for the top 1, top 2, and top 3 choices. As a benchmark, Table 3 also tabulates the accuracy of random guessing. For grade K1, the table shows that random guessing incorrectly predicts the top choice 97% of the time. The Lexicographic model incorrectly predicts the top choice 63% of the time, which means that it predicts the top choice (out of more than 30 options) correctly 27% of the time. For pairwise comparisons, random guessing by definition predicts wrongly 50% of the time. The Lexicographic model predicts incorrectly 30% of the time, while the MNL and MMNL models reduce the percentage of mistakes to 18%. A similar comparison holds for grade K2. In summary, for predicting individual choices, the MNL-based models are indistinguishable, and both outperform the Lexicographic model.

For predicting distribution of choices, the Lexicographic model is not much better than random guessing, and for the joint distribution of top 2 choices, it is even worse than guessing. These findings are shown in Panel B of Table 3. The Lexicographic model does not allow students to prefer a more distant school in the same tier to a closer school, if the continuing, sibling, and English Language Learner status of the student are the same at both schools. For these metrics, the performances of the MNL-based models are nearly identical, and both outperform random guessing and the Lexicographic model.

As a result of this analysis, we formulate the following hypotheses before choices are submitted in the new plan:

- For equilibrium forecasts, the MNL-based choice models would perform similarly to one another, and both would systematically outperform the Lexicographic model. For all models, the actual prediction error would be significantly larger than what one would have expected if one believed the models to be exactly correct.

- For choice forecasts, the comparison across choice models would be the same as the equilibrium forecasts. Moreover, the Lexicographic model would reasonably predict individual choices, but would perform poorly for the distribution of choices.

6 Comparing Forecasts and Prediction Errors

6.1 Equilibrium and Choice Forecasts

To illustrate the underlying calculations of the prediction exercise, Figure 5 plots the actual access to quality in each neighborhood in the first post-reform year (2014), as well as the predicted access to quality according to each choice model. For each prediction, the figure also contains the 95% prediction interval. The calculations for the aggregate measures of prediction accuracy are analogous to that of the back-test. THe predictions were published in our pre-analysis plan Pathak and Shi (2014) before the actual choice data were submitted.

Contrary to our first hypothesis, the MNL-based models outperform the Lexicographic model in predicting equilibrium outcomes only for grade K1, but not for grade K2. Table 4 shows that for grade K1, the MNL-based models exhibit a smaller RMSE than the Lexicographic model, with a significantly higher fraction of outcomes being within its 95% prediction interval. This follows the pattern observed in the back-test. However, for grade K2, the Lexicographic model exhibits similar RMSE as the MNL model, with slightly better prediction accuracy for two out of the three targets, namely: access to quality and the number of unassigned students. Moreover, for these two targets, the percentage of neighborhoods for which the outcome is within the 95% prediction interval is also higher in the Lexicographic model compared to the MNL-based models.[36]

We cannot reject the other hypotheses about equilibrium outcomes in Table 4. In all cases, the MNL and MMNL models exhibit nearly identical results, regardless of whether we consider the RMSE or the fraction of predictions within the 95% prediction interval. The expected RMSE is also similar between the two models. In addition, regardless of the model or the metric, the actual RMSE is higher than the expected RMSE, which shows that none of the models is absolutely accurate.

For choice forecasts, the results are consistent with our hypotheses, as they follow the pattern

[36] This phenomenon is not due to greater uncertainty in the Lexicographic model prediction. Column 5 of Table 4 shows the expected RMSE of Lexicographic is similar to that of the MNL-based models.

of those reported in the back-test. Table 5 shows that the MNL and MMNL models exhibit nearly identical performance, regardless of the grade and metric. Furthermore, the prediction error is smaller in the MNL-based models than in the Lexicographic model. The amount by which the MNL-based models outperform the Lexicographic model is also much higher for the distribution of choices than individual choices. This pattern was also present in the back-test.

6.2 Decomposing Prediction Errors

The inconsistent performance that we found of the MNL-based models in predicting equilibrium outcomes can be due to a variety of causes, including: 1) unexpected realization of applicant pool characteristics, 2) unexpected changes in choice pattern due to changes in the popularity of schools, 3) unexpected changes in choice patterns due to framing and other behavioral issues, and 4) unexpected changes in the number of choices ranked by students. The first source of error is related to the applicant pool forecast; the second source is related to our level of sophistication in accounting for time trends; the third source is related to the validity of the choice models themselves, and the last source is related to our simplifying assumption that everyone ranks up to ten options.

The largest source of error turns out to be in the applicant pool predictions. When we reproduce the prediction exercise for equilibrium outcomes using the characteristics of actual post-reform applicants, we find that the MNL-based choice models consistently outperform the Lexicographic model, as expected from our back-test. Table 6 shows that the RMSE of MNL-based models are smaller than the Lexicographic model for both grades and all outcomes of interest,[37] and the proportion of actual outcomes that fall within the prediction interval is higher. Furthermore, the MNL and mixed MNL models have nearly identical performance, as expected from the back-test.

Moreover, we find evidence that the choice patterns themselves remained stable across the reform, and that any issue with our assumption on the length of ranked-order list is not of first-order importance. The stability of choice patterns suggests that both family's underlying preferences for schools are not significantly affected by the new assignment system, and that unobservable school characteristics such as school quality remained relatively stable across the two years.[38] The

[37]For the number of unassigned students, all three models perform similarly in RMSE, but the prediction intervals from MNL-based models cover the actual outcome more often than the intervals from the Lexicographic model, despite a similar expected RMSE.

[38]We do not claim that the preference distributions of families and the unobservable characteristics of schools are statistically indistinguishable before and after the reform. Indeed, Table A1 and A2 document statistically significant

stability of family preferences imply that the salience effect of highlighting school tiers were limited. Table 7 reports on how prediction accuracy changes when we allow the prediction to use additional information from the post-reform dataset. In this table, we report the RMSE of the MNL-based models under the following assumptions:

- New Applicants with Old Demand Model: Using the actual set of post-reform applicants and their characteristics (instead of a simulated applicant pool from pre-reform), but fitting the choice model using pre-reform choice data. This is the same as in Table 6.
- New Applicants with Refit Demand Model: Using the actual set of applicants and their characteristics, and fitting the choice model also using post-reform choice data.[39]
- New Applicants with Refit Demand Model and Ranking Length: The same as above, except also using the actual number of choices ranked by each applicant.
- Sampling Actual Choices and Using Applicant Forecast: Using the predicted number of students from each neighborhood, but sampling students from the actual applicant pool and using the actual choices of these students. In predicting the number of students, we follow the sampling methodology in the original forecasts.

Comparing the prediction error from these assumptions shows the following:

1. The MNL-based models do isolate stable components of preferences in the sense that the model estimated from pre-reform data predicts outcomes just as well as the model estimated from post-reform data. The RMSE in Table 7 for "New Applicants with Old Demand Model" is similar to "New Applicants with Refit Demand Model."
2. The assumption about rank-order list length is not of first-order importance. When we control for the actual lengths of submitted rank-order lists, the prediction error only improves for access to quality, but not for distance to school and the number unassigned. In comparison, predictions of the applicant pool are first-order, as the RMSE improves significantly for every metric when we compare the original forecasts with a simulated applicant pool to the version using the actual applicant pool.

differences from 2013 and 2014 estimates. However, these distributions have not changed so much so as to affect the prediction accuracy of the equilibrium outcomes of interest, as shown in Table 7 and explained below.

[39] Table A1 and Table A2 contain the coefficient estimates.

3. Much of the overall error in the original forecast is due to predicting the wrong number of students from each neighborhood. This is seen in how large the prediction error is with Sampling Actual Choices Using Applicant Forecast.

Our findings suggest that the MNL-based models' unexpectedly poor performance in the original prediction exercise is primarily due to poor predictions of the applicant pool, rather than due to changes in choice patterns or in the lengths of rankings. Hence, choice models may effectively predict counterfactual outcomes, as long as there are accurate forecasts about auxiliary input variables.

The robustness of the MNL-based models across the reform is further supported by an analogous exercise that compares prediction accuracy for choice outcomes using models estimated from pre-reform and post-reform data. Note that using the MNL model estimated from post-reform choice data should always perform better than the model using pre-reform data since we are evaluating on post-reform data. However, Table 8 shows that the improvement from estimating using post-reform data is small: the choice patterns are stable enough so that it is reasonable to use a model estimated from pre-reform data to predict post-reform outcomes.

Finally, we examine the effect of how choices are presented. Figure 6 shows that the salience of tier information in the new presentation of choices has a small effect on the distribution of preferences: for grade K1, the actual percentage of students who ranked a Tier 1 school as top choice is not higher than predicted by the MNL model. For grade K2, the actual percentage is slightly higher, but only by a few percentage points.[40] The Lexicographic model, on the other hand, vastly overstates the salience effect of tiers, as it predicts that more than half of the students will rank a Tier 1 school first, when the actual percentage is less than 35%.[41]

6.3 Causes of Errors for Population Forecast

The errors in population forecasts resulted in the MNL-based models performing badly for grade K2 in predicting equilibrium outcomes. Table A3 shows that there are three major errors: (1) the number of continuing K2 students is much larger than predicted, (2) the number of new grade

[40] Hastings and Weinstein (2008) find in a study on the Charlotte-Mecklenburg school district that information cues in school choice may affect school market shares by about 5%, which is of the same order of magnitude as we find here.

[41] Even under the lexicographic model, students may rank a lower tier school first if he is a continuing student at the school, has an older sibling at that school, or is an ELL student wanting to go to a particular language program.

K1 and K2 students is significantly less than predicted, and (3) the proportion of grade K2 ELL students are smaller than predicted.

Were these errors foreseeable? It is difficult to comment on this with any level of rigor since nearly anything can seem foreseeable after the fact. Nevertheless, we give our best guesses below. We think that the first source of error was possibly foreseeable, as it is caused by misunderstanding how BPS assigns continuing students. We assumed that currently enrolled students who wish to continue are assigned the same program code for the next grade, but in reality BPS sometimes changes the program code when students change grades and our forecast did not capture these changes adequately. The second error is unexpected, as the number of applicants had been rising in previous years. The low number of applicants is due to either a break in the previous trend in the number of kindergarten-aged children in Boston or a greater substitution to school options outside of BPS, including charter and private schools or public schools in neighboring districts. If second effect is dominant (substitution to other systems), then a more complex structural model in which application to BPS is endogenous might have improved the predictions. Unfortunately, our data do not allow us to distinguish between these two alternatives.[42] The third discrepancy is driven by a simultaneous change in the test that BPS uses to determine eligibility for ELL programs, as the new test decreased the proportion of eligible students. This third change was done by the BPS Office of English Learners, which has little overlap with the office in charge of school assignment. The effect of this change was unforeseen by the BPS staff we were working with in the enrollment department, and was therefore would have been hard to foresee for any outside researcher.

7 Selecting Another Policy

While our analysis has focused on the absolute accuracy of the choice models, in this section, we consider whether BPS would have chosen a different choice plan given the prediction errors. Even if the prediction errors are large in an absolute sense, they may have not affected the relative ranking of alternative plans or BPS's policy decision.

[42]What would help us distinguish between the two explanation of reduced number of K2 new applicants in 2014 is reliable data on the number of appropriately aged children in Boston in 2013 and 2014. If there is a corresponding drop in the number of 5 year olds between these years, then reduced participation may be simply due to population trends. If there is no drop or an increase in the population counts, then we may conclude that the drop in BPS participation is driven by substitution to other school system.

The alternative choice plans we consider are the 2012-2013 school assignment reform proposals described in Pathak and Shi (2013). Most proposals partition the city into alternative zones, ranging from six to 23. Two proposals are variants of the Home-Based plan. The decision process that led Boston to adopt the Home-Based plan involved a compromise across several dimensions. But the effects on access and proximity were central, and the school board was also concerned about insufficient school capacity. We therefore evaluate the relative performance of other plans with respect to these three equilibrium targets.

Table 9 reports on access to quality for grade K1 for the Allston-Brighton neighborhood. Each entry of Panel A reports access to quality for eight plans for different choice models and applicant samples. Column 1, for example, shows that access to quality is highest under the 10 Zone plan, according to the MNL choice model estimated with post-reform choices and post-reform applicants. In contrast, access to quality is the lowest under the status quo. Since we cannot directly compare plans that were not implemented, column 1 serves as our reference point. The relative ranking across the eight plans is completely unchanged when we use the MNL model fit using pre-reform choice data, but applying the model on post-reform applicants. Panel B shows that, of the possible comparisons (e.g., Status Quo vs. Home Based A, Status Quo vs. Home Based B, Home Based A vs. Home Based B, etc.), there are no reversals of pairwise comparisons.[43]

How would the comparisons across plans be affected by prediction errors in applicant pool charateristics? Column 3 of Table 9 assesses this by forecasting access to quality based on pre-reform choices and applicants. This more closely mirrors the Pathak and Shi (2013) report. The forecast provides a more optimistic scenario for both versions of the Home-Based plan compared to the reference in column 1. Specifically, the Home-Based plans have the highest access to quality after the 10 Zone plan, but in column 1 they have the lowest access to quality after the status quo. In fact, there are reversals across 8 of 22 possible non-trivial pairwise comparisons of plans. This suggests that if Boston chose a plan based only on access to quality in the Allston-Brighton neighborhood, the MNL model forecast could have led to a different choice. However, this pattern is also present with the Lexicographic model, under which access is higher under the Home-Based plans than other alternatives. There are more reversals of pairwise comparisons under Lexicographic

[43]For this tabulation, we only consider comparisons where the difference in access is at least 1%, to avoid tallying trivial differences across plans.

than MNL.

Access to quality in a given neighborhood is not the only factor used to select among plans. We therefore report on how the ranking across plans changes under different choice models, aggregating across the three outcomes and 14 neighborhoods in Table 10. Only 5% of the pairwise comparisons across plan dimensions change when the MNL model is fit from pre-reform data compared to post-reform data, which supports our argument that choice patterns are stable across the reform. In other words, any effect due to information cues would not have significantly changed the relative ranking across plans. However, column 3 shows that there are larger reversals across the ranking of plans with the choice model fit using pre-reform choices and applied on the pre-reform applicant pool. As shown in the last row of the table, the relative rankings reverse on average 16% of the times, which is three times as high as when we had applied the choice model on the same post-reform applicant pool as in the point of reference. This shows that the relative ranking of alternative policies may change significantly due to errors in forecasting applicants. In other words, Boston may have chosen another plan had there been a better forecast of the auxiliary variables.

Whether the MNL forecast's susceptibility to errors in who applies undermines its value for decision-making depends on the performance of the alternative. Column 3 shows that errors in forecasting applicants do not erase the benefits of the MNL model compared to Lexicographic. Despite the errors in the forecast of the auxiliary variables, there are significantly fewer prediction reversals with the MNL model fitted from past data than the Lexicographic model, under which nearly one-third of pairwise comparisons across plans are reversed. The performance of the Lexicographic here is not much better than a random prediction, which would reverse one-half of comparisons. In summary, even though counterfactual comparisons of choice plans are sensitive to prediction errors in the auxiliary inputs, there is still value in using a discrete choice model instead of our ad hoc alternative for comparisons between plans.

8 Conclusions

This paper reports on an out-of-sample validation of discrete choice models of school demand. Forecasts from these models influenced a policy change that affected thousands of Boston families. We made predictions prior to the policy change, so as to make sure that the forecasts are truly ex ante, free from the biases of the researcher's ex post rationalizations. Since we observe choices participants made in the new policy, we also conduct a decomposition of sources of prediction error.

We find that, once we control for changes in the environment outside of the structural model, the choice models are reasonably accurate compared to expectations set by back-testing. Both the MNL and mixed MNL choice model significantly outperform the Lexicographic model, when using the actual applicants. Moreover, the MNL-based models perform similarly when refit with post-reform data, suggesting that the preference distribution measured by the choice model is stable even with a large change in choice sets and how choices are framed. We also find that the MNL model's performance is similar to the mixed MNL model, a fact foreshadowed by the back-tests. The richness of the data we have on individual characteristics likely reduces the potential benefit of the more flexible and computationally-intensive specification.

The scenario in which an analyst has access to the actual participants under the new policy allows us to focus attention on choice model performance. Yet this hypothetical scenario does not correspond to any real-world forecasting problem. Without the actual participants, the magnitude of the error from the applicant forecast is so large for grade K2 that it undermines the benefit of the MNL model. In fact, the prediction error from the Lexicographic specification is smaller for several forecast targets compared to the MNL model. Our decomposition shows that the Lexicographic's superior performance in grade K2 is driven by errors in the applicant and choice forecasts counteracting one another. The error in the MNL forecast is also large enough to change the ranking of several other alternative policies, and may have led the city to pick a different plan. However, the negative effects of errors in the applicant forecast do not erase the benefit of discrete choice modeling: despite the presence of the errors in auxiliary inputs, the model fitted from pre-reform data still reproduces the majority of counterfactual comparisons across plans, and does so much more consistently than the Lexicographic alternative.

In absolute terms, there is still substantial scope to improve the demand model predictions. An open question is whether a more principled approach to variable selection in the choice models would have led to further improvements. It's also possible that alternative non-discrete choice approaches than the Lexicographic model would have had higher performance.

Another open question is whether one can construct in our setting a reduced form approach to forecast equilibrium outcomes based on changes in choice sets, without the need to model individual student choices. The difficulty of a purely reduced form approach in our setting is two fold: finding suitable features of choice sets using which to directly forecast outcome, and finding sufficient exogenous variation in the data so as to identify the impact of each feature on the outcomes of interest. Finding suitable features is difficult because there are many possible combinations[44] of choice sets across neighborhoods and it is unclear how to compress this information into a manageable number of features.[45] Finding sufficient variation is difficult because the choice sets were largely constant in Boston from 1988 to 2013, and variation across neighborhoods are confounded by spatial differences in demographics and school qualities.

Discrete choice models have widespread applications in economics beyond school demand. Our setting and policy change show possibilities for scenarios where substitution among choices is central. While standard choice models may succeed in predicting choice behavior, there can still be significant unforeseen prediction error for policy relevant outcomes due to changes in the environment that are outside of the model. Difficulty predicting these auxiliary inputs likely plays a large role in other applications.

[44]With 80 schools, the number of potential choice sets for one neighborhood is $2^{80} \approx 1.2\times10^{24}$. Assuming everyone in the same neighborhood receives the same choice set. The number of combinations of choice sets is $2^{80\times14} \approx 1.3\times10^{336}$.

[45]To capture competition, it is important to define features based not only on a student's own choice set, but also the choice sets of others. For example, if I have the best school in my choice set and no one else does, then I have perfect access to this school. However, if everyone else also has this school in their choice set, and the school capacity is small, then my access to the school would be low.

A Estimating the Mixed MNL Choice Model

Unlike in the MNL model, the log likelihood function associated with the mixed MNL (MMNL) model is difficult to evaluate directly since it involves multi-dimensional integrals with no closed form formula. Hence, we estimate the MMNL model using Markov Chain Monte Carlo (MCMC) instead of maximum likelihood.

Train (2003) reviews the basic framework to estimate MMNL models using MCMC. The framework is based on Gibbs sampling and the Metropolis-Hasting algorithm. However, our setting has many fixed coefficients since we have a fixed effect for each of the 80 or so schools.. It is known that the simple Metropolis Hastings with random walk proposals does not perform well when estimating a vector of many dimensions (Katafygiotis and Zuev, 2008), especially if the dimensions are correlated. We therefore modify the framework to use Metropolis-Within-Gibbs (MWG), which samples blocks of coordinates iteratively (rather than all coordinates at once), and Hamiltonian Monte Carlo (HMC), which incorporates gradient information for directions to sample. We describe these methods in greater detail in Section A.2.

A.1 Specifying the Likelihood Function

The first step of applying MCMC techniques is specifying the full likelihood function of observing the data given the model parameters. An equivalent representation of the MMNL model from Section 3.3.2 is as follows. Let the vector of characteristics $x_{ij} = (x_{ijr}, x_{ijf})$, where x_{ijr} corresponds to the first L components, which represent the terms with random coefficients, and x_{ijf} the last $K - L$ components, which have fixed coefficients. Define coefficient vector $\beta = (\beta_r, \beta_f)$ similarly. The latent utilities are as follows.

$$u_{ij} = \delta_{s(j)} + \beta_f \cdot x_{ijf} + \gamma_i \cdot x_{ijr} + \epsilon_{ij}, \tag{17}$$

$$\gamma_i \sim \mathcal{N}(\beta_r, \Sigma), \tag{18}$$

$$\epsilon_{ij} \sim \text{Gumbel}(0, 1), \tag{19}$$

The set of parameters to be estimated is (δ, β, Σ). In order for the model to be well-specified, we normalize the last component of δ to be zero. Moreover, the covariance matrix Σ can be written in the block diagonal form

$$\Sigma = \begin{pmatrix} \Sigma_1 & & \\ & \Sigma_2 & \\ & & \Sigma_3 \end{pmatrix}, \tag{20}$$

where Σ_1, Σ_2, and Σ_3 are 1×1, 1×1 and 3×3 symmetric positive definite matrices.

The data to fit these parameters are the observed choices of every student along with the observed characteristics vector x_{ij}. Suppose that student i makes m_i choices, and let the chosen programs from best to worst be $y_{i1}, y_{i2}, \cdots, y_{im_i}$.

The likelihood function can be expressed as follows. Given γ_i, the conditional likelihood is

$$\phi_i(\delta, \beta_f | \gamma_i) = \prod_{c=1}^{m_i} \frac{\exp(\delta_{s(y_{ic})} + \beta_f \cdot x_{iy_{ic}f} + \gamma_i \cdot x_{iy_{ic}r})}{\sum_{d=c}^{m_i} \exp(\delta_{s(y_{id})} + \beta_f \cdot x_{iy_{id}f} + \gamma_i \cdot x_{iy_{id}r})}. \tag{21}$$

This is the MNL likelihood function. The full likelihood function incorporating all the data is

$$\Phi(\delta, \beta_f, \beta_r, \Sigma) = \prod_{i=1}^{n} \int_{\mathbb{R}^5} \phi_i(\delta, \beta_f | \gamma_i) \exp(-\frac{1}{2}\Sigma^{-1} \|\gamma_i - \beta_r\|^2) d\gamma_i. \tag{22}$$

Here, n is the number of students; recall that the random coefficients γ_i each has five dimensions.)

The goal is to sample in proportion to the posterior likelihood function Φ. Because Φ is complex, we do this by MCMC. As a detour, we will give an overview of MCMC and the specific techniques we use. Readers familiar with these techniques can jump to Section A.3.

A.2 Overview of the MCMC procedure

The idea behind Markov Chain Monte Carlo (MCMC) is to sample from a distribution by constructing a Markov chain whose unique stationary distribution is the desired distribution of interest. If the chain is easy to simulate and fast-mixing, meaning that it converges quickly to the stationary distribution, then we can sample efficiently by simulating the chain, after throwing out a so-called "burn-in" period at the beginning before the chain has converged.

The workhorses of MCMC are Gibbs sampling and Metropolis-Hasting. Gibbs sampling is used when the desired distribution can be factored into several marginal distributions that are easier to sample. For example, to sample from a joint distribution on x, y, and z, one might iteratively sample one variable at a time conditional on the other ones. We initialize x^0, y^0 and z^0 arbitrarily.

For each $t \geq 1$, sample iteratively from the following conditional distributions:

$$\begin{aligned} x^t &\mid y^{t-1}, z^{t-1} \\ y^t &\mid x^t, z^{t-1} \\ z^t &\mid x^t, y^t \end{aligned} \tag{23}$$

After a sufficient number S of samples, and after throwing out the initial burn-in of B samples, $\{(x^t, y^t, z^t) : B < t \leq S\}$ would approximate samples from the desired posterior distribution, although successive samples are correlated. One can remove the serial correlation by either sampling independently from this set or keeping only samples in which t is a multiple of Δ, where Δ is a sufficiently large positive integer.

Metropolis-Hasting is a technique used to sample from an arbitrary distribution with given likelihood function $L(x)$. There are many variants, but the common idea is to use a proposal distribution that is easy to sample from and reject certain samples to get the likelihood ratios to be correct. The proposal distribution may depend on the current iterate x. Let transition probability density be $T(y|x)$; this is the probability density of proposing y given that the current sample is x. In order to obtain the correct likelihoods, we can only accept a fraction of the samples proposed and must reject the others. The probability that we accept proposal y given the previous iterate being x is

$$A(y|x) = \min(1, \frac{L(y)T(x|y)}{L(x)T(y|x)}). \tag{24}$$

Note that if $T(y|x)$ is proportional to $L(y)$, then the acceptance probability is always 1 as the proposal distribution already matches the target. Otherwise, the above formula is tuned so that the following identity, called "detailed balance" in the literature, holds:

$$L(x)T(y|x)A(y|x) = L(y)T(x|y)A(x|y). \tag{25}$$

This equation guarantees that the desired likelihood function $L(x)$ is proportional to a stationary distribution of the Markov chain induced by the proposal and acceptance process. Furthermore, if the chain is ergodic, which is true for example if the proposal distribution has full support, then the stationary distribution is unique.

The sampling procedure is then to initialize x^0 arbitrarily, and for each $t \geq 1$

1. Draw y according to $T(y|x^{t-1})$.
2. Set $x^t = \begin{cases} y & \text{with prob. } A(y|x^{t-1}), \\ x^{t-1} & \text{otherwise.} \end{cases}$

By iterating this many times and discarding sufficient burn-in samples, we arrive at the desired distribution.

Because of the flexibility in the proposal distributions, the above techniques have many variants. The goal is to find a proposal distribution that balances ease of sampling with locally approximating the target distribution. Without easy sampling, each step would take too long; if it is too far from the target distribution, then the acceptance probabilities would be very low and the chain may get stuck at a certain iterate for a very long time. In the following sections we present the three variants we use: Random Walk Metropolis (RWM), Metropolis-Within-Gibbs (MWG), and Hamiltonian Monte Carlo (HMC).

A.2.1 Random Walk Metropolis (RWM)

This method is the easiest to sample from, as it uses a simple random walk to propose the next value: if the current iterate is x, it proposes $y = x + \epsilon$, where ϵ is multivariate normal distributed, $\epsilon \sim \texttt{Normal}(0, \rho I)$, where I is the identity matrix, and ρ is a scale parameter. Other covariance matrices can also be used instead of the identity but it must be the same for every x. The scale parameter is tuned to match the overall variance of the desired distribution. Too small a ρ will produce too much serial correlation; too large a ρ and acceptance probability might be near zero so the chain may get stuck. We tune ρ by multiplying it up or down so that the average acceptance ratio since last tuning is between 0.4 and 0.6, which is the ball park value suggested by the literature.[46] The number of steps we wait before tuning increases exponentially, so that between the conclusion of our burn-in sample until our last iteration there is no tuning.

This method performs well when the target distribution does not have too many dimensions and the scale along various dimensions is approximately the same. However, with more dimensions, it becomes exponentially harder to guess the right direction, and the method may take very long to

[46] See Roberts, Gelman, and Gilks (1997).

converge; when there are dimensions that are at very different scales, then there may be no ρ that is good for all dimensions.

A.2.2 Metropolis Within Gibbs (MWG)

Metropolis Within Gibbs is a simple extension of RWM that allows sub-blocks of coordinates to have different scales. It is simple to sample each sub-block iteratively, conditional on the others, much like running several RWM f Gibbs sampling framework. This method also reduces the number of dimensions sampled at each step. The drawback is that more samples are needed.

Precisely speaking, instead of sampling all dimensions of vector x simultaneously, write it in terms of sub-vectors $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}$. Each sub-vector may represent several coordinates. Initialize x^0 arbitrarily and for $t \geq 1$, sample

$$\begin{aligned} x_1^t &\quad | \quad x_2^{t-1}, \cdots x_k^{t-1} \\ x_2^t &\quad | \quad x_1^t, x_3^{t-1}, \cdots x_k^{t-1} \\ &\cdots \\ x_k^t &\quad | \quad x_1^t, \cdots x_{k-1}^t \end{aligned} \tag{26}$$

Each of the above is sampled using RWM, perhaps with different scale parameters for different sub-vectors. In each Gibbs iteration, for each of the variables, we only take one step of Metropolis-Hasting, which involves one proposal and possible acceptance. Because of detailed balance, embedding Metropolis-Hasting into Gibbs sampling in this way also works.

A.2.3 Hamiltonian Monte Carlo (HMC)

This method uses the gradient of the log likelihood function to inform the proposals, which can significantly improve the acceptance probabilities in high dimensions. The drawback is that each iteration is slower as several gradient calls are needed. The method, motivated by Hamiltonian dynamics in physics, models the current iterate x as a location vector and treats the negative log likelihood function as an energy potential. In each step, the method samples a random momentum vector and simulates the trajectory of the object by discretizing time and alternatively updating the momentum using the potential function and updating the position using the momentum. To make detailed balance work out, the first and last steps of simulation are half-steps. Precisely speaking, let the gradient of the log likelihood function be $G(x) = \nabla(\log(L(x)))$. Let ϵ and Δ be tuning parameters representing the discretization in time and the number of steps to simulate respectively. The proposal is based on the following pseudocode taken from a description of HMC by Neal (2011):

Algorithm 1 Pseudocode for one step of HMC

```
Function HMC_STEP(x):
Draw momentum p_0 ~ Normal(0, I).
Initialize y = x, p = p_0.
Update p = p + εG(y)/2.
for Δ − 1 iterations do
    Update y = y + εp
    Update p = p + εG(y).
end for
Update y = y + εp.
Update p = p + εG(y)/2.
return { y   with prob. A(y|x) = min(1, L(y)/L(x) exp((||p_0||^2 − ||p||^2)/2))
       { x   otherwise
```

Note that the chance of proposing y given x is simply the chance of drawing momentum p_0. Moreover, by the reversibility of the intermediate steps of discrete simulation, if we started at y and drew a momentum of $-p$ (where p is the final momentum vector in HMC_STEP), then the proposal would be x. This implies that

$$\frac{T(y|x)}{T(x|y)} = \frac{\exp(-\frac{1}{2}\|p_0\|^2)}{\exp(-\frac{1}{2}\|-p\|^2)}, \tag{27}$$

which implies that

$$\frac{T(y|x)A(y|x)}{T(x|y)A(x|y)} = \frac{\exp(-\frac{1}{2}\|p_0\|^2)}{\exp(-\frac{1}{2}\|-p\|^2)} \frac{L(y)}{L(x)} \exp(\frac{\|p_0\|^2 - \|p\|^2}{2}) = \frac{L(y)}{L(x)}. \tag{28}$$

So detailed balance holds, and the following is a valid Metropolis-Hasting sampler: Initialize x^0 arbitrarily. For $t \geq 1$, set $x^t = \text{HMC\_STEP}(x^{t-1})$.

One can show that as the time discretization $\epsilon \to 0$, for any fixed total simulation time $\epsilon\Delta$, the acceptance probability goes to 1. Hence, we would like ϵ to be small enough so the chain does not get stuck and $\epsilon\Delta$ large enough so that successive samples are not too serially correlated. In practice, we fix $\Delta = 20$ and tune ρ so that the empirical acceptance rate since last tuning is between 0.5 and 0.8. As before, we increase the interval between tuning times exponentially so that no tuning happens in the sample we keep (after burn-in and before the last iteration). Another detail is that to prevent cases in which $\epsilon\Delta$ is exactly what makes the proposal y go back to original point x, instead of using a constant ϵ, we draw $\tilde{\epsilon} \sim \texttt{Uniform}(0.85\epsilon, 1.15\epsilon)$ before each call to HMC_STEP, and use $\tilde{\epsilon}$ as the step size throughout that call. Because this distribution is a-priori fixed, we preserve detailed balance. The above techniques are described by Neal (2011) as best practices when implementing HMC.

A.3 The MCMC Sampler

Our MCMC procedure is based on the one in Train (2003) but breaks the fixed coefficient estimation into two steps, one step using Hamiltonian Monte Carlo (HMC) and the other Metropolis Within Gibbs (MWG). We use HMC to estimate the school fixed effects and MWG to estimate the other fixed coefficients. These techniques allow us to accommodate the large number of school fixed effects and the unequal scales across the other fixed coefficients.

To sample from the full likelihood function $\Phi(\delta, \beta_f, \beta_r, \Sigma)$ (Equation 22), we initialize δ^0, β_f^0, β_r^0, Σ_1^0, Σ_2^0, Σ_3^0 arbitrarily. For each $t \geq 1$, we do a few layers of Gibbs sampling. In some of the layers we embed a form of Metropolis-Hasting, but in each Gibbs iteration we only take one step of Metropolis-Hasting, much as it is in MWG. Furthermore, let T be a parameter indicating how long we wait before tuning. We initialize T to be 1 and increase this parameter multiplicatively, so that tuning becomes exponentially less frequent. For $t \geq 1$, each MCMC step is as follows:

1. Draw $\gamma_i^t | \delta^{t-1}, \beta_f^{t-1}, \beta_r^{t-1}, \Sigma^{t-1}$. This is done using one iteration of RWM with likelihood function
$$L(x) = \phi_i(\delta^{t-1}, \beta_f^{t-1}, x) \exp(-\frac{1}{2}(\Sigma^{t-1})^{-1} \|x - \beta_r^{t-1}\|^2) \tag{29}$$
and starting value γ_i^{t-1}. (See Equation 21 for definition of ϕ_i.) We initialize $\rho = 0.05$ and initially to tune for each i every $\texttt{Uniform}(1000T, 1500T)$ steps.

2. Draw $\beta_r^t | \gamma_i^t, \Sigma^{t-1}$. This is sampling from $\texttt{Normal}(\frac{1}{n}\sum_{i=1}^n \gamma_i^t, \frac{1}{m}\Sigma^{t-1})$.

3. Draw $\Sigma^t | \gamma_i^t, \beta_r^t$. This can be done as follows: for $l \in \{1, 2, 3\}$, let $\mathbf{C}_l^t$ be the covariance matrix

of the lth block of γ_i^t assuming mean as in the lth block of β_r^t. (Recall that the random coefficients are organized into three blocks, with "ell match" being the first block, "walk zone" being the second, and "distance," "mcas," and "% white/asian" being the third.) Let k_l be the number of variables in the lth block and let n be the number of students. Draw Σ_l^t according to the Inverse Wishart Distribution with degree of freedom $\nu = k_l + n$ and scale matrix $\Psi = k_l I_{l\times l} + n\mathbf{C}_l^t$.

4. Draw $\delta^t|\gamma_i^t, \beta_f^{t-1}$. This is done using one step of HMC with likelihood function

$$L(x) = \prod_{i=1}^{n} \phi_i(x, \beta_f^{t-1}|\gamma_i^t), \tag{30}$$

and constraining the last component to be zero. We initialize $\epsilon = 0.015$, and $\Delta = 20$. We tune every $1000T$ steps.

5. Draw $\beta_f^t|\gamma_i^t, \delta^t$. This is done using one iteration of MWG with likelihood function

$$L(x) = \prod_{i=1}^{n} \phi_i(\delta^t, x|\gamma_i^t). \tag{31}$$

We break the fixed coefficients β_f into 6 subvectors: 1) "continuing;" 2) "sibling;" 3) "ell language match;" 4) "distance*black/hispanic" and "distance*income est."; 5) "mcas*black" and "mcas*income est."; 6) "% white/asian*black/hispanic" and "% white/asian*income est." We initialize the scales ρ for each subvector to be .5, .5, .1, .1, .5, and .5 respectively. We tune every $\texttt{Uniform}(100T, 150T)$ steps.

We run these steps 1,000,000 times, increasing the tuning interval parameter T by a factor of 1.2 every 5000 iterations. We throw out the first 500,000 iterations as burn-in. Note that no tuning happens in the interval we keep. This ensures the correctness of the Markov chain in this period.

For a robustness check, we re-ran this procedure six times, each time with different initial values, and we found nearly identical results each time.

B Computing Equilibrium Forecasts

All post-reform equilibrium forecasts are computed by averaging the results of 1000 iterations of the following sequence of steps.

1. Sample applicant pool X according to the assumptions described in Section 3.3.4. More details are given in Section 4.2 of the Part I report, Pathak and Shi (2015).

2. Sample choice model parameters:

   - For the Lexicographic model, we skip this step since the model does not have parameters.
   - For the MNL model, we sample

   $$(\delta, \beta) \sim N(\mu, \Sigma), \tag{32}$$

     where μ is the maximum likelihood estimate of the fixed effect δ and coefficients β, and Σ is the inverse of the Hessian of the log-liklihood function evaluated at μ.
   - For the MMNL model, we sample (δ, β, Σ) from the posterior distribution from MCMC and independently sample for each student i the individual coefficients $\gamma_i \sim N(\beta_r, \Sigma)$.

3. For each student, compute a relative ranking of all options for which he is eligible within his choice set, truncating to the top 10 choices. (This corresponds to the $Y|X$, using the notation introduced in Section 1.) For the MNL and MMNL models, this involves independently sampling idiosyncratic taste shocks $\epsilon_{ij} \sim$ Gumbel$(0, 1)$ for every student i and eligible option j. We also sample a lottery number l_i for each student i, $l_i \sim Uniform(0, 1)$.

4. Compute the assignment using the deferred acceptance algorithm described in Section 2 using the following inputs:

   - The simulated choice rankings from the previous step.
   - The program capacities imputed from the round one assignment from the previous year. (The "previous year" is 2013 for the calculation of post-reform forecasts, and 2012 for the back-test.)

- The following priority structure: define the priority π_{ij} of student i for program j to be (the higher priority the better)

$$\pi_{ij} = Boost_{ij} + l_i, \tag{33}$$

$$Boost_{ij} = 8Continuing_{ij} + 4PresentSchool_{ij} + 2Sibling_{ij} + SameSide_{ij}, \tag{34}$$

  where the variables on the right hand side of (34) are binary indicator variables for whether the student is a continuing student for program j, a continuing student for another program in the same school as program j, has a sibling in the school of program j, or is on the same side of the East Boston bridge as the school housing program j. For the back-test, the $Boost_{ij}$ follows Equation 37, which is explained later in this section.

5. Compute the equilibrium outcome of interest for each of the fourteen neighborhoods:

- Access to quality: Let the set of students assigned to school j be denoted I_j, and define

$$z_j = \begin{cases} \min_{i \in I_j} \pi_{ij} & \text{if school } j \text{ is full,} \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

  This is an estimate of the minimum priority needed to get into school j, given the generated preferences and priorities of other students. The estimate is based on the large market approximation of Azevedo and Leshno (2016). Define the access of student i to school j to be the probability that his lottery number is high enough for his priority to be higher than the cutoff of z_j,

$$Access_{ij} = \max(\min(Boost_{ij} + 1 - z_j, 1), 0), \tag{36}$$

  and the student's access to quality as the maximum $Access_{ij}$ over all program j in his menu from a Tier 1 or 2 school. The final result is the average of the access to quality estimates for every student i living within the neighborhood.

- Distance: compute the average walking distance for an assigned student from the neighborhood to his assigned school. The walking distance is from Google Maps API, based on the student's home address and the school's address. For students for whom we do not have a home address, we use the centroid of the geocode where the student lives as a proxy.

- Unassigned: compute the number of students from the neighborhood who are not assigned.

In each of the 1000 iterations, we compute for each neighborhood a scalar estimate for each of the three equilibrium outcomes of interest. The final forecast is the average of these 1000 values. The estimated 95% prediction intervals are from the empirical 2.5 and 97.5 percentiles of these 1000 values.

In computing the actual outcome, only Steps 4 and 5 are needed. Instead of the simulated values from steps 1-3, we use the actual applicant pool X^*, the actual choices Y^*, and the actual lottery number l_i for each student i. As a result, only one iteration is needed.

The pre-reform forecasts (from the back-testing exercise) are computed similarly, except that Step 4 above is altered to account for the different priority structure. Instead of the same-side priorities above, the pre-reform assignment plan uses walk-zone priorities, which only apply to 50% of the seats. The exact implementation is as follows. Each program j is split into two bins of equal size, j_1 and j_2. Bin j_1 is called the walk-zone bin and j_2 is the open bin. If program capacity is odd, then the walk-zone bin has one additional seat. Student preferences are augmented to be over the bins, so that for the same program, every student prefers the walk-zone bin over the open bin, but the relative preference between programs is as before. Priorities are now computed for every student i and every bin. For a walk-zone bin j_1 of program j, the priority boost is

$$Boost_{ij_1} = 8Continuing_{ij} + 4PresentSchool_{ij} + 2Sibling_{ij} + WalkZone_{ij}, \tag{37}$$

where $WalkZone_{ij}$ is a binary indicator variable for whether student i lives in the walk-zone of the school housing program j. For an open bin j_2, the boost is as above except without the $WalkZone_{ij}$ term. Given these preferences over bins and student priorities, we compute student assignments using the deferred acceptance algorithm. For access to quality, we define each student's access to each bin using the analog of equation (36) for bins and define a student's access to quality by finding the maximum access to an eligible quality bin, which is defined to be a bin of a program from a

Tier 1 or 2 school in the student's menu.

C Evaluating Choice Forecasts

Using the notation of Section 4.2, the quantities that need to be computed to evaluate choice forecasts for a given choice model are as follows.

1. Best prediction $\hat{Y}_{ik}$ for the set of top k choices of student i, where $k \in \{1, 2, 3\}$.

2. For each $k \in \{1, 2, 3\}$, market share s_{hj} of top k choices from this neighborhood that is for a program in school j.

3. For each tuple of schools (s_j, s_l), the proportion p_{jl} of students who ranked at least two choices and ranked school j first and l second.

4. For each tuple of programs (j, l), best prediction $\hat{z}_i(j, l)$ for whether student i prefers program j over program l.

Items 1-3 can be computed using many samples of the permutation of top 3 choices, (y_{i1}, y_{i2}, y_{i3}), for each student i. For $\hat{Y}_{ik}$ this is because, due to how the percentage of mistakes in Top k choices is defined, we have by linearity of expectations that the optimal deterministic prediction $\hat{Y}_{ik}$, if we believe the choice model to be correct, is simply the top k most commonly occurring options in the set $\{y_{i1}, \cdots, y_{ik}\}$. For s_{hl} and p_{jl}, having many samples of the permutation of top 3 choices suffices since the empirical market shares and empirical proportions are unbiased estimates of the true values. The details for each choice model are as follows.

- For the Lexicographic model, one sample of (y_{i1}, y_{i2}, y_{i3}) for each student i suffices since the model is deterministic.

- For the MNL model, we sample 5000 independent draws of model parameters $(\delta, \beta) \sim N(\mu, \Sigma)$, where μ is the maximum likelihood estimate and Σ is the inverse of the Hessian of the log-likelihood function at μ. For each draw of (δ, β), and for each student i and program j, we produce 200 independent draws of $\epsilon_{ij} \sim \text{Gumbel}(0, 1)$ and use these to simulate rankings.

Hence, for each student, we have 1,000,000 samples of (y_{i1}, y_{i2}, y_{i3}) that are almost independent of one another.[47]

- For the MMNL model, we use the same recipe as above: we produce 5000 independent samples of the model parameters (δ, β, Σ) from the MCMC posterior, and for each of these samples and each student i, we produce an independent draw of individual coefficients $\gamma_i \sim N(\beta_r, \Sigma)$. For each of the 5000 combinations of (δ, β, γ), we produce 200 draws of ϵ_{ij} for each student i and program j as before and compute 1,000,000 almost independent samples of (y_{i1}, y_{i2}, y_{i3}).

Item 4 can be computed easily for the Lexicographic model. For the MNL-based methods, the desired quantity $\hat{z}_i(j,l)$ has the following form:

$$\hat{z}_i(j,l) = \begin{cases} 1 & \text{if } \mathbb{P}(u_{ij} \geq u_{il}) \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

Define $\bar{u}_{ij} = u_{ij} - \epsilon_{ij}$. This is student i's utility for program j without counting his idiosyncratic taste shock ϵ_{ij}. Define $\bar{u}_{il}$ similarly. Observe that for the MNL model,

$$\mathbb{P}(u_{ij} \geq u_{il}) = \mathbb{E}\left[\frac{\exp(\bar{u}_{ij})}{\exp(\bar{u}_{ij}) + \exp(\bar{u}_{il})}\right], \tag{39}$$

where the expectation is taken over the randomness in the parameters β and δ, which affect $\bar{u}_{ij}$ and $\bar{u}_{il}$. Hence, we can estimate the above quantity using the 5000 independent samples of model parameters β and δ from the previous calculations for items 1-3. For the MMNL model, the same technique can be applied except that the expectation in Equation (39) is over β, δ, and γ_i, and we use the 5000 independent samples of (δ, β, γ) from before.

Another benchmark we use in evaluating choice forecasts is random guessing, in which case the choice ranking y_i is assumed to be a uniformly random permutation of options within student i's menu. For the metrics on individual choice, we do not need to explicitly sample but can instead explicitly write formulae for computing the relevant quantities. Let $|S_i|$ be the number of options

[47] They are not completely independent because 5000 draws of (δ, β) are shared across students and across each 200 draws of ϵ_{ij}. The completely independent alternative would be to produce one million independent draws of (δ, β) for each student, which is computationally expensive and we doubt would change the results.

in student i's menu.

$$\% \text{ Mistakes in Top } k \text{ Choices} = 1 - k/|S_i|, \tag{40}$$

$$\% \text{ Mistakes in Pairwise Comparisons} = 0.5. \tag{41}$$

For the metrics on distribution of choices, we can compute the top k market shares simply by distributing each student's market share uniformly among his available options and averaging over students of each neighborhood. For the joint distribution of top two choices, we assume that every ordered pair of distinct options is equally likely and apply the linearity of expectations and average across students.

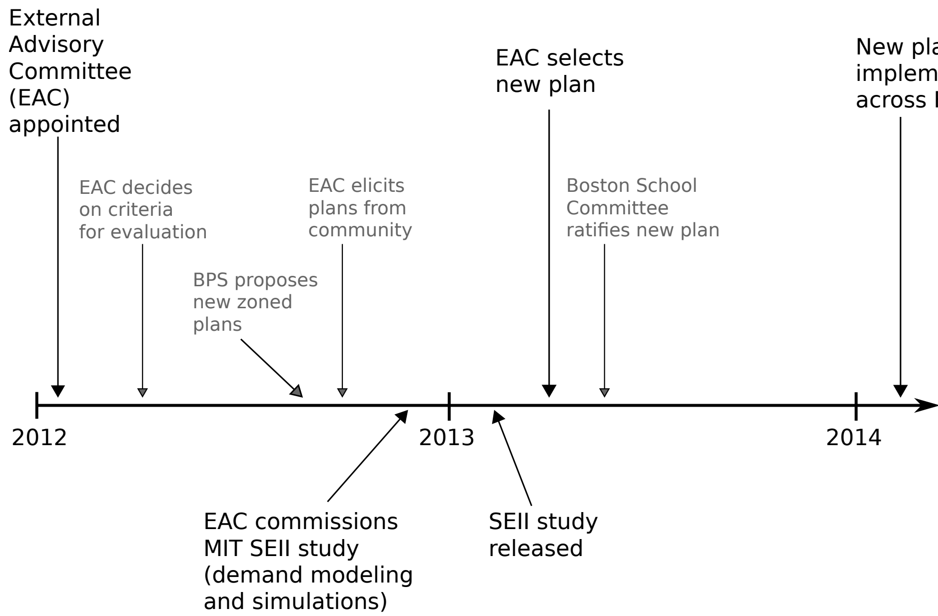


Figure 1: Timeline of Policy Reform

On the top of the timeline, we mark milestones in the Boston school assignment reform that culminated in the implementation of the Home Based plan in 2014. On the bottom of the timeline, we mark the timelines of the Pathak and Shi (2013) study from the MIT School Effectiveness and Inequality Initiative (SEII) that influenced the decision. The SEII study was based on discrete choice modeling.

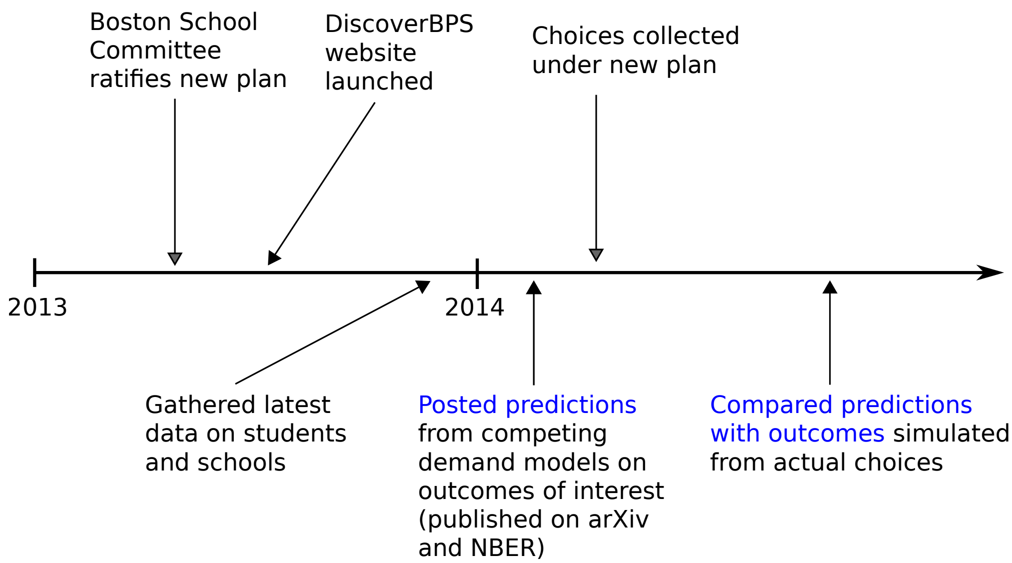


Figure 2: Timeline of this Research Project

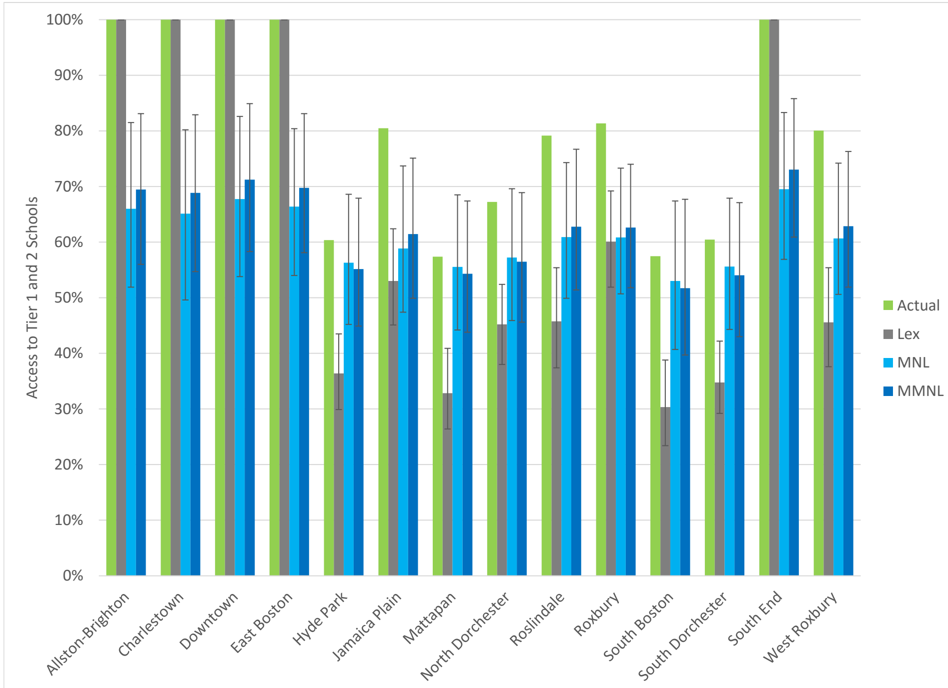
On the top of the timeline, we mark milestones in the implementation of the new assignment plan. On the bottom of the timeline, we mark the main steps of this project.

(A) Before reform (in 2013)

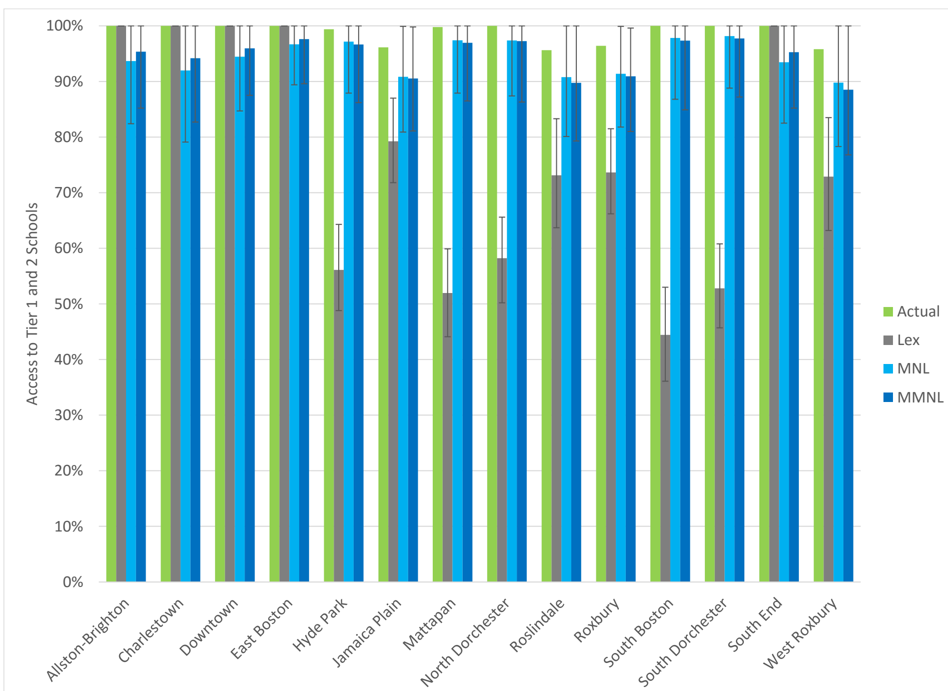
(B) After reform (in 2014)

FIGURE 3: Illustration of the Change in Choice Sets

Panel (a) shows the geographic zones under the Three Zone plan in 2013. The choice sets include all schools in a student's zone, as well as any additional schools within a one-mile walk-zone and a few city-wide schools. Panel (b) shows a portion of the web portal which generated the list of school options available under the Home Based plan in 2014 for each student. The list is generated based on the student's home address and various school characteristics, including the tier of the school and its distance to the student.



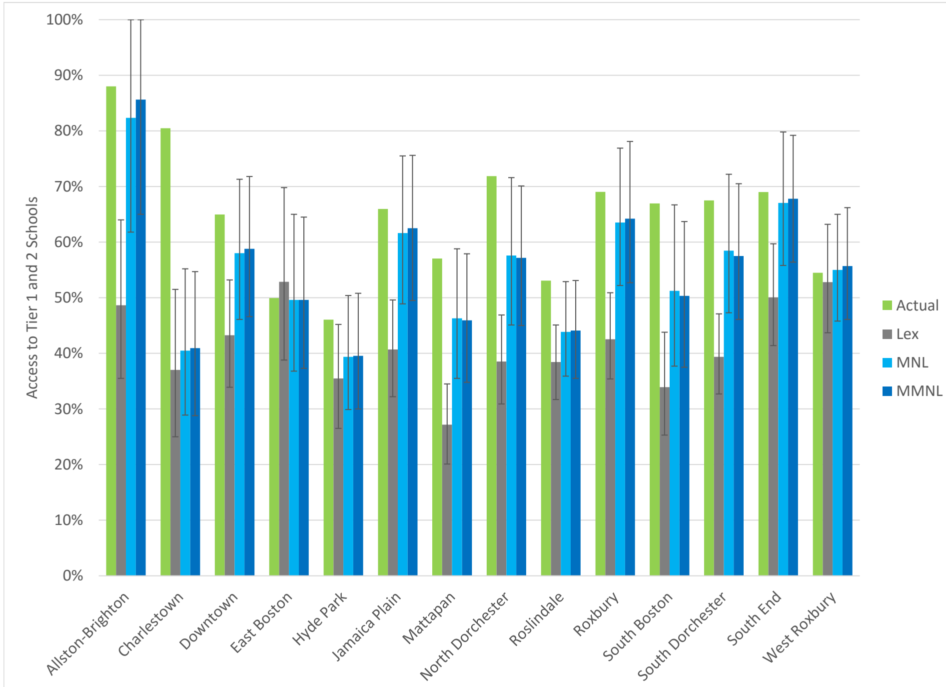
(A) Grade K1



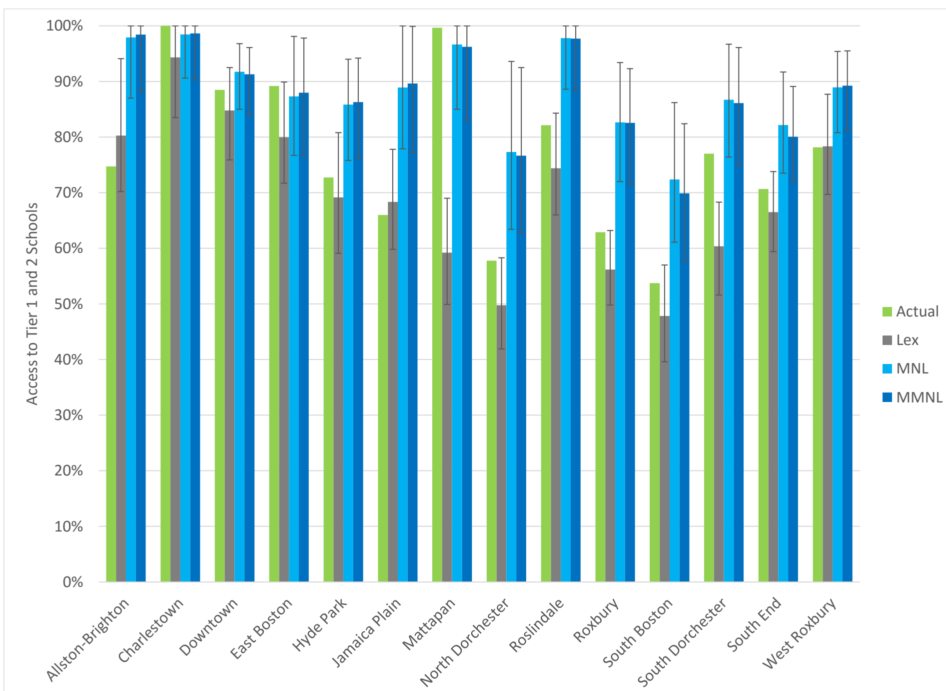
(B) Grade K2

FIGURE 4: Backtesting Access to Quality Predictions

This figure compares for each grade and neighborhood the actual access to quality one year before the reform (2013) and the predicted access to quality using each choice model based on data from two years before the reform (2012). Access to quality is defined as the average chance students from the neighborhood have of being assigned to a tier 1 or tier 2 school, supposing that the student ranks all such schools and ranks them first and supposing that all other students hold their preferences fixed. The three choice models are Lexicographic, Multinomial Logit (MNL) and Mixed MNL (MMNL). Whisker bars represent 95% prediction intervals.



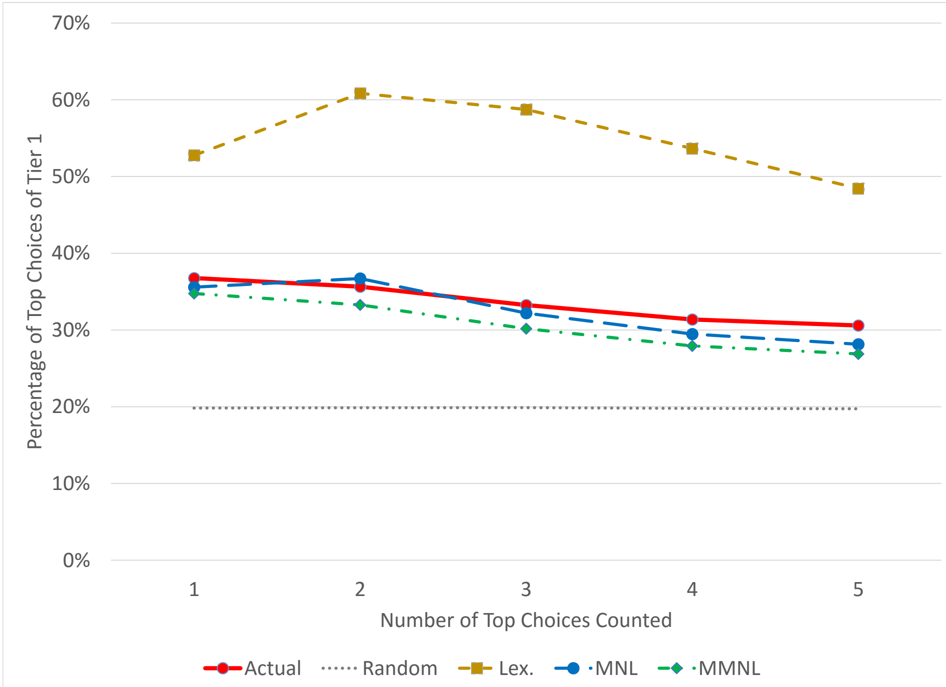
(A) Grade K1



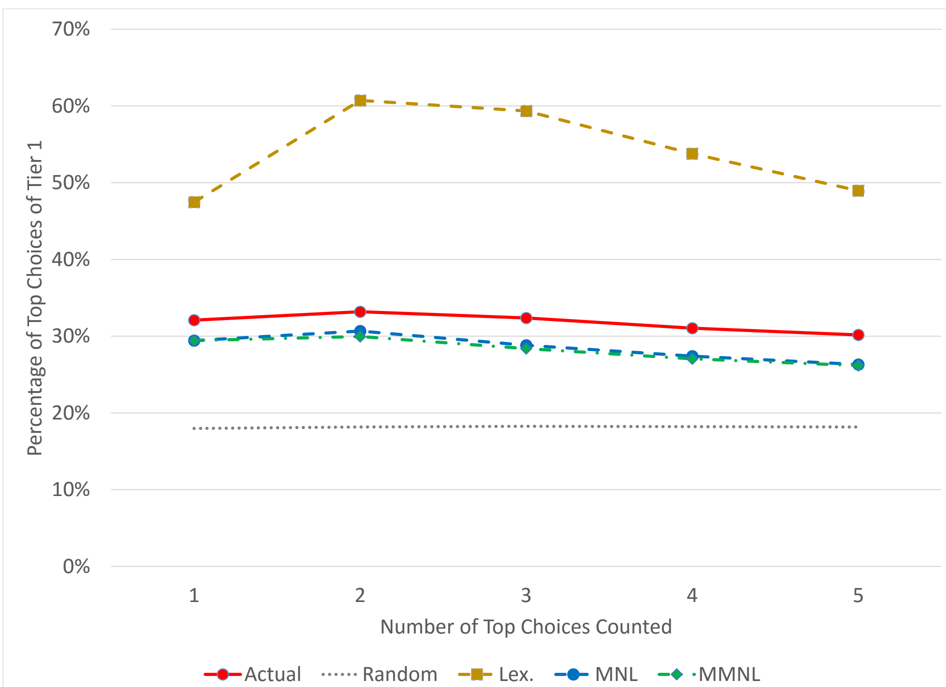
(B) Grade K2

FIGURE 5: Predicted vs. Actual Access to Quality

This figure compares for each grade and neighborhood the actual access to quality in the first year after the reform (2014) and the predicted access to quality using each choice model based on data from before the reform. Access to quality is defined as the average chance students from the neighborhood have of being assigned to a tier 1 or tier 2 school, supposing that the student ranks all such schools and ranks them first and supposing that all other students hold their preferences fixed. The three choice models are Lexicographic, Multinomial Logit (MNL) and Mixed MNL (MMNL). Whisker bars represent 95% prediction intervals.



(A) Grade K1



(B) Grade K2

Figure 6: Predicted vs. Actual Fraction of Top Choices Ranking Tier 1 Schools

This figure shows the actual percentage of top ranked choices that are for tier 1 schools in the first year after the reform (2014), and compares it with the predicted percentage using various choice models estimated from pre-reform data, but applied to the actual post-reform applicant characteristics. The actual percentages is shown in the solid line, while the choice models are dashed. The choice models compared are Lexcographic, Multinomial Logit (MNL), Mixed MNL (MMNL), and ranking programs uniformly randomly. In computing the percentage of top k choices (k ranging from 1 to 5), we average across all students who ranked at least k options.

Table 1. Comparison of Choice Sets

	Grade K1 (1)	Grade K2 (2)
	A: Applicants in New Assignment Plan	
Schools in New and Old Choice Sets	9.4	12.6
Schools Added to New Choice Sets	2.5	2.9
Schools Removed from Old Choice Sets	16.1	18.4
% of Top k Choices under New Choice Sets that were in Old Choice Sets		
Top 1	93%	95%
Top 3	91%	92%
Top 5	90%	91%
	B: Applicants in Old Assignment Plan	
% of Top k Choices under Old Choice Sets that are in New Choice Sets		
Top 1	84%	79%
Top 3	76%	73%
Top 5	68%	68%

Notes: This table compares choice sets between the old three-zone plan (in 2013) and the new home-based assignment plan (in 2014). Schools in New and Old Choice Sets is the average number of choices a student who applied in 2014 has that were also available in 2013 under the old plan. Schools Added to New Choice Sets is the average number of choices an applicant in the new plan has that were not available in the old choice set. Schools Removed from Old Choice Sets is the average number of choices an applicant in the new plan no longer has under the new plan, but would have had under the old plan. "% of Top k Choices under New Choice Sets that were in Old Choice Sets" reports whether highly ranked choices in the new plan were also available in the old plan. For each k (1, 3 or 5), we compute the percentage of top k choices of applicants under the new plan that were also possible options if those applicants were to apply a year earlier under the old plan. "% of Top k Choices under Old Choice Sets are in New Choice Sets" reports whether highly ranked choices under the old plan are still possible options in the new plan. For each k (1, 3 or 5), we compute the percentage of top k choices of applicants under the old plan that are still available options if those applicants were to apply a year later under the new plan.

Table 2. Backtesting Equilibrium Predictions Using Data from Two Years Prior to Predict One Year Prior

	Grade K1			Grade K2		
	RMSE	Exp. RMSE	% in 95% P.I.	RMSE	Exp. RMSE	% in 95% P.I.
	(1)	(2)	(3)	(4)	(5)	(6)
			A: Access to Quality			
Lexicographic	22%	(3%)	36%	31%	(3%)	36%
MNL	23%	(6%)	36%	5%	(4%)	100%
MMNL	20%	(6%)	36%	5%	(4%)	100%
			B: Distance (miles)			
Lexicographic	0.72	(0.25)	36%	0.46	(0.11)	29%
MNL	0.44	(0.18)	64%	0.26	(0.10)	64%
MMNL	0.40	(0.18)	64%	0.26	(0.10)	79%
			C: Unassigned			
Lexicographic	51	(17)	14%	26	(13)	57%
MNL	19	(17)	93%	19	(10)	71%
MMNL	18	(17)	93%	18	(11)	79%

Notes: This table reports backtesting results for equilibrium outcomes for the last year of the old assignment plan (in 2013) using choice models estimated from data two years prior (in 2012). In both years, we use the choice set from the old assignment plan (in 2013). For each of the 14 neighborhoods, access to quality is defined as the chance a random student from a neighborhood has of being assigned a tier 1 or tier 2 school, if the student were to rank all eligible programs in such schools first, holding fixed the choices of other students. Distance is the average Google-Maps walk distance between each assigned student from the neighborhood and his assigned school. Unassigned is the number of students unassigned from the neighborhood. Lexicographic, multinomial logit (MNL), and mixed MNL (MMNL) are the three choice models. For each grade, each choice model and each outcome of interest, we compute a 14-dimensional vector corresponding to our prediction for each of the 14 neighborhoods. The prediction is based on the average of 1000 independent simulations, accounting for uncertainty from sampling students, coefficients estimates, unobserved taste shocks, and lottery numbers. Columns 1 and 4 report the root mean squared error (RMSE), which is defined as the Euclidean distance between the 14-dimensional vector of predictions and the corresponding vector of actual outcomes. Columns 2 and 5 report the expected RMSE, which measures how large a RMSE we should anticipate from a random sample if the model were correct. Each independent simulation yields a 14-dimensional vector of predictions, which we call a prediction sample. The expected RMSE is estimated using the average Euclidean distance between each prediction sample and the sample mean. Columns 3 and 6 present another metric of how unexpected the RMSE is if the model were completely correct. % in 95% P.I. is the percentage of neighborhoods for which the actual outcome lies within the 95% prediction interval of the outcome, using the respective choice model. The prediction interval is estimated from the 2.5 and 97.5 percentiles of 1,000 simulations of each choice model.

Table 3. Backtesting Choice Predictions Using Data from Two Years Prior to Predict Choices from One Year Prior

	Grade K1				Grade K2			
	Random	Lexicographic	MNL	MMNL	Random	Lexicographic	MNL	MMNL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A: Individual Choices (% Mistakes)							
Top Choice	97%	63%	58%	58%	97%	37%	35%	34%
Top 2 Choices	94%	70%	59%	58%	94%	67%	57%	57%
Top 3 Choices	90%	69%	54%	54%	90%	67%	54%	54%
All Pairwise Comparisons	50%	30%	18%	18%	50%	16%	10%	10%
	B: Distribution of Choices (Statistical Distance)							
Market Shares by Neighborhood								
Top Choice	56%	46%	22%	21%	52%	26%	16%	15%
Top 2 Choices	52%	48%	19%	18%	52%	48%	20%	19%
Top 3 Choices	49%	49%	16%	16%	48%	51%	17%	16%
Joint Distribution of Top 2 Choices	64%	76%	45%	41%	67%	79%	51%	47%

Notes: This table reports backtesting results for choices for the last year of the old assignment plan (in 2013) using choice models estimated from data two years prior (in 2012). In both years, we use the choice set from the old assignment plan (in 2013). Panel A reports on individual choices of students, and Panel B reports on the distribution of student choices averaged across 14 Boston neighborhoods. Each column corresponds to a choice model: Random in columns 1 and 5 denotes uniformly random choices, Lexicographic in columns 2 and 6 denotes the lexicographic model, MNL in columns 3 and 7 denotes the multinomial logit model, and MMNL in columns 4 and 8 denotes the mixed MNL model. % Mistakes in Panel A uses each demand model's best guess of the student's choice and reports the fraction of incorrect guesses. Top Choice is for the first choice. Top 2 Choices is for the unordered set of first and second choice, and we report the percentage of elements in this set that are wrongly predicted and average over students who ranked at least two options. Top 3 Choices reports the analog for the unordered set of first, second and third choice, averaging over students who ranked at least three choices. Pairwise is the set of pairwise comparisons of options implied by the student's actual ranking compared to the best guess of each comparison from each choice model. The first three rows of Panel B report the statistical distance (a.k.a. total variation distance) between the predicted distribution of neighborhood-level market shares and the actual distribution, averaged across the neighborhoods. Joint distribution of top 2 choices aggregates students across neighborhoods and compares the predicted joint probability distribution of the first and second choice of students who ranked at least two choices and the actual distribution.

Table 4. Accuracy of Equilibrium Predictions Compared to Actual Outcomes

	Grade K1			Grade K2		
	RMSE	Exp. RMSE	% in 95% P.I.	RMSE	Exp. RMSE	% in 95% P.I.
	(1)	(2)	(3)	(4)	(5)	(6)
			A: Access to Quality			
Lexicographic	26%	(5%)	14%	13%	(4%)	86%
MNL	13%	(6%)	71%	15%	(5%)	36%
MMNL	13%	(6%)	79%	14%	(5%)	36%
			B: Distance			
Lexicographic	0.34	(0.14)	50%	0.14	(0.09)	71%
MNL	0.19	(0.12)	57%	0.13	(0.07)	71%
MMNL	0.19	(0.12)	57%	0.14	(0.07)	71%
			C: Unassigned			
Lexicographic	30	(16)	57%	34	(9)	43%
MNL	22	(16)	86%	41	(7)	14%
MMNL	21	(17)	86%	40	(8)	14%

Notes. This table reports the accuracy of predictions under three choice models for equilibrium outcomes using data from 2013 (the last year of the old assignment plan) compared to data from 2014 (the first year of the new assignment plan). For each grade, each outcome of interest, each choice model, and each of the 14 neighborhoods, we compute the prediction error as the squared difference between the predicted outcome for this neighborhood (based on the demand model) with the actual outcome (based on the actual choices). Table 2 notes contain definitions of the prediction targets and the columns.

Table 5. Accuracy of Choice Predictions Compared to Actual Choices

	Grade K1				Grade K2			
	Random	Lexicographic	MNL	MMNL	Random	Lexicographic	MNL	MMNL
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A: Individual Choices (% Mistakes)							
Top Choice	93%	59%	54%	53%	93%	33%	32%	33%
Top 2 Choices	85%	62%	51%	51%	87%	60%	54%	55%
Top 3 Choices	78%	58%	47%	47%	80%	56%	50%	51%
All Pairwise Comparisons	50%	28%	23%	23%	50%	14%	12%	13%
	B: Distribution of Choices (Statistical Distance)							
Market Shares by Neighborhood								
Top Choice	47%	41%	20%	21%	45%	22%	15%	15%
Top 2 Choices	41%	43%	16%	16%	43%	48%	19%	20%
Top 3 Choices	37%	41%	13%	14%	38%	44%	15%	17%
Joint Distribution of Top 2 Choices	62%	72%	41%	41%	67%	75%	49%	48%

Notes: This table reports on the accuracy of choice predictions using data from 2013 (the last year of the old assignment plan) compared to data from 2014 (the first year of the new assignment plan). Table 3 notes define the prediction targets and the columns.

Table 6. Accuracy of Equilibrium Predictions Using Actual Applicants with Estimated Choices

	Grade K1			Grade K2		
	RMSE	Exp. RMSE	% in 95% P.I.	RMSE	Exp. RMSE	% in 95% P.I.
	(1)	(2)	(3)	(4)	(5)	(6)
			A: Access to Quality			
Lexicographic	22%	(2%)	7%	22%	(2%)	0%
MNL	5%	(3%)	64%	12%	(3%)	79%
MMNL	6%	(3%)	64%	12%	(3%)	79%
			B: Distance			
Lexicographic	0.21	(0.07)	43%	0.15	(0.03)	14%
MNL	0.15	(0.08)	57%	0.08	(0.04)	57%
MMNL	0.15	(0.08)	50%	0.09	(0.05)	71%
			C: Unassigned			
Lexicographic	8	(4)	50%	15	(3)	21%
MNL	7	(4)	79%	14	(4)	36%
MMNL	7	(4)	64%	15	(4)	43%

Notes: This table reports the accuracy of predictions under three choice models for equilibrium outcomes using data from 2013 (the last year of the old assignment plan) compared to data from 2014 (the first year of the new assignment plan) with the actual set of applicants. Unlike Table 4, which forecasts applicant characteristics using the 2013 applicant pool and demographic projections, the calculation here uses the actual set of applicants in 2014 and their characteristics. Choices are generated from demand model estimates fit from old data. Table 2 notes contain definitions of the prediction targets and the columns.

Table 7. Prediction Improvements Using Post-Reform Data

	Grade K1		Grade K2	
	MNL	MMNL	MNL	MMNL
	(1)	(2)	(3)	(4)
	A: Access to Quality			
Original Prediction	13%	13%	15%	14%
New Applicants with				
Old Demand Model	5%	6%	12%	12%
Refit Demand Model	7%	7%	13%	13%
Refit Demand Model + Ranking Length	3%	3%	7%	7%
Sampling Actual Choices Using Applicant Forecast	8%		10%	
	B: Distance			
Original Prediction	0.19	0.19	0.13	0.14
New Applicants with				
Old Demand Model	0.15	0.15	0.08	0.09
Refit Demand Model	0.16	0.15	0.07	0.07
Refit Demand Model + Ranking Length	0.18	0.18	0.09	0.10
Sampling Actual Choices Using Applicant Forecast	0.08		0.07	
	C: Unassigned			
Original Prediction	22	21	41	40
New Applicants with				
Old Demand Model	7	7	14	15
Refit Demand Model	7	7	14	14
Refit Demand Model + Ranking Length	6	6	10	10
Sampling Actual Choices Using Applicant Forecast	22		15	

Notes: This table compares the accuracy of predictions from Table 4 using additional information from the new assignment plan. Each cell entry is the RMSE of the prediction error. Table 2 notes contain definitions of the prediction targets. Original Prediction is reproduced from columns 1 and 4 of Table 4. New Applicants with Old Demand Model uses new applicants in 2014, their characteristics, and predicted choices from the demand model fit in 2013, following columns 1 and 5 of Table 5. New Applicants with Refit Demand Model uses new applicants in 2014 and predicted choices from demand model refit in 2014. New Applicants with Refit Demand Model + Ranking Length uses new applicants in 2014, predicted choices from demand model refit in 2014, and the actual number of choices ranked by each new applicant in 2014. Sampling Actual Choices Using Applicant Forecast does not use demand-model predicted choices. It is computed by considering continuing and non-continuing students separately. Continuing students are already registered in BPS in a lower grade. Non-continuing students are new to the system. We predict the set of continuing students using the same methodology as in the original prediction and assume each applicant ranks their previous program first. For non-continuing students, we use the same methodology as in the original prediction and sample actual choices in 2014 with replacement.

Table 8. Accuracy of Choice Predictions from Refit Demand Models

	Demand Model	Grade K1		Grade K2	
	Fit Using Data	MNL	MMNL	MNL	MMNL
		(1)	(2)	(3)	(4)
		A: Individual Choices (% Mistakes)			
Top Choice	Old	54%	53%	32%	33%
	New	49%	49%	31%	30%
Top 2 Choices	Old	51%	51%	54%	55%
	New	50%	49%	51%	51%
Top 3 Choices	Old	47%	47%	50%	51%
	New	45%	45%	48%	48%
All Pairwise Comparisons	Old	23%	23%	12%	13%
	New	21%	21%	11%	11%
		B: Distribution of Choices (Statistical Distance)			
Market Shares by Neighborhood					
Top Choice	Old	20%	21%	15%	15%
	New	18%	17%	12%	11%
Top 2 Choices	Old	16%	16%	19%	20%
	New	14%	13%	15%	14%
Top 3 Choices	Old	13%	14%	15%	17%
	New	11%	10%	11%	11%
Joint Distribution of Top 2 Choices	Old	41%	41%	49%	48%
	New	39%	37%	45%	43%

Notes: This table compares the accuracy of choice predictions from choice models fitted using 2013 data (the last year of the old assignment plan) and choice models fitted using 2014 data (the first year of the new assignment plan). Accuracy is evaluated compared to the actual choices of students in 2014. Table format follows Table 7, except we include an additional row for each outcome specifying the source year for the data used to fit the demand model. We consider only the multinomial-logit (MNL) model (columns 1 and 3) and the mixed MNL (MMNL) model (columns 2 and 4). Table 3 notes contain definitions of the prediction targets.

Table 9. Reversals of Comparisons of Counterfactual Predictions for Access to Quality in Grade K1 for the Neighborhood Allston-Brighton under Various Simulation Assumptions

Choice Model (Year of Fitting)	MNL (2014)	MNL (2013)		Lexicographic
Applicant Pool	2014	2014	2013	2013
	(1)	(2)	(3)	(4)
A. Counterfactual Predictions				
Status Quo (3 Zone)	72.0%	77.7%	84.3%	100%
Home Based A	77.5%	82.9%	96.3%	55.0%
Home Based B	79.1%	84.5%	97.8%	57.0%
6 Zone	87.3%	91.3%	94.5%	54.3%
9 Zone	86.5%	90.7%	94.2%	54.4%
10 Zone	98.4%	99.8%	100.0%	64.9%
11 Zone	86.4%	90.5%	94.2%	54.2%
23 Zone	86.7%	91.3%	94.6%	57.6%
B. Reversal of Pairwise Comparisons				
# of Non-Trivial Comparisons	(Point of reference)	22	22	25
# of Reversals of Non-Trivial Comparisons		0	8	14
Percentage of Reversals		0%	36%	56%

Notes. In panel A, we report the point predictions for access to quality in grade K1 for the neighborhood Allston-Brighton under various proposed plans. Each row corresponds to a plan proposed during the 2012-2013 Boston student assignment reform, with each plan representing a different set of choice menus and priorities. The first row is the pre-reform status quo, and the second is the plan chosen after the reform. The third row is a variant of the plan in the second row, except with more choices. The remaining plans represent alternative partitioning of Boston into assignment zones. Each column specifies the choice model and the applicant pool used in the simulations. Column 1 uses the multinomial-logit (MNL) choice model fitted from post-reform choices using the post-reform applicant pool in 2014. Column 2 uses the MNL model fitted from pre-reform choices from 2013 but still simulated using the post-reform applicant pool. Column 3 is similar to Column 2 but uses the pre-reform applicant pool from 2013. Column 4 uses the lexicographic choice model and the pre-reform applicant pool.

In Panel B, we measure how much columns 2 through 4 in Panel A differ from panel 1 in terms of the relative rankings of access to quality across plans. Since column 1 is the point of reference, it is left blank in panel B. Consider first the comparison between columns 1 and 2 of Panel A, which are reported in column 2 of Panel B. Since there are 8 plans, there are 28 comparisons. Each comparison corresponds to a pair of rows from Panel A, and we call the comparison "trivial" if the access to quality predictions in these two rows are within an addititve difference of 1.0% of one another in both columns 1 and 2. For example, the comparison between the 11 and 23 Zone plans is trivial, but the comparison between the Status Quo and the Home Based A plan is not. Row 1 in column 2 of Panel B reports the number of non-trivial comparisons between columns 1 and 2 of panel A. Row 2 of Panel B reports the how many non-trivial comparisons are reversed, which means that the columns differ in which plan results in a higher access to quality. For example, for columns 1 and 3 of Panel A, the comparison between the Home Based A plan and the 23 zone plan is reversed, but between the Home Based A plan and the Status Quo is not. Row 3 of Panel C reports the ratio between the previous two rows expressed as a percentage. As a benchmark, if the columns agree exactly on the relative rankings across plans, then the percentage of reversals is 0.

Table 10. Percentage of Reversals of Comparisons of Counterfactual Predictions Under Alternative Simulation Assumptions

Choice Model (Year of Fitting)	MNL (2014)	MNL (2013)		Lexicographic
Applicant Pool	2014	2014	2013	2013
	(1)	(2)	(3)	(4)
Access to Quality				
K1		13%	17%	32%
K2		4%	23%	35%
Distance				
K1	(Point of	2%	7%	18%
K2	reference)	1%	11%	24%
Unassigned				
K1		5%	20%	44%
K2		5%	21%	38%
Overall		5%	16%	32%

Notes: This table reports the aggregate result of the analysis in Table 9 on the percentage of reversals of pairwise comparisons of counterfactual predictions, when averaged across the 14 neighborhoods and performed for each of the three equilibrium moments of interest. See the notes for Table 9 for a description of the columns as well as the eight proposed plans compared. See the notes for Table 3 for a description of the three forecast targets. The numbers in the first row correspond to an analysis that is analogous to that in the last row of Panel B of Table 9 for all 14 neighborhoods, instead of just Allston-Brighton, and reporting the average across neighborhoods. The second row is similar, except for grade K2. The next four rows are for different moments of interest, but the analysis is analogous. Recall from the notes of Table 9 that the analysis requires a definition of "non-trivial" difference between given pair of plans, and that the threshold for a non-trivial difference in access to quality is set to an additive difference of 1.0%. For distance, this threshold is set to 0.01 miles. For unassigned, this is set to 0.5 students/neighborhood. The last row reports the unweighted average of the first six rows, and corresponds to an aggregate measure of how much relative rankings of counterfactual predictions are different across simulation assumptions.

Table A1. MNL and MMNL Coefficient Estimates

	MNL			MMNL		
	2012	2013	2014	2012	2013	2014
	(1)	(2)	(3)	(4)	(5)	(6)
distance	-0.365***	-0.403***	-0.557***	-0.638***	-0.674***	-0.793***
	(0.014)	(0.015)	(0.028)	(0.037)	(0.039)	(0.045)
continuing	4.027***	4.354***	4.369***	4.777***	4.966***	5.201***
	(0.052)	(0.054)	(0.064)	(0.069)	(0.068)	(0.085)
sibling	2.104***	2.102***	2.619***	2.478***	2.451***	3.089***
	(0.037)	(0.038)	(0.048)	(0.045)	(0.045)	(0.060)
walk zone	0.500***	0.399***	0.133***	0.339***	0.185***	0.053***
	(0.019)	(0.020)	(0.023)	(0.028)	(0.028)	(0.029)
ell program x ell student	1.548***	1.211***	0.543***	1.892***	1.311***	0.614***
	(0.035)	(0.040)	(0.045)	(0.058)	(0.059)	(0.061)
ell program language match x ell	0.606***	0.672***	0.802***	0.610***	0.967***	0.989***
student	(0.043)	(0.049)	(0.062)	(0.052)	(0.060)	(0.078)
distance x black/hispanic	0.115***	0.114***	0.216***	0.188***	0.183***	0.268***
	(0.010)	(0.011)	(0.019)	(0.024)	(0.024)	(0.031)
distance x block group income	-0.262***	-0.296***	-0.274***	-0.295***	-0.343***	-0.337***
	(0.021)	(0.023)	(0.039)	(0.049)	(0.052)	(0.062)
mcas x black	-0.874***	-1.062***	-0.901***	-1.100***	-1.371***	-1.283***
	(0.105)	(0.111)	(0.089)	(0.153)	(0.144)	(0.130)
mcas x block group income	0.424*	-0.906***	1.762***	1.065***	0.925***	2.388***
	(0.221)	(0.252)	(0.216)	(0.299)	(0.313)	(0.278)
% white/asian x black/hispanic	-2.581***	-2.666***	-2.984***	-3.732***	-3.861***	-4.000***
	(0.097)	(0.094)	(0.114)	(0.162)	(0.148)	(0.170)
% white/asian x block group income	1.982***	1.778***	1.052***	2.633***	2.217***	1.355***
	(0.211)	(0.219)	(0.249)	(0.322)	(0.311)	(0.351)

Notes: This table reports the estimated coefficients of the multinomial logit (MNL) and mixed MNL (MMNL) choice models. The year in each column corresponds to the source year for choice data. All models include a fixed effect for each school. distance is the Google Maps walking distance from the school to the student's home. continuing is a binary indicator variable for whether the student is continuing at the school from a previous grade. sibling is an indicator for whether the student has an older sibling at the school. walk zone is an indicator for whether the student is in the school's walk zone. ell program is an indicator for whether the program is for English language learners (ELL). ell student is an indicator whether the student is classified by the district as an English learner and thus eligible to ELL programs. ell program language match is an indicator for whether the program is an ELL program that targets students who speak a certain language and this language matches the student's home language. black/hispanic and black are indicators for the student's racial classification. mcas is the proportion of students at the school who scored "Advanced" or "Proficient" in the previous year's standardized test for math, averaging the proportions for grades 3,4, and 5. block group income is the medium household income of the census block group containing the centroid of the student's geocode of residence measured in hundreds of thousands of dollars. % white/asian is the proportion of the enrolled population at the school who are White or Asian. Standard errors are in parenthesis. Standard errors for MNL are computed using the Hessian matrix of the maximum likelihood at the point estimate of the coefficients. Standard errors for MMNL are computed using the sample standard deviation of the MCMC samples.

*significant at 10%; **significant at 5%; ***significant at 1%.

Table A2. Covariance Estimates for MMNL Model

	2012	2013	2014
	(1)	(2)	(3)
		A: Standard Deviations	
σ(ell program x ell student)	1.638***	1.358***	0.959***
	(0.058)	(0.063)	(0.068)
σ(walk zone)	0.981***	0.878***	0.703***
	(0.030)	(0.030)	(0.035)
σ(distance)	0.392***	0.409***	0.499***
	(0.011)	(0.011)	(0.016)
σ(mcas)	2.275***	2.121***	1.837***
	(0.093)	(0.101)	(0.083)
σ(%white/asian)	2.672***	2.512***	2.300***
	(0.093)	(0.106)	(0.106)
		B: Correlation Coefficients	
ρ(distance,mcas)	-0.232***	-0.285***	-0.134***
	(0.041)	(0.043)	(0.049)
ρ(distance, %white/asian)	-0.089**	-0.055	0.021
	(0.039)	(0.040)	(0.051)
ρ(mcas, %white/asian)	0.035	-0.110*	0.236***
	(0.056)	(0.061)	(0.068)

Notes: This table reports covariance matrix estimates for the random coefficients in the mixed multinomial logit (MMNL) model. The year in each column corresponds to the source year for choice data. The variables ell program, ell student, walk zone, distance, mcas, and % white/asian are defined in Table A1 notes. Panel A reports the square root of the variance of each random coefficient. Panel B reports the Pearson correlation coefficient of the three pairs of random coefficients for which we allow correlation. Standard errors of the estimates are in parenthesis, computed using the sample standard deviation of the MCMC samples.

*significant at 10%; **significant at 5%; ***significant at 1%.

Table A3. Prediction Error in Applicant Count and Demographics

		Predicted (1)	Std. Error (2)	Actual (3)
		A. Count of Applicants		
Grade K1	Continuing	92	7	158
	New	2652	177	2313
Grade K2	Continuing	1482	30	2051
	New	2196	153	1875
		B. Applicant Demographics		
ELL (Grade K1)	Yes	44.4%	1.0%	46.7%
	No	55.6%	0.7%	53.3%
ELL (Grade K2)	Yes	30.8%	0.7%	14.6%
	No	69.2%	0.7%	85.4%
Race	Hispanic	35.1%	0.6%	36.5%
	Black	28.8%	0.5%	28.0%
	White	22.6%	0.5%	22.9%
	Asian	8.4%	0.3%	7.9%
	Other	5.1%	0.3%	4.7%
Median Block Group Income	0-25K	16.9%	0.4%	17.5%
	25-50K	49.7%	0.6%	50.0%
	50-75K	20.7%	0.5%	20.2%
	75K+	12.7%	0.4%	12.3%
Neighborhood	Allston-Brighton	4.5%	0.3%	4.8%
	Charlestown	3.5%	0.2%	3.2%
	Downtown	3.7%	0.3%	3.4%
	East Boston	12.7%	0.7%	12.3%
	Hyde Park	6.3%	0.2%	6.4%
	Jamaica Plain	6.7%	0.4%	7.2%
	Mattapan	6.8%	0.3%	6.8%
	North Dorchester	5.3%	0.5%	5.6%
	Roslindale	8.5%	0.4%	8.1%
	Roxbury	13.6%	0.4%	14.1%
	South Boston	3.2%	0.2%	3.0%
	South Dorchester	13.2%	0.5%	13.6%
	South End	4.7%	0.2%	4.4%
	West Roxbury	7.3%	0.4%	7.2%

Notes: This table compares the predicted and actual new applicants across demographic categories. Column 1 reports the prediction for each category of students, and column 2 reports the standard deviation of the prediction. These are computed from the 1,000 simulated samples of applicant pools used for computing the equilibrium outcomes. Column 3 reports the actual number of students of each type. Column 1 reports the predicted percentage and column 2 the standard deviation of the prediction. The predictions are based on 2013 data (the last year of the old assignment plan). The numbers shown are the sample mean and standard deviations of the percentage of applicants of each category in the 1,000 simulation samples used for Table 4. Column 3 reports the actual percentages in the 2014 data (the first year of the new assignment plan). Panel A compares the predicted number of applicants to the actual number. Continuing students are those enrolled in BPS in the previous grade at the time of application. The remaining students are new applicants. Panel B reports applicant characteristics. ELL denotes whether the student is classified by BPS as eligible for English Language Learner programs. Race information is missing for students who applied but did not enroll in any school. Income and neighborhood information are based on centroid of student geocode. Median block group income refers to the median household income of the census block group in which the student resides, based on the 2010 census.

References

Abdulkadiroğlu, A., N. Agarwal, and P. Pathak (2015): "The Welfare Effects of Coordinated School Assignment: Evidence from the NYC High School Match," NBER Working Paper, 21046.

Abdulkadiroğlu, A., P. A. Pathak, and A. E. Roth (2009): "Strategy-proofness versus Efficiency in Matching with Indifferences: Redesigning the New York City High School Match," *American Economic Review*, 99(5), 1954–1978.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, and T. Sönmez (2005): "The Boston Public School Match," *American Economic Review, Papers and Proceedings*, 95, 368–371.

——— (2006): "Changing the Boston Public School Mechanism," Discussion paper, NBER WP 11965.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. Walters (2017): "Do Parents Value School Effectiveness?," NBER Working Paper, 23912.

Abdulkadiroğlu, A., and T. Sönmez (2003): "School Choice: A Mechanism Design Approach," *American Economic Review*, 93, 729–747.

Agarwal, N., and P. Somaini (2014): "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism," NBER Working Paper 20775.

Angrist, J., and J.-S. Pischke (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24(2), 3–30.

Ashenfelter, O., and D. Hosken (2008): "The Effects of Mergers on Consumers Prices: Evidence from Five Selected Case Studies," NBER Working Paper 13589.

Azevedo, E. M., and J. D. Leshno (2016): "A Supply and Demand Framework for Two-sided Matching Markets," *Journal of Political Economy*, 124(5), 1235–1268.

Berry, S., J. Levinsohn, and A. Pakes (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112(1), 68–105.

Burge, K. (2012): "Study Finds Inequalities in Schools' Zone Plans," Boston Globe, October 1.

Burgess, S., E. Greaves, A. Vignoles, and D. Wilson (2015): "What Parents Want: School Preferences and School Choice," *Economic Journal*, 125(587), 1262–1289.

Calsamiglia, C., C. Fu, and M. Guell (2017): "Structural Estimation of a Model of School Choices: the Boston Mechanism vs. Its Alternatives," Working paper, CEMFI.

Drolet, and Luce (2004): "The Rationalizing Effects of Cognitive Load on Response to Emotional Tradeoff Difficulty," *Journal of Consumer Research*, 31(1), 63–77.

Dubins, L. E., and D. A. Freedman (1981): "Machiavelli and the Gale-Shapley algorithm," *American Mathematical Monthly*, 88, 485–494.

Dur, U., S. D. Kominers, P. A. Pathak, and T. Sönmez (2016): "Reserve Design: Unintended Consequences and the Demise of Walk Zones in Boston," forthcoming, *Journal of Political Economy.*

Einav, L., and J. Levin (2010): "Empirical Industrial Organization: A Progress Report," *Journal of Economic Perspectives*, 24(2), 145–162.

Fishburn, P. (1974): "Lexicographic Orders, Utilities and Decision Rules: A Survey," *Management Science*, 20(11), 1442–1471.

Gale, D., and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69, 9–15.

Glazerman, S., and D. Dotter (2016): "Market Signals: Evidence on the Determinants and Consequences of School Choice from a Citywide Lottery," Mathematica Policy Research, June.

Goldstein, D. (2012): "Bostonians Committed to School Diversity Haven't Given Up on Busing," *The Atlantic*, October 10.

Handy, D. (2012): "Debate on Overhauling Boston Schools' Assignment System Continues," 90.9 *WBUR*, November 13.

Harris, D., and M. Larsen (2015): "What Schools Do Families Want (and Why?)," Technical Report, New Orleans, LA: New Orleans Education Research Alliance.

Hastings, J., T. J. Kane, and D. O. Staiger (2009): "Heterogeneous Preferences and the Efficacy of Public School Choice," Working paper, Brown University.

Hastings, J., and J. M. Weinstein (2008): "Information, School Choice and Academic Achievement: Evidence from Two Experiments," *Quarterly Journal of Economics*, 123(4), 1373–1414.

Hausman, J. A., and P. A. Ruud (1987): "Specifying and Testing Econometric Models for Rank-ordered Data," *Journal of Econometrics*, 34(1), 83–104.

He, Y. (2012): "Gaming the Boston Mechanism in Beijing," Working paper, Rice University.

Heckman, J. (2010): "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 2, 356–398.

Hurwicz, L. (1950): "Prediction and Least Squares," in *Statistical Inference in Dynamic Economic Models*, ed. by T. C. Koopmans. John Wiley & Sons.

Hwang, S. (2015): "A Robust Redesign of High School Match," University of Britsh Columbia.

Johnson, E. J., R. J. Meyer, and S. Ghose (1989): "When choice models fail: Compensatory models in negatively correlated environments," *Journal of Marketing Research*, 26(Aug), 255–290.

Kapor, A., C. Neilson, and S. Zimmerman (2017): "Hetereogneous Beliefs and School Choice Assignment Mechanisms," Working Paper, Princeton University.

Katafygiotis, L., and K. Zuev (2008): "Geometric insight into the challenges of solving high-dimensional reliability problems," *Probabilistic Engineering Mechanics*, 23(2–3), 208 – 218, 5th International Conference on Computational Stochastic Mechanics.

Keane, M., and K. Wolpin (2007): "Exploring the Usefulness of a Nonrandom Holdout Sample for Model Validation: Welfare Effects on Female Behavior," *International Economic Review*, 48(4), 1351–1378.

KLING, J. R., S. MULLAINATHAN, E. SHAFIR, L. C. VERMUELEN, AND M. V. WROBEL (2012): "Comparison Friction: Experimental Evidence from Medicare Drug Plans," *Quarterly Journal of Economics*, 127, 199–235.

KOHLI, R., AND K. JEDIDI (2007): "Representation and Inference of Lexicographic Preference Models and their Variants," *Marketing Science*, 26(3), 380–399.

LEVINSON, M. (2015): "The Ethics of Pandering in Boston Public Schools' School Assignment Plan," *Theory and Research in Education*, 13(1), 38–55, For an open access version of this article, please click here.

LUMSDAINE, R., J. STOCK, AND D. WISE (1992): "Three Models of Retirement: Computational Complexity vs. Predictive Validity," in *Topics in the Economics of Aging*, ed. by D. Wise. University of Chicago Press, Chicago.

MANZINI, P., AND M. MARIOTTI (2012): "Choice by lexicographic semiorders," *Theoretical Economics*, 7, 1–23.

MARSCHAK, J. (1953): "Economic Measurements for Policy and Prediction," in *Studies in Econometric Methods*, eds., Hood and Koopmans, New York Wiley, p. 1-26.

MCFADDEN, D. (1974): "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 3, 303–328.

——— (2001): "Economic Choices," *American Economic Review*, 91(3), 351–378.

MCFADDEN, D., F. REID, A. TALVITIE, M. JOHNSON, AND ASSOCIATES (1979): "Overview and Summary: Urban Travel Demand Forecasting Project," *Urban Travel Demand Forecasting Project*, Final Report, Vol. I. Institute of Transportation Studies, University of California Berkeley.

MCFADDEN, D., A. TALVITIE, AND ASSOCIATES (1977): "Validation of Disaggregate Travel Demand Models: Some Tests," *Urban Travel Demand Forecasting Project*, Final Report, Vol. V. Institute of Transportation Studies, University of California Berkeley.

Menino, T. (2012a): "Press Release," http://www.cityofboston.gov/news/default.aspx?id=5873 November 29.

——— (2012b): "State of the City Address," January 17 Available at http://www.cityofboston.gov/.

Misra, S., and H. Nair (2011): "A Structural Model of Sales-Force Compensation Dynamics: Estimation and Field Implementation," *Quantitative Marketing and Economics*, 9(3), 211–225.

Neal, R. (2011): "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, ed. by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, chap. 5. CRC Press.

Nevo, A., and M. Whinston (2010): "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference," *Journal of Economic Perspectives*, 24(2), 69–82.

Pathak, P., and P. Shi (2013): "Simulating Alternative School Choice Options in Boston," Working Paper, MIT.

Pathak, P. A., and P. Shi (2014): "Demand Modeling, Forecasting, and Counterfactuals, Part I," NBER Working Paper 19589.

——— (2015): "Demand Modeling, Forecasting, and Counterfactuals, Part I," Available at `http://arxiv.org/abs/1401.7359`.

Pathak, P. A., and T. Sönmez (2008): "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism," *American Economic Review*, 98(4), 1636–1652.

——— (2013): "School Admissions Reform in Chicago and England: Comparing Mechanisms by their Vulnerability to Manipulation," *American Economic Review*, 103(1), 80–106.

Payne, J. W., J. R. Bettman, and E. J. Johnson (1988): "Adaptive strategy selection in decision making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 534–552.

Peters, C. (2006): "Evaluating the Performance of Merger Simulation: Evidence from the US Airline Industry," *Journal of Law and Economics*, 49(2), 627–649.

Reiss, P. C., and F. A. Wolak (2007): *Handbook of Econometrics, Volume 6A*chap. Structural Econometric Modeling: Rationales and Examples from Industrial Organization, pp. 4277–4415. Elsevier.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997): "Weak convergence and optimal scaling of random walk Metropolis algorithms," *The Annals of Applied Probability*, 7(1), 110–120.

Roth, A. E. (1982): "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research*, 7, 617–628.

Ruijs, N., and H. Oosterbeek (2012): "School choice in Amsterdam. Which schools do parents prefer when school choice is free?," Working paper, Amsterdam.

Seelye, K. Q. (2012): "4 Decades after Clashes, Boston Again Debates School Busing," *New York Times*, October 4.

——— (2013): "No Division Required in This School Problem," *New York Times*, March 12.

Shi, P. (2013): "Closest Types: A Simple Non-Zone-Based Framework for School Choice," Working paper, MIT.

——— (2015): "Guiding School-choice Reform through Novel Applications of Operations Research," *Interfaces*, 45(2), 117–132.

Slovic, P. (1975): "Choice Between Equally Valued Alternatives," *Journal of Experimental Psychology: Human Perception Performance*, 1, 280–287.

Sönmez, T. (2013): "Bidding for Army Career Specialties: Improving the ROTC Branching Mechanism," *Journal of Political Economy*, 121(1), 186–219.

Thorngate, W. (1980): "Efficient Decision Heuristics," *Behavioral Science*, 25(May), 219–225.

Todd, P., and K. Wolpin (2006): "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, 96(5), 1384–1417.

Train, K. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, UK.

TVERSKY, A. (1969): "Intransitivity of Preferences," *Psychological Review*, 76(1).

——— (1972): "Elimination by Aspects: A Theory of Choice," *Psychological Review*, 79(4).

TVERSKY, A., S. SATTAH, AND P. SLOVIC (1988): "Contingent Weighting in Judgment and Choice," *Psychological Review*, 95, 371–384.

VAZNIS, J., AND T. ANDERSEN (2012): "Plans Upend Boston School Assignments," *Boston Globe*, September 25.

WALTERS, C. R. (2014): "The Demand for Effective Charter Schools," NBER Working Paper 20640.

WISE, D. A. (1985): "Behavioral Model versus Experimentation: The Effects of Housing Subsidies on Rent," In *Methods of Operations Research 50*, ed. Peter Brucker and R. Pauly, 441-489, Koningsten: Verlag Anton Hain.

YEE, M., E. DAHAN, J. HAUSER, AND J. ORLIN (2007): "Greedoid-Based Noncompensatory Inference," *Marketing Science*, 26(4), 532–549.