Estimating Institutional Quality with Instruments: Three Essays and Applications in Education and Healthcare

by

Peter D. Hull

B.A., Wesleyan University (2010)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Peter D. Hull, MMXVII. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature redacted

Author .....
Department of Economics
May 15, 2017

Signature redacted

Certified by .....
Joshua D. Angrist
Ford Professor of Economics
Thesis Supervisor

Signature redacted

Certified by .....
Amy N. Finkelstein
John and Jennie S. MacDonald Professor of Economics
Thesis Supervisor

Signature redacted

Certified by ....
Parag A. Pathak
Jane B. Carlton and Dennis W. Carlton Professor of Microeconomics
Thesis Supervisor

Signature redacted

Accepted by .....
Ricardo J. Caballero
Ford International Professor of Economics
Chairman, Departmental Committee on Graduate Studies

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUN 14 2017
LIBRARIES
ARCHIVES

MIT Libraries

77 Massachusetts Avenue
Cambridge, MA 02139
http://libraries.mit.edu/ask

DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

The images contained in this document are of the best quality available.

Estimating Institutional Quality with Instruments: Three Essays and Applications in Education and Healthcare

by

Peter D. Hull

Submitted to the Department of Economics
on May 15, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis develops and applies instrumental variable (IV) techniques to estimate average causal effects in settings with multiple unordered treatments. When an instrument shifts individuals across several distinct margins, classical estimators may not recover causally-interpretable parameters, particularly when effects are heterogeneous and there are fewer instruments than treatment choices. The first chapter of this thesis establishes minimal effect heterogeneity restrictions that permit IV identification of multiple local average treatment effects (LATEs) in such scenarios, using interactions of the instrument and stratifying controls. Under weaker conditional homogeneity assumptions, a novel non-parametric weighting scheme identifies these LATEs. I use this framework to estimate the economic returns to GED certification in a sample that includes individuals who would otherwise obtain a traditional high school diploma, as well as those who would otherwise drop out. The theoretical results may also offer a solution to endogenous attrition bias in randomized trials; I illustrate this in a re-analysis of the 2008 Oregon Health Insurance Experiment.

Identification via stratifying interactions may be less feasible with a large number of treatments, though in such settings a researcher may be willing to forego unbiased estimation in favor of a low mean squared error (MSE) across the set of causal estimates. Indeed, policymakers, particularly in the spheres of education and healthcare, often make such trade-offs in forming observational measures of the average effectiveness, or "quality," of the public institutions they regulate. Conventional school value-added models (VAMs), for example, use empirical Bayes methods to shrink school-average test scores towards their grand mean after regression-adjusting for student demographics and past achievement. Under a standard selection-on-observables assumption, these VAMs produce minimum-MSE predictions of true school quality – predictions that increasingly drive school accountability and restructuring policies. The second chapter of this thesis (joint with Joshua Angrist, Parag Pathak, and Christopher Walters), shows how school admission lottery instruments can be used to test VAM validity, quantify the effects of selection bias, and improve conventional school quality predictions. In doing so, we provide a general framework for optimally combining observational and quasi-experimental estimators of multiple treatment effects. An application to Boston middle school quality suggests a failure of selection-on-observables and systematic bias in typical VAM estimates. Nevertheless, we find that the relative magnitude of this selection bias is modest, and policy decisions based on these VAMs are still likely to generate substantial achievement gains. Hybrid quality estimates that incorporate lottery information lead to further reductions in MSE and generate larger gains to students.

An important limitation of the approaches developed in the first two chapters is the assumption of homogeneous causal effects. A constant effects framework is, in particular, difficult to impose on quality measures derived from limited dependent variables, such as the 30-day mortality indicators used to estimate hospital quality in conventional risk adjustment models (RAMs). Moreover, this restriction rules out both institutional comparative advantage and Roy selection-on-gains, two powerful economic forces likely to arise in many settings, including healthcare. The third chapter of this thesis develops alternative IV techniques for estimating institutional quality which allow for nonlinear causal response functions, institutional specialization, and Roy selection. I use these techniques to estimate the quality of a large set of U.S. hospitals by their causal effects on emergency Medicare patient mortality, combining conventional RAM predictions and IV quality estimates derived from quasi-experimental ambulance company referral. I find that most emergency care market exhibit positive unobserved selection-on-gains. This non-random sorting obscures important correlations between quality and hospital ownership, volume, and average spending; incorporating quasi-experimental data in Medicare reimbursement policy tends to magnify existing transfers while leaving the kinds of hospitals that benefit from performance-based payment schemes unchanged. The results moreover highlight limitations of quality-based guidance policies in situations with significant positive Roy selection, an important but largely overlooked issue in prevailing quality-based regulation. Taken together, the three chapters of this thesis thus reveal a broad scope for IV-based estimates of multiple-treatment models for both research and policymaking.

JEL Classification: C26, I21, I11

Thesis Supervisor: Joshua D. Angrist
Title: Ford Professor of Economics

Thesis Supervisor: Amy N. Finkelstein
Title: John and Jennie S. MacDonald Professor of Economics

Thesis Supervisor: Parag A. Pathak
Title: Jane B. Carlton and Dennis W. Carlton Professor of Microeconomics

Acknowledgments

This thesis would not have been possible without the constant guidance and support of my MIT advisers Josh Angrist, Amy Finkelstein, and Parag Pathak, as well as the many instances of unofficial advising by Chris Walters of UC Berkeley. Along with their unbounded supply of knowledge and insight, I am grateful for my advisers' patient and thoughtful mentoring, their limitless generosity of time and attention, and their constant collegiality and optimism. I could not have asked for better teachers, role models, and collaborators.

Sincere thanks are also due to MIT faculty members Alberto Abadie, Nikhil Agarwal, Isaiah Andrews, David Autor, Victor Chernozhukov, Joe Doyle, Bob Gibbons, Jon Gruber, Anna Mikusheva, Whitney Newey, Jim Poterba, and Heidi Williams, as well as to my fellow graduate students Michael Abrahams, Alex Bartik, Aicha Ben Dhia, Vivek Bhattacharya, Sydnee Caldwell, Mayara Felix, John Firth, Bill Goulding, Nick Hagerty, Greg Howard, Sally Hudson, Donghee Jo, Tetsuya Kaji, Gabriel Kreindler, Jack Liebersohn, Ernest Liu, Rachael Meager, Yusuke Narita, Scott Nelson, Arianna Ornaghi, Bryan Perry, Brendan Price, Miikka Rokkanen, Lauren Russell, and Elizabeth Setren. I am also deeply grateful for financial support from the National Institute on Aging (grant #TC2-AG000186) and the Spencer Foundation (#201600065), as well as to the staff of the National Bureau of Economic Research, MIT, and the School Effectiveness and Inequality Initiative.

Last, but not least, I thank my parents, Gretchen and Jonathan, my sisters, Catharine and Margaret, and my close friends Gavin Alexander, Becky Eidelman, Tamar Glatman-Zaretsky, Damien Lally, Rebecca Martin, Todd Rosenthal, Adam Schlesinger, and Alexandre Staples for their constant supply of love, support, and welcome distractions over these past five years. You kept me (relatively) sane, against all odds, and never failed to brighten my day.

Thank you all so very, very much.

* * *

This thesis is dedicated to the memory of Cara Anne Nickolaus, a wise friend and a kind soul.

Contents

1 IsoLATEing: Identifying Counterfactual-specific Treatment Effects with Cross-stratum Comparisons **11**

1.1 Introduction 11

1.2 Theoretical Framework 14

1.2.1 IV Identification 18

1.2.2 Relaxing Independence and Homogeneity 22

1.3 Applications 25

1.3.1 The Returns to GED Certification 25

1.3.2 Differential Attrition in an RCT 28

1.4 Conclusions 31

1.5 Figures and Tables 33

1.6 Econometric Appendix 37

2 Leveraging Lotteries for School Value-added: Testing and Estimation **43**

2.1 Introduction 43

2.2 Setting and Data 47

2.2.1 Boston Public Schools 47

2.2.2 Data and Descriptive Statistics 48

2.3 Value-added Framework 49

2.4 Validating Conventional VAMs 51

2.4.1 Test Procedure 51

2.4.2 Test Results 53

2.4.3 Heterogeneity vs. Bias . . . 56
2.5 The Distribution of School Effectiveness . . . 58
2.5.1 A Random Coefficients Lottery Model . . . 58
2.5.2 Simulated Minimum Distance Estimation . . . 61
2.5.3 Empirical Bayes Posteriors . . . 62
2.6 Parameter Estimates . . . 65
2.6.1 Hyperparameters . . . 65
2.6.2 School Characteristics, Value-added, and Bias . . . 67
2.7 Policy Simulations . . . 68
2.7.1 Mean Squared Error . . . 69
2.7.2 Consequences of School Closure . . . 70
2.8 Conclusions . . . 72
2.9 Figures and Tables . . . 73
2.10 Data Appendix . . . 87
2.11 Econometric Appendix . . . 89
2.12 Supplementary Results . . . 100
2.13 Appendix Figures and Tables . . . 105

3 Estimating Hospital Quality with Quasi-experimental Data 117
3.1 Introduction . . . 117
3.2 Quality identification . . . 121
3.2.1 The Quasi-experimental Setting . . . 121
3.2.2 Non-parametric Identification . . . 122
3.2.3 Semi-parametric Quality Estimation . . . 124
3.3 Estimating Hospital Quality . . . 132
3.3.1 Data and RAMs . . . 132
3.3.2 Estimation . . . 134
3.3.3 Posteriors . . . 138
3.4 Results . . . 141
3.4.1 Hospital Quality and Patient Sorting . . . 141
3.4.2 Policy Consequences of RAM Bias . . . 144

3.5 Conclusions 148
3.6 Figures and Tables 151
3.7 Data Appendix 164
3.8 Econometric Appendix 166
3.9 Supplementary Results 171
3.10 Appendix Figures and Tables 174

Chapter 1

IsoLATEing: Identifying Counterfactual-specific Treatment Effects with Cross-stratum Comparisons

1.1 Introduction

What are the labor market returns to passing a high school equivalency test, such as the U.S. General Educational Development (GED) exam? As with many economic questions, a likely answer is "it depends." In recent years the inherent heterogeneity of causal effects has become a central consideration in applied research. In a seminal contribution, Imbens and Angrist (1994) show that an instrumental variables (IV) regression on a single binary treatment variable may estimate average causal effects for "compliers" – those who are induced into treatment by receipt of the instrument. IV identification of these local average treatment effects (LATEs) is ensured by a binary instrument that is as good as randomly assigned, monotone in its effect on treatment receipt, and excludable

For valuable feedback on this chapter, I thank Joshua Angrist, Kirill Borusyak, Mayara Felix, Amy Finkelstein, Sally Hudson, Jack Liebersohn, Kathleen Mullen, Parag Pathak, Miikka Rokkanen, Chris Taber, Chris Walters, and seminar participants from the University of Calgary and the Bank of Canada. I also gratefully acknowledge financial support from the National Institute on Aging (grant #T32-AG000186)

from potential outcome realizations.

When treatment effects vary, it is of natural interest to characterize their heterogeneity. A straightforward decomposition stratifies individuals along a dimension that is unaffected by and independent of the instrument. When the Imbens and Angrist (1994) assumptions hold within such strata, stratum-specific LATEs are identified by conditional IV regressions. For example, in a randomized trial in which offers for a particular program are assigned via lottery, stratification on pre-randomization controls can reveal differential effects of the program for compliers with different baseline characteristics.

Often, however, the most important dimensions of heterogeneity are not directly revealed by a baseline stratification. Suppose some individuals are caused to take the GED by a plausibly exogenous decrease in passing standards and subsequently may earn different wages in adulthood. Such compliers may be drawn from meaningfully different counterfactual levels of education: for some the alternative to the GED may be to drop out of high school, while others may see an easier GED as a lower-cost substitute to a high school diploma. Under the LATE assumptions, an IV regression with a single GED treatment channel identifies a causally-interpretable weighted average of effects for these two types of individuals, but this parameter may be difficult to interpret *economically*. Namely, if labor markets tend to reward workers for higher levels of educational achievement, the overall LATE may mix together potentially large positive and negative effects, resulting in a weighted average that could be either positive or negative depending on the complier mix. Except in very special situations, baseline measures are unlikely to provide enough information to perfectly separate individuals by their counterfactual educational attainment.

In some settings it may be difficult even to give *causal* interpretation to effects averaged over different treatment counterfactuals. In an extreme case, the instrument may move compliers from a state in which outcomes, measured by a survey or otherwise voluntarily provided, are completely unobserved to the researcher. This leads to the well-known problem of differential attrition: restricting analyses to individuals with *ex post* valid outcomes is likely to introduce selection bias into an otherwise gold-standard randomized design, while IV estimates obtained in the full sample do not identify meaningful causal parameters. Here isolating effects for a subset of compliers – those who would contribute outcomes even when untreated – is of first-order concern, yet as in the GED example potential attritors are unlikely to be perfectly identified and removed by baseline characteristics alone.

In this chapter I explore ways in which baseline stratifications, while not able to completely separate different complier groups, may nevertheless be useful for disentangling treatment effects by their counterfactual state. The basic strategy is intuitive: if there exists a stratification across which the composition of compliers with different alternatives varies but, on average, causal effects do not, differences in stratum-specific reduced-form effects may be attributed to differences in complier shares in such a way that identifies a LATE for treatment relative to each counterfactual. This intuition motivates an IV regression with multiple endogenous variables instrumented by interactions of the original instrument with stratum indicators. The requirement that average complier treatment effects be mean-independent of the stratification may be too strong in practice, however. In general I show that IV can inform bounds on particular linear combinations of counterfactual-specific effects, and that the approach may be generalized to settings where average cross-stratum heterogeneity is captured by a rich set of controls. In this case, I propose a non-parametric weighting estimator to identify the multiple LATEs.

Interacting an instrument with covariates to identify coefficients on multiple endogenous variables has a long history in economics; the results here extend the usual constant-effects framework to a minimal set of assumptions that allow for treatment effect heterogeneity while not imposing further behavioral or statistical restrictions.[1] In two closely related settings, Behaghel et al. (2013) consider non-parametric identification of multiple causal channels given an independently-assigned instrument for each channel, while Kirkebøen et al. (2016) show that counterfactual-specific LATEs may be recovered by IV regression when a researcher is able to directly observe and stratify on each individual's most-preferred alternative to treatment. This chapter offers an alternative approach for when only a single quasi-experiment is available and the counterfactual treatment status of each individual is unknown. All three techniques fit within the general principal stratification framework of Frangakis and Rubin (2002), which can be thought to extend the three behavioral groups – always-takers, never-takers, and compliers – of the original LATE theorem to allow for multiple causal channels.[2] Finally, in an extension the IV results I propose using covariates and a non-parametric weighting scheme to account for heterogeneity in average causal effects, an approach similar to that

[1] Wooldridge (2002) discusses identification of simultaneous equation models by nonlinear transformations of the instrument and predetermined covariates. Abdulkadiroğlu et al. (2016) and Cohodes (2015) are two recent examples of this approach in a treatment effects framework.

[2] In another example of this framework, Feller et al. (2014) estimate differential effects of early childhood interventions across alternative care programs using a parametric assumption on the distribution of potential outcomes. Kline and Walters (2016) and the analysis in chapter 3 explore alternative procedures using semi-parametric Roy selection models.

of Angrist and Fernandez-Val (2013) and Angrist and Rokkanen (2016) for LATE extrapolation across different quasi-experiments and within regression discontinuity designs, respectively.

I develop the main theoretical results in section 1.2 in the context of the motivating GED example, and in section 1.3 I illustrate identification with a selection model similar to the one developed by Heckman and Urzúa (2010). I then apply the theory to two settings. First, I exploit a plausibly exogenous policy change that differentially affected GED passing standards in five U.S. states to replicate the findings of Heckman et al. (2012) that an easier GED exam decreases high school completion rates. Furthermore, I find that non-GED students that are older at the time of the change are more likely to drop out than to finish high school. Leveraging an assumption that an individual's age in the year of reform is not systematically related to her returns-to-schooling profile in adulthood, I use a cohort stratification and instrumented difference-in-differences approach to jointly estimate average GED wage gains for those who would otherwise drop out from high school and those who would otherwise graduate. Although the extent of identifying variation in this application is modest and the estimation is correspondingly imprecise, the resulting point estimates are remarkably similar to the parameters of Heckman and Urzúa's original structural model.

Finally, I return in section 1.3 to the issue of non-random attrition in randomized control trials. Rather than restricting analyses to the subset of individuals who contribute outcomes *ex post*, I propose using a pre-randomization stratification to isolate causal effects for compliers that would always provide survey outcomes. One promising choice of strata uses the common surveying practice of limited intensive follow-up. Since in practice second-round intensive surveying is often random, this stratification is likely uncorrelated with the distribution of complier treatment effects. Moreover to the extent further follow-up attempts are successful, average response rates will vary by surveying intensity, generating plausibly exogenous cross-stratum variation in complier shares. I use this logic to estimate the effects of Medicaid enrollment in the Oregon Health Insurance Experiment. Despite evidence of significant differential attrition, the results confirm robustness of the original Finkelstein et al. (2012) estimates for a variety of financial, health, and medical care outcomes.

1.2 Theoretical Framework

For each individual in a population, suppose we observe a Bernoulli instrument Z, an outcome Y, a dummy covariate X, and a variable T which can equal either 1, a, or b. Here $T = 1$ indicates an

individual in treatment, while someone with $T = a$ or $T = b$ is said to be in one of two possible untreated states, or "fallbacks."[3] Indicators for being in a fallback state are given by A and B, respectively; treatment is then indicated by $D = 1 - A - B$. As a stylized example, we may imagine Z indicates a quasi-experimental reduction in a student's GED passing standards, Y denotes her adult earnings, and $D = 1$ if the individual becomes GED-certified. Uncertified individuals may either be high school dropouts ($A = 1$) or have a traditional high school diploma ($B = 1$).

As in Rubin (1974), causal effects are defined in terms of potential outcomes. Potential treatment and fallback states when $Z = z \in \{0, 1\}$ are written D_z, A_z, and B_z, while Y_{zt} denotes potential realizations of Y when the instrument takes on the value z and the treatment status is $t \in \{1, a, b\}$. Potential outcomes and assignments are assumed to be independent across individuals, satisfying the usual stable unit treatment value assumption, and may be linked to observed outcomes and fallback states in the following way:

$$Y = Y_1 - A(Y_1 - Y_a) - B(Y_1 - Y_a) \tag{1.1}$$

$$= Y^* + (1 - A)(Y_1 - Y_a) + (1 - B)(Y_1 - Y_b) \tag{1.2}$$

$$A = A_0 + Z(A_1 - A_0) \tag{1.3}$$

$$B = B_0 + Z(B_1 - B_0), \tag{1.4}$$

where $Y^* = Y_a + Y_b - Y_1$.

We start with three restrictions on the set of latent variables:

Assumption 1.1 *Independence*: $((Y_{z1}, Y_{za}, Y_{zb}, A_z, B_z)_{z=0,1}) \perp\!\!\!\perp Z \mid X$

Assumption 1.2 *Exclusion*: $Pr(Y_{0t} = Y_{1t}) = 1$, for each $t \in \{1, a, b\}$

Assumption 1.3 *Monotonicity*: $Pr(A_1 \leq A_0) = Pr(B_1 \leq B_0) = 1$.

In the GED example, Assumption 1.1 states that the variation in passing standards captured by Z is as good as randomly assigned with respect to potential outcomes, within strata defined by X. Conditional independence is sufficient for identification of the reduced-form causal effects of Z on Y, A, and B in each stratum. Interpretation of the earnings effect by way of schooling T requires an

[3] It is straightforward to state the following assumptions and prove identification results in the general case of n untreated states and n distinct elements in the support of X. This chapter restricts analysis to the specific case where $n = 2$ for ease of notation and exposition.

exclusion restriction (Assumption 1.2), which defines the single-indexed potential outcomes $Y_t = Y_{zt}$ for each t. Finally, we assume the effect of the instrument on treatment status is monotone, in the sense that no individual is induced to either untreated state by Z. Monotonicity is central to LATE identification and is naturally assumed in many contexts; in the stylized example it implies that no student is led to drop out or complete high school when it is easier to obtain a GED, which may be thought of as a revealed preference restriction.[4]

When Assumption 1.3 holds we may categorize individuals as one of four types by their potential treatment and fallback states:

	$D_1 = 0$	$D_1 = 1$
$D_0 = 0$	**1. Never-takers** ($A_1 = 1$, $A_0 = 1$ $B_1 = 0$, and $B_0 = 0$, or $A_1 = 0$, $A_0 = 0$, $B_1 = 1$, and $B_0 = 1$)	**2. a-compliers** ($A_1 = 0$, $A_0 = 1$, $B_1 = 0$, and $B_0 = 0$) **3. b-compliers** ($A_1 = 0$, $A_0 = 0$, $B_1 = 0$, and $B_0 = 1$)
$D_0 = 1$		**4. Always-takers** ($A_1 = 0$, $A_0 = 0$, $B_1 = 0$, and $B_0 = 0$)

In the GED example, never-takers are those who would either always drop out of high school or always obtain a traditional diploma, regardless of the testing environment, while always-takers are students that obtain a GED even when it is difficult to pass. Compliers are individuals who switch to a GED when the test becomes easier, and may either drop out of (a-compliers) or complete high school (b-compliers) when passing standards tighten. Note that a-compliers are those with $A_1 < A_0$ while b-compliers have $B_1 < B_0$.

Although stated with expanded notation, it is straightforward to verify that Assumptions 1.1-1.3 are equivalent to those typically used to analyze causal effects with a single treatment channel. The payoff to the more elaborate setup is that we can now write the conditional LATE identified by such analyses as an average of two fallback-specific LATEs for each of the now-differentiated a-complier and b-complier sub-populations. Specifically, we have the following:

[4]Students who would have otherwise completed high school may be led to *attempt* the GED when passing requirements are low but not actually pass and perhaps drop out instead. If passing forecast errors were systematically related to potential outcomes this may violate Assumption 1.3, though idiosyncratic monotonicity violations may be accommodated by extensions along the lines of de Chaisemartin (forthcoming) in the single-treatment case.

Lemma 1.1 : Consider the IV regression of Y on D, instrumented by Z and conditional on X. Suppose $Pr(D_1 > D_0|X) > 0$. Then under Assumptions 1.1-1.3 the endogenous regressor coefficient identifies

$$E[Y_1 - Y_a|A_1 < A_0, X]\omega(X) + E[Y_1 - Y_b|B_1 < B_0, X](1 - \omega(X)), \tag{1.5}$$

where

$$\omega(X) = \frac{Pr(A_1 < A_0|X)}{Pr(A_1 < A_0|X) + Pr(B_1 < B_0|X)}. \tag{1.6}$$

The proof of Lemma 1.1, derived in the econometric appendix along with all other propositions, uses the equivalence of Assumptions 1.1-1.3 to the assumptions of Imbens and Angrist (1994) in order to write the conditional IV estimand as a local average treatment effect. With two observable fallbacks to treatment, this LATE may in turn be written as the weighted average of average causal effects of the two complier groups, with weights equal to their population shares among all compliers. Thus, an IV regression with a single GED treatment channel weights together the LATE for students with a dropout counterfactual and the LATE for students who would have otherwise completed high school; here we are interested in extracting these two causal effects from the overall average.

Note that while the model has been formulated in terms of multiple fallback states, we could equivalently write Assumption 1.3 with the inequalities reversed and prove Lemma 1.1 and subsequent results with A and B corresponding to distinct treatment states and D representing a single fallback.[5] In this context it is worth noting that while Assumptions 1.1-1.3 ensure excludability of Z from Y given D (since Z is assumed excludable from Y given T, and Assumption 1.3 rules out the remaining possibility of Z shifting individuals across fallback states when $D = 0$), they do not rule out violations of exclusion when either A or B is considered in isolation. This is because an individual not observed in state a may still be induced into state b by the instrument if she is a b-complier, and similarly for b. Thus the theory derived here can also be thought of as addressing cases of known exclusion restriction violations, where the instrument affects untreated (from the

[5]Identification with multiple treatment states is quite similar to the literature on multiple "mediators." See Reardon and Raudenbush (2013) for related theory and Kling et al. (2007) and Pinto (2015) for applications with the Moving to Opportunity experiment.

perspective of one state) individuals by shifting them to another observable treatment state.[6]

Finally, it's worth noting that the multiple-treatment version of Assumptions 1.1-1.3 is closely related to the multiple "encouragement design" of Behaghel et al. (2013), who consider two treatment states as well as an instrument that takes on three values, $\tilde{Z} \in \{1, a, b\}$. With $Z = \mathbf{1}[\tilde{Z} = 1]$ and the underlying realization of the multi-valued instrument attributed to individual heterogeneity, Assumptions 1.1-1.3 are implied by their framework.[7] Under similar assumptions, Kirkebøen et al. (2016) discuss identification of fallback-specific LATEs in cases when the econometrician is able to directly measure A_0 and B_0 for all individuals. The identification results derived here can be thought of as an alternative method for the more general situation in which only one valid instrument for treatment is available, and where A_0 and B_0 (e.g. each student's potential schooling level when they do not take the GED) are not observable. Instead, the current approach requires a much weaker stratification that only generates some variation in the average unobserved counterfactual across the population.

1.2.1 IV Identification

Intuition for the main theoretical result can be seen in equations (1.5) and (1.6). Independence of the instrument ensures identification of the conditional first stage causal effect of Z on $1 - A$ and on $1 - B$, and the monotonicity assumption implies these are equal to $Pr(A_1 < A_0|X)$ and $Pr(B_1 < B_0|X)$, respectively. Thus the complier shares underlying the weighting scheme in Lemma 1.1 are identified. If the two counterfactual-specific LATEs are the same across the strata, equations (1.5) and (1.6) then constitute a system of two equations (one for each stratum) and two unknowns (the two counterfactual-specific LATEs), and identification is achieved as long as the two equations are not perfectly collinear. An IV regression with two endogenous variables instrumented by Z and the interaction of Z with X exactly implements this intuition.

[6]The multiple-treatment analogue of Assumptions 1.1-1.3 can also be placed in the ordered treatment model studied by Angrist and Imbens (1995). Namely, with $S = A + 2B$ and potentials defined accordingly, the ordered treatment setting may be modeled by Assumption 1.1, Assumption 1.2, and a modified Assumption 1.3 that only requires $Pr(S_1 \geq S_0) = Pr(A_1 + 2B_1 \geq A_0 + 2B_0) = 1$ rather than the stronger condition that $Pr(A_1 \geq A_0) = 1$ and $Pr(B_1 \geq B_0) = 1$.

[7]The present approach to the multiple-treatment setting uses an IV regression with two instruments and two endogenous variables, as does Behaghel et al. (2013). However the instruments proposed here will not satisfy their identifying assumptions in general, so that the approach is indeed distinct. In the special case where a-compliers and b-compliers are completely separated by the stratification the Behaghel et al. (2013) assumptions *will* be satisfied, and Proposition 1.1 may be proved in a manner similar to their result. This case is quite far afield from the main motivation of this paper, however, as it would in fact involve a baseline stratification that perfectly distinguishes the two causal channels of interest, as in Kirkebøen et al. (2016).

Formally, consider the just-identified, two-treatment IV system:

$$Y = \mu^y + \alpha(1-A) + \beta(1-B) + \gamma^y X + \epsilon^y \tag{1.7}$$

$$1 - A = \mu^a + \pi^a Z + \rho^a (Z \times X) + \gamma^a X + \epsilon^a \tag{1.8}$$

$$1 - B = \mu^b + \pi^b Z + \rho^b (Z \times X) + \gamma^b X + \epsilon^b .. \tag{1.9}$$

To economize on notation, define

$$\alpha(x) = E[Y_1 - Y_a | A_1 < A_0, X = x] \tag{1.10}$$

$$\beta(x) = E[Y_1 - Y_b | B_1 < B_0, X = x] \tag{1.11}$$

$$f_a(x) = Pr(A_1 < A_0 | X = x) \tag{1.12}$$

$$f_b(x) = Pr(B_1 < B_0 | X = x). \tag{1.13}$$

Here $\alpha(x)$ and $\beta(x)$ are stratum- and counterfactual-specific LATEs, while $f_a(x)$ and $f_b(x)$ denote the corresponding shares of a- and b-compliers. We then have the following result:

Proposition 1.1 : Suppose the matrix of complier shares,

$$\Pi = \begin{bmatrix} Pr(A_1 < A_0 | X = 0) & Pr(B_1 < B_0 | X = 0) \\ Pr(A_1 < A_0 | X = 1) & Pr(B_1 < B_0 | X = 1) \end{bmatrix}, \tag{1.14}$$

is nonsingular. Then under Assumptions 1.1-1.3 the endogenous regressor coefficients in equation (1.7) identify:

$$\alpha = \omega\alpha(0) + (1-\omega)\alpha(1) + \delta_a(\beta(0) - \beta(1)) \tag{1.15}$$

$$\beta = (1-\omega)\beta(0) + \omega\beta(1) + \delta_b(\alpha(0) - \alpha(1)), \tag{1.16}$$

where

$$\delta_a = \left(\frac{f_a(0)}{f_b(0)} - \frac{f_a(1)}{f_b(1)} \right)^{-1} \tag{1.17}$$

$$\delta_b = \left(\frac{f_b(0)}{f_a(0)} - \frac{f_b(1)}{f_a(1)} \right)^{-1}, \tag{1.18}$$

and

$$\omega = \left(1 - \frac{f_a(1)}{f_b(1)} / \frac{f_a(0)}{f_b(0)}\right)^{-1}. \tag{1.19}$$

Proposition 1.1 extends the Imbens and Angrist (1994) interpretation of IV with heterogeneous effects to regressions on multiple endogenous variables. This is achieved by interacting a single instrument with a stratification that induces variation in the composition of a- and b-compliers (so that the first-stage matrix Π is invertible and the IV rank requirement is satisfied). Although similarly derived, equations (1.15)-(1.18) are, however, not as easily interpreted as in the single-treatment case. For example, the IV coefficient on $1 - A$ is a weighted average of average causal effects from a-compliers in the two strata plus a "bias" term reflecting the gap in average causal effects of b-compliers across strata. Since the coefficient on this term is identified by components of the first stage matrix Π, with a bounded outcome one could identify bounds on the average $\omega\alpha(0)+(1-\omega)\alpha(1)$, with more narrow intervals given by stratifications where the ratio $f_a(X)/f_b(X)$ has more variation. Note, however, that since these ratios are always positive given Assumption 1.3, the parameter ω will be contained in $(-\infty, 0) \cup (1, \infty)$ so that its weighting scheme is never convex.

As can be seen in equations (1.15)-(1.16), a special case in which one may wish to estimate a regression of the form of equations (1.7)-(1.9) is when treatment effects are constant. More generally, α and β are easily interpreted causal parameters when the chosen stratification is mean-independent of the average treatment effect for the two groups of compliers. That is, suppose in addition to Assumptions 1.1-1.3 we have:

Assumption 1.4 *LATE homogeneity*: $E[Y_1 - Y_a | A_1 < A_0, X]$ and $E[Y_1 - Y_b | B_1 < B_0, X]$ do not depend on X

The form of equations (1.15)-(1.16) then makes the following result immediate:

Corollary to Proposition 1.1 : Suppose Π is of full rank. Then under Assumptions 1.1-1.4 the endogenous regressor coefficients in equation (1.7) identify $\alpha = E[Y_1 - Y_a | A_1 < A_0]$ and $\beta = E[Y_1 - Y_b | B_1 < B_0]$.

With LATE homogeneity, therefore, the multiple endogenous variable IV regression correctly deconvolutes the weighted average of fallback-specific LATEs given by Lemma 1.1.[8] In the stylized example, α and β identify the average returns to GED certification for individuals induced to the GED from a dropout and high school completion counterfactual, respectively, provided these do not vary systematically with the covariate X.[9]

It is instructive to consider what kinds of data-generating processes may accommodate Assumptions 1.1-1.4. Consider a classical selection model of an individual deciding between the treatment and alternative states to maximize her state-specific latent utility ν_{ti}:

$$A_i = \mathbf{1}[\nu_{ai} \geq \nu_{bi}, \nu_{ai} \geq \nu_{1i}] \tag{1.20}$$

$$B_i = \mathbf{1}[\nu_{bi} \geq \nu_{ai}, \nu_{bi} \geq \nu_{1i}] \tag{1.21}$$

$$D_i = \mathbf{1}[\nu_{1i} \geq \nu_{ai}, \nu_{1i} \geq \nu_{bi}]. \tag{1.22}$$

To satisfy Assumption 1.3 it is sufficient to have, for unidimensional η_i,

$$\nu_{1i} = h(X_i, Z_i, \eta_i), \tag{1.23}$$

such that $h(x, 1, \eta_i) \geq h(x, 0, \eta_i)$ almost-surely for $x = 0, 1$. Exclusion and LATE homogeneity then hold if potential outcomes may be written

$$Y_{ti} = \mu_t + \gamma X_i + \epsilon_{ti}, \tag{1.24}$$

such that

$$E[\epsilon_{1i} - \epsilon_{ai} | A_{1i} < A_{0i}, X_i] = E[\epsilon_{ai} - \epsilon_{ai} | A_{1i} < A_{0i}] \tag{1.25}$$

$$E[\epsilon_{1i} - \epsilon_{bi} | B_{1i} < B_{0i}, X_i] = E[\epsilon_{ai} - \epsilon_{bi} | B_{1i} < B_{0i}], \tag{1.26}$$

[8] It is straightforward to extend Proposition 1.1 to consider multi-valued X and the over-identified IV regression instrumented by multiple stratum interactions. When Assumption 1.4 holds across all values in the support of X the two LATEs will be identified by any such regression, just as in the constant effects case. A test of overidentifying restrictions would thus be valid for jointly testing Assumptions 1.1-1.4.

[9] Proposition 1.1 also provides a way to indirectly validate Assumption 1.4 given an exogenous control G thought to be correlated with individual treatment effects. For example, estimating equations (1.7)-(1.9) by setting $Y = G \times A$ yields a second-stage coefficient on $1 - B$ of $\delta_b(E[G|A_1 < A_0, X = 1] - E[G|A_1 < A_0, X = 0])$ under Assumptions 1.1-1.3, since then $\beta(0) = \beta(1) = 0$. One could therefore test whether the control G systematically varies for a-compliers across the X-stratification (and likewise for b-compliers). It is straightforward to generalize this test along the lines of the following section.

while Assumption 1.1 holds if the vector of structural disturbances $(\eta_i, \nu_{ai}, \nu_{b1}, \epsilon_{1i}, \epsilon_{ai}, \epsilon_{bi})'$ is independent of Z_i, conditional on X_i. Note that in writing equations (1.24)-(1.26) we are neither assuming that the stratum indicator X_i is excludable from the structural outcome equation (1.24) nor that it is independent of its error ϵ_{ti} (in which case X_i may itself be thought of as an instrument); rather, LATE homogeneity asserts that X_i enters the outcome equation in an additively-separable way, and that *differences* in the residual determinants of Y_{ti} are mean-independent of X_i in the compliant sub-populations. Section 1.3 illustrates identification through a parametric example of such a model.

1.2.2 Relaxing Independence and Homogeneity

That cross-stratum comparisons are informative for a common pair of average treatment effects is essential for their identification by IV. For any given application, which stratification is most likely to maintain an independent instrument and homogeneous LATEs while still producing first-stage variation depends on the specific context. Helpfully, as with intent-to-treat effect and conventional LATE identification (Hirano et al., 2003; Abadie, 2003), this approach may be extended to settings where Assumptions 1.1 and 1.4 only hold conditional on a rich set of predetermined covariates. The strategy is again intuitive: one could imagine running conditional versions of the IV regression (1.7)-(1.9) at each point in the support of a discretely-valued control W. When conditional cross-stratum comparisons identify conditional fallback-specific LATEs, averaging the resulting coefficients over the marginal complier distribution of W will recover population LATEs. Such a procedure is conceptually possible yet likely infeasible when W is continuous or takes on many discrete values.[10] I next outline an alternative, more flexible implementation of this basic idea for a generic vector of controls.

We start by considering the conditional analogues of Assumptions 1.1 and 1.4:

Assumption 1.1′ $((Y_{z1}, Y_{za}, Y_{zb}, A_z, B_z)_{z=0,1}) \perp\!\!\!\perp Z \mid W, X$

Assumption 1.4′ $E[Y_1 - Y_a | A_1 < A_0, W, X]$ and $E[Y_1 - Y_b | B_1 < B_0, W, X]$ do not depend on X

Here Assumption 1.1′ only requires the instrument Z to be as good as randomly assigned once

[10] As Hirano et al. (2003) note, a related issue is whether standard asymptotic theory adequately approximates the sampling distributions of such manually-reweighted estimators. See Robins and Ritov (1997) and Angrist and Hahn (2004) for a discussion of this problem.

potential confounders in W and X are held fixed, while Assumption 1.4′ allows for arbitrary cross-stratum heterogeneity in average complier treatment effects that is captured non-parametrically by W. We then have the following result:

Proposition 1.2 : Suppose $Pr(Z = 1|W, X)$ and $Pr(X = 1|W)$ are bounded away from zero and one and that the matrix of conditional complier shares

$$\Pi(W) = \begin{bmatrix} Pr(A_1 < A_0|W, X = 0) & Pr(B_1 < B_0|W, X = 0) \\ Pr(A_1 < A_0|W, X = 1) & Pr(B_1 < B_0|W, X = 1) \end{bmatrix} \tag{1.27}$$

is nonsingular with probability one. Define

$$\lambda = \frac{E[Z|W, X] - Z}{E[Z|W, X](1 - E[Z|W, X])}, \tag{1.28}$$

and

$$\mu_a = \frac{E[\lambda AX|W] - E[\lambda A|W]X}{E[X|W](1 - E[X|W])} \tag{1.29}$$

$$\mu_b = \frac{E[\lambda BX|W] - E[\lambda B|W]X}{E[X|W](1 - E[X|W])}. \tag{1.30}$$

Then, under Assumptions 1.1′, 1.2, 1.3, and 1.4′,

$$E[Y_1 - Y_a|A_1 < A_0] = E\left[\frac{E[\lambda A|W]}{E[\lambda A]}\frac{\lambda\mu_b}{E[\lambda\mu_b A|W]}Y\right] \tag{1.31}$$

and

$$E[Y_1 - Y_b|B_1 < B_0] = E\left[\frac{E[\lambda B|W]}{E[\lambda B]}\frac{\lambda\mu_a}{E[\lambda\mu_a B|W]}Y\right]. \tag{1.32}$$

Furthermore these weighting schemes for Y are non-parametrically identified by the conditional expectations $E[X|W]$, $E[Z|W, X]$, $E[A|W, X, Z]$, and $E[B|W, X, Z]$.

The proof of Proposition 1.2 shows that conditional-on-W versions of, for example, the coefficient on $1 - A$ in equation (1.7) can be written as the ratio of $E[\lambda\mu_b Y|W]$ to $E[\lambda\mu_b A|W]$. Averaging this ratio over the marginal distribution of W for a-compliers (using $E[\lambda A|W]/E[\lambda A]$ weights) thus

identifies $E[Y_1 - Y_a|A_1 < A_0]$. For these results to hold the conditional IV estimand must be well-defined along the support of W, so that both $Z|W, X$ and $X|W$ must be almost-surely stochastic and the conditional first-stage matrix $\Pi(W)$ must be always-surely invertible. Thus while W should be general enough to make Z ignorable and stratum-specific LATEs homogeneous, it must still allow for the kind of cross-stratum variation in complier shares underlying the basic IV approach.

As in the unconditional case, we can write a model consistent with Assumptions 1.1′ and 1.4′ by adding covariates to the structural treatment and potential outcome equations:

$$\nu_{1i} = h(W_i, X_i, Z_i, \eta_i) \tag{1.33}$$

$$Y_{ti} = f(W_i, X_i) + g_t(W_i) + \epsilon_{ti}, \tag{1.34}$$

where now differences in ϵ_{ti} are assumed to be mean-independent of X_i given W_i, the vector of structural errors is independent of Z_i given X_i and W_i, and the function $h(w, x, z, \eta_i)$ is almost-surely monotone in z given W_i and X_i. The conditional LATEs that are weighted together by Proposition 1.2 are then

$$E[Y_1 - Y_a|A_1 < A_0, W] = g_1(W) - g_a(W) + E[\epsilon_1 - \epsilon_a|A_1 < A_0, W] \tag{1.35}$$

$$E[Y_1 - Y_b|B_1 < B_0, W] = g_1(W) - g_b(W) + E[\epsilon_1 - \epsilon_b|B_1 < B_0, W]. \tag{1.36}$$

Proposition 1.2 motivates a tractable non-parametric estimation procedure for recovering the unconditional LATEs when Assumptions 1.1′, 1.2, 1.3, and 1.4′ hold. Namely, a researcher may in a first step flexibly approximate four conditional probability functions: $E[X|W]$, $E[Z|W, X]$, $E[A|W, X, Z]$, and $E[B|W, X, Z]$. The econometric appendix shows how these can then be used to form sample analogues of λ, μ_a, μ_b, $E[\lambda A|W]$, $E[\lambda B|W]$, $E[\lambda\mu_b A|W]$, and $E[\lambda\mu_a B|W]$, and thus of the weighting schemes in equations (1.31) and (1.32). Of course, unlike with the IV procedure in Proposition 1.1, inference for this multi-step estimator will in general be non-standard and somewhat more involved. In practice, under usual regularity conditions, finite-sample first-order approximations to the asymptotic distribution of the estimator may be based on either bootstrap procedures or analytic equations derived via the approaches of Andrews (1991) and Newey (1994a, 1994b).

1.3 Applications

1.3.1 The Returns to GED Certification

In 1997 the GED Testing Service required all U.S. states to meet new passing score requirements. Prior to this reform, five states – Louisiana, Mississippi, Nebraska, New Mexico, and Texas – awarded GEDs to students that obtained either a minimum score of 40 (out of a possible 80) on each of five standardized sub-tests *or* an average score of 45 across all sub-tests, while starting January 1st, 1997 both criteria were required nationwide. In a difference-in-differences design, Heckman et al. (2012) show that this increase in test difficulty, plausibly exogenous from the perspective of current students, significantly increased the share of high school graduates in affected states. The authors further show that the effect was concentrated among students who were older at the time of the policy change and were thus likely less constrained in their ability to drop out of high school when facing a more stringent GED passing threshold. To the extent an individual's age at the time of reform was not directly priced in the relative labor market returns she faced in subsequent decades, a stratification that filters the differential reduced-form effect of the policy across birth cohorts through differential rates of high school completion may be used to separate the overall causal effect into counterfactual-specific effects along the lines of Proposition 1.1.

I first illustrate this approach with a stylized model of degree choice and subsequent earnings. Heckman and Urzúa (2010) use such a selection model to demonstrate identification of multiple GED effects under large support or parametric conditions; I extend their simulated data-generating process to accommodate a stratification scheme consistent with Assumptions 1.1-1.4. Here Y denotes an individual's log hourly earnings in adulthood, D indicates GED certification, while A and B indicate the two GED fallbacks of dropping out and completing high school, respectively. The stratification X indicates an individual's age (either 16 or 17) when a quasi-experimental reduction in GED passing standards, Z, is announced. An appendix section contains a full description of how these variables are generated from draws of latent correlated factors in the Heckman and Urzúa (2010) parameterization.

Population first-stage and reduced-form coefficients from an IV regression of Y on D are reported in Panel A of Table 1.1. An exogenous decrease in GED passing standards increases the share of GED-certified students by 5.5 percentage points and decreases hourly earnings by an average of 1.2%, figures quite consistent with Heckman and Urzúa's original model. By monotonicity the

former represents the total share of a- and b-compliers in the population; from Lemma 1.1 the ratio of reduced-form to first-stage effects, -0.209, is the overall average complier GED effect. Both effects may be decomposed into strata-specific first-stage and reduced-form moments, displayed in Panel B. Compliers in the dropout-constrained ($X = 0$) subsample are more likely to obtain a high school diploma when untreated, while older compliers (with $X = 1$) are more likely to drop out in response to stricter GED passing standards. The model parameterizes adult wages such that students tend to see gains when shifted to the GED from a dropout counterfactual and losses when the GED replaces a high school diploma. Consequently, the reduced form effect of an easier GED exam is higher when $X = 1$ than when $X = 0$. Population coefficients in the two-treatment IV regression, $\alpha = 0.277$ and $\beta = -0.396$, are obtained by inverting the first stage matrix in columns 2 and 3 and multiplying by the reduced form vector in column 4. Since the model satisfies LATE homogeneity – an individual's cohort X is allowed to affect the level of her wages but not the returns-to-schooling frontier – by Proposition 1.1 these represent counterfactual-specific local average treatment effects of GED certification on adult earnings.

I next compare the performance of IV estimators in the simulated model with real-world estimates of the returns to GED certification using quasi-experimental variation in GED passing standards from the 1997 reform. I first construct a sample of 22,923 individuals born in the U.S. in 1978-1979 and in 1981-1982 who report positive earnings and hours worked in the 2013 American Community Survey and that completed at least two years of high school.[11] These individuals were of age 16 and 17 in either the year 1995 (one year prior to the mandated GED score change) or 1998 (one year after the change) and were likely to face differential GED costs while in high school. As in Heckman et al. (2012) I exclude the actual year of the policy change, which occurred in the middle of the academic calendar, for analytic clarity.[12] Trends in the rate of GED attainment and in log hourly earnings across birth cohorts and by birth states are plotted in Figure 1.1. The proportion of GED-certified individuals born in the five affected states declines sharply for the later

[11] Self-employed individuals and college-educated individuals are also excluded for ease of interpretation. Hourly wages are constructed by dividing annual wage and salary income by the product of the usual number of weeks worked within a year and the number of hours worked per week. The latter is imputed from categories reported in the 2013 ACS using the average number of hours reported within the same categories in the 2007 ACS, the last year in which underlying hours were reported. All reported results are robust to modest deviations in these sample construction choices.

[12] In their application, Heckman et al. (2012) also drop states that were otherwise affected by the policy change but that had already required candidates to meet both a minimum and mean score requirement, while showing robustness to the choice of control group. To increase power I use all states other than the five affected by the "and/or" scoring change as controls, though the results are similar without the already-required states.

birth cohorts, from 19.5 to 14.5 percentage points, while rates from other states show only a modest decrease. Importantly, similar comparisons between earlier birth cohorts not affected by the policy show almost no difference in certification trends, supporting the claim of Heckman et al. (2012) that such difference-in-differences comparisons may be causal. As in the calibrated model, an increase in passing standards led to a decline in GED completion by around 5 percentage points, with only a negligible overall increase in subsequent labor market earnings; these estimates are plotted in Figure 1.2.

Under assumptions analogous to those of Imbens and Angrist (1994), the ratio of difference-in-differences effects of the policy change on earnings to effects on GED completion identifies the average return to the GED for policy compliers (Hudson et al., 2015). Assumptions 1.1-1.3 may similarly be extended to accommodate this instrumented difference-in-differences framework, in which case the LATE explicitly corresponds to a weighted average of compliers with different non-GED schooling counterfactuals. Difference-in-differences estimates of the effect of the policy on high school dropout and completion rates are 2.3 and 2.7 percentage points, respectively. By monotonicity this suggests that among the 5% complier population, 46% would have dropped out under the stricter GED testing regime, while 54% would have completed high school instead. Consistent with Heckman et al. (2012) the dropout counterfactual appears concentrated in the older ($X = 1$) stratum, with 17 year-olds seeing a 4.3 percentage point increase in the probability of dropping out, compared with only 0.1 percentage points among 16 year-olds.

IV estimates of α and β using the 1997 reform are reported in column 1 of Table 1.2.[13] The overall effect of GED certification on log hourly wages for all compliers is estimated at -0.12, but as in the calibrated model an IV regression with two endogenous variables suggests a heterogeneous underlying story. Compliers who are induced to the GED from a dropout counterfactual appear to see an average increase in hourly wages of 15%, while those who are drawn from a high school diploma are estimated to take a massive average 35% cut in their hourly earnings. While this application and its estimates are intended as illustrative (indeed, inference based on birth-state clusters with a treatment group of only five states yields standard errors that fail to reject a large range of possible estimates), it is quite striking how closely they resemble the corresponding moments of the model parameterized to match the Heckman and Urzúa (2010) priors, reproduced

[13]Due to the small and likely weak sources of identifying variation in this example, I report bias-adjusted (Fuller) 2SLS estimates of these parameters, though unadjusted estimates are essentially the same. All regressions control for state of birth and residency.

in column 2. Monte carlo replications of similarly-powered IV regressions reported in column 3 also closely track the real-world LATE estimates.

1.3.2 Differential Attrition in an RCT

Distinguishing between multiple treatment alternatives can be of first-order importance in a randomized control trial with imperfect follow-up. Suppose program offers Z are randomly assigned to an initial population who may then choose whether or not to comply with the treatment *and* whether or not to report subsequent outcomes, Y. Individuals can then be said to select between three possible states: being treated and reporting outcomes (D), not being treated and reporting outcomes (A), and not reporting outcomes (B). Since outcomes are only measured in states D and A (suppose the researcher arbitrarily sets $Y = 0$ for anyone with $B = 1$), the estimable local average treatment effect in the entire sample,

$$E[Y_1 - Y_0|D_1 > D_0] = E[Y_1 - Y_a|A_1 < A_0]\omega + E[Y_1|B_1 < B_0](1 - \omega) \tag{1.37}$$

$$\text{for} \quad \omega = \frac{Pr(A_1 < A_0)}{Pr(A_1 < A_0) + Pr(B_1 < B_0)}, \tag{1.38}$$

is not a weighted average of causal treatment effects on the latent, potentially unreported outcome whenever there are any compliers with B as a fallback (that is, when $Pr(B_1 < B_0) \neq 0$). Facing such endogenous attrition, researchers often choose to conduct their analyses on a restricted sub-sample of individuals that report outcomes, in the hope of identifying causal effects. Such a procedure, however, is also unlikely to be easily interpreted when $Pr(B_1 < B_0) \neq 0$, as conditioning on *ex post* outcomes ($B_Z = 0$) will then in general introduce correlation between potential outcomes and the unconditionally-independent instrument.[14]

As the form of equations (1.37) and (1.38) suggests, the differential attrition problem can be mapped to the multiple-counterfactual setting of Assumptions 1.1-1.3. For a given baseline stratification X, independence of Z from potential outcomes (Assumption 1.1) is ensured by virtue of the randomized design, and in many settings a program offer is likely to have no direct effect on latent outcomes and to not deter program participation. Assumptions 1.2 and 1.3 would then be satisfied

[14]Common approaches to the differential attrition problem include parametric sample selection modeling (Gronau, 1974; Heckman, 1976) and partial non-parametric identification of causal effects (Lee, 2009; Behaghel et al., 2009; Engberg et al., 2014). Methods involving Bayesian inference (Little and Rubin, 1987) and covariate re-weighting (Frölich and Huber, 2014) have also been proposed under different assumptions than those considered here.

provided that (1) Z has no direct effect on attrition behavior given treatment status and (2) the effect of assignment on attrition through treatment is monotone. These primitive assumptions that place the differential attrition problem within the general setting considered here are the same as those used to estimate non-parametric bounds on causal parameters by the methods of Lee (2009) and Behaghel et al. (2009).[15]

To solve the differential attrition problem with Proposition 1.1, we require an appropriate pre-randomization stratification that induces variation in response behavior while maintaining LATE homogeneity. One candidate exploits the practice of randomized intensive follow-up, a common surveying technique that is often recommended when attrition rates are large (e.g. Duflo et al., 2008). Specifically, suppose that upon initially measuring outcomes a researcher selects among the attritors a random fraction p for additional follow-up attempts. Denote this set and another random fraction p of initial responders by $X = 1$ and let $X = 0$ for all other individuals.[16] Since X is highly correlated with an individual's probability of facing more intensive follow-up, the $X = 1$ strata is likely to contain a relatively larger proportion of untreated compliers with an observed outcome. Moreover, since X is randomly assigned in the population, LATEs for both types of compliers will be the same across strata, provided the additional follow-up attempt draws second-round responses from individuals representative of the pool of initial non-responders.[17]

Importantly, however, Assumption 1.4 is not guaranteed by randomized intensive follow-up *per se*; researchers hoping to use Propositions 1.1 or 1.2 to resolve differential attrition concerns should carefully consider and perhaps design their intensive surveying schemes in order to maximize the plausibility of this approach. As a simplistic but instructive example, suppose a researcher randomly assigns offers for a job-training program and initially conducts phone interviews on employment outcomes (e.g. whether or not a person has successfully found a new job) throughout the day. As treated individuals may be more likely to be employed, the offer may have an effect (through treatment) on the probability an individual will be home to answer the survey: these people will have $B_1 < B_0$. However, suppose the exact timing of follow-up interviews is as good as random

[15]Note that monotonicity of response behavior with respect to the instrument is central to the latent index framework most commonly used to study and assess general selection bias (Angrist, 1997).

[16]It will, by Rao-Blackwell logic, in fact be more efficient to let $X = p$ for all initial responders, rather than employing a randomized estimator. I follow this approach in the empirical application below.

[17]In an unpublished manuscript, DiNardo et al. (2006) discuss parametric and semi-parametric methods of using randomized intensive follow-up to overcome differential attrition in estimating intent-to-treat effects. The current approach is distinct from their approach, both in the interest in local average treatment effects and the specific way the intensive follow-up scheme is used for identification.

with respect to working hours (perhaps due to alphabetical or other quasi-random queuing of survey attempts), and that the random second round of interviews occurs in similar fashion on a subsequent day. In this case individuals in the intensive stratum of X will face a higher probability of being home when surveyed (on either day one or day two), but those successfully interviewed on the second day will not vary systematically from those interviewed in the initial round. Assumption 1.4 would then hold, and Proposition 1.1 may be used to overcome the differential attrition problem. Such an approach would, however, likely fail under alternative, non-randomized follow-up interview attempts.

I follow this method to estimate the effects of Medicaid on survey outcomes in the first year following a lottery of roughly 90,000 low-income adults in Oregon. Finkelstein et al. (2012) discuss the setting for the Oregon Health Insurance Experiment, which selected around 35,000 individuals over eight lottery drawings from March through September 2008. Selected individuals became eligible for enrollment in OHP Standard, a comprehensive Medicaid program, and roughly 30% of lottery winners successfully enrolled. In addition to administrative hospital discharge data, Finkelstein et al. (2012) collected outcomes by a mail survey, distributed one year later in the summer of 2009, and found evidence that Medicaid increased health care utilization, decreased out-of-pocket expenditure and debt, and improved overall health among survey responders. The relatively low rate of response (at 50%) and moderate imbalance (at around 2 percentage points) in the probability of response by eligibility status, however, suggest caution in interpreting these restricted IV estimates.[18]

I use the Finkelstein et al. (2012) public-use database to replicate the authors' main survey analysis sample. For illustrative simplicity I restrict attention to the largest experimental stratum, consisting of 9,770 single-person household members in the seventh survey wave. Attrition appears to be a more serious issue in this sub-sample, with an overall response rate of only 42% and with eligible individuals roughly 4 percentage points less likely to respond to any survey question.[19] As in the main sample, 30% of initial non-respondents were selected for additional follow-up attempts

[18]Finkelstein et al. (2012) address attrition concerns by showing balance in eligibility status across baseline covariates in the survey respondent sub-sample. The authors also construct Lee (2009) bounds for intent-to-treat effects, finding generally robust results for health care use and financial strain outcomes while not able to reject the null of no effect on self-reported health.

[19]Although equations (1.37)-(1.38) describe a model in which Z makes attrition less likely, Proposition 1 may also be used to recover causal effects when the instrument monotonically increases (through treatment) the probability of non-response. In either case using $1 - A = 1 - (1 - C)R$ and $1 - B = R$ as the two endogenous variables in the two-instrument IV regression (where C indicates treatment receipt and R denotes survey response) will identify the average causal effect of treatment among compliers who always respond by the coefficient on $1 - A$.

by mail and phone. The average yield on such intensive surveying was around 22%, suggesting a strong contrast across the stratification scheme described above. I let the stratum indicator $X = 1$ for those designated for intensive follow-up and for a proportionate random sample of initial respondents. The endogenous variables A, B, and D are constructed from survey response and treatment indicators as outlined above.

IV estimates of the effect of Medicaid enrollment on a variety of health, financial, and medical care outcomes are reported in Table 1.3. As in Finkelstein et al. (2012), column 1 reports "restricted" IV estimates from specifications with a single treatment variable D, estimated over the subsample of individuals with successfully recorded outcomes.[20] Column 2 instead reports estimates of the coefficient on $1 - A$ in IV regressions of the form of equations (1.7)-(1.9) from the full experimental sample. By Proposition 1.1, these represent local average treatment effects for compliers who would always provide survey outcomes. Interestingly, the two-treatment IV specification yields point estimates quite close to those obtained by the restricted single-treatment model across virtually every outcome. Although the former is generally less precisely-estimated, the two are highly correlated, so that estimated differences (reported in column 3 of Table 1.3) are tightly distributed around zero. This suggests that, despite apparent endogenous attrition, the estimates reported in Finkelstein et al. (2012) serve as fairly reliable measures of true causal effects of Medicaid enrollment.

1.4 Conclusions

Although originally formulated within the context of additive, constant-effects causal models, the method of instrumental variables is often found to be robust to deviations from such parametric frameworks. Indeed, IV estimation of treatment effects has often clarified the minimal assumptions needed for causal interpretation in a heterogeneous world. This chapter adds to this tradition by extending the theoretical framework of Imbens and Angrist (1994) to settings where more than one causal channel is needed to answer an economic or causal question but only one quasi-experiment is available. The ease by which Proposition 1.1 may be applied, using an estimator with statistical properties familiar to most applied researchers, is readily apparent in the above empirical applications. More involved, though still tractable estimation may be used to relax the key identifying

[20] I follow Finkelstein et al. (2012) in weighting all restricted IV estimates by the inverse probability of being included in the intensive follow-up group. In practice this has little effect on the restricted IV point estimates.

assumptions given sufficiently rich controls. As the discussion in Section 1.3.2 illustrates, in some cases one may be able to increase the plausibility of the key LATE homogeneity assumption by a carefully-constructed surveying design. This suggests new tools for overcoming the fundamental issue of differential attrition in randomized program evaluation.

1.5 Figures and Tables

Figure 1.1: Trends in GED Attainment and Log Hourly Earnings by the 1997 "and/or" GED Scoring Change

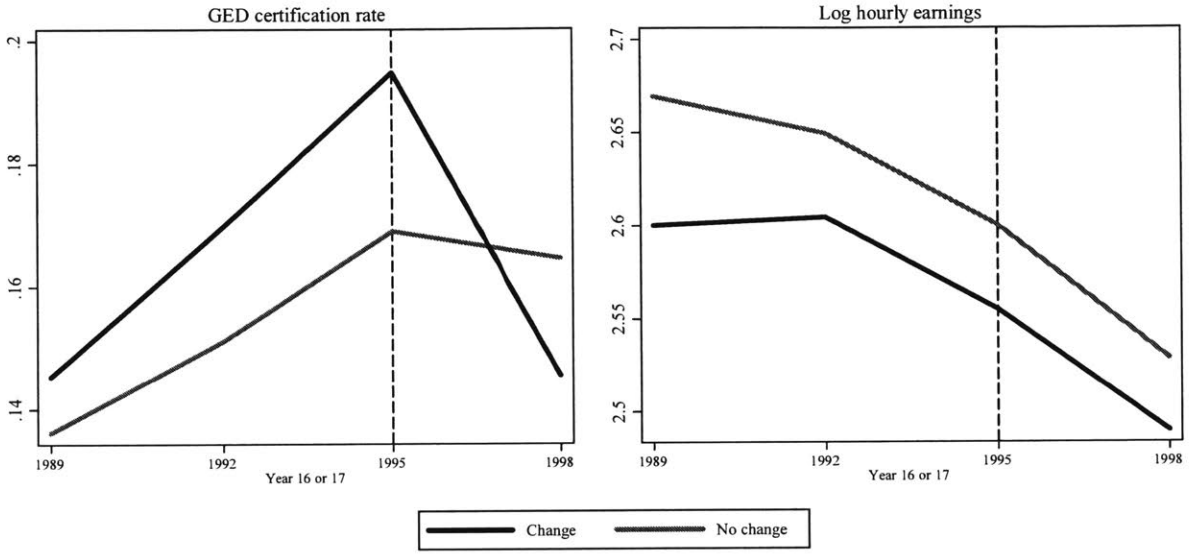
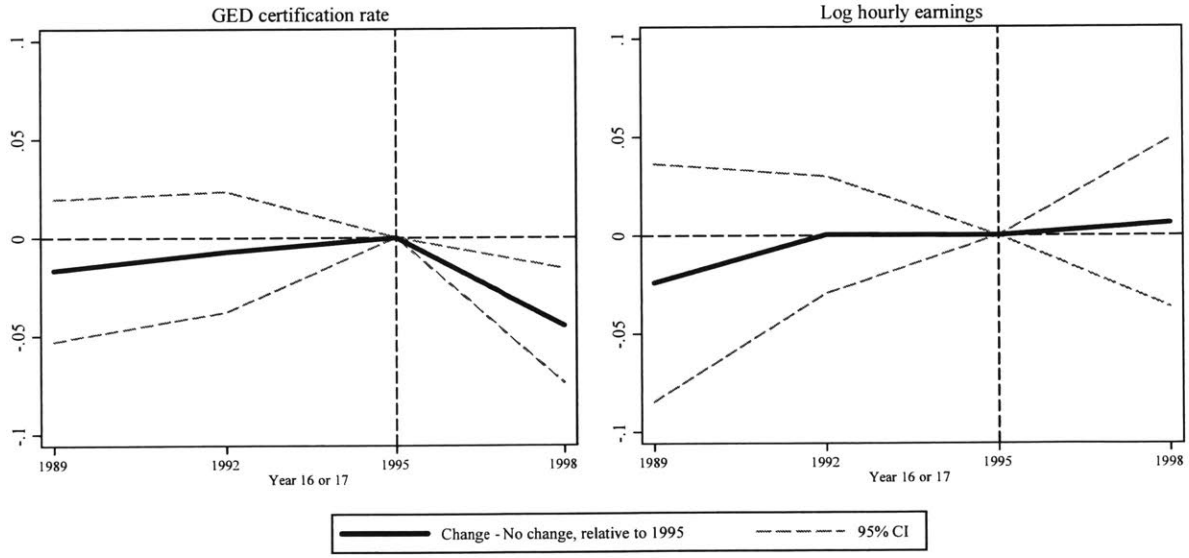


Figure 1.2: Difference-in-differences Estimates of the Effect of the 1997 "and/or" Scoring Change



Notes: Figure 1.1 plots average GED certification rates and 2013 log hourly earnings, by birth year, for a sample of employed individuals born in states that were and were not required to eliminate the "and/or" scoring option on the GED in 1997. Figure 1.2 plots growth in these variables relative to the cohort that was age 16 or 17 in 1995. 95% confidence intervals are based on robust standard errors that cluster by birth state.

Table 1.1: Simulated First-stage and Reduced-form Effects of a GED Scoring Change

	First stage (complier shares)			Reduced form
	D (all compliers)	1-A (dropout counterfactual)	1-B (high school diploma counterfactual)	Y (log hourly earnings)
	(1)	(2)	(3)	(4)
		A. Single-treatment IV		
Full sample	0.055			-0.012
		B. Two-treatment IV		
Dropout-constrained stratum (X=0)		0.005	0.050	-0.019
Unconstrained stratum (X=1)		0.073	0.029	0.009

Notes: This table reports moments from the simulated model of GED effects inspired by Heckman and Urzúa (2010) and described in the text. Column 4 reports reduced-form effects of the instrument, a decrease in GED passing standards by 0.75 standard deviations, on log hourly earnings in the full sample (Panel A) and in two subsamples differentiated by the average difficulty of dropping out of high school (Panel B). Column 1 reports first-stage effects of the instrument on an indicator for completing the GED in the full sample, while columns 2 and 3 report first-stage effects on 1-A and 1-B in each subsample, where A indicates a student dropping out of high school and B indicates high school completion. The single-treatment IV coefficient is -0.209, the ratio of reduced-form to first-stage effects in Panel A. The two-treatment (isoLATE) IV coefficients are 0.277 and -0.396, the inverse of the first-stage matrix in Panel B post-multiplied by the reduced-form vector.

Table 1.2: Estimated and Simulated Returns to GED Certification

	ACS data	Data calibrated to the Heckman and Urzúa (2010) model	
	isoLATE estimates	Population LATEs	isoLATE monte carlo
	(1)	(2)	(3)
		A. Single-treatment IV	
All compliers	-0.121	-0.209	-0.207
	(0.268)		(0.109)
		B. Two-treatment IV	
Dropout counterfactual compliers	0.150	0.277	0.257
	(0.506)		(0.418)
High school diploma counterfactual compliers	-0.347	-0.396	-0.392
	(0.208)		(0.203)

Notes: Column 1 of this table reports estimates of local average treatment effects in a sample of 22,923 employed individuals who were either 16 or 17 in either 1995 or 1998. The outcome is 2013 log hourly earnings. A cohort indicator and state of birth and residency indicators are included as controls, with the interaction of cohort and an indicator for being born in a state subject to a "and/or" score change in 1997 as the excluded instrument. The isoLATE stratification is by those born in the earlier vs. later year of their cohort. Column 2 reports corresponding moments of the model parameterized according to Heckman and Urzua (2010) and described in the text, while column 3 reports average IV estimates of these moments from 500 monte carlo replications of the two-treatment IV specification (N=100,000). Robust standard errors, clustered by birth state, are reported in parentheses in column 1; estimate standard deviations are reported in parentheses in column 3.

Table 1.3: Estimated Medicaid Effects from the Oregon Health Insurance Experiment

	Estimation		Difference
	Restricted IV	isoLATE IV	
	(1)	(2)	(3)
A. Healthcare access			
Have usual place of clinic-based care	0.335	0.397	-0.062
	(0.073)	(0.140)	(0.096)
Have personal doctor	0.264	0.184	0.080
	(0.069)	(0.147)	(0.113)
Got all needed medical care, last six	0.266	0.215	0.051
	(0.061)	(0.120)	(0.088)
Got all needed drugs, last six months	0.242	0.199	0.043
	(0.054)	(0.096)	(0.069)
Didn't use ER for nonemergency, last	-0.037	-0.080	0.043
six months	(0.040)	(0.086)	(0.064)
B. Healthcare utilization			
Using prescription drugs currently	-0.040	-0.107	0.067
	(0.078)	(0.185)	(0.145)
Outpatient visits, last six months	0.199	0.159	0.040
	(0.066)	(0.120)	(0.075)
ER visits, last six months	0.038	0.043	-0.005
	(0.063)	(0.104)	(0.054)
Inpatient hospital admissions, last	0.046	0.085	-0.039
six months	(0.041)	(0.088)	(0.065)
C. Financial strain			
Any out of pocket medical expensis,	-0.204	-0.223	0.019
	(0.069)	(0.119)	(0.068)
Owe money for medical expenses	-0.257	-0.253	-0.005
	(0.068)	(0.117)	(0.067)
Borrowed/skipped bills to pay medical	-0.196	-0.155	-0.041
bills, last six months	(0.066)	(0.118)	(0.075)
Refused treatment because of medical	-0.015	-0.005	-0.010
debt, last six months	(0.039)	(0.056)	(0.027)
D. Health outcomes			
Heath good/very good/excellent	0.225	0.192	0.033
	(0.071)	(0.129)	(0.079)
Health not poor	0.113	0.148	-0.035
	(0.046)	(0.092)	(0.063)
Health same or better, last six months	0.225	0.225	0.000
	(0.063)	(0.107)	(0.059)

Notes: This table reports 2SLS estimates of the effects of Medicaid using randomized Medicaid offers from the Oregon Health Insurance Experiment as instruments. Columns 1 and 4 use a single treatment variable, restrict estimation to those individuals with valid survey responses for each outcome, and weight by the inverse probability of intensive follow-up, as in Finkelstein et al. (2012). Columns 1 and 4 are estimated with two endogenous variables as described in the text using the full sample of 9,770 single-person households in the 7th survey wave. Robust standard errors are reported in parentheses.

1.6 Econometric Appendix

Proof of Lemma 1.1

Consider the reduced-form regression of Y on Z, conditional on X. By the excludability of Z given treatment and fallback status (Assumption 2), we can write

$$Y = Y^* + (1 - A)(Y_1 - Y_a) + (1 - B)(Y_1 - Y_b),$$

where $Y^* = Y_a + Y_b - Y_1$. Thus the slope of this regression identifies:

$$\begin{aligned}
E[Y|Z = 1, X] - E[Y|Z = 0, X] =& E[Y^* + (1 - A)(Y_1 - Y_a) + (1 - B)(Y_1 - Y_b)|Z = 1, X] \\
& - E[Y^* + (1 - A)(Y_1 - Y_a) + (1 - B)(Y_1 - Y_b)|Z = 0, X] \\
=& E[Y^*|Z = 1, X] - E[Y^*|Z = 0, X] \\
& + E[(1 - A_1)(Y_1 - Y_a)|Z = 1, X] - E[(1 - A_0)(Y_1 - Y_a)|Z = 0, X] \\
& + E[(1 - B_1)(Y_1 - Y_b)|Z = 1, X] - E[(1 - B_0)(Y_1 - Y_b)|Z = 0, X] \\
=& E[(Y_1 - Y_a)(A_0 - A_1)|X] + E[(Y_1 - Y_b)(B_0 - B_1)|X] \\
=& E[Y_1 - Y_a|A_1 < A_0, X]Pr(A_1 < A_0|X) \\
& + E[Y_1 - Y_b|B_1 < B_0, X]Pr(B_1 < B_0|X),
\end{aligned}$$

where the third equality follows by independence of Z given X (Assumption 1.1) and the fourth by monotonicity (Assumption 1.3). Furthermore, the conditional first-stage regression of D on Z is

$$\begin{aligned}
E[D|Z = 1, X] - E[D|Z = 0, X] &= E[D_1 - D_0|X] \\
&= E[(1 - A_1 - B_1) - (1 - A_0 - B_0)|X] \\
&= E[A_0 - A_1|X] + E[B_0 - B_1|X] \\
&= Pr(A_1 < A_0|X) + Pr(B_1 < B_0|X).
\end{aligned}$$

This again follows by Assumptions 1.1 and 1.3. The conditional IV coefficient on D is the ratio of reduced-form to first-stage expressions, completing the proof □

Proof of Proposition 1.1

The proof to Lemma 1.1 shows that, under Assumptions 1.1-1.3, the conditional reduced form is

$$E[Y|Z=1,X] - E[Y|Z=0,X] = \alpha(X)f_a(X) + \beta(X)f_b(X),$$

while the conditional first stages are

$$E[1-A|Z=1,X] - E[1-A|Z=0,X] = f_a(X)$$
$$E[1-B|Z=1,X] - E[1-B|Z=0,X] = f_b(X).$$

Consider the multiple-endogenous variable IV regression of equations (1.7)-(1.9). Let $\mathbf{Y}$ denote a vector of observations of Y, $\mathbf{X}$ a matrix of observations of $1-A$ and $1-B$, and $\mathbf{Z}$ a matrix of observations of Z and ZX. The endogenous regressor coefficients satisfy:

$$\begin{aligned}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} &= p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right)\\ &= p\lim\left(((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right),\end{aligned}$$

where $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{X}}$ are matricies of residuals from regressing $\mathbf{Z}$ and $\mathbf{X}$ on X and a constant. By above,

$$p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}}\right) = \begin{bmatrix}f_a(0) & f_b(0)\\ f_a(1) & f_b(1)\end{bmatrix} = \Pi$$
$$p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right) = \begin{bmatrix}\alpha(0)f_a(0)+\beta(0)f_b(0)\\ \alpha(1)f_a(1)+\beta(1)f_b(1)\end{bmatrix}.$$

When Π is invertible, the continuous mapping theorem and Slutsky's theorem imply:

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix}f_a(0) & f_b(0)\\ f_a(1) & f_b(1)\end{bmatrix}^{-1}\begin{bmatrix}\alpha(0)f_a(0)+\beta(0)f_b(0)\\ \alpha(1)f_a(1)+\beta(1)f_b(1)\end{bmatrix}.$$

Proposition 1.1 follows from simplifying this expression. □

Proof of Proposition 1.2

Define, for each $V \in \{A, B, Y\}$,S

$$\begin{aligned}
\delta^V_{W,X} &= E[V|Z=0,W,X] - E[V|Z=1,W,X] \\
&= \frac{E[(1-Z)V|W,X]}{1-E[Z|W,X]} - \frac{E[ZV|W,X]}{E[Z|W,X]} \\
&= E\left[\frac{E[Z|W,X]-Z}{E[Z|W,X](1-E[Z|W,X])}V|W,X\right] \\
&= E[\lambda V|W,X],
\end{aligned}$$

and note that by Assumptions 1.1′ and 1.2 we have

$$\begin{aligned}
\delta^A_{W,X} &= E[A_0 - A_1|W,X] \\
\delta^B_{W,X} &= E[B_0 - B_1|W,X] \\
\delta^Y_{W,X} &= E[(Y_a - Y_1)(A_0 - A_1) + (Y_b - Y_1)(B_0 - B_1)|W,X].
\end{aligned}$$

Next define the random vector

$$\begin{bmatrix} \alpha_W \\ \beta_W \end{bmatrix} = \begin{bmatrix} \delta^A_{W,0} & \delta^B_{W,0} \\ \delta^A_{W,1} & \delta^B_{W,1} \end{bmatrix}^{-1} \begin{bmatrix} \delta^Y_{W,0} \\ \delta^Y_{W,1} \end{bmatrix}.$$

Here α_W and β_W are conditional analogues of the multiple endogenous variable IV specification used in Proposition 1.1. Following the same steps as there, we may show that under the assumptions

$$\begin{aligned}
\alpha_W &= \frac{\delta^Y_{W,0}\delta^B_{W,1} - \delta^Y_{W,1}\delta^B_{W,0}}{\delta^A_{W,0}\delta^B_{W,1} - \delta^A_{W,1}\delta^B_{W,0}} \\
&= E[Y_1 - Y_a|A_1 < A_0, W],
\end{aligned}$$

so that, by the Law of Iterated Expectations,

$$\begin{aligned}
E[Y_1 - Y_a|A_1 < A_0] &= E\left[\frac{E[(Y_1 - Y_a)(A_1 - A_0)|W]}{Pr(A_1 < A_0)}\right] \\
&= E\left[\frac{Pr(A_1 < A_0|W)}{Pr(A_1 < A_0)}\alpha_W\right].
\end{aligned}$$

Finally, note that we can write

$$\begin{aligned}\delta^V_{W,0}\delta^B_{W,1} - \delta^V_{W,1}\delta^B_{W,0} &= E[\lambda V|W, X=0]E[\lambda B|W, X=1] - E[\lambda V|W, X=1]E[\lambda B|W, X=0] \\ &= \frac{E[\lambda V(1-X)|W]E[\lambda BX|W]}{(1-E[X|W])E[X|W]} - \frac{E[\lambda VX|W]E[\lambda B(1-X)|W]}{E[X|W](1-E[X|W])} \\ &= E\left[\lambda \frac{E[\lambda BX|W] - E[\lambda B|W]X}{E[X|W](1-E[X|W])} V|W\right] \\ &= E\left[\lambda \mu_b V|W\right],\end{aligned}$$

and

$$\begin{aligned}Pr(A_1 < A_0|W) &= E[E[A_0 - A_1|W, X]|W] \\ &= E[E[\lambda \cdot A|W, X]|W] \\ &= E[\lambda A|W].\end{aligned}$$

Thus, once again applying the Law of Iterated Expectations,

$$E[Y_1 - Y_a|A_1 < A_0] = E\left[\frac{E[\lambda A|W]}{E[\lambda A]} \frac{\lambda \mu_b Y}{E\left[\lambda \mu_b A|W\right]}\right]$$

The same steps show the result for $E[Y_1 - Y_b|B_1 < B_0]$.

Note that the function $\lambda(w, x, z)$ generating $\lambda = \lambda(W, X, Z)$ is identified by the conditional expectation function $E[Z|W, X]$ and that

$$\begin{aligned}E[\lambda A|W=w] = \sum_{x=0,1}\sum_{z=0,1} &\lambda(w,x,z)E[A|W=w, X=x, Z=z] \\ &\times Pr(Z=z|W=w, X=x)Pr(X=x|W=w)\end{aligned}$$

and similarly for $E[\lambda B|W]$. Moreover,

$$E[\lambda BX|W=w] = \sum_{z=0,1} \lambda(w,1,z)E[B|W=w, X=1, Z=z]Pr(Z=z|W=w, X=1)E[X|W=w].$$

Thus both the weights $E[\lambda A|W=w]/E[\lambda A]$ and the function $\mu_b(w,x)$ generating $\mu_b = \mu_b(W,X)$ are identified by the conditional expectation functions $E[X|W]$, $E[Z|W,X]$, $E[A|W,X,Z]$, and

$E[B|W, X, Z]$. Finally, note that

$$E[\lambda\mu_b A|W = w] = \sum_{x=0,1}\sum_{z=0,1} \lambda(w, x, z)\mu_a(w, z)E[A|W = w, X = x, Z = z] \times Pr(Z = z|X = x, W = w)Pr(X = x|W = w)$$

We can thus form sample analogues of the weighting schemes identifying $E[Y_1 - Y_a|A_1 < A_0]$ from non-parametric estimates of these conditional expectation functions. The same result follows for $E[Y_1 - Y_b|B_1 < B_0]$. □

GED Selection Model

Section 1.3.1 simulates data on educational attainment and labor market returns using a model inspired by Heckman and Urzúa (2010). Potential log hourly earnings are given by

$$Y_{ti} = \mu_t + \gamma X_i + \epsilon_{it}, \tag{1.39}$$

where $X_i = 1$ is a cohort indicator and $t \in \{1, a, b\}$ indexes the individual's educational status: GED-certified, high school dropout, or traditional high school graduate. Individuals observe the schooling environment and chooses the alternative t that maximizes ν_{ti}, where

$$\nu_{1i} = \Phi(\pi_1\tilde{Z}_{1i} - \eta_{1i}) \tag{1.40}$$

$$\nu_{ai} = \Phi(\pi_a\tilde{Z}_{ai} - \eta_{ai})\mathbf{1}[X_i \geq \xi_i] \tag{1.41}$$

$$\nu_{bi} = \Phi(\pi_b\tilde{Z}_{bi} - \eta_{bi}), \tag{1.42}$$

and where $\Phi(\cdot)$ denotes the normal CDF. That is, individuals choose the schooling level that gives them the highest latent utility, subject to the constraint that some may not be allowed to drop out of high school by virtue of being too young ($\mathbf{1}[X_i \leq \xi_i]$). To simulate the model, I let $(\tilde{Z}_{1i}, \tilde{Z}_{ai}, \tilde{Z}_{bi}) \sim N(\mu_Z, \Sigma_Z)$ and $(\epsilon_{1i}, \epsilon_{ai}, \epsilon_{bi}, \eta_{1i}, \eta_{ai}, \eta_{bi}) \sim N(0, \Sigma_{\epsilon\nu})$ where

$$\Sigma_Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$$\Sigma_{\epsilon\nu} = \begin{bmatrix} 0.64 & 0.16 & 0.16 & 0.024 & -0.32 & 0.016 \\ 0.16 & 1 & 0.2 & 0.02 & -0.3 & 0.01 \\ 0.16 & 0.2 & 1 & 0.02 & -0.4 & 0.04 \\ 0.024 & 0.02 & 0.02 & 1 & 0.6 & 0.1 \\ -0.32 & -0.3 & -0.4 & 0.6 & 1 & 0.2 \\ 0.016 & 0.01 & 0.04 & 0.1 & 0.2 & 1 \end{bmatrix},$$

and where $(\mu_1, \mu_a, \mu_b) = (0.3, 0.1, 0.7)$ and $(\pi_1, \pi_a, \pi_b) = (0.2, 0.3, 0.1)$. With $X_i = \xi_i = 0$, this model is the same as the one in Heckman and Urzúa (2010). To generate cross-strata first stage variation I let $\xi \sim N(0.5, 0.025)$ and draw X uniformly with probability 0.5. Setting $\gamma = 0.2$ allows an individual's cohort to affect the level of her adulthood wages, but not her relative returns to schooling. As in Heckman and Urzúa (2010), I apply Proposition 1.1 with an instrument Z_i that represents an exogenous increase in $\tilde{Z}_{1i}$ by 0.75 standard deviations.

Chapter 2

Leveraging Lotteries for School Value-added: Testing and Estimation

(with Joshua Angrist, Parag Pathak, and Christopher Walters)

2.1 Introduction

Public school districts increasingly use value-added models (VAMs) to assess teacher and school effectiveness. Conventional VAM estimates compare test scores across classrooms or schools after regression-adjusting for students' demographic characteristics and earlier scores. Achievement differences remaining after adjustment are attributed to differences in teacher or school quality.

A version of this chapter appeared in the May 2017 issue of *The Quarterly Journal of Economics*, published by Oxford University Press; for permission to reuse its figures, please contact journals.permissions@oup.com. We gratefully acknowledge funding from the National Science Foundation, the Laura and John Arnold Foundation, and the Spencer Foundation, and are indebted to SEII research managers Annice Correia and Eryn Heying for invaluable help and support. Thanks also go to Isaiah Andrews, Guido Imbens, Pat Kline, Rick Mansfield, Chris Nielson, Stephen Raudenbush, Jesse Rothstein, Doug Staiger, and seminar participants at the 2014 All California Labor Economics Conference, the APPAM Fall 2014 research conference, the 2014 and 2017 AEFP meeting, the 2015, 2016, and 2017 ASSA annual meetings, the 2015 SOLE/EALE annual meeting, the 2015 NBER Summer Institute, the Federal Reserve Bank of New York, the 2015 Becker/Friedman Applied Microeconomics Conference, the University of Chicago Committee on Education Workshop, Brown University, the University of Chicago Workshop on Quantitative Research Methods, UC Merced, the University of Arizona, Northwestern University, and the SREE Spring 2017 Conference for suggestions and comments.

Some districts use estimates of teacher value-added to guide personnel decisions, while others use VAMs to generate "report cards" that allow parents to compare schools.[1] Value-added estimation is a high-stakes statistical exercise: low VAM estimates can lead to school closures and teacher dismissals, while a growing body of evidence suggests the near-term achievement gains produced by effective teachers and schools translate into improved outcomes in adulthood (see, e.g., Chetty et al. (2011) and Chetty et al. (2014b) for teachers and Angrist et al. (2016a) and Dobbie and Fryer (2015) for schools).

Because the stakes are so high, the use of VAM estimates for teacher and school assessment remains controversial. Critics note that VAM estimates may be misleading if the available control variables are inadequate to ensure *ceteris paribus* comparisons. VAM estimates may also reflect considerable sampling error. The accuracy of teacher value-added models is the focus of a large and expanding body of research. This work demonstrates that teacher VAM estimates have predictive value, but has yet to generate a consensus on the substantive importance of bias or guidelines for "best practice" VAM estimation (Kane and Staiger, 2008; Rothstein, 2010; Koedel and Betts, 2011; Kinsler, 2012; Kane et al., 2013; Chetty et al., 2014a; Chetty et al., 2016; Rothstein, forthcoming; Chetty et al., forthcoming). While the social significance of school-level VAMs is similar to that of teacher VAMs, validation of VAMs for schools has received less attention.

The proliferation of partially-randomized urban school assignment systems provides a new tool for measuring school value-added. Centralized assignment mechanisms based on the theory of market design, including those used in Boston, Chicago, Denver, New Orleans, and New York, use information on parents' preferences over schools and schools' priorities over students to allocate scarce admission offers. These matching algorithms typically employ random sequence numbers to distinguish between students with the same priorities, thereby creating stratified student assignment lotteries. Similarly, independently-run charter schools often use admissions lotteries when oversubscribed. Scholars increasingly use these lotteries to identify causal effects of enrollment in various school sectors, including charter schools, pilot schools, small high schools, and magnet schools (Cullen et al., 2006; Hastings and Weinstein, 2008; Abdulkadiroğlu et al., 2011; Angrist et al., 2013; Bloom and Unterman, 2014; Deming, 2014). Lottery-based estimation of individual

[1]The Education Commission of the States noted that, as of 2016, fourteen states – Alabama, Arizona, Florida, Indiana, Louisiana, Maine, Mississippi, New Mexico, North Carolina, Ohio, Oklahoma, Texas, Utah, and Virginia – were issuing letter-grade report cards with grades determined at least in part by adjusted standardized test scores (`http://www.ecs.org/html/educationissues/accountability/stacc_intro.asp`).

school value-added is less common, however, reflecting the fact that lottery samples for many schools are small, while other schools are under-subscribed.

This chapter develops econometric methods that leverage school admissions lotteries for VAM testing and estimation, accounting for the partial coverage of lottery data. Our first contribution is the formulation of a new lottery-based test of conventional VAMs. This test builds on recent experimental and quasi-experimental VAM validation strategies, including the work of Kane and Staiger (2008), Deutsch (2013), Kane et al. (2013), Chetty et al. (2014a) and Deming (2014). In contrast with earlier studies, which implicitly look at average-across-schools validity in a test with one degree of freedom, ours is an over-identification test that looks at each of the orthogonality restrictions generated by a set of lottery instruments. Intuitively, the test developed here asks whether conventional VAM estimates correctly predict the effect of randomized admission at every school that has a lottery, as well as predicting an overall average effect. Our test of VAM validity parallels the classical over-identification test, since the latter can be described either as testing instrument-error orthogonality or as a comparison of alternative just-identified IV estimates that should be the same under the null hypothesis.[2]

Application of this test to data from Boston reveals moderate but statistically significant bias in conventional VAM estimates. This finding notwithstanding, conventional VAM estimates may nevertheless provide a useful guide to school quality if the degree of bias is modest. To assess the practical value of VAM estimates, we develop and estimate a hierarchical random coefficients model that describes the joint distribution of value-added, VAM bias, and lottery compliance across schools. The model is estimated via a simulated minimum distance procedure that matches moments of the distribution of conventional VAM estimates, lottery reduced forms, and first stages to those predicted by the random coefficients structure. Estimates of the model indicate substantial variation in both causal value-added and selection bias across schools. Nevertheless, the estimated joint distribution of these parameters implies that conventional VAM estimates are highly correlated with school effectiveness.

A second contribution of our study is to use the random coefficients framework and lottery variation to improve conventional VAM estimates. Our approach builds on previous estimation strategies that trade reduced bias for increased variance (Morris, 1983; Judge and Mittlehammer, 2004; Mittlehammer and Judge, 2005; Judge and Mittlehammer, 2007). Specifically, we compute

[2]The theory behind VAM over-identification testing is sketched in Angrist et al. (2016b).

empirical Bayes (EB) hybrid posterior predictions that optimally combine relatively imprecise but unbiased lottery-based estimates with biased but relatively precise conventional VAM estimates. Importantly, our approach makes efficient use of the available lottery information without requiring a lottery for every school. Hybrid estimates for under-subscribed schools are improved by information on the *distribution* of bias contributed by schools with oversubscribed lotteries. The hybrid estimation procedure generates estimates that, while still biased, have lower mean squared error than conventional VAM estimates. Our framework provides a general recipe for combining non-experimental and quasi-experimental estimators and may therefore be useful in other settings.[3]

Finally, we quantify the consequences of bias in conventional VAM estimates and the payoff to hybrid estimation using a Monte Carlo simulation calibrated to our Boston estimates. Simulation results show that policy decisions based on conventional estimates that control for baseline test scores or measure score growth are likely to boost achievement. For example, replacing the lowest-ranked Boston school with an average school is predicted to generate a gain of 0.24 test score standard deviations (σ) for affected students, roughly two-thirds of the benefit obtained when true value-added is used to rank schools (0.37σ). Hybrid estimates are highly correlated with conventional estimates (the rank correlation is 0.74), and hybrid estimation generates modest additional gains, reducing mean squared error by 30% and increasing the benefits of school closure policies by 0.08σ (33%). Conventional school VAMs would therefore appear to provide a useful guide for policy-makers, while hybrid estimators generate worthwhile improvements in policy targeting.

The next section describes the Boston data used for VAM testing and estimation, and Section 2.3 describes the conventional value-added framework as applied to these data. Section 2.4 derives our VAM validation test and discusses test implementation and results. Section 2.5 outlines the random coefficients model and empirical Bayes approach to hybrid estimation, while Section 2.6 reports estimates of the model's hyperparameters and the resulting posterior predictions of value-added. Section 2.7 discusses policy simulations. Finally, Section 2.8 concludes with remarks on how the framework developed here might be used in other settings.

[3]These settings include the analysis of teacher, hospital, doctor, firm, and neighborhood effects, as in Chetty, Friedman, and Rockoff (2014a, 2014b), Chandra et al. (2015), Fletcher et al. (2014), Card et al. (2013), and Chetty and Hendren (2016). Chetty and Hendren combine observational and quasi-experimental estimates of neighborhood effects using a procedure discussed in Section 2.5, below.

2.2 Setting and Data

2.2.1 Boston Public Schools

Boston public school students can choose from a diverse set of enrollment options, including traditional Boston Public School (BPS) district schools, charter schools, and pilot schools. As in most districts, Boston's charter schools are publicly funded but free to operate within the confines of their charters. For the most part, charter staff are not covered by collective bargaining agreements or other BPS regulations.[4] Boston's pilot school sector arose as a union-supported alternative to charter schools, developed jointly by the BPS district and the Boston Teachers Union. Pilot schools are part of the district but typically have more control over their budgets, scheduling, and curriculum than do traditional public schools. On the other hand, pilot school teachers work under collective bargaining provisions similar to those in force at traditional public schools.

Applicants to traditional public and pilot schools rank between three and ten schools as the first step in a centralized match (students not finishing elementary or middle school who are happy to stay where they are need not participate in the match). Applicants are then assigned to schools via a student-proposing deferred acceptance mechanism, as described in Abdulkadiroğlu et al. (2006). This mechanism combines student preferences with a strict priority ranking over students for each school. Priorities are determined by whether an applicant is already enrolled at the school and therefore guaranteed admission, has a sibling enrolled at the school, or lives in the school's walk-zone. Ties within these coarse priority groups are broken by random sequence numbers, which we refer to as lottery numbers. In an evaluation of the pilot sector exploiting this centralized random assignment scheme, Abdulkadiroğlu et al. (2011) find mostly small and statistically insignificant effects of pilot school attendance relative to the traditional public school sector.

In contrast with the centralized match that assigns seats at traditional and pilot schools, charter applicants apply to individual charter schools separately before the fall of the school year in which they hope to enter. By Massachusetts law, oversubscribed charter schools must select students through public admissions lotteries, with the exception of applicants with siblings already enrolled in the charter, who are guaranteed seats. Charter offers and centralized assignment offers are made independently; students applying to the charter sector can receive multiple offers. In practice,

[4]Boston's charter sector includes both "Commonwealth" charters, which are authorized by the state and operate as independent school districts, and "in-district" charters, which are authorized and overseen by the Boston School Committee.

some Boston charter schools offer all of their applicants seats, while others fail to retain complete information on past admissions lotteries. Studies based on charter lotteries show that Boston charter schools boost test scores and increase four-year college attendance (see, for example, Abdulkadiroğlu et al. (2011) and Angrist et al. (2016a)).

2.2.2 Data and Descriptive Statistics

The data analyzed here consist of a sample of roughly 28,000 sixth-grade students attending 51 Boston traditional, pilot, and charter schools in the 2006/7 through 2013/14 school years. In Boston, sixth grade marks the first grade of middle school, so most rising sixth graders participate in the centralized match.

Baseline test scores come from fifth grade Massachusetts Comprehensive Assessment System (MCAS) tests in math and English Language Arts (ELA), while outcomes are measured in sixth, seventh and eighth grades. Test scores are standardized to have mean zero and unit variance in the population of Boston charter, pilot, and traditional public schools, separately by subject, grade, and year. Other variables used in the empirical analysis include school enrollment, race, sex, subsidized lunch eligibility, special education status, English-language learner status, and suspensions and absences. The data appendix describes the administrative files used to construct the working extract.

Our analysis combines data from the centralized traditional and pilot match with lottery data from individual charter schools. The BPS lottery instruments code offers at applicants' first choice (highest ranked) middle schools in the match. In particular, BPS lottery offers indicate applicants whose lottery numbers are below the highest (worst) number offered a seat at their first-choice school, among those in the same priority group. Conditional on application year, first-choice school, and an applicant's priority at that school (what we call the assignment strata), offers of seats at a first choice are randomly assigned. Charter lottery instruments indicate offers made on the night of the admissions lottery at each charter school. These offers are randomly assigned for non-siblings conditional on the target school and application year.[5]

The schools and students analyzed here are described in Table 2.1. We exclude schools serving

[5]For a much smaller group of applicants, the centralized BPS mechanism induces random tie-breaking for lower-ranked school choices. The use of tie-breaking from these choices generates complications beyond the scope of this chapter; see Abdulkadiroğlu et al. (forthcoming) for a comprehensive analysis of empirical strategies that exploit centralized assignment.

fewer than 25 sixth graders in each year, leaving a total of 25 traditional public schools, 9 pilot schools, and 17 charter schools. Of these, 37 schools have sixth grade as a primary entry point and 28 (16 traditional, 7 pilot, and 5 charter) had at least 50 students subject to random sixth grade assignment. Applicants to these 28 schools constitute our lottery sample. Conventional ordinary least squares (OLS) value-added models are estimated in a sample of 27,864 Boston sixth graders with complete baseline, demographic, and outcome information; 8,718 of these students are also in the lottery sample.

About 77% of Boston sixth graders enroll at schools with usable lotteries, and, as can be seen in the descriptive statistics reported in Table 2.2, demographic characteristics for this group are comparable to those of the full BPS population. Columns 3 and 4 of Table 2.2 report characteristics of the subset of students subject to randomized lottery assignment. Lotteried students are slightly more likely to be African American and to qualify for a subsidized lunch, and somewhat less likely to be white or to have been suspended or absent in fifth grade. Table 2.2 also documents the comparability of students offered not offered seats in a lottery. These results, reported in columns 5-7, compare the baseline characteristics of lottery winners and losers, controlling for assignment strata. Consistent with conditional random assignment of offers, estimated differences by offer status are small and not significantly different from zero, both overall and within school sectors.[6]

2.3 Value-added Framework

As in earlier investigations of school value-added, the analysis here builds on a constant-effects causal model. This reflects a basic premise of the VAM framework: internally valid treatment effects from earlier years and cohorts are presumed to have predictive value for future cohorts. Student i's potential test score at school j, denoted Y_{ij}, is therefore written as the sum of two non-interacting components, specifically:

$$Y_{ij} = \mu_j + a_i, \tag{2.1}$$

[6]Lottery estimates may be biased by selective sample attrition. As shown in Appendix Table 2.A1, follow-up data are available for 81% of lottery applicants, while sample retention is 2.8 percentage points higher for lottery winners than for losers, a difference driven by traditional public school lotteries. Table 2.2 shows that that baseline characteristics are balanced in the sample with follow-up scores, so the modest differential attrition documented in Table 2.A1 seems unlikely to affect the results reported here.

where μ_j is the mean potential outcome at school j and a_i is student i's "ability," or latent achievement potential. This additively-separable model implies that causal effects are the same for all students. The constant effects framework focuses attention on the possibility of selection bias in VAM estimates rather than treatment effect heterogeneity (though we explore heterogeneity as well).

A dummy variable, D_{ij}, is used to indicate whether student i attended school j in sixth grade. The observed sixth-grade outcome for student i can therefore be written

$$Y_i = Y_{i0} + \sum_{j=1}^{J} (Y_{ij} - Y_{i0}) D_{ij}$$

$$= \mu_0 + \sum_{j=1}^{J} \beta_j D_{ij} + a_i. \tag{2.2}$$

The parameter $\beta_j \equiv \mu_j - \mu_0$ measures the causal effect of school j relative to an omitted reference school with index value 0. In other words, β_j is school j's value-added.

Conventional value-added models use regression methods to mitigate selection bias. Write

$$a_i = X_i'\gamma + \epsilon_i, \tag{2.3}$$

for the regression of a_i on a vector of controls, X_i, which includes lagged test scores. Note that $E[X_i\epsilon_i] = 0$ by definition of γ. This decomposition implies that observed outcomes can be written

$$Y_i = \mu_0 + \sum_{j=1}^{J} \beta_j D_{ij} + X_i'\gamma + \epsilon_i. \tag{2.4}$$

It bears emphasizing that equation (2.4) is a causal model: ϵ_i is defined so as to be orthogonal to X_i, but need not be uncorrelated with the school attendance indicators, D_{ij}.

We are interested in how OLS regression estimates compare with the causal parameters in equation (2.4). We therefore define population regression coefficients in a model with the same conditioning variables:

$$Y_i = \alpha_0 + \sum_{j=1}^{J} \alpha_j D_{ij} + X_i'\Gamma + v_i. \tag{2.5}$$

This is a population projection, so the residuals in this model, v_i, are necessarily orthogonal to all

right-hand-side variables, including the school attendance dummies.

Regression model (2.5) has a causal interpretation when the parameters in this equation coincide with those in the causal model, equation (2.4). This in turn requires that school choices be unrelated to the unobserved component of student ability, an assumption that can be expressed as:

$$E\left[\epsilon_i | D_{ij}\right] = 0; \; j = 1, ..., J. \tag{2.6}$$

Restriction (2.6), sometimes called "selection-on-observables," means that $\alpha_j = \beta_j$ for each school. In practice, of course, regression estimates need not have a causal interpretation; rather, they may be biased. This possibility is represented by writing

$$\alpha_j = \beta_j + b_j,$$

where school j's bias parameter b_j is the difference between its regression and causal parameter.

2.4 Validating Conventional VAMs

2.4.1 Test Procedure

The variation in school attendance generated by oversubscribed admission lotteries allows us to assess the causal interpretation of conventional VAM estimates. A vector of dummy variables, $Z_i = (Z_{i1}, .., Z_{iL})'$, indicates lottery offers to student i for seats at L oversubscribed schools. Offers at school ℓ are randomly assigned conditional on a set of lottery-specific stratifying variables, $C_{i\ell}$. These variables include an indicator for applicants to school ℓ and possibly other variables such as application cohort and walk-zone status. The vector $C_i = (C_{i1}', .., C_{iL}')'$ collects these variables for all lotteries. The models used here also add the OLS VAM controls (X_i in equation (2.5)) to the vector C_i to increase precision.

We assume that lottery offers are (conditionally) mean-independent of student ability. Thus,

$$E[\epsilon_i | C_i, Z_i] = \lambda_0 + C_i'\lambda_c, \tag{2.7}$$

for a set of parameters λ_0 and λ_c. This implies that admission offers are valid instruments for

school attendance after controlling for lottery assignment strata, an assumption that underlies recent lottery-based analyses of school effectiveness (Cullen et al., 2006; Abdulkadiroğlu et al., 2011; Deming et al., 2014).

With fewer lotteries than schools (that is, when $L < J$), the restrictions in (2.7) are insufficient to identify the parameters of the causal model, equation (2.4). Even so, these restrictions can be used to test the selection-on-observables assumption. Equations (2.6) and (2.7) imply that $L + J$ orthogonality conditions are available to identify J school effects, β_j. The resulting L overidentifying restrictions generate an over-identification test of the sort widely used with instrumental variables (IV) estimators.

To describe the over-identification test statistic, let Z denote the $N \times L$ matrix of lottery offers for a sample of N students, and let C denote the corresponding matrix of stratifying variables, with associated projection matrix $P_C = C(C'C)^{-1}C'$ and annihilator matrix $M_C = I - P_C$. The Lagrange multiplier (LM) over-identification test statistic associated with two-stage least squares (2SLS) models estimated assuming homoskedasticity can be written:

$$\hat{T} = \frac{\hat{\epsilon}' P_{\tilde{Z}} \hat{\epsilon}}{\hat{\sigma}^2_{\hat{\epsilon}}}, \tag{2.8}$$

where $P_{\tilde{Z}} = M_C Z(Z'M_C Z)^{-1}Z'M_C$ is the lottery offer projection matrix after partialling out randomization strata, $\hat{\epsilon}$ is an $N \times 1$ vector of OLS VAM residuals (since OLS and 2SLS coincide when the set of D_{ij} is in the instrument list), and $\hat{\sigma}^2_{\hat{\epsilon}} = \hat{\epsilon}' M_C \hat{\epsilon}/N$ is an estimate of the residual variance of ϵ_i partialling out strata effects. Under the joint null hypothesis described by selection-on-observables and lottery exclusion (equations 2.6 and 2.7), the statistic $\hat{T}$ has an asymptotic χ^2_L distribution.[7]

A simple decomposition of $\hat{T}$ reveals an important connection with classical overidentification tests and previously used VAM validity tests. Let $\hat{Y}_i$ denote the fitted values generated by OLS VAM estimation (computed from regression model (2.5)), and let Y and $\hat{Y}$ denote $N \times 1$ vectors

[7]The test statistic in equation (2.8) is derived assuming homoskedastic errors. An analogous test allowing heteroskedasticity uses a White (1980) robust covariance matrix to test the hypothesis that coefficients on lottery offers equal zero in a regression of ϵ_i on Z_i and C_i.

collecting individual Y_i and $\hat{Y}_i$. Our LM statistic can then be written

$$\begin{aligned}\hat{T} &= \frac{((Y-\hat{\varphi}\hat{Y})+(\hat{\varphi}-1)\hat{Y})'P_{\tilde{Z}}((Y-\hat{\varphi}\hat{Y})+(\hat{\varphi}-1)\hat{Y})}{\hat{\sigma}^2_{\tilde{\epsilon}}} \\ &= \frac{(\hat{\varphi}-1)^2}{\hat{\sigma}^2_{\tilde{\epsilon}}(\hat{Y}'P_{\tilde{Z}}\hat{Y})^{-1}} + \frac{(Y-\hat{\varphi}\hat{Y})'P_{\tilde{Z}}(Y-\hat{\varphi}\hat{Y})}{\hat{\sigma}^2_{\tilde{\epsilon}}}. \end{aligned} \tag{2.9}$$

Here, the scalar $\hat{\varphi} = (\hat{Y}'P_{\tilde{Z}}\hat{Y})^{-1}\hat{Y}'P_{\tilde{Z}}Y$ is the 2SLS estimate from a model that uses lottery offers as instruments in an equation with Y_i on the left and $\hat{Y}_i$, treated as endogenous, on the right. Equation (2.9) shows that the omnibus test statistic $\hat{T}$ combines two terms. The first is a one-degree-of-freedom Wald-type test statistic for $\hat{\varphi} = 1$ (note that the denominator of this term estimates the asymptotic variance of $\hat{\varphi}$). The second is the Sargan (1958) statistic for testing the $L-1$ overidentifying restrictions generated by the availability of L instruments to estimate $\hat{\varphi}$.[8]

In what follows, the estimate $\hat{\varphi}$ is called a "forecast coefficient." This connects $\hat{T}$ with tests of "forecast bias" implemented in previous VAM validation efforts (Kane and Staiger, 2008; Chetty et al., 2014a). These earlier tests similarly ask whether the coefficient on predicted value-added equals one in IV procedures relating outcomes to VAM fitted values (though the details sometimes differ). Forecast bias arises when VAM estimates for a group of schools are off the mark, a failure of average predictive validity. Importantly, the omnibus test statistic, $\hat{T}$, checks more than forecast bias: this statistic asks whether *each* oversubscribed lottery generates score gains commensurate with the gains predicted by an OLS VAM.

2.4.2 Test Results

The conventional VAM setup assessed here includes two value-added specifications. The first, referred to as the "lagged score" model, includes indicators for sex, race, subsidized lunch eligibility, special education status, English-language learner status, and counts of baseline absences and suspensions, along with cubic functions of baseline math and ELA test scores. Specifications of this type are at the heart of the econometric literature on value-added models (Kane et al., 2008; Rothstein, 2010; Chetty et al., 2014a). The second, a "gains" specification, uses grade-to-grade score changes as the outcome variable and includes all controls from the lagged score model except base-

[8] Angrist et al. (2016b) interpret VAM validity tests using the moment-based theory of specification testing developed by Newey (1985) and Newey and West (1987). In practice, Wald and LM test statistics typically use different variance estimators in the denominator.

line test scores. This model is motivated by widely-used accountability policies that measure test score growth.[9] As in Rothstein (2009), we benchmark the extent of cross-school ability differences using an "uncontrolled" model that adjusts only for year effects. Although the uncontrolled model almost certainly provides a poor measure of school value-added, many districts distribute school report cards based on unadjusted test score levels.[10]

Figure 2.1 summarizes the value-added estimates generated by sixth-grade math scores. We focus on math scores because value-added for math appears to be more variable across schools than value-added for ELA (bias tests for ELA, presented in Appendix Table A.II, yield similar results). Each bar in Figure 2.1 reports an estimated standard deviation of α_j across schools, expressed in test score standard deviation units and adjusted for estimation error.[11] Adding controls for demographic variables and previous scores reduces the standard deviation of α_j from 0.5σ in the uncontrolled model to about 0.2σ in the lagged score and gains models. This shows that observed student characteristics explain a substantial portion of the variation in school averages. The last three bars in Figure 2.1 report standard deviations of within-sector value-added, constructed using residuals from regressions of $\hat{\alpha}_j$ on dummies for schools in the charter and pilot sectors. Controlling for sector effects reduces variation in α_j, reflecting sizable differences in average conventional value-added across sectors.

Table 2.3 summarizes test results for sixth grade math VAMs in Panel A. The first row shows the forecast coefficient, $\hat{\varphi}$. The estimator used here is the optimal IV procedure for heteroskedastic models described by White (1982). The second row reports first stage F-statistics measuring the strength of the relationship between lottery offers and predicted value-added. With a weak first stage, forecast coefficient estimates may be biased towards the corresponding OLS estimand, that is, the coefficient from a regression of test scores on VAM fitted values. In simple models, this regression coefficient must equal one, so a weak first stage makes a test of the forecast coefficient

[9]The gains specification can be given a theoretical foundation as follows: suppose that human capital in grade g, denoted A_{ig}, equals lagged human capital plus school quality, so that $A_{ig} = A_{ig-1} + q_{ig}$ where $q_{ig} = \sum_j \beta_j D_{ij} + \eta_{ig}$ and η_{ig} is a random component independent of school choice. Suppose further that test scores are noisy proxies for human capital, so that $Y_{ig} = A_{ig} + \nu_{ig}$ where ν_{ig} is classical measurement error. Finally, suppose that school choice in grade g is determined solely by A_{ig-1} and variables unrelated to achievement. Then a lagged score model that controls for Y_{ig-1} generates biased estimates, but a gains model with $Y_{ig} - Y_{ig-1}$ as the outcome variable measures value-added correctly.

[10]For example, California's School Accountability Report Cards list school proficiency levels (see `http://www.sarconline.org`), while Massachusetts' school and district profiles provide information on proficiency levels and test score growth (see `http://profiles.doe.mass.edu`).

[11]The estimated standard deviations plotted in the figure are given by $\hat{\sigma}_\alpha = (\frac{1}{J}\sum_j[(\hat{\alpha}_j - \hat{\mu}_\alpha)^2 - SE(\hat{\alpha}_j)^2])^{1/2}$, where $\hat{\mu}_\alpha$ is mean value-added and $SE(\hat{\alpha}_j)$ is the standard error of $\hat{\alpha}_j$.

equaling one less likely to reject.[12] First-stage F-statistics for the sixth grade lagged score and gains models are close to 30, suggesting finite-sample bias is not an issue in the full lottery sample. First-stage strength is more marginal, however, when charter lotteries are omitted.

Table 2.3 also reports p-values for three VAM validity tests. The first is for forecast bias, that is, the null hypothesis that the forecast coefficient equals one. The second tests the associated set of overidentifying restrictions, which require that just-identified IV estimates of the forecast coefficient be the same for each lottery instrument, though not necessarily equal to one. The third "omnibus test" combines these restrictions.

On average, VAM fitted values predict the score gains generated by random assignment remarkably well. This can be seen in columns 1 and 2 of Table 2.3, which show that the lagged score and gains specifications generate forecast coefficients for sixth graders equal to 0.86 and 0.95; the former is only marginally statistically different from one ($p = 0.07$), while the second has $p = 0.55$. At the same time, the over-identification and omnibus tests reject for both models.[13]

The source of these rejections can be seen in Figure 2.2, which plots reduced form estimates of the effects of lottery offers on test scores against corresponding first-stage effects of lottery offers on conventional VAM fitted values for sixth grade math. Each panel also shows a line through the origin with slope equal to the forecast coefficient reported in Table 2.3 (plotted as a solid line), along with a dashed 45-degree line. In other words, Figure 2.2 gives a visual representation of the forecast coefficient: VAM models that satisfy equation (2.6) should generate points along the 45-degree line, with deviations due solely to sampling error. Though the lines of best fit have slopes close to one, points for many lotteries are farther from the diagonal than sampling variance alone would lead us to expect. Earlier validation strategies focus on forecast coefficients, ignoring overidentifying restrictions. Figure 2.2 shows that such strategies may fail to detect substantial deviations between conventional VAM predictions and reduced form lottery effects for individual lotteries.

Figure 2.2 also suggests that a good portion of conventional VAM estimates' predictive power for Boston schools comes from charter school lotteries, which contribute large first stage and reduced form effects. The relationship between OLS value-added and lottery estimates is weaker in the

[12] When estimated in the same sample with no additional controls, OLS regressions on OLS fitted values necessarily produce coefficients of one. In practice, the specification used here to test VAM differs from the model producing fitted values in that it also controls for lottery strata and excludes non-lotteried students.

[13] As a point of comparison, Angrist et al. (2016b) report tests of VAM validity in the Charlotte-Mecklenberg lottery data analyzed by Deming (2014). There, as well, the forecast coefficient is close to one, while the omnibus test generates a p-value of 0.02.

traditional public and pilot school sectors. This is confirmed in columns 3 and 4 of Table 2.3, which report results of VAM bias tests for sets of instruments that exclude charter lotteries. At 0.55 and 0.68, estimated forecast coefficients from traditional public and pilot lotteries are farther from one than the coefficients computed using all lotteries. Although removal of charter lotteries reduces precision, omnibus tests computed without them also reject at the 1-percent level.[14]

Finally, Panel B of Table 2.3 reports test results combining data from sixth through eighth grade. As in Abdulkadiroğlu et al. (2011) and Dobbie and Fryer (2013), school effects on seventh and eighth grade scores are modeled as linear in the number of years spent in each school. In a linear constant effects framework, regressions of test score outcomes on baseline controls and years of enrollment in each school recover causal school effects in the absence of sorting on unobserved ability. The omnibus VAM validity test in this case regresses residuals from the multi-grade (stacked) model on sixth grade lottery offers, while the forecast coefficient is generated by using lottery offers to instrument OLS VAM fitted values from the multi-grade model. The omnibus test results show clear rejections in the multi-grade set-up as well as for sixth grade only, in spite of the fact that the first-stage F-statistics here are noticeably lower.

2.4.3 Heterogeneity vs. Bias

The omnibus test results reported in Table 2.3 suggest conventional VAM estimates fail to predict the effects of lottery offers perfectly. This is consistent with bias in OLS VAMs. In a world of heterogeneous causal effects, however, these rejections need not reflect selection bias. Rather, these results might signal divergence between the local average treatment effects (LATEs) identified by lottery instruments and possibly more representative effects captured by OLS (Imbens and Angrist, 1994; Angrist et al., 1996). Moreover, with unrestricted potential outcomes, even internally valid OLS VAM estimates (that is, those satisfying selection-on-observables) capture weighted average causal effects that need not match average effects for the entire sample of students attending particular schools (Angrist, 1998).

Three analyses shed light on the distinction between heterogeneity and bias. The first is a set of

[14]The first stage F-statistics for the specifications without charter lotteries are 11.2 and 9.3, suggesting weak instruments might be a problem in thse models. It is encouraging, therefore, that limited information maximum likelihood (LIML) forecast coefficient estimates are virtually the same as the estimates reported in Table 2.3. A related concern is whether the heteroskedastic-robust standard errors and test statistics used in Table 2.3 are misleading due to common school-year shocks (as suggested by Kane and Staiger (2002) for teachers). Reassuringly, cluster-robust test results are also similar to those in Table 2.3.

bias tests using OLS VAM specifications that allow school effects to differ across covariate-defined subsamples (e.g. special education students or those with low levels of baseline achievement). This approach accounts for variation in school effects across covariate cells that may be weighted differently by IV and OLS. The second analysis tests for bias in OLS VAMs estimated in the lottery sample. This asks whether differences between IV and OLS are caused by differences between students subject to lottery assignment and the general student population. The final analysis estimates OLS VAM separately for applicants who respond to lottery offers ("compliers") and for other groups in the sample of lottery applicants.

Estimates by subgroup, reported in Panel A of Table 2.4 for the OLS VAM sample, consistently generate rejections in omnibus tests of VAM validity. Column 2 shows test results computed using models that allow VAM estimates to differ by year, thereby accommodating "drift" in school effects over time (Chetty et al. (2014a) document such drift in teacher value-added); columns 3-5 show results for subgroups defined by subsidized lunch eligibility, special education status, and baseline test score terciles; and column 6 reports results from models that allow value-added to differ across cells constructed by fully interacting race, sex, subsidized lunch eligibility, special education, English-language learner status, and baseline score tercile. The forecast coefficients and omnibus test statistics generated by each of these subgroup schemes are similar to those for the full sample. Moreover, as can be seen in Panel B of Table IV, test results for models that use only the lottery sample for OLS VAM estimation are also similar to the full sample results. This suggests that rejection of the omnibus test is not driven by differences in OLS VAM between students subject to random assignment and the general population.[15]

Lottery-based IV estimates identify average causal effects for compliers, that is, for lottery applicants whose attendance choices shift in response to random offers, rather than for the full population of students that enroll in a particular school. To investigate the link between lottery compliance and treatment effects, we predict value-added at the target school for individual lottery applicant using covariate-specific OLS estimates from the model in column 6 of Table 2.4 (estimated in the lottery sample). Maintaining the hypothesis of OLS VAM validity, we allow for the possibility

[15] In a subset of the data used here, Walters (2014) documents a link between the propensity to apply to Boston charter schools and the causal effect of charter school attendance. This finding is not at odds with our constant effects assumption because Walters studies the effects of charter schools relative to a heterogeneous mix of traditional public schools, while we allow a distinct effect for every traditional public school. The effect heterogeneity uncovered by Walters may reflect variation in the quality of fallback public school options across charter applicants. Consistent with this possibility, Walters demonstrates that the relationship between charter application choices and causal effects is driven primarily by heterogeneity in outcomes at fallback traditional public schools.

that heterogeneous effects are reflected in a set of covariate-specific estimates. These predictions are then used to compare imputed average value-added for compliers to imputed average value-added for "never-takers" (those who decline lottery offers) and "always-takers" (those who enroll in the target school even when denied an offer) in each lottery. Averages for the three lottery compliance groups are estimated using methods described in the econometric appendix.

Figure 2.3 shows that imputed OLS value-added estimates for compliers, always-takers, and never-takers are similar. Formal tests for equality fail to reject the hypotheses that predicted effects for compliers equal predicted effects for always-takers ($p = 0.80$) or never-takers ($p = 0.33$). This suggests that lottery compliance is not a major source of treatment effect heterogeneity, though we cannot rule out unobserved differences between compliers and other groups.

2.5 The Distribution of School Effectiveness

The test results in Table 2.3 suggest conventional VAM estimates are biased. At the same time, OLS VAM estimates tend to predict lottery effects on average, with estimated forecast coefficients close to one. OLS estimates would therefore seem to be useful even if imperfect. This section develops a hybrid estimation strategy that combines lottery and OLS estimates in an effort to quantify the bias in conventional VAMs and produce more accurate value-added estimates.

2.5.1 A Random Coefficients Lottery Model

The hybrid estimation strategy uses a random coefficients model to describe the joint distribution of value-added, bias, and lottery compliance across schools. The model is built on a set of OLS, lottery reduced form, and first stage estimates. The OLS estimates come from equation (2.5), while the lottery reduced form and first stage equations are:

$$Y_i = \tau_0 + C_i'\tau_c + Z_i'\rho + u_i, \tag{2.10}$$

$$D_{ij} = \phi_{0j} + C_i'\phi_{cj} + Z_i'\pi_j + \eta_{ij};\ j = 1, ..., J.$$

Note that Z_i is the vector of all lottery admission offers, $Z_{i\ell}$ for $\ell = 1, \ldots, L$. Assumption (2.7) implies that the reduced form effect of admission in lottery ℓ is given by

$$\rho_\ell = \sum_{j=1}^{J} \pi_{\ell j} \beta_j,$$

where ρ_ℓ and $\pi_{\ell j}$ are the elements of ρ and π_j corresponding to $Z_{i\ell}$. This expression shows that the lottery at school ℓ identifies a linear combination of value-added parameters, with coefficients $\pi_{\ell j}$ equal to the shares of students shifted into or out of each school by the ℓth lottery offer.

OLS VAM, lottery reduced form, and lottery first stage estimates are modeled as noisy measures of school-specific parameters, which are in turn modeled as draws from a distribution of random coefficients in a larger population of schools. Specifically, we have:

$$\begin{aligned} \hat{\alpha}_j &= \beta_j + b_j + e_j^{\alpha}, \\ \hat{\rho}_\ell &= \sum_j \pi_{\ell j} \beta_j + e_\ell^{\rho}, \\ \hat{\pi}_{\ell j} &= \pi_{\ell j} + e_{\ell j}^{\pi}; \end{aligned} \tag{2.11}$$

where $e_j^{\alpha}, e_\ell^{\rho}$ and $e_{\ell j}^{\pi}$ are mean-zero estimation errors that vanish as the sample for each school and lottery tends to infinity. Subject to the usual asymptotic approximations, these errors are normally distributed with a known covariance structure. Table 2.1 shows that the OLS and lottery estimation samples used here typically include hundreds of students per school, so the use of asymptotic results seems justified.

The second level of the model treats the school-specific parameters β_j, b_j, and $\{\pi_{\ell j}\}_{\ell=1}^{L}$ as draws from a joint distribution of causal effects, bias, and lottery compliance behavior. The effect of admission at school ℓ on the probability of attending this school is parameterized as

$$\pi_{\ell\ell} = \frac{\exp(\delta_\ell)}{1 + \exp(\delta_\ell)}, \tag{2.12}$$

where the parameter δ_ℓ can be viewed as the mean utility in a binary logit model predicting student compliance with a random offer of a seat at school ℓ. Likewise, the effect of an offer to attend school

$\ell \neq j$ on attendance at school j is modeled as

$$\pi_{\ell j} = -\pi_{\ell\ell} \times \frac{\exp(\xi_j + \nu_{\ell j})}{1 + \sum_{k \neq \ell} \exp(\xi_k + \nu_{\ell k})}. \tag{2.13}$$

In this expression, the quantity $\xi_j + \nu_{\ell j}$ is the mean utility for school j in a multinomial logit model predicting alternative school choices among students that comply with offers made in lottery ℓ. The parameter ξ_j allows for the possibility that some schools are systematically more or less likely to serve as fallback options for lottery losers, while $\nu_{\ell j}$ is a random utility shock specific to school j in the lottery at school ℓ. The parametrization in (2.12) and (2.13) ensures that lottery offers increase the probability of enrollment at the target school and reduce enrollment probabilities at other schools, and that effects on all probabilities are between zero and one in absolute value.

Each school is characterized by a vector of four parameters: a value-added coefficient, β_j; a selection bias term, b_j; an offer compliance utility, δ_j; and a mean fallback utility, ξ_j. These are modeled as draws from a prior distribution in a hierarchical Bayesian framework. A key assumption in this framework is that the distribution of VAM bias is the same for schools with and without oversubscribed lotteries. This assumption allows the model to "borrow" information from schools with lotteries and generate posterior predictions for non-lottery schools that account for bias in conventional VAM estimates. Importantly, however, we allow for the possibility that average value-added may differ between schools with and without lotteries. Section 2.6.2 investigates the empirical relationship between oversubscription and bias.

Let Q_j denote an indicator for whether quasi-experimental lottery data are available for school j. School-specific parameters are modeled as draws from a conditional multivariate normal distribution:

$$(\beta_j, b_j, \delta_j, \xi_j)' | Q_j \sim N\left((\beta_0 + \beta_Q Q_j, b_0, \delta_0, \xi_0)', \Sigma\right). \tag{2.14}$$

The parameter β_Q captures the possibility that average value-added differs for schools with lotteries. The matrix Σ describes the variances and covariances of value-added, bias, and first stage utility parameters, and is assumed to be the same for lottery and non-lottery schools. Finally, lottery and school-specific utility shocks are also modeled as conditionally normal:

$$\nu_{\ell j} | Q_j \sim N\left(0, \sigma_\nu^2\right). \tag{2.15}$$

The vector $\theta = (\beta_0, \beta_Q, b_0, \delta_0, \xi_0, vec(\Sigma)', \sigma_\nu^2)'$ collects the hyperparameters governing the prior distribution of school-specific parameters. Our empirical Bayes framework first estimates these hyperparameters and then uses the estimated prior distribution to compute posterior value-added predictions for individual schools. Some of the specifications considered below extend the setup outlined here to allow the prior mean vector $(\beta_0, b_0, \delta_0, \xi_0)$ to vary across Boston's school sectors (traditional, charter, and pilot).

2.5.2 Simulated Minimum Distance Estimation

We estimate hyperparameters by simulated minimum distance (SMD), a variant of the method of simulated moments (McFadden, 1989). SMD focuses on moments that are determined by the parameters of interest, choosing hyperparameters to minimize deviations between sample moments and the corresponding model-based predictions. Our SMD implementation uses means, variances, and covariances of functions of the OLS value-added estimates, $\hat{\alpha}_j$, lottery reduced forms, $\hat{\rho}_\ell$, and first stage coefficients, $\hat{\pi}_{\ell j}$. For example, one moment to be fit is the average $\hat{\alpha}_j$ across schools; another is the cross-school variance of the $\hat{\alpha}_j$. Other moments are means and variances of reduced form and first stage estimates across lotteries. The econometric appendix lists the full set of moments used for SMD estimation.

The fact that the moments in this context are complicated nonlinear functions of the hyperparameters motivates a simulation approach. For example, the mean reduced form is $E[\rho_\ell] = \sum_j E[\pi_{\ell j}\beta_j]$. This is the expectation of the product of normally distributed random variables (the β_j) with ratios (the elements of π_j) involving correlated log-normals, a moment for which no analytical expression is readily available. Moments are therefore simulated by fixing a value of θ and drawing a vector of school-level parameters using equations (2.14) and (2.15). Likewise, the simulation draws a vector of the estimation errors in (2.11) from the joint asymptotic distribution of the OLS, reduced form and first stage estimates. The parameter and estimation draws are combined to generate a simulated vector of parameter estimates for the given value of θ. Finally, these are used to construct a set of model-based predicted moments. The SMD estimator minimizes a quadratic form that weights differences between predicted moments and the corresponding moments observed in the data. As described in Appendix B.2, the SMD estimates reported here are generated by a two-step procedure with an efficient weighting matrix in the second step.

2.5.3 Empirical Bayes Posteriors

Studies of teacher and school value-added typically employ EB strategies that shrink noisy teacher- and school-specific value-added estimates towards the grand mean, reducing mean squared error (see, e.g., Kane et al. (2008) and Jacob and Lefgren (2008)). In a conventional VAM model where OLS estimates are presumed unbiased, the posterior mean value-added for school j is

$$E\left[\alpha_j|\hat{\alpha}_j\right] = \left(\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + Var(e_j^\alpha)}\right)\hat{\alpha}_j + \left(1 - \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + Var(e_j^\alpha)}\right)\alpha_0, \tag{2.16}$$

where α_0 and σ_α^2 are the mean and variance of the conventional OLS VAM parameters, α_j. An EB posterior mean plugs estimates of these hyperparameters into (2.16).

Our setup extends this idea to a scenario where the estimated $\hat{\alpha}_j$ may be biased but lotteries are available to reduce this bias. The price for bias reduction is a loss of precision: because lottery estimates use only the variation generated by random assignment, they are less precise than the corresponding OLS estimates. Moreover, because some schools are under-subscribed, there are also fewer lottery instruments than schools and a VAM is not identified using lotteries alone. Even so, in the spirit of the combination estimators discussed by Judge and Mittlehammer (2004, 2007), our empirical Bayes approach trades off the advantages and disadvantages of OLS and lottery estimates to construct minimum mean squared error (MMSE) estimates of value-added.

To see how this trade-off works, suppose the first stage parameters, $\pi_{\ell j}$, are known rather than estimated (equivalently, $e_{\ell j}^\pi = 0\ \forall \ell, j$). Let Π denote the $L \times J$ matrix of these parameters, and let β, $\hat{\alpha}$ and $\hat{\rho}$ denote vectors collecting β_j, $\hat{\alpha}_j$ and $\hat{\rho}_\ell$. The econometric appendix shows that the posterior distribution for β in this case is multivariate normal with mean:

$$E\left[\beta|\hat{\alpha}, \hat{\rho}\right] = W_\alpha(\hat{\alpha} - b_0\iota) + W_\rho\hat{\rho} + (I - W_\alpha - W_\rho\Pi)\beta_0\iota, \tag{2.17}$$

where ι is a $J \times 1$ vector of ones. Posterior mean value-added is a linear combination of OLS estimates net of mean bias, $(\hat{\alpha} - b_0\iota)$, lottery reduced form estimates, $\hat{\rho}$, and mean value-added, $\beta_0\iota$. The weighting matrices, W_α and W_ρ, are functions of the first stage parameters and the covariance matrix of estimation error, value-added, and bias. Expressions for these matrices appear in the econometric appendix. As with conventional EB posteriors, an empirical Bayes version of the posterior mean plugs first-step estimates of b_0, β_0, W_α, and W_ρ into equation (2.17).

In addition to postulating a known first stage, suppose also that all schools are oversubscribed, so that $L = J$. In this case, the first stage matrix, Π, is square; if it is also full rank, the causal effects of all schools are identified using lotteries alone. Lottery-based value-added estimates may then be computed by indirect least squares as $\hat{\beta} = \Pi^{-1}\hat{\rho}$, and the posterior mean in equation (2.17) becomes

$$E\left[\beta|\hat{\alpha},\hat{\beta}\right] = W_\alpha(\hat{\alpha} - b_0\iota) + W_\beta\hat{\beta} + (I - W_\alpha - W_\beta)\beta_0\iota, \tag{2.18}$$

for $W_\beta = W_\rho\Pi$. This expression shows that when a lottery-based value-added model is identified, the posterior mean for value-added is a matrix-weighted average of three quantities: quasi-experimental IV estimates, conventional OLS estimates net of mean bias, and prior mean value-added, with weights (that sum to the identity matrix) chosen to minimize mean squared error.

In related work, Chetty and Hendren (2016) combine noisy quasi-experimental estimates of neighborhood effects based on movers with precise averages of permanent resident outcomes to generate optimal forecasts of neighborhood causal effects. A further special case of equation (2.18) illuminates the link between this approach and ours. Suppose the estimation error in OLS estimates is negligible ($Var(e_j^\alpha) = 0$), and that IV estimation error, e_j^β, is uncorrelated across schools. Appendix B.3 shows that under these simplifying assumptions, the jth element of equation (2.18) becomes

$$E\left[\beta_j|\hat{\alpha},\hat{\beta}\right] = \left(\frac{\sigma_\beta^2(1-R^2)}{Var(e_j^\beta)+\sigma_\beta^2(1-R^2)}\right)\hat{\beta}_j + \left(1 - \frac{\sigma_\beta^2(1-R^2)}{Var(e_j^\beta)+\sigma_\beta^2(1-R^2)}\right)\left(r_\alpha(\hat{\alpha}_j - b_0) + (1-r_\alpha)\beta_0\right), \tag{2.19}$$

where σ_β^2 is the variance of β_j, $r_\alpha = Cov(\beta_j,\alpha_j)/Var(\alpha_j)$ is the slope (also known as the reliability ratio) from a regression of causal value-added on OLS value-added, and R^2 is the R-squared from this regression. This expression coincides with equation (9) in Chetty and Hendren (2016) and can also be seen to be the same as the canonical empirical Bayes shrinkage formula expressed by equation (1.5) of Morris (1983).[16]

In practice, some schools are under-subscribed, so IV estimates of individual school value-added cannot be computed. Nevertheless, equation (2.17) shows that predictions at schools without lotteries can be improved using lottery information from other schools. Lottery reduced form parameters embed information for all fallback schools, including those without lotteries. This is a

[16]The connection with Morris (1983) can be made by observing that when $\hat{\alpha}_j = \alpha_j$, the term $r_\alpha(\hat{\alpha}_j - b_0) + (1-r_\alpha)\beta_0$ is the fitted value from a regression of β_j on α_j.

consequence of the relationship described by equation (2.11), which shows that the reduced form for any school with a lottery depends on the value-added of all other schools that applicants to this school might attend. Specifically, as long as $\pi_{\ell j} \neq 0$, the reduced form for lottery ℓ contains information that can be used to improve the posterior prediction of β_j. The test results in columns 2 and 5 of Table 2.5 show that estimates of $\pi_{\ell j}$ are significantly different from zero (at the 5% level) for 12 of the 22 under-subscribed schools in our sample. The ten schools not on this list have primary entry grades other than sixth. In other words, oversubscribed sixth grade lotteries contribute information on all schools with sixth grade entry.

Finally, equation (2.17) also reveals how knowledge of conventional VAM bias can be used to improve posterior predictions even for schools that are never lottery fallbacks. The econometric appendix shows that the posterior mean for β_j gives no weight to $\hat{\rho}$ when $\pi_{\ell j} = 0$ and $Cov(e_j^{\alpha}, e_{\ell}^{\rho}) = 0$ across all lotteries, ℓ. In this case the posterior mean for β_j simplifies to

$$E\left[\beta_j|\hat{\alpha}, \hat{\rho}\right] = r_{\alpha}(\hat{\alpha}_j - b_0) + (1 - r_{\alpha})\,\beta_0. \tag{2.20}$$

Even without a lottery at school j, predictions based on equation (2.20) improve upon the conventional VAM posterior given by equation (2.16). The improvement here comes from the fact that the schools with lotteries provide information that can be used to determine the reliability of conventional VAM estimates.[17]

Equations (2.17) through (2.20) are pedagogical formulas derived assuming first stage parameters are known. With an estimated first stage, the posterior distribution for value-added does not have a closed form. Although the posterior mean for the general case can be approximated using Markov Chain Monte Carlo (MCMC) methods, with a high-dimensional random coefficient vector, MCMC may be sensitive to starting values or other tuning parameters. We therefore report EB posterior modes (as in Chamberlain and Imbens (2004); these are also known as maximum *a posteriori* estimates). The posterior mode is relatively easily calculated, and coincides with the posterior mean when value-added is normally distributed, as in the fixed first stage case (see the

[17] Using the fact that $\alpha_j = \beta_j + b_j$, equation (2.16) can be written to look more like equation (2.20):

$$E\left[\alpha_j|\hat{\alpha}_j\right] = r_{\alpha}\left(\frac{\sigma_{\beta}^2+\sigma_b^2+2\sigma_{\beta b}}{\sigma_{\beta}^2+\sigma_{\beta b}}\right)(\hat{\alpha}_j - b_0) + \left(1 - r_{\alpha}\left(\frac{\sigma_{\beta}^2+\sigma_b^2+2\sigma_{\beta b}}{\sigma_{\beta}^2+\sigma_{\beta b}}\right)\right)\beta_0 + b_0,$$

This formulation of the conventional EB estimand adds bias, b_0, to a weighted average of bias-corrected OLS and global mean value-added.

econometric appendix for details). As a practical matter, the posterior modes for value-added turn out to be similar to the weighted averages generated by equation (2.17) under the fixed first stage assumption, with a correlation across schools of 0.95 in the lagged score model (see Appendix Figure 2.A1).

2.6 Parameter Estimates

2.6.1 Hyperparameters

The SMD procedure for estimating hyperparameters takes as input a set of lottery reduced form and first stage estimates, along with conventional VAM estimates for each value-added model. The lottery estimates come from regressions of test scores and school attendance indicators (the set of D_{ij}) on lottery offer dummies (Z_i), with controls C_i for randomization strata and the baseline covariates from the lagged score VAM specification (strata controls are necessary for instrument validity, while baseline covariates increase precision). Combining the lottery estimates with OLS estimates of the α_j generates hyperparameter estimates for a particular value-added model.

As can be seen in columns 1-3 of Table 2.6, the hyperparameter estimates reveal substantial variation in both causal value-added and selection bias across schools. The standard deviation of value-added, σ_β, is similar across specifications, ranging from about 0.20σ in the uncontrolled specification to 0.22σ in the lagged score and gains models. This stability is reassuring: the control variables that distinguish these models should not change the underlying distribution of causal school effectiveness if our estimation procedure works as we hope.

In contrast with the relatively stable estimates of σ_β, the estimated standard deviation of bias, σ_b, shrinks from 0.50σ with no controls to under 0.2σ in the lagged score and gains specifications. In other words, controlling for observed student characteristics and past scores reduces bias in conventional value-added estimates markedly. On the other hand, the estimated standard deviations of bias are statistically significant for all models, implying that controls for demographic variables and baseline achievement are not sufficient to produce unbiased comparisons. Columns 2 and 3 of Table 2.6 show that the standard deviations of bias in the lagged score and gains models equal 0.18σ and 0.17σ, slightly smaller than the standard deviation of causal value-added.[18]

[18]Rothstein (2009) assesses for bias in teacher VAMs using Granger-type causality tests that regress lagged test scores on future teacher dummies. Like our random coefficients model, these tests generate estimates of the standard

Earlier work on school effectiveness explores differences between Boston's charter, pilot, and traditional public sectors (Abdulkadiroğlu et al., 2011; Angrist et al., 2016a). These estimates show large charter school treatment effects in Boston, a finding that suggests accounting for sector differences may improve the predictive accuracy of school value-added models. Columns 4 and 5 of Table 2.6 therefore report estimates of lagged score and gains models in which the means of the random coefficients depend on school sector (Appendix Table 2.A3 reports the complete set of parameter estimates for the lagged score model). Consistent with earlier findings, models with sector effects show that average charter school value-added exceeds traditional public school value-added by roughly 0.4σ. Estimated differences in value-added between pilot and traditional public schools are smaller and statistically insignificant. By contrast, bias seems unrelated to sector, implying that conventional VAM models with demographic and lagged achievement controls accurately reproduce lottery-based comparisons of the charter, pilot and traditional sectors (this is also consistent with the findings of Abdulkadiroğlu et al. (2011). The estimates of σ_β and σ_b show that sector effects reduce cross-school variation in both value-added and bias by about 20-25 percent. The large charter effect on value-added notwithstanding, most of the variation in middle school quality in Boston is within sectors rather than between.

Estimated covariances between β_j and b_j, denoted $\sigma_{\beta b}$, are negative and mostly statistically significant, a result that can be seen in the third row of Table 2.6. A negative covariance between value-added and bias suggests that conditional on demographics and past achievement, students with higher ability tend to enroll in schools with lower value-added. Conventional VAMs therefore overestimate the effectiveness of low-quality schools and underestimate the effectiveness of high-quality schools. Estimates of β_Q, the lottery school value-added shifter, are close to zero in models without sector effects, and positive but small when sector effects are included. The estimate of β_Q for the lagged score model is statistically significant, implying that schools with lotteries are slightly more effective than under-subscribed schools in the same sector.

Studies of teacher value-added use the reliability ratio, $r_\alpha = \frac{Cov(\alpha_j, \beta_j)}{Var(\alpha_j)}$, as a summary measure of the predictive value of VAM estimates (Chetty et al., 2014a; Rothstein, forthcoming).[19] The fourth row of Table 2.6 reports model-based estimates of this parameter. The estimated reliability

deviation of bias in VAM estimates.

[19]Chetty et al. (2014a) use this parameter to define "forecast bias," equal to $1 - r_\alpha$. We use the term "reliability" here to distinguish between r_α and the forecast coefficient, $\hat{\varphi}$. The econometric appendix discusses the relationship between $\hat{\varphi}$ and r_α.

of the uncontrolled specification equals 0.08 with a standard error of 0.20, implying that school average test scores are only weakly related to school effectiveness. Reliability ratios in the lagged score and gains models equal 0.64 and 0.75 in models without sector effects, and 0.69 and 0.78 in models with sector effects. Consistent with the test results in Section 2.4, these estimates show that conventional VAM estimates are strongly, but not perfectly, linked to causal school quality.

2.6.2 School Characteristics, Value-added, and Bias

The individual school value-added posterior modes generated by our hybrid estimation strategy are positively correlated with conventional posterior means that ignore bias in OLS value-added estimates. This is evident in Figure 2.4, which plots hybrid modes against posterior means from conventional value-added models. Rank correlations in the lagged score and gains models are 0.79 and 0.74. The relationship between conventional and hybrid posteriors is weaker for lottery schools (indicated by filled markers) than for schools without lotteries: rank correlations for these two groups equal 0.60 and 0.90 in the gains model. This reflects the fact that lotteries are more informative about causal effects for schools with randomized admission. Importantly, although hybrid and conventional posteriors are strongly correlated, hybrid estimation changes some schools' ranks, so accountability decisions may be improved using the hybrid estimates.

Hybrid estimation generates posterior modes for bias as well as value-added. The value-added and bias posteriors therefore permit an exploration of the associations between school characteristics, causal value-added and bias. Table 2.7 reports coefficients from regressions of posterior modes for bias and value-added on school characteristics, with and without controls for sector. As can be seen in columns 1 and 3, students that appear more advantaged (as measured by baseline scores and special education status, for example) tend to enroll in schools with higher value-added, but this pattern is largely explained by the higher likelihood that these students enroll in charter schools. By contrast, column 4 shows that VAM bias within sectors is more positive for schools with more advantaged students, including those with higher average baseline test scores, fewer black students, fewer special education students, and fewer students eligible for subsidized lunches. The correlation of bias with baseline scores is noteworthy: although we see positive selection into the Boston charter sector, the popular impression that good schools have good peers is driven mostly by selection bias.

A key assumption underlying the hybrid approach is that the distribution of bias in conventional VAM estimates is unrelated to lottery oversubscription. This assumption implicitly restricts

the relationship between student ability and school enrollment patterns. For example, it requires that students who enroll in more and less popular schools have similar ability conditional on demographic variables and lagged achievement. Evidence in support of this assumption comes from the relationships between oversubscription rates, posterior bias estimates, and baseline scores.

As can be seen in the upper panel of Figure 2.5, posterior bias estimates are uncorrelated with the extent of oversubscription among lottery schools. Specifically, a regression of predicted bias from the lagged score model on the log of the oversubscription rate yields a slope coefficient of -0.02 with a standard error of 0.06.[20] The weak relationship between bias and the *degree* of oversubscription apparent in the figure is consistent with the hypothesis that bias distributions are similar for schools where lottery information is and is not available. Note also that this finding is not a mechanical consequence of assumptions imposed by the hybrid model, since the model ignores the degree of oversubscription within the lottery sample.

Recall that Table 2.2 shows that baseline scores and other observed characteristics are similar for students enrolled at schools with and without lotteries. The bottom of Panel B of explores this pattern further by showing that oversubscription rates are uncorrelated with average baseline scores at oversubscribed schools. A regression of average baseline scores on log oversubscription produces a coefficient of -0.03 with a standard error of 0.10. This finding, which does not rely on estimates from the model, shows that the observed ability of enrolled students is unrelated to lottery oversubscription within the lottery sample. We might therefore expect unobserved ability to be unrelated to oversubscription as well. Both panels of Figure 2.5 support the assumption postulating similar bias distributions for schools that are more and less heavily oversubscribed.[21]

2.7 Policy Simulations

We use a calibrated Monte Carlo simulation to gauge the accuracy and value of VAM estimates for decision-making. The simulation draws values of causal value-added, bias, and lottery first stage

[20] The oversubscription rate is defined as the ratio of the annual average number of lottery applicants to the average number of seats for charter schools, and the ratio of the average number of first-choice applicants to the average number of seats for traditional and pilot schools.

[21] An appendix section investigates the sensitivity of policy simulation results to violations of this assumption. These results show that hybrid estimation generates substantial gains even when the difference in mean bias between lottery and non-lottery schools is on the order of 0.2σ.

parameters from the estimated distributions underlying Table 2.6.[22] Estimation errors are also drawn from their joint asymptotic distribution and are combined with parameter draws to construct simulated OLS, reduced form and first stage estimates. These simulated estimates are then used to re-estimate the random coefficients model and construct conventional and hybrid EB posterior predictions. Each simulation therefore replicates the information available to a policymaker or parent, armed with both conventional and hybrid estimates, in a world calibrated to our model.

2.7.1 Mean Squared Error

Our first statistic for model assessment is root mean squared error (RMSE). Conventional VAMs generate value-added estimates of school quality with an RMSE far below that of a naive uncontrolled benchmark. This can be seen in Figure 2.6, which compares RMSE across specifications and estimation procedures. RMSE in the uncontrolled model is about 0.5σ, falling to around 0.18σ and 0.17σ in the lagged score and gains VAMs. Adjustments for past scores and other student demographics eliminate a good portion of the bias in uncontrolled estimates.

The RMSE of hybrid estimates is impressively stable across specifications, starting at 0.17σ in an uncontrolled benchmark model and falling to 0.14σ in the lagged score and gains models. With sector effects included, hybrid estimation reduces RMSE from 0.15σ to about 0.12σ in the lagged score model and from 0.14σ to about 0.10σ in the gains model. The relatively stable hybrid RMSE shows how the hybrid estimator manages to reduce bias even when non-lottery estimates are badly biased. Although the largest bias mitigation seen in the figure comes from controlling for covariates, hybrid estimation reduces RMSE by a further 20-30%.

Not surprisingly, the RMSE reduction yielded by the hybrid estimator reflects reduced bias at the cost of increased sampling variance. This can be seen by writing the mean squared error of an estimator, β_j^*, as

$$E\left[\left(\beta_j^* - \beta_j\right)^2\right] = E\left[Var\left(\beta_j^* | \beta_j\right)\right] + \sigma_{b*}^2 \tag{2.21}$$

where $\sigma_{b*}^2 = E\left[\left(E\left[\beta_j^* | \beta_j\right] - \beta_j\right)^2\right]$ is average bias squared and the expectation treats the value-added parameters, β_j, as random. Dark and light shading in Figure 2.6 shows the proportions of MSE due to bias and variance. OLS VAMs are precisely estimated: sampling variance contributes

[22] Simulation results for seventh and eighth grade, reported in Appendix Tables 2.A4 and 2.A5, yield conclusions similar to those for sixth grade. These and other supplementary simulation results are discussed in the appendix.

only a small part of their overall MSE. Hybrid estimation reduces MSE, while also increasing the proportion of error due to sampling variance to around 30%. This reflects the tradeoff motivating the hybrid approach: hybrid posteriors leverage lottery estimates to reduce bias in exchange for increased sampling variance relative to conventional VAMs.[23]

2.7.2 Consequences of School Closure

Massachusetts' school accountability framework uses value-added measures to guide decisions about school closures, restructuring, and expansion. A stylized description of these decisions is that they replace weak schools with those judged to be stronger on the basis of value-added estimates. We therefore simulate the achievement consequences of closing the lowest-ranked district school (traditional or pilot) and sending its students to new schools with better estimated value-added.

This analysis ignores possible transition effects such as disruption due to school closure, peer effects from changes in school composition, and other factors that might inhibit replication of successful schools. The results should nevertheless provide a rough guide to the potential consequences of VAM-based policy decisions. Quasi-experimental analyses of charter takeovers and other school reconstitution efforts in Boston, New Orleans, and Houston have shown large gains when low-performing schools are replaced by schools operating according to pedagogical principles seen to be effective elsewhere (Fryer, 2014; Abdulkadiroğlu et al., 2016). This suggests transitional consequences are dominated by longer-run determinants of school quality, at least for modest policy interventions of the sort considered here.

The potential for VAMs to guide decision-making is highlighted by the first row of Table 2.8, which shows the score gains produced by decisions based on true value-added. Closing the worst school and replacing it with an average school boosts achievement by 0.37σ, while more targeted replacement policies generate even larger gains. Consistent with the high RMSE of uncontrolled estimates, however, Table 2.8 also shows that policies based on uncontrolled test score levels generate only small gains. For example, replacing the lowest-scoring district school with an average school is predicted to increase scores for affected students by 0.06σ on average. Likewise, a policy that replaces the lowest-ranked school with an average top quintile school generates a gain of 0.10σ. These small effects reflect the large variation in bias evident for the uncontrolled model in Table

[23]Appendix Table 2.A6 shows that hybrid estimates generate forecast coefficients close to one in both the lagged score and gains specifications, with or without charter lotteries. The hybrid estimates also pass the overidentification and omnibus specification tests.

2.6: closure decisions based on average test scores target schools with many low achievers rather than low value-added. The bias in uncontrolled VAM estimates also leads to a wide dispersion of simulated closure effects, with a cross-simulation standard deviation (reported in brackets) of 0.2σ.

In contrast, closure and replacement decisions based on conventional lagged score and gains models yield substantial achievement gains. For instance, replacing the lowest-ranked school with an average school boosts scores by an average of 0.24σ when rankings are based on the gains specification. This is 65% of the corresponding benefit generated by a policy that ranks schools by true value-added. Hybrid estimation increases these gains to 0.32σ, an improvement of over 30% relative to the conventional model. This incremental effect closes roughly half the gap between conventional estimates and the maximum possible impact.

The effects of VAM-based policies and the incremental benefits of using lotteries grow when value-added predictions are used to choose expansion schools in addition to closures. In the gains specification, for example, replacing the lowest-ranked school with a typical top-quintile school generates an average improvement of 0.39σ when conventional posteriors are used to estimate VAM and an improvement of 0.53σ when rankings are based on hybrid predictions. The hybrid approach also reduces the uncertainty associated with VAM-based policies by doing a better job of finding reliably good replacement schools.

The largest gains seen in Table 2.8 result from a policy that replaces the lowest-ranking traditional or pilot school with a charter school. This mirrors Boston's ongoing in-district charter conversion policy experiment (Abdulkadiroğlu et al., 2016). Reflecting the large difference in mean value-added between charter and district schools, charter conversion generates significant gains regardless of how value-added is estimated. Accurate value-added estimation increases the efficacy of charter conversion, however: selecting schools for conversion based on the lagged score model rather than the uncontrolled model boosts the effect of charter expansion from 0.28σ to 0.58σ, while the hybrid estimator increases this to 0.67σ, close to the maximum possible gain of 0.71σ.

The results in Table 2.8 show that, even when VAM estimates are imperfect, they predict causal value-added well enough to be useful for policy. For example, causal value-added is more than 0.2σ below-average for schools ranked at the bottom by the conventional lagged score and gains specifications. As can be seen in Table 2.6, this represents roughly a full standard deviation in the distribution of true school effectiveness. Value-added for low-ranked schools is even more negative when rankings are based on hybrid estimates. Replacement schools may not be the very

worst schools in the district, but are likely to be much worse than average, so policies that replace them with schools predicted to do better generate large average gains.[24]

2.8 Conclusions

School districts increasingly rely on regression-based value-added models to gauge and report on school quality. This chapter leverages admissions lotteries to test and improve conventional VAM estimates of school value-added. An application of our approach to data from Boston suggests that conventional value-added estimates for Boston's schools are biased. Nevertheless, policy simulations show that accountability decisions based on estimated VAM are likely to boost achievement. A hybrid estimation procedure that combines conventional and lottery-based estimates generates predictions that, while still biased, achieve lower mean squared error and improved policy targeting relative to conventional VAMs.

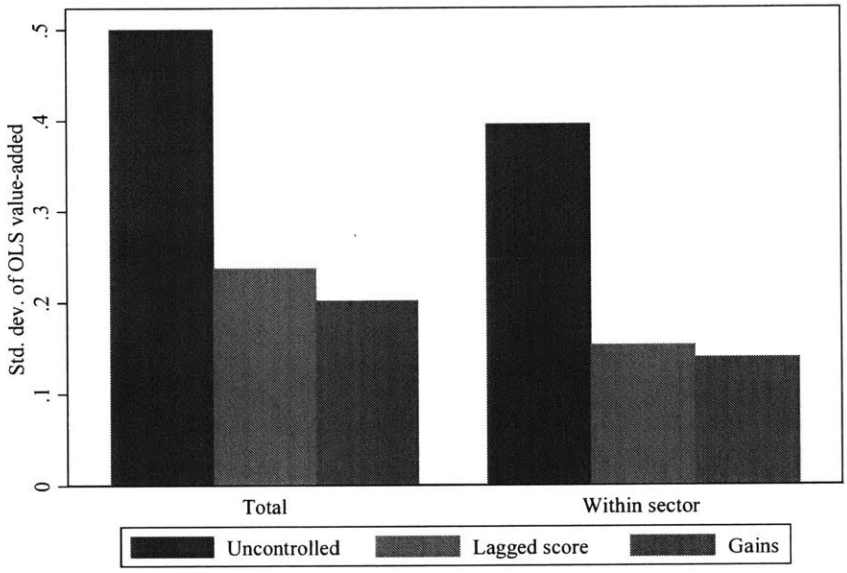
Hybrid school value-added estimation requires some kind of lottery-based admissions scheme, such as those increasingly used for student assignment in many of America's large urban districts. As our analysis of Boston's multiple-offer charter sector shows, however, admissions need not be centralized for lotteries to be of value. The utility of hybrid estimation in other cities will vary with the extent of lottery coverage, but results for Boston show hybrid estimation remains useful even when lottery data are missing for many schools. Our approach also ignores effect heterogeneity linked to school choices, a limitation that might matter in settings with more specialized schools and very heterogeneous student populations.

The methods developed here may be adapted to combine quasi-experimental and non-experimental estimators in other settings. Candidates for this extension include the quantification of teacher, doctor, hospital, firm, and neighborhood effects. Assignment lotteries in these settings are rare, but our hybrid estimation strategy may be used to exploit other sources of quasi-experimental variation. A hybrid approach to testing and estimation is likely to be fruitful in any context where a set of credible quasi-experiments is available to benchmark a larger set of non-experimental comparisons.

[24]The simulations in Table 2.8 predict the consequences of decisions based on the eight years of data in our sample. Districts often estimate value-added over shorter time periods. To gauge the effects of using four years of data, Appendix Table 2.A7 reports simulation results that double sampling variance. This produces results which are qualitatively similar to those from the full sample, with slightly smaller closure effects. Appendix Table 2.A3 reports estimates from a model (described in the supplementary results appendix) that allows value-added and bias to vary by year. These estimates suggest a limited role for idiosyncratic temporal variation in VAM parameters.

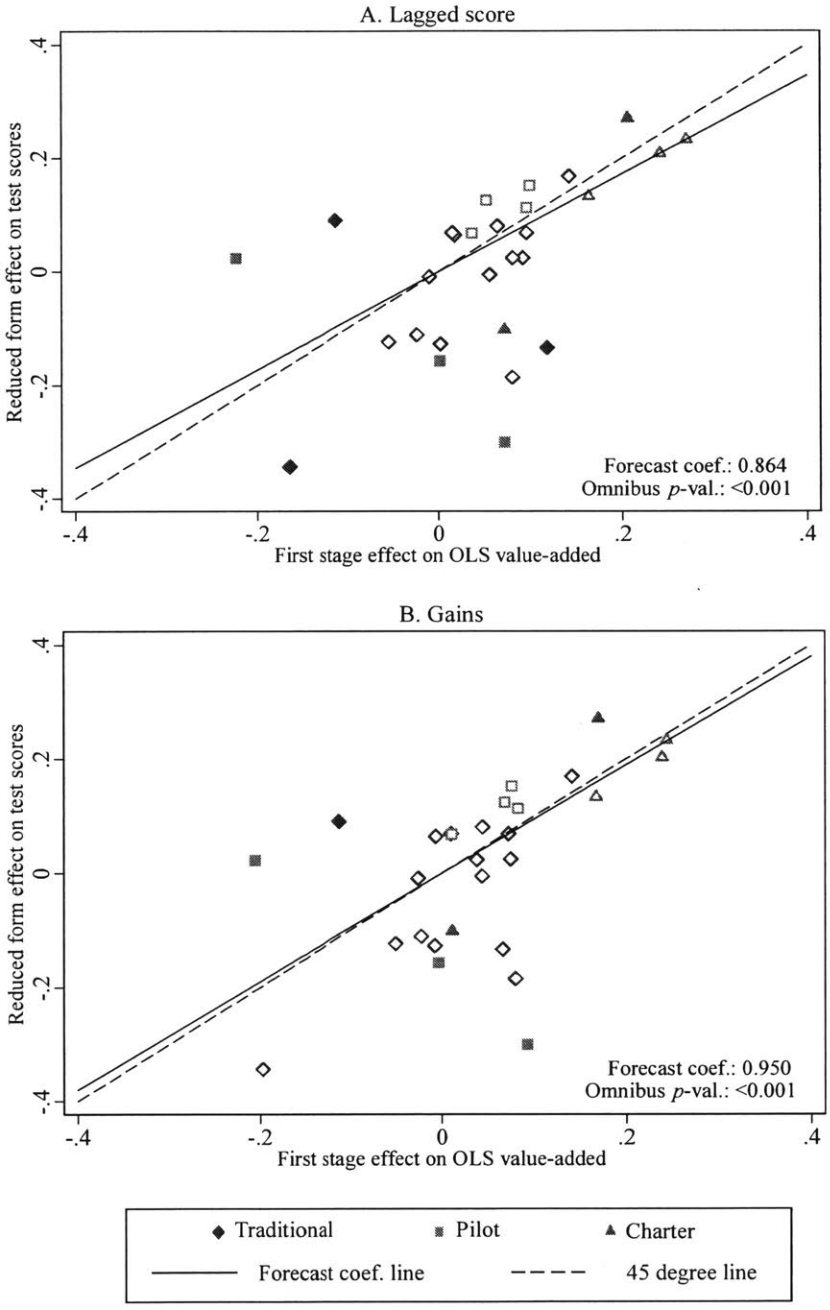
2.9 Figures and Tables

Figure 2.1: Standard Deviations of School Effects from OLS Value-added Models



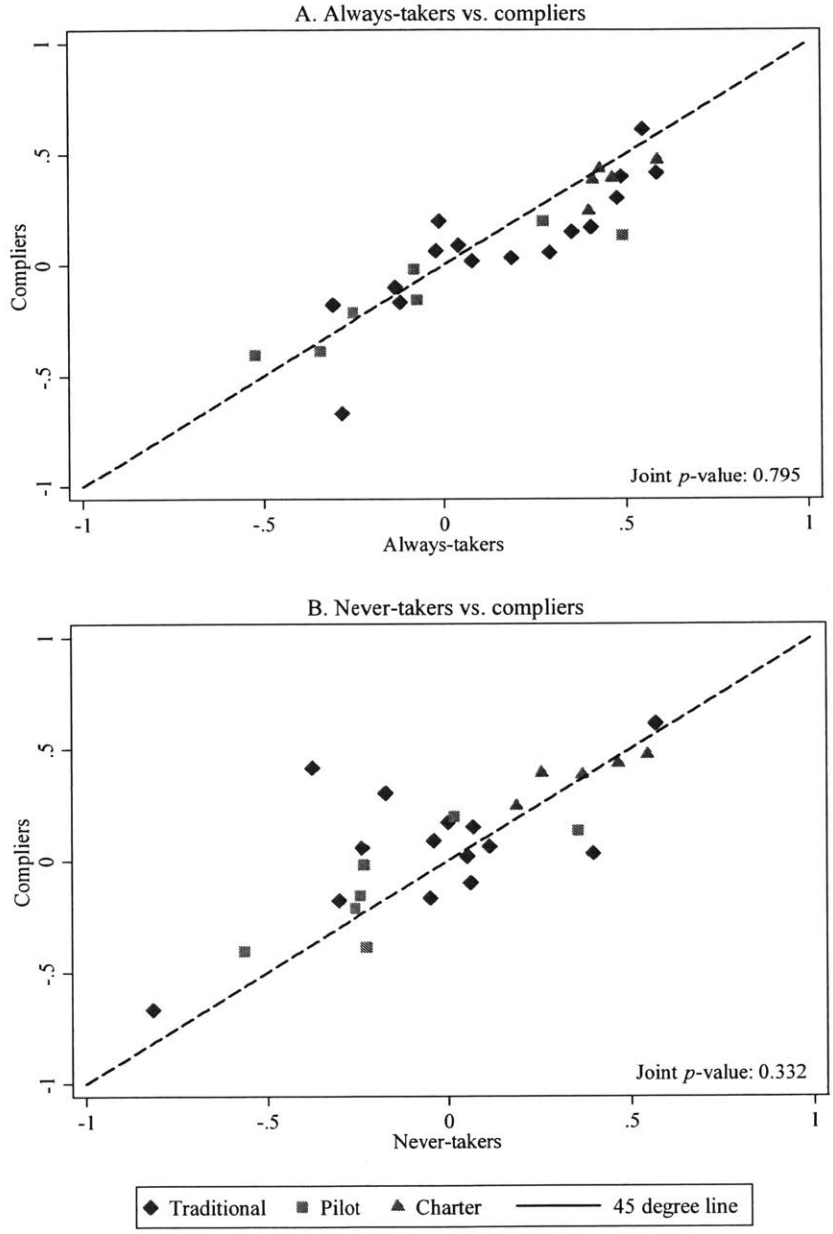
Notes: This figure compares standard deviations of school effects from alternative OLS value-added models. The notes to Table 2.3 describe the controls included in the lagged score and gains models; the uncontrolled model includes only year effects. The variance of OLS value-added is obtained by subtracting the average squared standard error from the sample variance of value-added estimates. Within-sector variances are obtained by first regressing value-added estimates on charter and pilot dummies, then subtracting the average squared standard error from the sample variance of residuals.

Figure 2.2: Visual Instrumental Variables Tests for Bias



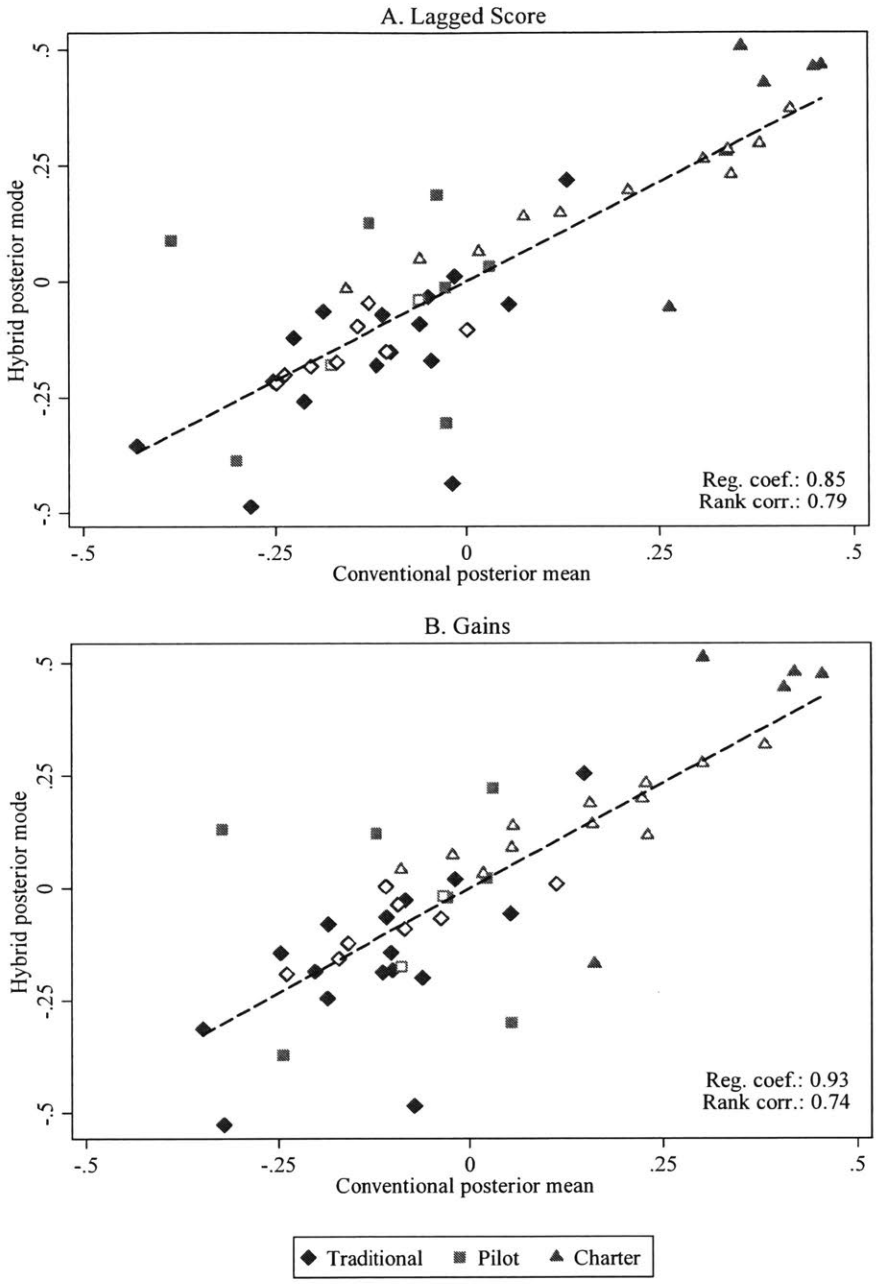
Notes: This figure plots sixth grade lottery reduced form estimates against value-added first stages from each of 28 school admission lotteries. The notes to Table 2.3 describe the underlying models. Filled markers indicate reduced form and first stage estimates that are significantly different from each other at the 10% level. The solid lines have slopes equal to the forecast coefficients in Table 2.3, while dashed lines indicate the 45-degree line. Omnibus p-values are for the over-identification test statistic described in Section 2.4.1.

Figure 2.3: Comparisons of Conventional Value-added by Lottery Compliance



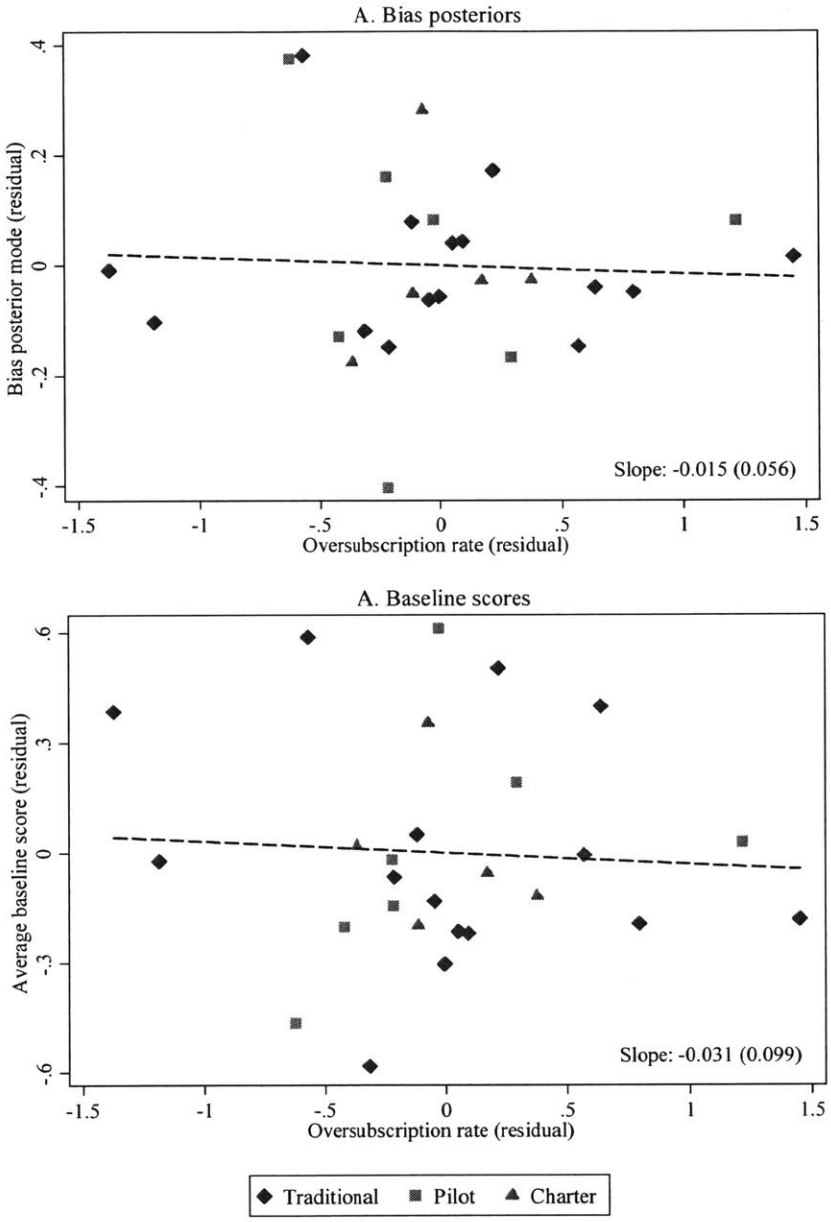
Notes: This figure compares OLS estimates of average value-added for admission lottery compliers to estimates for always- and never-takers in each of 28 school lotteries. OLS estimates come from a lagged-score VAM that allows school effects to differ across the subgroups used in column 6 of Table 2.4, estimated in the lottery sample. Complier, always-taker, and never-taker means are estimated using methods described in Appendix B. *P*-values are for joint tests of complier and always-/never-taker equality across all schools. The *p*-value for a test that pools the estimates in both panels is 0.289.

Figure 2.4: Empirical Bayes Posterior Predictions of School Value-added



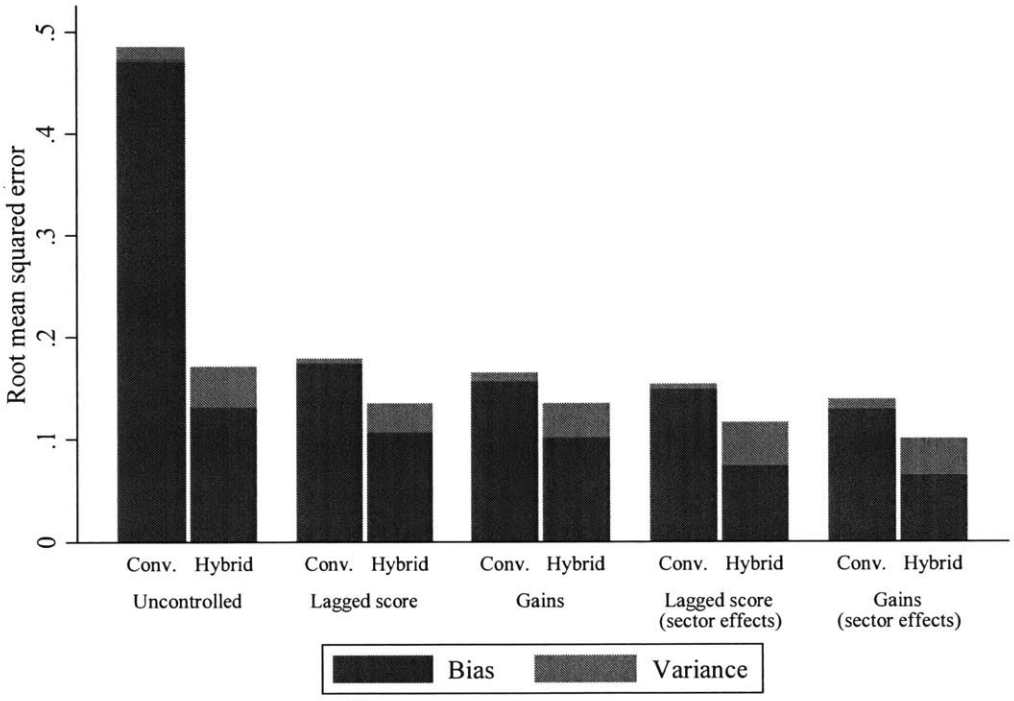
Notes: This figure plots empirical Bayes posterior modes of value-added from the hybrid model against posterior means based on OLS value-added. Posterior modes are computed by maximizing the sum of the log-likelihood of the OLS, reduced form, and first stage estimates conditional on all school-specific parameters plus the log-likelihood of these parameters given the estimated random coefficient distribution. Conventional posteriors shrink OLS estimates towards the mean in proportion to one minus the signal-to-noise ratio. Filled markers indicate lottery schools. Dashes indicate OLS regression lines linking the two sets of estimates.

Figure 2.5: Relationship Between Oversubscription and Bias for Lottery Schools



Notes: Panel A of this figure plots posterior mode predictions of bias in sixth grade math VAMs against oversubscription rates for schools with admission lotteries. The oversubscription rate is defined as the log of the ratio of the average number of first-choice applicants (for traditional and pilot schools) or the average number of total applicants (for charters) to the average number of available seats for each admission grade. Bias modes come from the lagged score model with sector effects. Panel B plots school average baseline math and ELA scores against oversubscription rates. Points in the figure are residuals from regressions of bias modes, mean baseline scores and oversubscription rates on pilot and charter indicators. Dashes indicate OLS regression lines.

Figure 2.6: Root Mean Squared Error for Value-added Posterior Predictions



Notes: This figure plots root mean squared error (RMSE) for posterior predictions of sixth grade math value-added. Conventional predictions are posterior means constructed from OLS value-added estimates. Hybrid predictions are posterior modes constructed from OLS and lottery estimates. The total height of each bar indicates RMSE. Dark bars display shares of mean squared error due to bias, and light bars display shares due to variance. RMSE is calculated from 500 simulated samples drawn from the data generating processes implied by the estimates in Table 2.6. The random coefficients model is re-estimated in each simulated sample.

Table 2.1: Boston Students and Schools

Enrollment				Enrollment			
OLS sample	Lottery sample	6th grade entry?	Lottery school?	OLS sample	Lottery sample	6th grade entry?	Lottery school?
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Traditional public schools (25)				B. Pilot schools (9)			
1,095	79	Y	Y	538	310	Y	Y
1,025	445	Y	Y	1,260	433	Y	Y
1,713	1,084	Y	Y	585	296	Y	Y
547	218	Y	Y	78	5	Y	
217	46	Y		453	46	Y	Y
1,354	581	Y	Y	380	67	Y	Y
263	44	Y		242	179	Y	Y
1,637	492	Y	Y	558	73	Y	Y
472	104	Y		18	12		
1,238	591	Y	Y	C. Charter schools (17)			
537	11			738	406	Y	Y
331	35	Y	Y	361	23		
335	82	Y		357	215	Y	
952	232	Y	Y	393	332	Y	Y
294	71	Y	Y	338	16		
333	90	Y		511	115	Y	Y
766	243	Y	Y	71	8		
372	47	Y	Y	300	23		
137	14	Y		389	342	Y	Y
1,091	225	Y	Y	654	34		
1,086	127	Y	Y	45	3		
577	104	Y	Y	53	2		
622	61	Y		415	305	Y	Y
906	270	Y	Y	70	6		
267	19			104	23		
				701	92		
				85	37		

Notes: This table counts the students included in each school in the OLS value-added and lottery samples. The sample covers cohorts attending sixth grade in Boston between the 2006-2007 and 2013-2014 school years. Columns 3 and 7 indicate schools for which sixth grade is the primary entry grade, while columns 4 and 8 indicate whether a school has enough students subject to random admission variation to be included in the lottery sample. Total numbers of schools in each sector appear in parentheses in the school type headings.

Table 2.2: Descriptive Statistics

	Means				Lottery offer balance			
	OLS sample		Lottery sample					
Baseline covariate	All students (1)	Lottery school students (2)	All students (3)	Lottery school students (4)	All lotteries (5)	Traditional (6)	Pilot (7)	Charter (8)
Hispanic	0.345	0.342	0.354	0.361	-0.017	-0.007	0.003	-0.006
					(0.013)	(0.017)	(0.033)	(0.018)
Black	0.410	0.394	0.485	0.468	-0.011	-0.005	-0.052	-0.009
					(0.014)	(0.018)	(0.034)	(0.020)
White	0.122	0.125	0.072	0.078	0.010	0.006	0.005	0.009
					(0.007)	(0.008)	(0.015)	(0.010)
Female	0.490	0.487	0.504	0.502	0.017	0.034	-0.013	-0.025
					(0.014)	(0.019)	(0.037)	(0.020)
Subsidized lunch	0.806	0.811	0.830	0.831	0.020	0.020	0.006	-0.005
					(0.010)	(0.013)	(0.026)	(0.016)
Special education	0.208	0.214	0.195	0.196	0.006	-0.003	-0.022	0.015
					(0.011)	(0.013)	(0.030)	(0.016)
English-language learner	0.205	0.224	0.206	0.214	0.006	-0.001	0.018	0.004
					(0.011)	(0.014)	(0.027)	(0.016)
Suspensions	0.093	0.073	0.076	0.070	-0.025	-0.025	0.009	-0.016
					(0.016)	(0.023)	(0.025)	(0.017)
Absences	1.710	1.567	1.534	1.466	-0.087	-0.138	-0.092	0.110
					(0.095)	(0.080)	(0.260)	(0.167)
Math score	0.058	0.053	0.004	0.016	0.022	-0.026	0.080	0.036
					(0.024)	(0.030)	(0.061)	(0.035)
ELA score	0.030	0.006	0.013	0.016	0.035	0.045	0.060	0.013
					(0.025)	(0.030)	(0.061)	(0.036)
N	27,864	21,446	8,718	7,748	8,718	4,849	1,303	3,655

Notes: This table reports sample means and investigates balance of random lottery offers. Column 1 shows mean characteristics for all Boston sixth graders enrolled between the 2006/7 and 2013/14 school years, and column 2 shows means for students enrolled at schools that have randomized entrance lotteries in at least one year. Columns 3 and 4 report mean characteristics for students subject to random lottery assignment. Columns 5-8 report coefficients from regressions of baseline characteristics on lottery offers, controlling for assignment strata. Robust standard errors are reported in parenthenses.

Table 2.3: Tests for Bias in Conventional Value-added Models

	All lotteries		Excluding charter lotteries	
	Lagged score	Gains	Lagged score	Gains
	(1)	(2)	(3)	(4)
	A. Sixth grade			
Forecast coefficient (φ)	0.864	0.950	0.549	0.677
	(0.075)	(0.084)	(0.164)	(0.193)
First stage F-statistic	29.6	26.6	11.2	9.3
p-values:				
Forecast bias	0.071	0.554	0.006	0.095
Overidentification	0.003	0.006	0.043	0.052
Omnibus test χ^2 statistic (d.f.)	77.7 (28)	72.1 (28)	48.0 (23)	41.7 (23)
p-value	<0.001	<0.001	<0.001	0.010
N	8,718		6,162	
	B. All middle school grades			
Forecast coefficient (φ)	0.880	0.924	0.683	0.726
	(0.055)	(0.060)	(0.124)	(0.133)
First stage F-statistic	14.7	15.0	7.6	7.8
p-values:				
Forecast bias	0.028	0.204	0.011	0.039
Overidentification	0.011	0.011	0.062	0.065
Omnibus test χ^2 statistic (d.f.)	172.8 (75)	167.0 (75)	111.6 (60)	107.9 (60)
p-value	<0.001	<0.001	<0.001	<0.001
N	20,935		15,027	

Notes: This table reports the results of tests for bias in conventional value-added models (VAMs) for sixth through eighth grade math scores. The lagged score VAM includes cubic polynomials in baseline math and ELA scores, along with indicators for application year, sex, race, subsidized lunch, special education, limited-English proficiency, and counts of baseline absences and suspensions. The gains VAM drops the lagged score controls and uses score growth from baseline as the outcome. Seventh and eighth grade VAMs measure exposure to each school using total years of enrollment since the lottery. Forecast coefficients are from instrumental variables regressions of test scores on fitted values from conventional VAMs, instrumenting fitted values with lottery offer indicators. All models are estimated via an asymptotically efficient GMM procedure and control for assignment strata fixed effects, demographic variables, and lagged scores. The forecast bias test checks whether the forecast coefficient equals one, and the overidentificiation test checks the IV model's overidentifying restrictions. The omnibus test combines forecast bias and overidentifying restrictions. Panel A uses sixth grade math scores, while Panel B stacks outcomes from sixth through eighth grade. Standard errors and test statistics in Panel B cluster by student. Columns 3 and 4 exclude charter school lotteries.

Table 2.4: Robustness of Sixth Grade Bias Tests to Effect Heterogeneity

Value-added model		Baseline VAM specification	VAM estimated by subgroup: Baseline year	Subsidized lunch	Special education	Baseline score tercile	Interacted groups
		(1)	(2)	(3)	(4)	(5)	(6)
		A. VAM estimated on the OLS sample					
Lagged score	Forecast coefficient (φ)	0.864	0.916	0.849	0.863	0.866	0.930
		(0.075)	(0.072)	(0.075)	(0.074)	(0.075)	(0.061)
	Omnibus test $\chi^2(28)$ statistic	77.7	68.2	82.8	79.0	83.3	73.4
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Gains	Forecast coefficient (φ)	0.950	1.016	0.944	0.955	0.891	0.953
		(0.084)	(0.082)	(0.083)	(0.083)	(0.079)	(0.065)
	Omnibus test $\chi^2(28)$ statistic	72.1	65.7	74.4	72.4	80.9	66.6
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
		B. VAM estimated on the lottery sample					
Lagged score	Forecast coefficient (φ)	0.868	0.962	0.851	0.872	0.873	0.934
		(0.070)	(0.068)	(0.069)	(0.070)	(0.070)	(0.052)
	Omnibus test $\chi^2(28)$ statistic	62.3	51.8	67.9	63.5	65.9	56.3
	p-value	<0.001	0.004	<0.001	<0.001	<0.001	0.001
Gains	Forecast coefficient (φ)	0.926	1.035	0.912	0.937	0.890	0.941
		(0.077)	(0.077)	(0.076)	(0.077)	(0.073)	(0.055)
	Omnibus test $\chi^2(28)$ statistic	57.8	50.2	60.1	58.4	61.1	42.2
	p-value	<0.001	0.006	<0.001	<0.001	<0.001	0.041

Notes: This table reports lottery-based tests for bias in school value-added models that allow for effect heterogeneity by baseline characteristics. The notes to Table 2.3 describe the underlying models. Panel A shows results for the full OLS sample, while Panel B shows results for the lottery subsample. Column 1 reports estimates that do not allow effect heterogeneity, while columns 2-6 are from models allowing value-added to differ across groups defined by the covariates in the column headings. The covariates used to define groups in column 6 are race, gender, subsidized lunch, special education, English language learner status, and baseline score terciles based on average fifth grade math and ELA test scores in the OLS sample.

Table 2.5: Fallback Status of Schools Without Sixth Grade Lotteries

Lottery students with fallback enrollment	p-value: not a lottery fallback	Sixth grade entry?	Lottery students with fallback enrollment	p-value: not a lottery fallback	Sixth grade entry?
(1)	(2)	(3)	(4)	(5)	(6)
	A. Traditional publics			C. Charters	
39	0.013	Y	320	<0.001	Y
36	0.018	Y	11	0.080	
113	<0.001	Y	16	0.427	
79	<0.001	Y	16	0.204	
94	<0.001	Y	24	0.724	
21	0.045	Y	42	0.145	
60	0.006	Y	3	0.111	
12	0.016		2	0.390	
	B. Pilots		10	0.257	
5	0.033	Y	33	0.010	
15	0.169		112	<0.001	
			34	0.378	

Notes: This table reports p-values for tests of whether each non-lottery school in the OLS sample serves as a fallback for one of the 28 lottery schools. Columns 1 and 4 count the number of students in the lottery sample who are observed enrolling in the undersubscribed school when not given an offer. Columns 2 and 5 report p-values from tests of the hypothesis that the undersubscribed school's first stage coefficients are zero in all lotteries with such students. Columns 3 and 6 indicate whether sixth grade is a school's primary entry point. First stage regressions control for assignment strata indicators, demographic variables, and lagged test scores.

Table 2.6: Estimates of the Joint Distribution of Causal Value-added and VAM Bias for Sixth Grade Math Scores

		Models without sector effects			Models with sector effects	
		Uncontrolled	Lagged score	Gains	Lagged score	Gains
	Parameters	(1)	(2)	(3)	(4)	(5)
σ_β	Std. dev. of causal VA	0.195	0.220	0.222	0.171	0.170
		(0.024)	(0.021)	(0.023)	(0.028)	(0.023)
σ_b	Std. dev. of VAM bias	0.501	0.182	0.166	0.148	0.133
		(0.061)	(0.048)	(0.048)	(0.029)	(0.030)
$\sigma_{\beta b}$	Covariance of	-0.018	-0.014	-0.017	-0.016	-0.013
	VA and bias	(0.010)	(0.003)	(0.004)	(0.006)	(0.003)
r_α	Regression of VA	0.078	0.644	0.753	0.694	0.783
	on OLS (reliability ratio)	(0.204)	(0.066)	(0.072)	(0.152)	(0.122)
VA shifters	Charter				0.426	0.396
					(0.104)	(0.106)
	Pilot				0.130	0.111
					(0.129)	(0.129)
	Lottery school (β_Q)	0.040	-0.024	-0.033	0.104	0.066
		(0.127)	(0.061)	(0.054)	(0.042)	(0.041)
Bias shifters	Charter				-0.005	-0.063
					(0.103)	(0.099)
	Pilot				-0.121	-0.089
					(0.124)	(0.121)
	χ^2 statistic (d.f.):	10.9 (7)	10.8 (7)	9.1 (7)	9.0 (13)	6.0 (13)
	Overid. p-value:	0.145	0.147	0.247	0.773	0.946

Notes: This table reports simulated minimum distance estimates of parameters of the joint distribution of causal school value-added and OLS bias. The moments used in estimation are functions of OLS value-added, lottery reduced form, and first stage estimates, as described in Appendix B. Uncontrolled estimates come from an OLS regression that controls only for year effects. The notes to Table 2.3 describe the underlying models. Simulated moments are computed from 500 samples constructed by drawing school-specific parameters from the random coefficient distribution along with estimation errors based on the asymptotic covariance matrix of the estimates. The estimates in columns 4 and 5 are from models allowing the means of the random coefficients distribution to depend on school sector. Moments are weighted by an estimate of the inverse covariance matrix of the moment conditions, calculated from a first-step estimate using an identity weighting matrix. The weighting matrix is produced using 1,000 simulations, drawn independently from the samples used to simulate the moments.

Table 2.7: Correlates of Posterior Value-added and VAM Bias

	Overall		Within-sector	
	Value-added	Bias	Value-added	Bias
School characteristic	(1)	(2)	(3)	(4)
Fraction black	0.158	-0.208	-0.050	-0.217
	(0.143)	(0.075)	(0.083)	(0.073)
Fraction hispanic	0.065	0.031	0.268	0.048
	(0.201)	(0.105)	(0.127)	(0.112)
Fraction subsidized lunch	-0.132	-0.452	0.085	-0.474
	(0.306)	(0.181)	(0.203)	(0.164)
Fraction special education	-0.997	-0.501	0.009	-0.508
	(0.330)	(0.157)	(0.316)	(0.217)
Fraction English-language learne	-0.542	-0.135	0.297	-0.092
	(0.247)	(0.221)	(0.243)	(0.254)
Mean baseline math score	0.157	0.143	0.012	0.145
	(0.088)	(0.051)	(0.070)	(0.047)
Mean baseline ELA score	0.201	0.135	0.039	0.138
	(0.085)	(0.060)	(0.074)	(0.060)
Charter and pilot controls?			Y	Y

Notes: This table reports coefficients from regressions of empirical Bayes posterior modes for causal value-added and bias on school characteristics. Columns 1 and 2 show coefficients from bivariate regressions, while columns 3 and 4 show coefficients from regressions controlling for charter and pilot indicators. Posterior modes come from the lagged score model with sector effects for sixth grade math scores. Robust standard errors are reported in parentheses.

Table 2.8: Consequences of Closing the Lowest-ranked District School for Affected Students

		Replacement school			
		Average district school	Average above-median school	Average top-quintile school	Average charter school
Model	Posterior method	(1)	(2)	(3)	(4)
	True value-added	0.370	0.507	0.610	0.711
		[0.080]	[0.089]	[0.094]	[0.094]
Uncontrolled	Conventional	0.056	0.078	0.095	0.280
		[0.191]	[0.197]	[0.204]	[0.198]
	Hybrid	0.153	0.223	0.259	0.377
		[0.143]	[0.156]	[0.169]	[0.151]
Lagged score	Conventional	0.226	0.307	0.367	0.577
		[0.159]	[0.168]	[0.176]	[0.165]
	Hybrid	0.315	0.437	0.529	0.665
		[0.131]	[0.141]	[0.147]	[0.145]
Gains	Conventional	0.240	0.327	0.391	0.580
		[0.148]	[0.156]	[0.163]	[0.153]
	Hybrid	0.316	0.434	0.525	0.657
		[0.115]	[0.126]	[0.136]	[0.128]

Notes: This table reports simulated consequences of closing the lowest-ranked BPS district school based on value-added predictions. The reported impacts are average effects on test scores for students at the closed school. Standard deviations of these effects across simulations appear in brackets. The scenario in column 1 replaces the lowest-ranked district school with an average district school. Column 2 replaces the lowest-ranked school with an average above-median district school, and column 3 uses an average top-quintile district school. Column 4 replaces the lowest-ranked district school with an average charter school. Conventional empirical Bayes posteriors are means conditional on OLS estimates only, while hybrid posteriors are modes conditional on OLS and lottery estimates. All models include sector effects. Statistics are based on 500 simulated samples, and the random coefficients model is re-estimated in each sample.

2.10 Data Appendix

The administrative data used for this project come from student demographic and attendance information in the Massachusetts Student Information Management System (SIMS), standardized student test scores from the Massachusetts Comprehensive Assessment System (MCAS) database, Boston charter school admission lottery records, and information from the centralized BPS student assignment system. We describe each data source and our cleaning and matching process in detail below; the construction of our main analysis file closely follows that of previous studies, in particular Abdulkadiroğlu et al. (2011).

Student enrollment, demographics, and test scores

The Massachusetts SIMS contains snapshots of all students in a public school in Massachusetts in October and at the end of each school year. These records contain demographic information on students, their current schools, their residence, and their attendance. We work with SIMS files for the 2005-2006 through the 2013-2014 school years and limit the sample to students enrolled in a Boston school over this period. Schools are classified as charters by the Massachusetts Department of Elementary and Secondary Education website (`http://www.profiles.doe.mass.edu`), and as pilots by the Boston pilot school network website (`http://www.ccebos.org/pilotschools/schools.html`). All remaining Boston schools are considered traditional public schools for the purposes of this study.

Enrollment in the SIMS is grade-specific. When a student repeats grades, we retain the first school a student attended in that grade. We then record students attending multiple schools in a given school year as enrolled in the school for which the attendance duration is longest, with duration ties broken randomly. This results in a unique student panel across grades; for the purposes of this study we restrict focus to sixth grade students enrolled from 2006/7 to 2013/14, using their fifth grade information for baseline controls. These controls include indicators for student race (Hispanic, black, white, Asian, and other race), sex, free- or reduced-price lunch eligibility, special education status, and English-language learner status, as well as counts of the number of days a student was suspended or truant over the school year. Suspension data are unavailable in the SIMS starting in the 2012-2013 school year; we include an indicator for students missing this baseline information whenever suspensions are used.

Our primary outcomes for measuring school value-added are sixth, seventh and eighth grade standardized test scores from the Massachusetts Comprehensive Assessment System (MCAS) database. MCAS math and ELA scores are standardized by grade and year to a combined BPS and Boston charter school reference population. MCAS outcome scores are merged to SIMS data via a state-assigned unique student identifier. We also merge baseline (fifth grade) MCAS scores for each student in our sample (fifth grade MCAS information is available starting in the 2005-2006 school year).

Charter school lotteries

The lottery analysis uses records for five of the six Boston middle school charters with sixth grade admission for the 2006/7 through the 2013/14 academic year. These schools are Academy of the Pacific Rim, Boston Preparatory, MATCH Charter Public Middle School, Roxbury Preparatory, and UP Academy Boston. The remaining school, Smith Leadership Academy, has declined to participate in our studies. For each school and each oversubscribed year we obtain a list of names of students eligible for entry by lottery, as well as information on whether each student was offered a seat on lottery night. Students are marked as ineligible if they submit an incomplete or late application; we also exclude students with a sibling currently enrolled in the school, as they are guaranteed admission. For UP Boston, which is an in-district charter school, students applying from outside of BPS are placed in a lower lottery priority group.

A student is coded as receiving a charter admission offer if he or she is offered a seat on lottery night. These offers are randomly assigned within strata defined by school, application year, and, in the case of UP Boston, BPS priority group. Students are retained the first year they apply to a charter school. We match the set of charter offers and randomization strata to state data by student name, grade, and application year; 97% of charter lottery applicants are successfully matched.

The BPS mechanism

We obtained a complete record of student-submitted preferences, school priorities, random tie-breaking sequence numbers, and assignments from the BPS deferred-acceptance mechanism, for seats in fall 2006 through fall 2013. For each application year, students are classified by priority (given by whether the applicant has an enrolled sibling and by walk-zone residence, a 1.5 mile radius)

at schools that they rank first. Students guaranteed admission by virtue of current enrollment, as well as certain other students with guaranteed or nonstandard priorities, are not subject to random assignment (see Abdulkadiroğlu et al. (2006) for a complete description of priorities in BPS). We construct indicators for whether an applying student was offered a seat; these offers are randomly assigned within strata defined by school, application year, and priority. We drop all schools with fewer than 50 students subject to conditionally-random admission, and match offers and randomization strata to state data via a BPS unique student identifier. Students are retained the first year they enter the BPS mechanism for sixth grade entry.

Sample Selection

We restrict attention to Boston public schools with at least 25 sixth grade students enrolled in each year of operation from 2006/7 to 2013/14. In our merged analysis file this leaves 51 schools (see Table 2.1). Students enrolled at these schools are retained if they were enrolled in Boston in both fifth and sixth grade, if their baseline demographic, attendance, and test score information is available, and if we observe their sixth grade MCAS test scores. These restrictions leave a total of 27,864 Boston students, summarized in detail in Table 2.2. Of these, 8,718 students are subject to quasi-experimental variation in sixth grade admission at 28 schools, either from a charter school lottery or from assignment by the BPS mechanism.

2.11 Econometric Appendix

Comparison of Compliance Groups

Figure 2.3 compares average predicted value-added for lottery compliers, always-takers and never-takers. Predicted value-added comes from a version of equation (5) that interacts the school dummies D_{ij} with race, gender, subsidized lunch, special education, English language learner status, and baseline score terciles. For a student with covariates X_i, this interacted VAM yields an estimate of a covariate-specific value-added parameter $\alpha_j(X_i)$ for each school j.

The arguments in Abadie (2003) imply that $E[\alpha_j(X_i)\kappa_{ij}^a]/E[\kappa_{ij}^a]$ equals the average value-added of school j for always-takers in lottery j, where $\kappa_{ij}^a = D_{ij}(1-Z_{ij})/(1-E[Z_{ij}|C_{ij}])$. Averages of value-added for never-taker and compliers are similarly given by $E[\alpha_j(X_i)\kappa_{ij}^n]/E[\kappa_{ij}^n]$ and $E[\alpha_j(X_i)(1 -$

$\kappa_{ij}^a - \kappa_{ij}^n)]/E[1 - \kappa_{ij}^a - \kappa_{ij}^n]$, for $\kappa_{ij}^n = (1 - D_{ij})Z_{ij}/E[Z_{ij}|C_{ij}]$. We construct points in Figure 2.3 based on the sample analogues of these quantities, using a saturated model for lottery strata to estimate $E[Z_{ij}|C_{ij}]$.

To adjust for first-step error in the estimation of $\alpha_j(X_i)$ and $E[Z_{ij}|C_{ij}]$, inference in Figure 2.3 uses a Bayesian bootstrap procedure (Rubin, 1981). The Bayesian bootstrap smooths bootstrap samples by reweighting rather than resampling observations, preventing the omission of small lottery strata that would occasionally be dropped in a standard nonparametric bootstrap. The Bayesian bootstrap used here is implemented by drawing vectors of Dirichlet$(1, ..., 1)$ weights, then re-estimating the interacted VAM and recomputing predicted value-added for compliers, always-takers and never-takers, weighting all moments with the Dirichlet weights. Inference for differences in means between compliance groups are based on the bootstrap covariance matrix of these differences across trials.

Simulated Minimum Distance

We estimate Bayesian hyperparameters via simulated minimum distance (SMD). The vector of parameters to be estimated is

$$\theta = \left(\alpha_0, \beta_0, \beta_Q, \delta_0, \xi_0, \Sigma, \sigma_\nu^2\right)'.$$

These parameters are estimated by fitting means, variances, and covariances of OLS value-added, lottery reduced form, and first stage estimates. The complete vector of observed estimates is

$$\hat{\Omega} = (\hat{\alpha}_1, ..., \hat{\alpha}_J, \hat{\rho}_1, ..., \hat{\rho}_L, \hat{\pi}_{11}, ..., \hat{\pi}_{L1}, ..., \hat{\pi}_{LJ})'.$$

Let $\Omega = (\alpha_1, ..., \pi_{LJ})'$ denote the probability limits of these estimates. Assume that the sampling distribution of $\hat{\Omega}$ is well approximated by asymptotic theory, so that

$$\hat{\Omega} \sim N(\Omega, V_e),$$

where V_e is a covariance matrix derived from conventional asymptotics. This requires within-school and within-lottery samples to be large enough for asymptotic approximations to be accurate. Under this assumption and the distributional assumptions in equations (2.12) through (2.15), values of Ω and $\hat{\Omega}$ can be simulated for any value of θ. We use this procedure to generate simulated data sets, and estimate θ by minimizing the distance between simulated and observed moments.

Our estimation procedure targets the following first moments:

$$\hat{m}_1 = \frac{1}{J}\sum_j \hat{\alpha}_j,$$

$$\hat{m}_2 = \frac{1}{L}\sum_j Q_j\hat{\alpha}_j,$$

$$\hat{m}_3 = \frac{1}{L}\sum_\ell \hat{\rho}_\ell,$$

$$\hat{m}_4 = \frac{1}{L}\sum_\ell \left(\frac{\hat{\rho}_\ell}{\hat{\pi}_{\ell\ell}}\right)$$

$$\hat{m}_5 = \frac{1}{L}\sum_\ell \hat{\psi}_\ell$$

$$\hat{m}_6 = \frac{1}{L}\sum_\ell \hat{\pi}_{\ell\ell},$$

$$\hat{m}_7 = -\frac{1}{J}\sum_j \frac{1}{L-Q_j}\sum_{\ell\neq j} \hat{\pi}_{\ell j},$$

$$\hat{m}_8 = -\frac{1}{J}\sum_\ell \frac{1}{L-Q_j}\sum_{j\neq\ell} \frac{\hat{\pi}_{\ell j}}{\hat{\pi}_{\ell\ell}},$$

$$\hat{m}_9 = \frac{1}{L}\sum_\ell \left[\frac{(\hat{\pi}_{\ell\ell})^2}{\sum_k (\hat{\pi}_{kk})^2}\right]\cdot\left(\frac{\hat{\rho}_\ell}{\hat{\psi}_\ell}\right).$$

$\hat{m}_1$ is the mean OLS coefficient, which provides information about $\beta_0 + b_0$, the sum of mean value-added and mean bias. $\hat{m}_2$ is the mean OLS coefficient among lottery schools, which helps to identify β_Q, the difference in value-added between lottery and non-lottery schools. $\hat{m}_3$ and $\hat{m}_4$ are the mean reduced form and IV coefficients, which provide information about β_0. $\hat{m}_5$ is the mean of a "pseudo-reduced form" prediction that uses OLS value-added estimates, given by $\hat{\psi}_\ell = \sum_j \hat{\pi}_{\ell j}\hat{\alpha}_j$. $\hat{m}_6$ is the mean first stage across lotteries, which can be used to estimate δ_0. $\hat{m}_7$ is the average fallback probability across lotteries, and $\hat{m}_8$ is the average ratio of this probability to the first stage, which gives the share of compliers drawn from included schools. These two moments help to estimate ξ_0, the mean fallback utility for included schools relative to the omitted school. $\hat{m}_9$ is the average ratio of the lottery reduced form to the pseudo-reduced form (the forecast coefficient). We weight this average by the squared lottery first stage to avoid unstable ratios caused by small first stages. This moment yields information about the variance of b_j, the bias in conventional value-added estimates, along with the correlation between β_j and b_j.

The next seven moments are variances of parameter estimates:

$$\hat{m}_{10} = \frac{1}{J}\sum_j (\hat{\alpha}_j - \bar{\alpha})^2,$$

$$\hat{m}_{11} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\rho}_\ell - \bar{\rho})^2,$$

$$\hat{m}_{12} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\psi}_\ell - \bar{\psi})^2,$$

$$\hat{m}_{13} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\pi}_{\ell\ell} - \bar{\pi}_{own})^2,$$

$$\hat{m}_{14} = \tfrac{1}{J}\textstyle\sum_j \left[\left(\frac{1}{L-Q_j}\sum_{\ell\neq j} \hat{\pi}_{\ell j}\right) - \bar{\pi}_{other}\right]^2,$$

$$\hat{m}_{15} = \tfrac{1}{J}\textstyle\sum_j \left[\left(\frac{1}{L-Q_j}\sum_{\ell\neq j} \frac{\hat{\pi}_{\ell j}}{\hat{\pi}_{\ell\ell}}\right) - \bar{s}_{other}\right]^2,$$

$$\hat{m}_{16} = \tfrac{1}{J(L-1)}\textstyle\sum_j \sum_{\ell\neq j} (\hat{\pi}_{\ell j} - \bar{\pi}_j)^2.$$

Here $\bar{\alpha}$ indicates the sample average of the α_j, and similarly for other variables. $\hat{m}_{10}$ is the variance of conventional value-added estimates across schools, which depends on the variances of value-added and bias as well as their covariance. $\hat{m}_{11}$ and $\hat{m}_{12}$ are variances of the lottery reduced form and predicted reduced form, which contain additional information about the joint distribution of value-added and bias. $\hat{m}_{13}$ is the variance of the first stage across lotteries, which helps to identify the variance of δ_j. $\hat{m}_{14}$ computes the mean share of students drawn from each school across lotteries, then takes the variance of this mean share across schools. This is the between-school variance in fallback probabilities. $\hat{m}_{15}$ is the variance of the mean share of compliers drawn from a particular school; $\bar{s}_{other}$ is the mean of this variable. These two moments yield information about the variances of ξ_j and $\nu_{\ell j}$, which govern heterogeneity in fallback probabilities. $\hat{m}_{16}$ computes the variance of fallback shares across lotteries at every school, then averages across schools. This is the average within-school variance in fallback probabilities. This moment helps to separate the variance of ξ_j, the school-specific mean fallback utility, from σ_ν^2, the variance of idiosyncratic school-by-lottery utility shocks.

Finally, we match six covariances:

$$\hat{m}_{17} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\rho}_\ell - \bar{\rho})(\hat{\alpha}_\ell - \bar{\alpha}),$$

$$\hat{m}_{18} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\rho}_\ell - \bar{\rho})\left(\hat{\psi}_\ell - \bar{\psi}\right),$$

$$\hat{m}_{19} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\rho}_\ell - \bar{\rho})(\hat{\pi}_{\ell\ell} - \bar{\pi}_{own}),$$

$$\hat{m}_{20} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\alpha}_\ell - \bar{\alpha})(\hat{\pi}_{\ell\ell} - \bar{\pi}_{own}),$$

$$\hat{m}_{21} = \tfrac{1}{L}\textstyle\sum_\ell (\hat{\rho}_\ell - \bar{\rho})\left[\left(\frac{1}{L-1}\sum_{k\neq\ell} \hat{\pi}_{k\ell}\right) - \bar{\pi}_{other}\right],$$

$$\hat{m}_{22} = \frac{1}{J}\sum_j (\hat{\alpha}_j - \bar{\alpha}) \left[\left(\frac{1}{L-Q_j} \sum_{\ell \neq j} \hat{\pi}_{\ell j} \right) - \bar{\pi}_{other} \right],$$

$$\hat{m}_{23} = \frac{1}{L}\sum_\ell (\hat{\pi}_{\ell\ell} - \bar{\pi}_{own}) \left[\left(\frac{1}{L-1} \sum_{k \neq \ell} \hat{\pi}_{k\ell} \right) - \bar{\pi}_{other} \right].$$

$\hat{m}_{17}$ and $\hat{m}_{18}$ are covariances of the reduced form with conventional value-added and the pseudo-reduced form, which help to identify variation in bias, as well as the covariance between bias and value-added. $\hat{m}_{19}$ is the covariance between reduced forms and first stages, which is informative about the covariance between β_j and δ_j. $\hat{m}_{20}$ is the covariance of conventional value-added and the first stage, which helps to identify the covariance between b_j and δ_j. $\hat{m}_{21}$ is the covariance of the reduced form and average fallback probability, which helps to identify the covariance of β_j and ξ_j. $\hat{m}_{22}$ is the covariance of OLS value-added with the average fallback probability, which depends on the covariance between b_j and ξ_j. $\hat{m}_{23}$ is the covariance of a school's first stage and average fallback probability, which provides information about the covariance of ξ_j and δ_j.

There are 16 elements of θ and 23 moments, so the model has seven overidentifying restrictions. Models that include charter and pilot school effects add sector-specific values of $\hat{m}_1$, $\hat{m}_3$, $\hat{m}_4$, $\hat{m}_5$, $\hat{m}_6$, $\hat{m}_7$ and $\hat{m}_8$, yielding 24 parameters and 37 moments. Let $\hat{m}$ be the vector of all observed moments, and let $\tilde{m}(\theta)$ be the corresponding vector of simulated predictions. The simulated minimum distance estimator with weighting matrix A is

$$\hat{\theta}_{SMD}(A) = \arg\min_\theta \; J\,(\hat{m} - \tilde{m}(\theta))' A\,(\hat{m} - \tilde{m}(\theta)).$$

The set of simulation draws used to construct $\tilde{m}(\theta)$ is held constant throughout the optimization. For each evaluation of the objective function the vector θ is used to transform these draws to have the appropriate distributions.

We produce a first-step estimate of θ with an identity weighting matrix, then use this estimate to compute a model-based covariance matrix by simulation. Altonji and Segal (1996) show that estimation error in the weighting matrix can generate finite-sample bias in two-step optimal minimum distance estimates. This bias is caused by correlation between the observations used to compute the moment conditions and those used to construct the weighting matrix. We therefore compute the model-based weighting matrix using a second set of simulation draws independent of the draws used to compute the moments. The weighting matrix is given by

$$\hat{A} = \left[J \cdot \frac{1}{R} \sum_r \left(\tilde{m}^r \left(\hat{\theta}_{SMD}(I) \right) - \bar{m} \right) \left(\tilde{m}^r \left(\hat{\theta}_{SMD}(I) \right) - \bar{m} \right)' \right]^{-1},$$

where r indexes a second independent set of $R = 1,000$ simulation draws and $\bar{m}$ is the mean of the simulated moments. An efficient two-step estimate is given by $\hat{\theta}_{SMD}\left(\hat{A}\right)$.

Under standard regularity conditions the minimized SMD criterion function follows a χ^2 distribution (Sargan, 1958; Hansen, 1982):

$$J\left(\hat{m} - \tilde{m}\left(\hat{\theta}_{SMD}(\hat{A})\right)\right)' \hat{A}\left(\hat{m} - \tilde{m}\left(\hat{\theta}_{SMD}(\hat{A})\right)\right) \sim \chi^2_q,$$

where q is the difference between the number of moments and the number of parameters to be estimated. Table VI reports this J-statistic.

Empirical Bayes Posteriors with a Known First Stage

With a know first stage matrix, Π, the posterior distribution for β_j and b_j can be derived analytically. In matrix form, the model can be written

$$\hat{\alpha} = \beta + b + e_\alpha,$$

$$\hat{\rho} = \Pi\beta + e_\rho,$$

$$(e'_\alpha, e'_\rho)|\beta, b \sim N(0, V_e),$$

$$(\beta', b')' \sim N\left((\iota'\beta_0, \iota' b_0)', V_\Theta\right),$$

where we have set $\beta_Q = 0$. The posterior density for the random coefficients $\Theta = (\beta, b)$ conditional on the observed estimates $\hat{\Omega} = (\hat{\alpha}, \hat{\rho})$ is given by

$$f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\theta\right) = \frac{f_{\hat{\Omega}|\Theta}\left(\hat{\Omega}|\Theta\right) f_\Theta\left(\Theta;\theta\right)}{f_{\hat{\Omega}}\left(\hat{\Omega};\theta\right)}.$$

The estimation errors and random coefficients are normally distributed, so we can write

$$-2\log f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\theta\right) = \left((\hat{\alpha} - \beta - b)', (\hat{\rho} - \Pi\beta)'\right) \begin{bmatrix} v_{\alpha\alpha} & v_{\alpha\rho} \\ v'_{\alpha\rho} & v_{\rho\rho} \end{bmatrix} \begin{pmatrix} \hat{\alpha} - \beta - b \\ \hat{\rho} - \Pi\beta \end{pmatrix}$$

$$+ \left((\beta - \beta_0\iota)', (b - b_0\iota)'\right) \begin{bmatrix} v_{\beta\beta} & v_{\beta b} \\ v'_{\beta b} & v_{bb} \end{bmatrix} \begin{pmatrix} \beta - \beta_0\iota \\ b - b_0\iota \end{pmatrix} + C_1,$$

where $v_{\alpha\alpha}$, $v_{\alpha\rho}$ and $v_{\rho\rho}$ are blocks of V_e^{-1}; $v_{\beta\beta}$, $v_{\beta b}$ and v_{bb} are blocks of V_Θ^{-1}; and C_1 is a constant that does not depend on Θ.

Rearranging this expression yields

$$-2\log f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\theta\right) = ((\beta-\beta^*)',(b-b^*)')\begin{bmatrix} v^*_{\beta\beta} & v^*_{\beta b} \\ v^{*\prime}_{\beta b} & v^*_{bb}\end{bmatrix}\begin{pmatrix}\beta-\beta^* \\ b-b^*\end{pmatrix} + C_2, \tag{2.22}$$

where C_2 is another constant. The parameters of this expression are

$$v^*_{\beta\beta} = v_{\alpha\alpha} + \Pi' v'_{\alpha\rho} + v_{\alpha\rho}\Pi + \Pi' v_{\rho\rho}\Pi + v_{\beta\beta},$$

$$v^*_{\beta b} = v_{\alpha\alpha} + \Pi' v'_{\alpha\rho} + v_{\beta b},$$

$$v^*_{bb} = v_{\alpha\alpha} + v_{bb},$$

and

$$\beta^* = W_\alpha(\hat{\alpha} - b_0\iota) + W_\rho\hat{\rho} + (I - W_\alpha - W_\rho\Pi)\beta_0\iota$$

with

$$W_\alpha = B^{-1}((v_{\alpha\alpha}+v_{bb})(v_{\alpha\alpha}+\Pi' v'_{\alpha\rho}+v_{\beta b})^{-1}(v_{\alpha\alpha}+\Pi' v'_{\alpha\rho}) - v_{\alpha\alpha}),$$

$$W_\rho = B^{-1}((v_{\alpha\alpha}+v_{bb})(v_{\alpha\alpha}+\Pi' v'_{\alpha\rho}+v_{\beta b})^{-1}(v_{\alpha\rho}+\Pi' v_{\rho\rho}) - v_{\alpha\rho}),$$

$$B = (v_{\alpha\alpha}+v_{bb})(v_{\alpha\alpha}+\Pi' v'_{\alpha\rho}+v_{\beta b})^{-1}(v_{\alpha\alpha}+\Pi' v'_{\alpha\rho}+v_{\alpha\rho}\Pi+\Pi' v_{\rho\rho}\Pi+v_{\beta\beta}) - (v_{\alpha\alpha}+v_{\alpha\rho}\Pi+v'_{\beta b}).$$

Equation (2.22) implies that the posterior for (β, b) is normal:

$$(\beta', b')'|\hat{\alpha},\hat{\rho} \sim N\left((\beta^{*\prime}, b^{*\prime})', V^*\right),$$

with

$$V^* = \begin{bmatrix} v^*_{\beta\beta} & v^*_{\beta b} \\ v^{*\prime}_{\beta b} & v^*_{bb}\end{bmatrix}^{-1}.$$

An empirical Bayes version of the posterior mean β^* is formed by plugging $\hat{\theta}_{SMD}$ and an estimate of V_e into the expressions for W_α and W_ρ.

Section 2.5.3 gives three special cases of the posterior mean. The first is when Π is invertible. Equation (2.18) is obtained by defining $W_\beta = W_\rho \Pi$ and substituting $W_\beta \Pi^{-1}$ for W_ρ in (2.17). The second special case adds the conditions that $Var(e_\alpha) = 0$ (α_j is known with certainty) and $Var(e_\beta) = \Pi^{-1} Var(e_\rho) \Pi^{-1\prime}$ is diagonal (sampling errors in IV estimates are independent across schools). In this case the only information in the sample about β_j comes from $(\alpha_j, \hat{\beta}_j)$ since β_j is uncorrelated with $\hat{\beta}_k$ and α_k for $k \neq j$. The vector $(\beta_j, \beta_j + b_j, \beta_j + e_j^\beta)$ is jointly normally distributed, so the posterior mean for β_j is the prediction from a linear regression of β_j on α_j and $\hat{\beta}_j$, given by:

$$\beta_j = \kappa_0 + \kappa_\alpha \alpha_j + \kappa_\beta \hat{\beta}_j + v_j.$$

Standard multivariate regression algebra implies the coefficients in this regression are

$$\kappa_\alpha = \frac{Var(\hat{\beta}_j) Cov(\alpha_j, \beta_j) - Cov(\hat{\beta}_j, \alpha_j) Cov(\beta_j, \hat{\beta}_j)}{Var(\alpha_j) Var(\hat{\beta}_j) - Cov(\alpha_j, \hat{\beta}_j)^2},$$

$$\kappa_\beta = \frac{Var(\alpha_j) Cov(\hat{\beta}_j, \beta_j) - Cov(\hat{\beta}_j, \alpha_j) Cov(\beta_j, \alpha_j)}{Var(\alpha_j) Var(\hat{\beta}_j) - Cov(\alpha_j, \hat{\beta}_j)^2},$$

$$\kappa_0 = E[\beta_j] - \kappa_\alpha E[\alpha_j] - \kappa_\beta E\left[\hat{\beta}_j\right].$$

Simplifying these expressions yields

$$\kappa_\alpha = \frac{Cov(\alpha_j, \beta_j)}{Var(\alpha_j)} \times \frac{Var(e_j^\beta)}{Var(e_j^\alpha) + \sigma_\beta^2 - \frac{Cov(\alpha_j, \hat{\beta}_j)^2}{Var(\alpha_j)}} = r_\alpha \times \frac{Var(e_j^\beta)}{Var(e_j^\beta) + \sigma_\beta^2 (1 - R^2)},$$

$$\kappa_\beta = \frac{\sigma_\beta^2 - \frac{Cov(\beta_j, \alpha_j)^2}{Var(\alpha_j)}}{Var(e_j^\alpha) + \sigma_\beta^2 - \frac{Cov(\alpha_j, \hat{\beta}_j)^2}{Var(\alpha_j)}} = \frac{\sigma_\beta^2 (1 - R^2)}{Var(e_j^\beta) + \sigma_\beta^2 (1 - R^2)},$$

$$\kappa_0 = (1 - \kappa_\alpha - \kappa_\beta)\beta_0 - \kappa_\alpha b_0 = (1 - \kappa_\beta)(1 - r_\alpha)\beta_0 - (1 - \kappa_\beta) r_\alpha b_0,$$

where $r_\alpha = Cov(\alpha_j, \beta_j)/Var(\alpha_j)$ and $R^2 = Cov(\alpha_j, \beta_j)^2 / (Var(\alpha_j) Var(\beta_j))$. These are the coefficients in equation (2.19).

The third special case is when lotteries provide no information about β_j (so $Cov(\hat{\rho}_\ell, \beta_j) = 0 \; \forall \ell$) and conventional VAM sampling errors are uncorrelated (so $Cov(e_j^\alpha, e_k^\alpha) = 0 \; \forall k \neq j$). In this case the posterior mean is simply the regression of β_j on $\hat{\alpha}_j$:

$$\beta_j = \tilde{\kappa}_0 + \tilde{\kappa}_\alpha \hat{\alpha}_j + \tilde{v}_j,$$

which has coefficients

$$\tilde{\kappa}_\alpha = \frac{Cov(\beta_j, \hat{\alpha}_j)}{Var(\hat{\alpha}_j)},$$

$$\tilde{\kappa}_0 = E[\beta_j] - \tilde{\kappa}_\alpha E[\alpha_j].$$

Simplifying these yields

$$\tilde{\kappa}_\alpha = \frac{Cov(\beta_j, \beta_j + b_j + e_j^\alpha)}{Var(\beta_j + b_j + e_j^\alpha)} = \frac{\sigma_\beta^2 + \sigma_{\beta b}}{\sigma_\beta^2 + \sigma_b^2 + 2\sigma_{\beta b} + Var(e_j^\alpha)},$$

$$\tilde{\kappa}_0 = (1 - \tilde{\kappa}_\alpha)\beta_0 - \tilde{\kappa}_\alpha b_0,$$

which are the coefficients in equation (2.20).

Empirical Bayes Posterior Modes

In practice the first stage matrix Π is unknown and must be estimated. The vector of unknown school-specific parameters is then

$$\Theta = (\beta_1, b_1, \delta_1, \xi_1, ..., \beta_J, b_J, \delta_J, \xi_J, \nu_{11}, ..., \nu_{LJ})'.$$

Up to a scaling constant, the posterior density for Θ conditional on the observed estimates $\hat{\Omega}$ (which now include the estimated $\hat{\pi}_{\ell j}$) and the prior parameters θ can be expressed

$$f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\theta\right) \propto \phi_m\left(\hat{\Omega} - \Omega(\Theta); V\right) \phi_m\left(\Theta - \bar{\Theta}(\theta); \Gamma(\theta)\right), \tag{2.23}$$

where

$$\bar{\Theta}(\theta) = (\beta_0 + \beta_Q, b_0, \delta_0, \xi_0, ...\beta_0, b_0, \delta_0, \xi_0, 0, ...0)',$$

$\phi_m(x; v)$ is the multivariate normal density function with mean zero and covariance matrix v, and

$$\Gamma(\theta) = \begin{bmatrix} I_J \otimes \Sigma & 0 \\ 0 & \sigma_\nu^2 I_{LJ} \end{bmatrix},$$

where I_J and I_{LJ} are identity matrices of dimension J and $L \times J$. Note that the probability limit of the vector of observed estimates, Ω, is a function of Θ, so we write $\Omega(\Theta)$.

As before we form an empirical Bayes posterior density by plugging $\hat{\theta}_{SMD}$ into equation (2.23) . The empirical Bayes posterior mean is

$$\Theta^*_{mean} = \int \Theta f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\hat{\theta}_{SMD}\right) d\Theta.$$

Since the first stage parameters $\pi_{\ell j}$ are nonlinear functions of δ and ξ, the density in equation (2.23) will not generally be normal. As a result the integral for the posterior mean does not have a closed form and it is not possible to sample directly from the posterior distribution. To avoid integration we instead work with the posterior mode:

$$\Theta^*_{mode} = \arg\max_{\Theta} \log\phi_m\left(\hat{\Omega} - \Omega(\Theta); V_e\right) + \log\phi_m\left(\Theta - \bar{\Theta}\left(\hat{\theta}_{SMD}\right); \Gamma\left(\hat{\theta}_{SMD}\right)\right).$$

The posterior mode coincides with the posterior mean in the fixed first stage case where the posterior distribution is normal. The mode is computationally convenient in the estimated first stage case, as it simply requires solving a regularized maximum likelihood problem.

We compare posterior modes for the β_j with conventional empirical Bayes posterior means based on OLS estimates of value-added. The conventional predictions are given by

$$\alpha^*_{j.} = \left(\frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_\alpha + Var(e^\alpha_j)}\right)\hat{\alpha}_j + \left(1 - \frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_\alpha + Var(e^\alpha_j)}\right)\hat{\alpha}_0, \tag{2.24}$$

where

$$\hat{\alpha}_0 = \frac{1}{J}\sum_j \hat{\alpha}_j,$$

$$\hat{\sigma}^2_\alpha = \frac{1}{J}\sum_j \left[(\hat{\alpha}_j - \hat{\mu}_\alpha)^2 - Var\left(e^\alpha_j\right)\right].$$

Models with sector effects replace $\hat{\alpha}_0$ in equation (2.24) with the regression predictions

$$\hat{\alpha}_{0j} = S_j'\left[\tfrac{1}{J}\sum_k S_k S_k'\right]^{-1}\left[\tfrac{1}{J}\sum_k S_k\hat{\alpha}_k\right],$$

where S_j is a vector including a constant and charter and pilot school indicators.

Relationship Between Forecast Coefficient and VAM Reliability

Here we derive the relationship between the probability limit of the IV forecast coefficient, φ, and the VAM reliability ratio, $r_\alpha = Cov(\beta_j, \alpha_j)/Var(\alpha_j)$. The IV model that generates φ is

$$Y_i = \Delta_0 + C_i'\Delta_c + \varphi\hat{Y}_i + \zeta_i.$$

The corresponding reduced form is

$$Y_i = \tau_0 + C_i'\tau_c + Z_i'\rho + u_i,$$

while the first stage is

$$\hat{Y}_i = \tilde{\tau}_0 + C_i'\tilde{\tau}_c + Z_i'\psi + \tilde{u}_i.$$

When C_i is a set of mutually exclusive and exhaustive indicator variables for participation in L lotteries, Theorem 4.5.1 in Angrist and Pischke (2009) implies that 2SLS estimation of this system yields the probability limit

$$\varphi = \sum_{\ell=1}^{L} \left(\frac{\omega_\ell}{\sum_{\ell'} \omega_{\ell'}} \right) \left(\frac{\rho_\ell}{\psi_\ell} \right),$$

where ρ_ℓ and ψ_ℓ are the elements of ρ and ψ corresponding to $Z_{i\ell}$, and

$$\omega_\ell = Pr\left[C_{i\ell} = 1\right] Var(Z_{i\ell}|C_{i\ell} = 1) \left(\psi_\ell\right)^2.$$

This equation shows that the forecast coefficient generated by an overidentified instrumental variables model equals a particular weighted average of lottery-specific forecast coefficients.

The equation for the forecast coefficient can be rewritten

$$\varphi = \frac{\sum_{\ell=1}^{L} \left(\frac{\tilde{\omega}_\ell}{\sum_k \tilde{\omega}_k} \right) \rho_\ell \psi_\ell}{\sum_{\ell=1}^{L} \left(\frac{\tilde{\omega}_\ell}{\sum_k \tilde{\omega}_k} \right) \left(\psi_\ell\right)^2},$$

where $\tilde{\omega}_\ell = Pr\left[C_{i\ell} = 1\right] Var\left(Z_{i\ell}|C_{i\ell} = 1\right)$. This expression shows that φ may be written as the coefficient from a weighted least squares regression through the origin of reduced form effects on test scores, ρ_ℓ, on first-stage effects on predicted value-added, ψ_ℓ.

In the notation of the random coefficients model, these reduced form and first stage effects are given by

$$\rho_\ell = \sum_{j=1}^{J} \pi_{\ell j} \beta_j,$$

$$\psi_\ell = \sum_{j=1}^{J} \pi_{\ell j} \alpha_j.$$

In a scenario with $E\left[\psi_\ell\right] = 0$ and the number of schools tending to infinity, we can then write

$$\varphi = \frac{Cov\left(\sqrt{\tilde{\omega}_\ell}\sum_{j=1}^{J}\pi_{\ell j}\beta_j, \sqrt{\tilde{\omega}_\ell}\sum_{j=1}^{J}\pi_{\ell j}\alpha_j\right)}{Var\left(\sqrt{\tilde{\omega}_\ell}\sum_{j=1}^{J}\pi_{\ell j}\alpha_j\right)}.$$

This expression shows that the forecast coefficient φ is a weighted regression of linear combinations of the β_j's on the same linear combinations of the α_j's. The weights depend on the size of each lottery and the offer rate, while the first stage coefficients that form the linear combinations depend on offer takeup rates and the distribution of fallback schools for lottery applicants. Although both φ and r_α measure the correlation of OLS and causal value-added, and coincide when the expectation of β_j given α_j is linear and the first stage parameters $\pi_{\ell j}$ are non-stochastic, in general these summary statistics should be expected to differ.

2.12 Supplementary Results

Results for Seventh and Eighth Grade

As for the bias tests reported in Table 2.3, school effects on seventh and eighth grade test scores are modeled as linear in the number of years spent in each school. The random coefficients framework of Section 2.5.1 is adapted to data from seventh grade by modifying the lottery first stage estimand as

$$\pi_{\ell\ell} = 2 \times \frac{\exp(\delta_\ell)}{1 + \exp(\delta_\ell)}.$$

The remaining equations describing the random coefficients model are unchanged. This specification guarantees that the effects of lottery offers on time spent in each school are less than two years in absolute value, which is the maximum potential attendance through seventh grade. Likewise, the model for eighth grade uses three years of potential attendance. The value-added and bias parameters, β_j and b_j, may then be interpreted as causal effects and VAM bias associated with one additional year of attendance at school j.

Appendix Table 2.A4 reports math hyperparameter estimates separately by grade. The results for seventh and eighth grade are qualitatively similar to those for sixth. Standard deviations of causal value-added in each grade are somewhat larger than the corresponding bias standard deviations, and covariances between value-added and bias are uniformly negative. Standard deviations

of annual school effects are smaller for the higher grades, which suggests there is some concavity in the relationship between achievement and years of exposure to a particular school. Similarly, differences in value-added between lottery and non-lottery schools and between charter and traditional schools are positive for seventh and eighth grade but smaller than these differences for sixth grade.

Appendix Table 2.A5 shows policy simulation results for school closure decisions based on seventh and eighth grade outcomes. The reported impacts are effects of one extra year spent at the replacement school rather than the closed school. Like the sixth grade results from Table 2.8, the simulations for higher grades show large gains associated with using conventional value-added models for accountability decisions. For example, replacing the lowest-performing district school according to the lagged score model with a typical top quintile school is predicted to generate an impact of 0.24σ per year on eighth grade scores, 63% of the gain attainable with knowledge of true value-added (0.38σ). Hybrid estimation boosts this effect to 0.29σ, a 22% improvement over the conventional model.

Models with Time-varying Value-added

The hyperparameter estimates reported in Table 2.6 are estimated under the assumption that causal school quality and bias are stable over time. Chetty et al. (2014a) document temporal instability in conventional teacher VAM estimates; a model that presumes constant school quality may be inappropriate if school value-added is similarly unstable.

To probe the stability of school value-added, we report estimates from a model allowing school effects to vary by year, fit to year-specific OLS and lottery estimates. This model is based on the specification

$$\beta_{jt} = \beta_j + \tilde{\beta}_{jt},$$

$$b_{jt} = b_j + \tilde{b}_{jt},$$

where (β_j, b_j) are joint normal as in equation (2.14), and $\tilde{\beta}_{jt}$ and $\tilde{b}_{jt}$ are *iid* uncorrelated normal shocks with mean zero and standard deviations $\sigma_{\tilde{\beta}}$ and $\sigma_{\tilde{b}}$. The first stage mean utility parameters δ_j and ξ_j are assumed stable over time, so changes in the first stage are captured by the idiosyncratic shocks $\nu_{\ell jt}$. The simulated minimum distance procedure uses time averages of the moments listed in the econometric appendix, and is augmented with variances of the year to year changes in $\hat{\alpha}_{jt}$,

$\hat{\tau}_{\ell t}$ and $\hat{\rho}_{\ell t}$ in order to estimate the standard deviations of the idiosyncratic value-added and bias shocks.

Minimum distance estimates from the model with time-varying value-added appear in Appendix Table 2.A8. These estimates suggest that the permanent components of value-added and bias are more important than the idiosyncratic components. Estimated standard deviations of the permanent component of value-added are between 0.17σ and 0.22σ across models, roughly similar to the corresponding estimates from Table 2.6. Estimated standard deviations of the idiosyncratic component are around 0.1σ in each model. Likewise, estimated standard deviations of the permanent component of bias equal 0.41σ, 0.21σ and 0.20σ, compared to 0.05σ, 0.07σ and 0.06σ. We cannot reject the null hypothesis that bias is constant over time at conventional levels. These results suggest that school value-added and bias are reasonably stable across years, so our preferred specifications use the more parsimonious model that abstracts from time variation.

Misclassification Results

Like many states and school districts, the Massachusetts Department of Elementary and Secondary Education implements an accountability scheme based on standardized tests. Massachusetts' Framework for School Accountability and Assistance places schools into five "levels" based on four-year histories of test score levels and changes. Schools in the bottom quintile of this measure are designated level 3 or higher. A subset of these schools are classified in levels 4 and 5, a designation that puts them at risk of restructuring or closure.[25] Appendix Table 2.A9 uses the simulations described in Section 2.7 to calculate the frequency of classification errors in accountability schemes of this sort.

Uncontrolled value-added estimates produce highly inaccurate school rankings. As can be seen in the second row of Table 2.A9, uncontrolled VAM misclassifies 86% of lowest decile schools, 71% of lowest quintile schools, and 59% of lowest tercile schools. These rates are not much better than the error rates for a policy that simply ranks schools randomly (90%, 80% and 67%, shown in the first row). Hybrid posterior modes that combine uncontrolled OLS and lottery estimates misclassify 73%, 45% and 36% of lowest decile, quintile and tercile schools. Although still high, these error rates represent a marked improvement on the rates produced by the conventional posterior mean

[25]The Massachusetts accountability system also uses information on graduation, dropout rates and from site visits to classify schools; see `http://www.doe.mass.edu/apa/sss/turnaround/level5/schools/FAQ.html` for details.

from an uncontrolled model.

Adding controls for demographics and previous achievement reduces misclassification rates based on both conventional and hybrid estimates. Conventional misclassification rates for lowest decile, quintile and tercile schools are 59%, 47% and 38% when rankings are based on estimates from the gains specification. In this model, hybrid estimation reduces classification error in the lowest decile from 59% to 41%, 31% fewer mistakes. The hybrid advantages in classifying lowest quintile and lowest tercile schools equal 38% and 39% in the gains specification. The pattern of classification improvement from the lagged score and gains specifications are broadly similar. For both the lagged score and gains models, hybrid estimation cuts mistakes in classifying upper and lower tercile schools to under one third.

The relationship between school rankings based on true and estimated value-added summarizes the predictive value of VAM estimates. Column 7 of Table 2.A9 reports coefficients from regressions of a school's rank in the causal value-added distribution on its rank in each estimated distribution. This rank coefficient increases from 0.15 in the uncontrolled conventional model to 0.61 in the conventional gains specification. Hybrid estimation boosts the rank coefficient for gains to 0.84. In other words, sufficiently controlled VAM estimates strongly predict relative value-added: a one-position increase in a school's VAM rank translates into an average increase of roughly 0.8 positions in the distribution of true school quality.

Sensitivity Analysis for Bias Assumption

As noted in Section 2.5, a key assumption underlying the hybrid approach is that bias distributions are the same for lottery and non-lottery schools. It is worth documenting the sensitivity of our results to violations of this assumption. To this end, we simulate versions of the model in which the parameters of the bias distribution differ for non-lottery schools. Policymakers in these simulations continue to presume that there is no difference between these groups.

As can be seen in Appendix Table 2.A10, realistic differences in bias distributions between lottery and non-lottery schools tend to modestly degrade the performance of hybrid posterior estimates. The second and third rows report results from simulations that set the mean bias, b_0, 0.2σ higher or lower for non-lottery schools. As shown in Table 2.6, these differences are roughly one standard deviation in the distribution of causal school quality, and more than one standard deviation in the distribution of bias. These changes cause root mean squared error for the hybrid estimator to grow

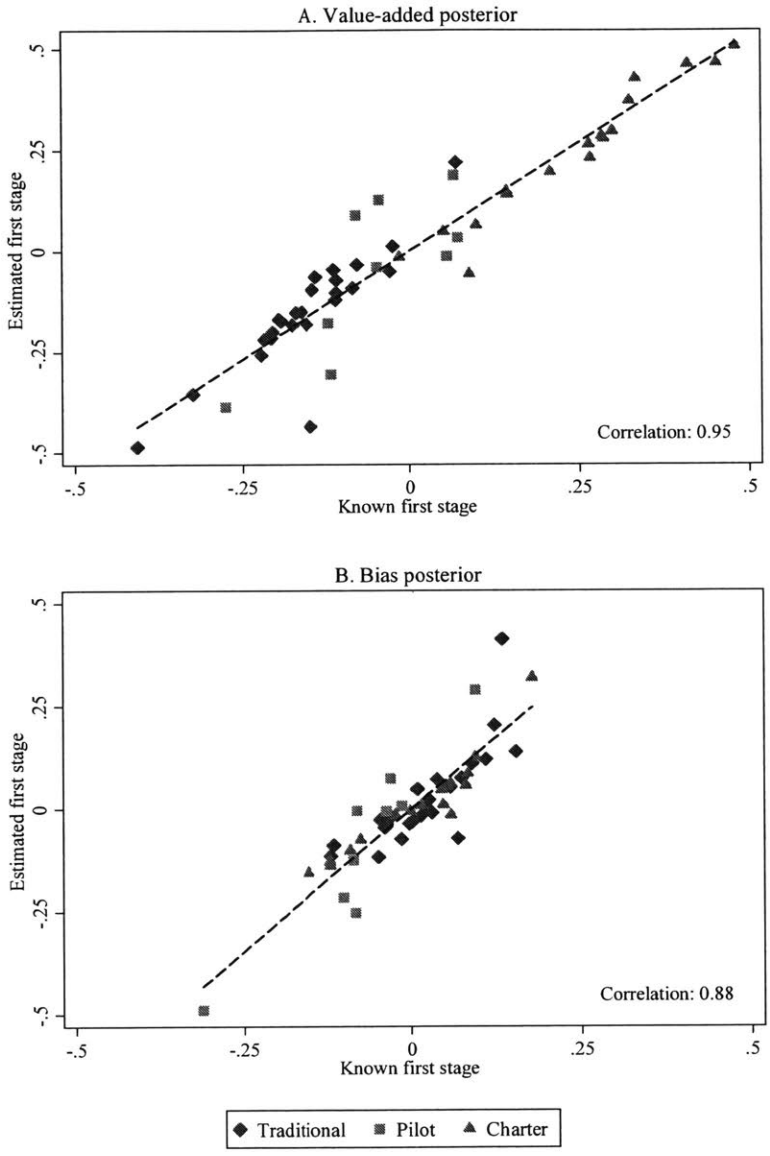
from 0.12σ to 0.15σ in the lagged score specification and 0.10σ to 0.14σ in the gains model.

The remaining rows of Table 2.A10 display the effects of changing the standard deviation of bias, σ_b, and the covariance between value-added and bias, $\sigma_{\beta b}$, for non-lottery schools. Doubling σ_b for schools without lotteries increases RMSE to 0.16σ and 0.14σ in the lagged score and gains models. As shown in the fifth row, cutting σ_b in half for non-lottery schools improves the performance of hybrid estimates; in this case, misspecification error is outweighted by the decline in the magnitude of bias for non-lottery schools. The final row shows that reversing the sign of $\sigma_{\beta b}$ for non-lottery schools increases RMSE to 0.13σ and 0.12σ in the lagged score and gains specifications.

Columns 2-5 of Table 2.A10 show that increases in RMSE due to misspecification of the bias distribution are accompanied by decreases in the expected benefits of school closure decisions based on the hybrid estimates. At the same time, these benefits remain substantial in all simulations. When σ_b is doubled for non-lottery schools, for example, a policy that uses hybrid gains estimates to replace the lowest-ranked school with an average school generates an average improvement of 0.24σ. The results in Table 2.A10 indicate that hybrid estimation is likely to be of value as long as differences in bias distributions between lottery and non-lottery schools are modest.

2.13 Appendix Figures and Tables

Figure 2.A1: Posterior Predictions With and Without a Known First Stage



Notes: This figure displays the correlation between posterior predictions of value-added and bias when the lottery first stage is treated as known vs. estimated. Estimates come from the lagged score value-added model with sector effects for sixth grade math scores. The horizontal axis in each panel displays posterior means computed under the assumption that there is no sampling error in the first stage coefficients. The vertical axis in each panel displays posterior modes accounting for estimation error in the first stage. Dashes show OLS lines of best fit.

Table 2.A1: Lottery Attrition

		Offer balance			
	Mean	All lotteries	Traditional	Pilot	Charter
	(1)	(2)	(3)	(4)	(5)
In lottery sample	0.813	0.028	0.036	-0.003	0.010
		(0.010)	(0.011)	(0.023)	(0.015)
N	10,718	10,718	5,589	1,512	4,867

Notes: This table reports the followup rate for the lottery sample and investigates differential attrition by lottery offer status. Column 1 shows the fraction of randomized lottery applicants that appear in the Boston sixth grade sample. Columns 2-5 report coefficients from regressions of an indicator for followup on lottery offers, controlling for lottery strata. Robust standard errors are reported in parenthenses.

Table 2.A2: Tests for Bias in ELA Value-added Models

	All lotteries		Excluding charter lotteries	
	Lagged score	Gains	Lagged score	Gains
	(1)	(2)	(3)	(4)
	A. Sixth grade			
Forecast coefficient (φ)	0.864	0.722	0.423	0.133
	(0.167)	(0.172)	(0.310)	(0.308)
First stage F-statistic	26.8	29.4	14.0	13.1
p-values:				
Forecast bias	0.416	0.105	0.063	0.005
Overidentification	0.039	0.007	0.157	0.127
Omnibus test χ^2 statistic (d.f.)	46.0 (28)	56.8 (28)	36.1 (23)	43.2 (23)
p-value	0.018	0.001	0.040	0.007
N	8,718		6,162	
	B. All middle school grades			
Forecast coefficient (φ)	0.969	0.924	0.699	0.550
	(0.101)	(0.107)	(0.193)	(0.191)
First stage F-statistic	11.3	12.3	6.7	6.6
p-values:				
Forecast bias	0.759	0.481	0.118	0.019
Overidentification	0.062	0.014	0.122	0.081
Omnibus test χ^2 statistic (d.f.)	119.0 (75)	137.1 (75)	80.7 (60)	92.9 (60)
p-value	<0.001	<0.001	0.039	0.004
N	20,935		15,027	

Notes: This table reports the results of tests for bias in conventional value-added models for sixth through eighth grade ELA scores. The notes to Table 2.3 describe the underlying models and test procedure. Standard errors, clustered by student, are reported in parentheses.

Table 2.A3: Joint Distribution of Value-added, Bias, and Lottery Compliance

	β_j	b_j	δ_j	ξ_j
	(1)	(2)	(3)	(4)
Standard deviation	0.171	0.148	0.764	0.864
	(0.028)	(0.029)	(0.131)	(0.584)
Covariance w/b_j	-0.016			
	(0.006)			
Covariance w/δ_j	0.009	0.043		
	(0.018)	(0.032)		
Covariance w/ξ_j	0.077	-0.102	-0.491	
	(0.029)	(0.056)	(0.151)	
Charter effect	0.426	-0.005	0.241	-1.934
	(0.104)	(0.103)	(0.387)	(0.425)
Pilot effect	0.130	-0.121	0.074	-0.479
	(0.129)	(0.124)	(0.312)	(0.434)
Std. dev. of ν_{ij}		1.566		
		(0.152)		

Notes: This table reports simulated minimum distance estimates of parameters governing the distribution of value-added, bias, and lottery compliance probabilities for the lagged score value-added model fit to sixth grade school attendance and math scores. The notes to Table 2.6 describe the estimation procedure.

Table 2.A4: Minimum Distance Estimates by Grade

		Sixth grade		Seventh grade		Eighth grade	
		Lagged score	Gains	Lagged score	Gains	Lagged score	Gains
		(1)	(2)	(3)	(4)	(5)	(6)
σ_β	Std. dev. of causal VA	0.171	0.170	0.137	0.120	0.109	0.101
		(0.028)	(0.023)	(0.022)	(0.021)	(0.019)	(0.018)
σ_b	Std. dev. of OLS bias	0.148	0.133	0.119	0.094	0.079	0.078
		(0.029)	(0.030)	(0.040)	(0.032)	(0.025)	(0.025)
$\sigma_{\beta b}$	Covariance of	-0.016	-0.013	-0.007	-0.006	-0.004	-0.005
	VA and bias	(0.006)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)
r_α	Regression of VA	0.694	0.783	0.625	0.747	0.771	0.798
	on OLS (reliability ratio)	(0.152)	(0.122)	(0.084)	(0.096)	(0.108)	(0.105)
VA shifters	Charter	0.426	0.396	0.210	0.192	0.145	0.133
		(0.104)	(0.106)	(0.091)	(0.081)	(0.075)	(0.073)
	Pilot	0.130	0.111	-0.039	-0.019	0.004	0.008
		(0.129)	(0.129)	(0.110)	(0.101)	(0.085)	(0.082)
	Lottery school (β_Q)	0.104	0.066	0.003	0.034	0.047	0.056
		(0.042)	(0.041)	(0.042)	(0.033)	(0.027)	(0.027)
Bias shifters	Charter	-0.005	-0.063	0.010	0.030	0.049	0.039
		(0.103)	(0.099)	(0.088)	(0.077)	(0.070)	(0.073)
	Pilot	-0.121	-0.089	0.009	0.062	-0.036	0.011
		(0.124)	(0.121)	(0.107)	(0.097)	(0.078)	(0.077)
	$\chi^2(13)$ statistic:	9.0	6.0	9.2	4.5	6.4	8.8
	Overid. p-value:	0.773	0.946	0.759	0.985	0.932	0.785

Notes: This table reports minimum distance estimates of parameters of the joint distribution of causal school value-added and OLS bias for each middle school grade. School exposure for seventh and eighth grade is measured as the number of years spent in each school. The notes to Table 2.6 describe the estimation procedure.

Table 2.A5: Per-year Effects of Closing the Lowest-ranked District School for Affected Children, by Grade

			Replacement school:			
Grade	Model	Posterior method	Average school (1)	Average above-median school (2)	Average top-quintile school (3)	Average charter school (4)
Seventh		True value-added	0.284	0.389	0.468	0.500
			[0.059]	[0.067]	[0.073]	[0.076]
	Lagged score	Conventional	0.187	0.253	0.299	0.403
			[0.108]	[0.116]	[0.119]	[0.116]
		Hybrid	0.225	0.301	0.356	0.441
			[0.101]	[0.108]	[0.113]	[0.112]
	Gains	Conventional	0.157	0.215	0.259	0.344
			[0.094]	[0.101]	[0.103]	[0.101]
		Hybrid	0.190	0.257	0.311	0.377
			[0.088]	[0.096]	[0.103]	[0.097]
Eighth		True value-added	0.229	0.316	0.384	0.369
			[0.049]	[0.055]	[0.059]	[0.062]
	Lagged score	Conventional	0.162	0.226	0.273	0.302
			[0.083]	[0.088]	[0.092]	[0.092]
		Hybrid	0.193	0.267	0.325	0.333
			[0.071]	[0.078]	[0.083]	[0.079]
	Gains	Conventional	0.142	0.199	0.241	0.268
			[0.080]	[0.084]	[0.088]	[0.087]
		Hybrid	0.171	0.238	0.293	0.297
			[0.079]	[0.088]	[0.093]	[0.087]

Notes: This table reports simulated test score impacts of closing the lowest-ranked BPS district school based on value-added predictions for seventh and eighth grade. The reported effects are average impacts of one year of attendance at the replacement school rather than the closed school. Standard deviations of these effects across simulations appear in brackets. The notes to Table 2.8 describe the simulation procedure.

Table 2.A6: Tests for Bias in Hybrid Value-added Posteriors

	All lotteries		Excluding charter lotteries	
	Lagged score	Gains	Lagged score	Gains
	(1)	(2)	(3)	(4)
Forecast coefficient (φ)	1.040	1.001	1.143	1.086
	(0.130)	(0.129)	(0.316)	(0.308)
First stage F-statistic	24.0	23.6	7.5	7.3
p-values:				
Forecast bias	0.760	0.994	0.651	0.780
Overidentification	0.849	0.710	0.808	0.783
Omnibus test χ^2 statistic (d.f.)	0.7 (28)	0.8 (28)	0.8 (23)	0.8 (23)
p-value	0.841	0.701	0.795	0.775
N	8,718		6,162	

Notes: This table reports the results of tests for bias in posterior value-added predictions for sixth grade math scores. Empirical Bayes posterior modes come from random coefficient models with sector effects. The notes to Table 2.3 describe the value-added models and test procedure. Robust standard errors are reported in parentheses.

Table 2.A7: Consequences of Closing the Lowest-ranked District School Based on Four Years of Data

		Replacement school			
Model	Posterior method	Average school (1)	Average above-median school (2)	Average top-quintile school (3)	Average charter school (4)
	True value-added	0.370	0.507	0.610	0.711
		[0.080]	[0.089]	[0.094]	[0.094]
Uncontrolled	Conventional	0.051	0.071	0.087	0.274
		[0.193]	[0.200]	[0.206]	[0.199]
	Hybrid	0.148	0.217	0.256	0.371
		[0.143]	[0.154]	[0.168]	[0.151]
Lagged score	Conventional	0.223	0.305	0.364	0.575
		[0.152]	[0.162]	[0.172]	[0.158]
	Hybrid	0.302	0.420	0.511	0.651
		[0.134]	[0.146]	[0.154]	[0.146]
Gains	Conventional	0.229	0.315	0.380	0.570
		[0.144]	[0.153]	[0.162]	[0.149]
	Hybrid	0.302	0.417	0.506	0.641
		[0.122]	[0.133]	[0.141]	[0.133]

Notes: This table reports simulated test score impacts of closing the lowest-ranked BPS district school based on value-added predictions computed from four years of data. These effects are computed by scaling up the covariance matrix of sampling errors underlying the simulations in Table 2.8 by a factor of two. The notes to Table 2.8 describe the simulation procedure.

Table 2.A8: Models with Time-varying Value-added and Bias

		Uncontrolled	Lagged score	Gains
		(1)	(2)	(3)
σ_β	Std. dev. of causal VA (permanent)	0.168	0.215	0.193
		(0.034)	(0.027)	(0.024)
	Std. dev. of causal VA (transitory)	0.103	0.084	0.091
		(0.050)	(0.043)	(0.040)
σ_b	Std. dev. of OLS bias (permanent)	0.414	0.214	0.199
		(0.018)	(0.025)	(0.023)
	Std. dev. of OLS bias (transitory)	0.046	0.066	0.066
		(0.164)	(0.062)	(0.064)
$\sigma_{\beta b}$	Covariance of VA and bias	-0.034	-0.037	-0.029
		(0.008)	(0.007)	(0.005)
r_α	Regression of VA on OLS (reliability ratio)	-0.041	0.510	0.431
		(0.189)	(0.182)	(0.147)
VA shifters	Charter	0.263	0.487	0.354
		(0.129)	(0.153)	(0.134)
	Pilot	-0.024	0.078	0.091
		(0.145)	(0.170)	(0.140)
	Lottery school (β_Q)	0.178	0.133	0.127
		(0.136)	(0.061)	(0.053)
Bias shifters	Charter	0.324	-0.043	0.020
		(0.179)	(0.151)	(0.138)
	Pilot	-0.151	-0.157	-0.117
		(0.204)	(0.170)	(0.139)
	$\chi^2(14)$ statistic:	15.5	12.1	15.5
	Overid. p- value:	0.347	0.597	0.343

Notes: This table reports simulated minimum distance estimates of parameters of the joint distribution of causal school value-added and OLS bias for sixth grade math scores from a model that allows school effects to vary by year. School value-added and bias are assumed to consist of a permanent component plus an independent and identically distributed transitory shock each year. The notes to Table 2.6 describe the estimation procedure.

Table 2.A9: Error Rates for Classification Decisions Among District Schools

Value-added model	Posterior method	Low-performing schools: Lowest decile (1)	Low-performing schools: quintile (2)	Low-performing schools: tercile (3)	High-performing schools: decile (4)	High-performing schools: quintile (5)	High-performing schools: tercile (6)	coefficient (7)
	Random	0.900	0.800	0.667	0.900	0.800	0.667	0.000
		[0.161]	[0.139]	[0.118]	[0.161]	[0.139]	[0.118]	[0.177]
Uncontrolled	Conventional	0.857	0.710	0.593	0.863	0.733	0.608	0.146
		[0.190]	[0.150]	[0.122]	[0.182]	[0.151]	[0.122]	[0.171]
	Hybrid	0.726	0.449	0.359	0.781	0.606	0.491	0.502
		[0.246]	[0.176]	[0.126]	[0.226]	[0.161]	[0.126]	[0.171]
Lagged score	Conventional	0.639	0.501	0.411	0.670	0.523	0.422	0.542
		[0.256]	[0.155]	[0.121]	[0.246]	[0.154]	[0.113]	[0.134]
	Hybrid	0.438	0.316	0.249	0.382	0.305	0.234	0.825
		[0.252]	[0.141]	[0.106]	[0.231]	[0.146]	[0.105]	[0.087]
Gains	Conventional	0.594	0.469	0.379	0.611	0.483	0.393	0.606
		[0.260]	[0.152]	[0.115]	[0.483]	[0.152]	[0.112]	[0.123]
	Hybrid	0.411	0.293	0.232	0.350	0.286	0.225	0.841
		[0.237]	[0.137]	[0.103]	[0.243]	[0.147]	[0.108]	[0.096]

Notes: This table reports simulated misclassification rates for policies based on empirical Bayes posterior predictions of value-added. The first row shows results for a system that ranks schools at random. Column 1 shows the fraction of district schools in the lowest decile of true sixth grade math value-added that are not classified in the lowest decile of estimated value-added for each model. Columns 2 and 3 report corresponding misclassification rates for the lowest quintile and tercile. Columns 4-6 report misclassification rates for schools in the highest decile, quintile and tercile of true value-added. Column 7 reports the coefficient from a regression of a school's rank in the true value-added distribution on its rank in the estimated distribution. Standard deviations of misclassification rates and rank coefficients across simulations appear in brackets.

Table 2.A10: Sensitivity of Hybrid Posteriors to Differences in Bias Distributions Between Lottery and Non-lottery Schools

			Closure effects			
Difference in bias distributions	Model	RMSE (1)	Average school (2)	Average above-median school (3)	Average top-quintile school (4)	Average charter school (5)
None	Lagged score	0.116	0.315	0.437	0.529	0.665
			[0.131]	[0.141]	[0.147]	[0.145]
	Gains	0.100	0.316	0.434	0.525	0.657
			[0.115]	[0.126]	[0.136]	[0.128]
Non-lottery b_0 0.2σ lower	Lagged score	0.146	0.295	0.408	0.511	0.647
			[0.135]	[0.151]	[0.157]	[0.143]
	Gains	0.140	0.283	0.391	0.493	0.626
			[0.124]	[0.136]	[0.138]	[0.133]
Non-lottery b_0 0.2σ higher	Lagged score	0.152	0.263	0.365	0.454	0.615
			[0.131]	[0.144]	[0.152]	[0.144]
	Gains	0.143	0.277	0.373	0.446	0.619
			[0.122]	[0.138]	[0.155]	[0.134]
Non-lottery σ_b twice as large	Lagged score	0.157	0.214	0.318	0.412	0.568
			[0.163]	[0.172]	[0.181]	[0.168]
	Gains	0.144	0.238	0.341	0.428	0.582
			[0.149]	[0.157]	[0.165]	[0.155]
Non-lottery σ_b half as large	Lagged score	0.100	0.337	0.462	0.553	0.692
			[0.125]	[0.137]	[0.144]	[0.132]
	Gains	0.092	0.334	0.454	0.543	0.679
			[0.111]	[0.123]	[0.127]	[0.116]
Non-lottery $\sigma_{\beta b}$ has opposite sign	Lagged score	0.129	0.306	0.423	0.508	0.660
			[0.138]	[0.151]	[0.159]	[0.143]
	Gains	0.118	0.290	0.407	0.492	0.634
			[0.127]	[0.135]	[0.140]	[0.134]

Notes: This table explores the sensitivity of simulated effects of closing the lowest-ranked district school to violations of the assumption that bias distributions are the same for lottery and non-lottery schools. Policymakers are assumed to make closure decisions based on hybrid posterior modes constructed under the incorrect assumption that there is no difference between these groups. The notes to Table 2.8 describe the simulation procedure.

Chapter 3

Estimating Hospital Quality with Quasi-experimental Data

3.1 Introduction

Outcome-based rankings of institutional quality draw interest in many settings, from school and teacher value-added to the lasting socioeconomic effects of residential, educational, and occupational choice.[1] In the U.S. these measures have begun to play an important policy role, particularly in education and healthcare. Hospitals with low risk-adjusted mortality rates, for example, are now rewarded with higher Medicare reimbursement rates, while providers with poor survival outcomes may be flagged as low-performers. Recent research has found that such quality-based policies shape both hospital incentives and patient admission patterns (Norton et al., 2016; Gupta, 2016; Dranove and Sfekas, 2008; Chandra et al., 2015).

To date, performance-based regulation has relied on observational quality estimators, such as value-added models (VAMs) in education and risk-adjustment models (RAMs) in health. These

For feedback on this chapter I thank Alberto Abadie, Nikhil Agarwal, Isaiah Andrews, Joshua Angrist, Joe Doyle, Amy Finkelstein, Bill Goulding, Jon Gruber, Nick Hagerty, Greg Howard, Jack Liebersohn, Bruce McGough, Rachael Meager, Yusuke Narita, Parag Pathak, Bryan Perry, Brendan Price, Evan Riehl, Doug Staiger, Chris Walters, and seminar participants at MIT and NBER. I especially thank Joe Doyle, Jon Gruber, John Graves, and Sam Kleiner for sharing code to construct the ambulance instruments; Maurice Dalton, Yunan Ji, Bryan Perry, and Jean Roth for their data expertise; and emergency service professionals Ben Artin, Mark Millet, Laura Segal, Julia Taylor, and Kevin Wickersham for answering my many institutional questions. I gratefully acknowledge funding from the National Institute on Aging (#T32-AG000186) and the Spencer Foundation (#201600065).

[1] See, for example, chapter 2, along with Chetty et al. (2014b), Chetty and Hendren (2016), Hoxby (2015), and Card et al. (2013) for estimates of the institutional effects of schools, teachers, neighborhoods, colleges, and firms.

methods leverage strong selection-on-observables assumptions: that, say, a patient's choice of hospital is as good as random conditional on a set of observed controls. When provider selection is correlated with potential health outcomes, hospital RAMs are prone to systematic bias, and supervisory policies can be distorted. RAM-based admission guidance programs may themselves be a source of this selection bias by encouraging the selection of high-ranked hospitals, as may other intrinsic factors like the medical expertise of a patient's ambulance driver or the non-random location of high-quality providers.

In principle, instrumental variable (IV) techniques offer a solution to selection bias, as in other settings. In practice, researchers hoping to exploit quasi-experimental variation in institutional choice face several methodological challenges. Linear IV methods, including those used in chapter 2 to reduce bias in school VAMs, typically depend on an assumption of constant causal effects – for example that switching from the highest- to the lowest-ranked hospital has the same health effect for all potential patients.[2] This rules out both institutional comparative advantage and selection-on-gains, two powerful economic forces that are likely important in many settings, including healthcare (Chandra and Staiger, 2007). Moreover, constant effect restrictions are inappropriate for modeling binary outcomes, including the 30-day survival indicators used in hospital RAMs.

This chapter develops a new approach for measuring institutional quality with nonlinear causal response functions, selection-on-gains, and quasi-experimental data. Usual nonlinear IV estimators use maximum likelihood methods that can be computationally intractable or require parametric assumptions that are difficult to assess or interpret. Even estimating a nonlinear first stage for institutional sorting requires solving a high-dimensional multinomial choice problem; for decades scholars have grappled with the practical difficulties of fitting these models without unrealistic restrictions on choice substitution patterns (Hausman and Wise, 1978; McFadden, 1989; McColloch and Rossi, 1994; Berry et al., 1995). In an application of state-of-the-art Markov-chain Monte Carlo techniques, Geweke et al. (2003) estimate the quality of 114 Los Angeles County hospitals with relative distance instruments, a multinomial probit model of hospital admissions, and a probit specification for the short-term mortality outcomes of elderly pneumonia patients. To evaluate

[2]Unlike with binary treatments, multi-dimensional linear IV has no local average treatment effect (LATE) interpretation except under strong assumptions (see, e.g., chapter 1, along with Behaghel et al., 2013; Kirkebøen et al., 2016; and Blackwell, forthcoming). Even in these cases, LATE-based quality measures are undesirable, as differences in complier populations could affect the rankings of institutions with the same average effectiveness. As formalized in section 3.2, quality differences in my framework reflect average treatment effects, though estimating other parameters, such as average treated effects on the treated, is also possible.

their likelihood further requires ex ante specification of independent priors for each of the model's 268 free parameters, along with several auxiliary functional form restrictions and calibrations. Characterizing the role of these parameterizations, versus the potentially-exogenous variation in hospital choice generated by the instruments, is far from straightforward.

Rather than fitting a fully-specified likelihood to data, my approach matches a sparse set of moments from a multi-dimensional Roy (1951) selection model to quantities identified by quasi-experimental instrument assignment. This yields a flexible framework, fully non-parametric given sufficiently-rich instrument variation, for estimating institutional effectiveness with comparative advantage and selection-on-gains. Distributional assumptions on the model's latent variables can be used to extrapolate from observed quasi-experimental quantities to structural parameters of interest with more limited variation. A minimum distance procedure easily implements this semi-parametric approach, even with many individuals, institutions, instruments, and covariates.

I use these methods to estimate hospital quality from a nationally representative sample of U.S. Medicare patients admitted for an emergency condition. Specifically, I fit a multivariate probit model for potential hospital admissions and 30-day survival outcomes using quasi-experimental variation in ambulance company assignment. In a recent paper, Doyle et al. (2015) propose ambulance company instruments as a more credible alternative to distance-based identification strategies, which may be biased by non-random hospital location (e.g. Hadley and Cunningham (2004)). They use ambulance referral variation to instrument for the average Medicare spending of a patient's hospital in linear mortality models, finding large returns to being treated by more intensive providers. My nonlinear approach instruments a patient's hospital directly, allowing for violations of the Doyle et al. (2015) exclusion restriction (that ambulance referral variation only affects patient outcomes by the average treatment intensity of the referred hospital), as well as heterogeneous treatment effects and Roy selection.

The initial analysis yields a set of noisy quality estimates for 1,041 U.S. hospitals with sufficient quasi-experimental data. As in other recent explorations of institutional quality (e.g. Chetty and Hendren (2016)), I use these estimates to fit a hierarchical linear model and compute empirical Bayes quality posteriors that optimally combine quasi-experimental estimates and observational RAM predictions. This procedure reduces overall mean squared prediction error and generates posteriors for the full set of U.S. hospitals in my analysis sample.

Quality posteriors reveal several important dimensions of hospital performance and patient

sorting. Higher-volume hospitals and those that spend more per Medicare patient appear to produce better average survival outcomes, while government-run hospitals are systematically lower-performing. For example, moving a patient to a provider that would on average increase her 30-day survival probability by one percentage point places her in a hospital with 1.9% higher spending and 4.3% higher Medicare patient volume. This is qualitatively similar to what earlier work has found when measuring quality by observational RAMs (Foster et al., 2013; Chandra et al., 2015; Doyle et al., 2015). However, consistent with a broad pattern of better hospitals attracting sicker patients, I show that the strength of these relationships is magnified when they are measured with quasi-experimental data. Comparing quality posteriors and observed survival rates, I moreover find robust evidence for hospital comparative advantage and positive Roy selection-on-gains, with patients admitting to more appropriate hospitals on average. This non-random sorting is only partly explained by differential hospital distance and generates systematic bias in observational RAMs.

To quantify the economic importance of selection bias, I use my estimates to simulate quality-based Medicare reimbursement and patient guidance policies. Ranking hospitals by quality posteriors instead of RAM predictions tends to magnify existing transfers across different types of hospitals rather than changing the distribution of policy winners and losers. Net subsidies paid to privately-owned and teaching hospitals, for example, increase by 9% and 15%. In simulations of quality-based admission policies, I find a typical patient has a 2.8 percentage point higher 30-day survival rate when choosing hospitals on the basis of RAM predictions, rather than admitting at random. Admission to hospitals with the highest quality posteriors yields larger survival rate improvements of between 3.3 and 4.5 percentage points. Nevertheless, the scope for health gains from quality-based admission policies is limited by the extent of positive Roy selection; moving a random patient from her selected hospital to the provider delivering the highest average quality-of-care would *decrease* expected survival by 11 percentage points. This highlights a general issue for performance-based guidance policies that is obscured by the usual constant-effect quality framework.

The remainder of this chapter is organized as follows: the next section develops a general method of moments approach for estimating institutional quality with instrumental variables and discusses non- and semi-parametric identification. I then outline the institutional setting for hospital quality and describe the Medicare analysis sample and estimation procedure in section 3.3. Next, section 3.4 discusses my findings on hospital quality, patient sorting, and the consequences of non-random sorting in performance-based healthcare policies. Section 3.5 concludes.

3.2 Quality identification

3.2.1 The Quasi-experimental Setting

Suppose we observe outcomes Y_i for each individual i attending one of many possible institutions $j = 1, \ldots, J$. We indicate institutional choice by a set of dummy variables D_{ij}, collected in the vector D_i. For example $D_{ij} = 1$ may denote patient i's admission to hospital j, while $Y_i = 1$ if she survives the first 30 days following admission. Corresponding to each institutional alternative is a potential outcome Y_{ij}; these are linked to observed outcomes by

$$Y_i = \sum\nolimits_j Y_{ij} D_{ij}. \tag{3.1}$$

Policymakers aim to rank institutions by quality, defined as $q_j = E[Y_{ij}]$. This represents the expected outcome from sending a random individual to institution j, so that institutional quality comparisons avoid any bias from non-random sorting that would cause Y_{ij} and D_{ij} to be correlated. Let $E[Y_{ij}|D_{ij} = 1] - E[Y_{ij}]$, the difference in average selected and potential outcomes, quantify this selection bias for institution j.

Along with choices and outcomes, suppose we observe an individual's assignment to a discretely-valued instrument Z_i. Without loss of generality we let Z_i be a vector of indicators $Z_{i\ell}$ for the set of L possible instrument values and denote vectors in the support of Z_i by z_ℓ. For example, in the hospital application, $Z_{i\ell} = 1$ (and $Z_i = z_\ell$) if ambulance company ℓ is dispatched to individual i. Attending institution j after being assigned to the ℓth instrument value generates latent utility $U_{ij}(z_\ell)$, and individuals choose the institution that maximizes these payoffs. Institutional selection is thus given by

$$D_{ij} = \mathbf{1}[U_{ij}(Z_i) \geq U_{ik}(Z_i), \forall k]. \tag{3.2}$$

Equations (3.1) and (3.2) structure the vector of observed outcomes, institutional choices, and instrument assignments, $(Y_i, D_i', Z_i')'$, by a generalized multi-dimensional Roy (1951) selection model (Heckman et al., 2008). This model asserts the existence of counterfactual outcomes Y_{ij} and latent utilities $U_{ij}(z_\ell)$ with a conventional stable unit treatment value assumption (Imbens and Rubin, 2015) and adopts an implicit exclusion restriction, that the instrument only affects outcomes

through the choice of institution. Importantly, the model does not limit the possibility of either institutional comparative advantage or endogenous selection on potential outcomes. The causal effects $Y_{ij} - Y_{ik}$ need not be constant across individuals, and potential outcomes may be correlated with the latent utilities governing institutional choice, generating "essential heterogeneity" in the language of Heckman et al. (2006).

A conditional independence assumption completes the quasi-experimental framework: that, given a set of auxiliary controls X_i, the instrument Z_i is as good as randomly assigned with respect to the vector of latent outcomes and utilities:

Assumption 3.1 (*Independence*): $\left((Y_{ij}, (U_{ij}(z_\ell))_{\ell=1,\ldots,L})_{j=1,\ldots,J}\right) \perp\!\!\!\perp Z_i \mid X_i$.

Quasi-random instrument assignment ensures that while institutional choice itself may be correlated with potential outcomes, there is variation in conditionally-exogenous factors $Z_{i\ell}$ that can affect sorting by changing the frontier of latent payoffs, $U_{ij}(Z_i)$. My framework leverages this variation with knowledge or first-step non-parametric estimation of the conditional expectation functions $p_\ell(X_i) = E[Z_{i\ell}|X_i]$. I refer to these as instrument "propensity scores" and maintain throughout an assumption of common support: that $p_\ell(X_i) > 0$ for each ℓ with probability one. All individuals thus face some non-zero risk of assignment to each of the L instrument values.

3.2.2 Non-parametric Identification

Quasi-experimental instrument assignment is a powerful restriction, sufficient for non-parametric estimation of certain moments of the model's latent variables, Y_{ij} and $U_{ij}(z_\ell)$. Namely, the following auxiliary result shows that Assumption 1 identifies both the first-stage shares of individuals who would choose each institution j if assigned to each instrument value ℓ (what I refer to as "choice probabilities") and the means of any function of potential outcomes for individuals who would select j under this assignment (termed "mean selected outcomes"):

Lemma 3.1 (*Identification of choice probabilities and mean selected outcomes*): Let $f(\cdot)$ be any measurable function of Y_i. Under Assumption 1,

$$Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) = E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right] \tag{3.3}$$

$$E[f(Y_{ij})|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k] = E\left[\frac{f(Y_i)D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right] / E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \tag{3.4}$$

Proof: See the econometric appendix.

Note that without controls (so that the instrument is unconditionally randomly assigned, as in a randomized control trial) choice probabilities and mean selected outcomes are given by the moments $E[D_{ij}|Z_{i\ell} = 1]$ and $E[f(Y_i)|D_{ij} = 1, Z_{i\ell} = 1]$. The formulas in Lemma 3.1 use the non-parametrically identified propensity scores to appropriately re-weight the data so that it mimics this idealized experimental setting.

Without further parameterizations of the model, equations (3.3) and (3.4) are enough to estimate institutional quality from rich quasi-experimental data. Intuitively, by varying the instrument $Z_{i\ell}$ and setting $f(Y_i) = Y_i$ we non-parametrically observe average outcomes at institution j across different groups of individuals for whom utility is maximized at j when $Z_i = z_\ell$. We can moreover rank these averages by the fraction that each group represents of the population, $Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)$. If the number of observed instrument values grows with the sample, we may expect to find assignments that bring this choice probability arbitrarily close to one; in the limit we could thus estimate the population $E[Y_{ij}] = q_j$ by constructing averages of estimated mean selected outcomes $E[Y_{ij}|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]$ that place more weight on z_ℓ with the highest choice probabilities.

Formally, given any consistent set of propensity score estimators $\hat{p}_\ell(\cdot)$, we have the following result:

Proposition 3.1 (*Local linear quality identification*): For each j, collect the set of choice probabilities $G_{j\ell} = Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)$ in the vector G_j. If the support of $G_j' Z_i$ has a supremum of 1, then, under Assumption 3.1, $\hat{q}_j \xrightarrow{p} q_j$ where given N independent, identically-distributed draws of $(Y_i, D_i', Z_i', X_i')'$,

$$\hat{q}_j = \arg_1 \min_{q,b} \sum\nolimits_{\ell:\hat{G}_{j\ell} \geq \hat{c}_j} \hat{w}_{j\ell} \left(\hat{H}_{j\ell} - q - b(1 - \hat{G}_{j\ell}) \right)^2, \tag{3.5}$$

for $\hat{G}_{j\ell} = \frac{1}{N} \sum_{i=1}^N \frac{D_{ij} Z_{i\ell}}{\hat{p}_\ell(X_i)}$, $\hat{H}_{j\ell} = \sum_{i=1}^N \frac{Y_i D_{ij} Z_{i\ell}}{\hat{p}_\ell(X_i)} / \sum_{i=1}^N \frac{D_{ij} Z_{i\ell}}{\hat{p}_\ell(X_i)}$, and where $\hat{c}_j$ and the $\hat{w}_{j\ell}$ are scalars with $\hat{c}_j \leq \max_\ell(\hat{G}_{j\ell})$, $\hat{c}_j \xrightarrow{p} 1$, $\hat{w}_{j\ell} > 0$, and $\sum_\ell \hat{w}_{j\ell} = 1$.

Proof: Let $\widehat{\ell^*}(j)$ be an arbitrary element from the set of instrument values ℓ maximizing the sample choice probabilities $\hat{G}_{j\ell}$. Under the assumptions, $\hat{G}_{j\widehat{\ell^*}(j)} \xrightarrow{p} 1$ and $\hat{H}_{j\widehat{\ell^*}(j)} \xrightarrow{p} E[Y_{ij}]$ by the Weak Law of Large Numbers. Thus $\hat{q}_j \xrightarrow{p} q_j$, provided the bandwidth $\hat{c}_j$ approaches 1 and the weights $\hat{w}_{j\ell}$ are convex. □

The local linear regression estimator $\hat{q}_j$ is consistent for institution j's quality when $G_j'Z_i$, the choice probability of the instrument assigned to individual i, has sufficiently large support. This result follows a broad literature on non-parametric identification of Roy models, including Heckman and Honore (1990), Lewbel (2007), and D'Haultfoeuille and Maurel (2013). In fact, the estimator in Lewbel (2007) is also consistent for q_j under a somewhat stronger support condition than the one used in Proposition 3.1.[3] Other estimators can be obtained by adding higher-order polynomials or other transformations of the regressor $1 - \hat{G}_{j\ell}$. Characterizing the optimal choice of weights, bandwidths, and local regressors for non-parametric quality estimation is left for future research.

3.2.3 Semi-parametric Quality Estimation

Limited variation in choice probabilities renders the estimator in Proposition 3.1 inconsistent. When institutional quality is not non-parametrically identified, further restrictions on the selection model can substitute for rich quasi-experimental data. Intuitively, there are parametrizations of the joint distribution of latent variables Y_{ij} and $U_{ij}(z_\ell)$ that make the moments identified by Lemma 3.1 functions of some finite-dimensional parameter vector, θ_0. A quasi-experimental design generating sufficient variation in the moments may pin down these structural parameters and thus the marginal means of latent Y_{ij} – that is, quality. A minimum distance procedure (Ferguson, 1958), which is computationally simple relative to earlier likelihood-based IV methods, implements this semi-parametric approach.

I first outline the proposed minimum distance quality estimator for a generic identified parameterization of the latent variables; I then establish and characterize identification for a particular multivariate probit specification which is later used to estimate hospital quality. Suppose for some known distribution function $F(\cdot)$ we have

$$\left((Y_{ij}, (U_{ij}(z_\ell))_{\ell=1,\ldots,L})_{j=1,\ldots,J}\right) \sim F(\theta_0), \tag{3.6}$$

so that the various choice probabilities and mean selected outcomes identified under Assumption 1 are also known functions of θ_0. Let $m(\cdot)$ be a vector collecting some subset of these functions and $\hat{m}$ be the sample analogues of the corresponding formulas of Y_i, D_i, Z_i, and $p_\ell(X_i)$ from Lemma

[3]Namely, note that we can write $D_{ij} = \mathbf{1}[0 \leq M_{ij}^* + V_{ij} \leq A_i^*]$ where for independent $M_i \sim U[0,1]$ and $g_j = \min_\ell G_{j\ell}$, we let $M_{ij}^* = -M_i + g_j$, $V_{ij} = G_j'Z_i - g_j$, and $A_i^* = 1 - M_i$. This corresponds to equation (1) in Lewbel (2007) and the support condition of that paper is satisfied if $G_j'Z_i$ continuously varies over $[p, 1]$.

3.1, constructed with some consistent non-parametric propensity score estimators $\hat{p}_\ell(\cdot)$. Under mild regularity conditions (see, e.g., Hirano et al. (2003)), we then have $\sqrt{N}(\hat{m} - m(\theta_0)) \Rightarrow N(0, Q)$, for some non-parametrically identified variance matrix Q. If the structural parameters in θ_0 are uniquely determined by the quasi-experimental variation in $m(\cdot)$, a consistent minimum distance estimator is then given by

$$\hat{\theta} = \arg\min_{\theta}(\hat{m} - m(\theta))'\hat{A}(\hat{m} - m(\theta)), \tag{3.7}$$

for some weight matrix $\hat{A}$. Furthermore, under the same conditions for the asymptotic normality of $\hat{m}$, we have

$$\sqrt{N}(\hat{\theta} - \theta_0) \Rightarrow N\left(0, (M'AM)^{-1}M'AQAM(M'AM)^{-1}\right) \tag{3.8}$$

where $M = \frac{\partial m(\theta)}{\partial \theta}|_{\theta_0}$ and $\hat{A} \xrightarrow{p} A$. As usual with such extremum estimators, the asymptotic variance of $\hat{\theta}$ is minimized by setting $\hat{A} = \hat{Q}^{-1}$ for some consistent variance estimator $\hat{Q} \xrightarrow{p} Q$, in which case $\sqrt{N}(\hat{\theta} - \theta_0) \Rightarrow N\left(0, (M'Q^{-1}M)^{-1}\right)$. Note that with Q non-parametrically identified, this estimator can be formed in a single step and its asymptotic variance is consistently estimated by $\left(\hat{M}'\hat{Q}^{-1}\hat{M}\right)^{-1}$ for $\hat{M} = \frac{\partial m(\theta)}{\partial \theta}|_{\hat{\theta}}$. The choice of quasi-experimental moment estimator $\hat{m}$ thus entirely determines the relative efficiency of both $\hat{\theta}$ and, applying the Delta method to the formulas implied by equation (3.6), the corresponding estimates of quality $E[Y_{ij}]$. When the model is overidentified, an omnibus specification test statistic can be formed from the estimator's minimized criterion function:

$$\hat{T} = N(\hat{m} - m(\hat{\theta}))'\hat{Q}^{-1}(\hat{m} - m(\hat{\theta})). \tag{3.9}$$

Under the joint null hypothesis of Assumption 3.1 and the correct specification of $F(\cdot)$, this statistic will have an asymptotic chi-squared distribution with degrees of freedom equal to the extent of overidentification.

Computing minimum distance estimates is relatively straightforward, even as the number of institutions J, instrument values L, and controls in X_i grows large. Each element of $\hat{m}$ is determined by one of $L - 1$ propensity scores which do not depend on the model's structural parameters and may be separately approximated by standard techniques (e.g. Geman and Hwang (1982)).

Given $\hat{m}$, evaluating the estimator's objective function requires computing at most $((D+1)J-1)L$ nonlinear functions for each candidate parameter vector θ, where D is the dimension of the outcome function $f(\cdot)$.[4] Importantly these functions do not depend on the data, so unlike with likelihood-based estimators the difficulty of the nonlinear computation does not increase with the sample size. In some cases, including the multivariate probit model considered below, $m(\theta)$ will take a form that is straightforward to evaluate by standard statistical software packages (see the econometric appendix). Simulation methods can solve more exotic parameterizations; again the fact that the simulated objects are non-stochastic makes this procedure fast relative to typical applications of the simulated minimum distance approach of McFadden (1989) and Pakes and Pollard (1989).

The separation of quasi-experimental data in $\hat{m}$ from the structural assumptions underlying $m(\theta)$ also helps establish and characterize identification of semi-parametric quality models. I illustrate this with a multivariate probit specification for the latent variables, which produces my benchmark hospital quality estimates. Let h_{ij} denote the latent health of emergency patient i upon admission to hospital j, and assume patients survive the first 30 days following admission when their health is above some arbitrary threshold, here normalized to zero:

$$Y_{ij} = \mathbf{1}[h_{ij} \geq 0]. \tag{3.10}$$

With the vector h_i collecting the J health indices, the observed outcome equation (3.1) becomes

$$Y_i = \mathbf{1}[h_i' D_i \geq 0]. \tag{3.11}$$

The random coefficients in h_i retain the feature of institutional comparative advantage from the general selection model: some individuals may be more likely to survive when moved from hospital j to hospital k, while for others such a move may result in worse health outcomes.

In my application, emergency patients are referred to hospitals by ambulance, with $Z_{i\ell}$ indicating the quasi-experimental assignment of ambulance company ℓ to patient i.[5] As shown in Doyle

[4] Namely, there are at most $(J-1)L$ linearly-independent choice probabilities and DJL mean selected outcomes.

[5] One could instead imagine using geographic instruments, such as indicators for a patient's home ZIP code, in place of the ambulance company design. Assumption 3.1 would then require a patient's location to be conditionally-independent from her latent health and admission utility at each hospital, as with the relative distance instruments used in Geweke et al. (2003). IV estimates would be biased if, for example, hospital quality is endogenously determined by local patient characteristics; Hadley and Cunningham (2004) offer evidence for this kind of non-random assignment. The importance of minimizing travel time for treating emergency conditions also brings into question the exclusion restriction for such models.

et al. (2015), differences in ambulance referral preferences may generate variation in hospital admissions. The multivariate probit specification structures this first-stage variation by a monotonicity assumption, as in the identification of local average treatment effects and related causal parameters (Imbens and Angrist, 1994; Heckman et al., 2006):

Assumption 3.2 (*Monotonicity*): $\forall \ell,m,j$, either $Pr(U_{ij}(z_\ell) \geq U_{ij}(z_m)) = 1$ or $Pr(U_{ij}(z_\ell) < U_{ij}(z_m)) = 1$.

To the extent ambulance companies have different preferences for referring to each hospital j, they are fixed over different subpopulations of patients when Assumption 3.2 holds. Indeed, monotonicity implies an additively-separable model for latent utility:

$$U_{ij}(z_\ell) = \pi_{j\ell} + \eta_{ij}. \tag{3.12}$$

In my application, $\pi_{j\ell} - \pi_{k\ell}$ represents ambulance company ℓ's relative preference for referring to hospital j over hospital k, while η_{ij} denotes the latent utility from admitting at hospital j for patient i, which may also reflect common ambulance company preferences. With the vector π_j collecting the $\pi_{j\ell}$ parameters, the admissions process in equation (3.2) becomes

$$D_{ij} = \mathbf{1}[\pi_j' Z_i + \eta_{ij} \geq \pi_k' Z_i + \eta_{ik}, \forall k]. \tag{3.13}$$

A final parametric assumption defines the multivariate probit specification, along with equations (3.10) and (3.12): joint-normality of latent health and utility,[6]

Assumption 3.3 (*Normality*): $(h_i', \eta_i')' \sim N(\mu, \Sigma)$.

Many parameterizations of the model will be observationally equivalent under Assumptions 3.1-3.3 for any amount of quasi-experimental data. Namely, without loss of generality we can normalize $E[\eta_i] = 0$, $Var(h_{ij}) = 1$, $\forall j$, and $Var(\eta_i) = I_J$, where I_x is an identity matrix of size x, and restrict attention to the vector of relative utilities $U_{ij}(z_\ell) - U_{i\bar{j}}(z_\ell)$ for a fixed reference hospital $\bar{j}$. The relevant structural parameter vector θ_0 then consists of J quality index coefficients $\beta_j = E[h_{ij}] = \Phi^{-1}(q_j)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $J(J-1)$

[6]Note that under joint-normality a patient's utility from care can be written as a linear function of potential health, as in the classic Grossman (1972) healthcare demand model, with an independent normal error term.

health-utility correlations $\rho_{jk} = Corr(h_{ij}\eta_{ik} - \eta_{ij})$, and $(J-1)L$ relative ambulance company preferences $\pi_{j\ell} - \pi_{\bar{j}\ell}$, for a total of $J^2 + (J-1)L$ parameters.[7]

For Bernoulli outcomes, quasi-experimental ambulance company assignment offers at most $(2J-1)L$ linearly-independent moments identified by Lemma 3.1, for any choice of $f(\cdot)$. The order condition for identifying θ_0 is thus satisfied with $L \geq J$ (in my setting, as many ambulance companies as hospitals) and the rank condition holds when ambulance company preferences are unique:

Proposition 3.2 (*Multivariate probit identification*): In the multivariate probit model, suppose Π, the $J \times L$ matrix of preference parameters $\pi_{j\ell}$, has no redundant columns and that Assumption 3.1 holds. Then all quality parameters q_j are identified if $L \geq J$.

Proof: For each instrument value ℓ, the J choice probabilities identified by Lemma 3.1 are uniquely determined by $J-1$ relative preferences $\pi_{j\ell} - \pi_{\bar{j}\ell}$ under Assumptions 3.2-3.3. With these parameters solved, the L mean selected outcomes for each institution j are determined by one quality parameter q_j and $J-1$ correlations ρ_{jk}, and uniquely so when the columns of Π are unique. Identification thus follows if $L \geq J$. □

Estimating each institution's quality by Proposition 3.2 will generally use the full set of choice probabilities. In practice, with many small institutions or rare instrument value assignments, some of the associated $E[D_{ij}Z_{i\ell}/p_\ell(X_i)]$ may be poorly approximated in finite samples, potentially rendering all quality estimates unreliable. This concern is particularly relevant in my hospital application: the distribution of hospital volume in administrative claims data is right-skewed, with many small providers.[8] A more attractive estimation approach leverages alternative-specific instruments of the kind traditionally found in multinomial choice applications (Keane, 1992). Suppose we can partition the instrument vector Z_i into J subvectors Z_{ij} whereby moving across different values in the support of Z_{ij} only affects the latent utility generated by institution j and not any other alternatives. This would be the case in the stylized hospital quality example if each ambulance company

[7] In general the cross-institution health correlations will not be identified, nor are they necessary for quality identification.

[8] The difficulty of estimating hospital quality models due to the presence of small providers is well-known: both federal policymakers and Geweke et al. (2003) remove patients admitted to low-volume hospitals from their analysis samples, though this practice likely induces selection bias. The separable identification result I provide in Proposition 3.3 overcomes this issue without endogenous sample selection.

has at most one preferred hospital – for example, the one based closest to company offices – but otherwise has no preferences that would differentially shift patients between other local hospitals. In this case, the following result shows we may separately identify the quality of each hospital using only a subset of the choice probabilities:

Proposition 3.3 (*Multivariate probit identification with alternative-specific instruments*): For a given j in the multivariate probit model, suppose $\pi_{k\ell} = \bar{\pi}_k$ for all instrument values ℓ in the support of an alternative-specific instrument vector Z_{ij} and all $k \neq j$. Then q_j is identified under Assumption 3.1 if the subvector of π_j corresponding to Z_{ij} has $L_j \geq J$ distinct values.

Proof: Under the assumptions the $J-1$ relative preference parameters $\bar{\pi}_k$ are identified by J choice probabilities involving D_{ik} for $k \neq j$ and any $Z_{i\ell}$ in Z_{ij}. With these known, the $2L_j$ choice probabilities and mean selected outcomes involving D_{ij} and the $Z_{i\ell}$ in Z_{ij} are uniquely determined by institution j's quality q_j, $J-1$ correlations ρ_{jk}, and L_j relative preferences $\pi_{j\ell}$, when the latter are non-redundant. □

Alternative-specific instruments thus provide a method for estimating the quality of only a subset of institutions for which choice probabilities and mean selected outcomes are likely to be well estimated, leaving the quality of other hospitals with less-rich quasi-experimental data underidentified.

Minimum distance quality estimators based on results like Propositions 3.2 and 3.3 use a low-dimensional parameterization of the distribution of potential outcomes and latent utility to extrapolate from a discrete set of non-parametric instrumental variable moments to the structural parameters of interest. This is in the spirit of Brinch et al. (forthcoming), who directly parameterize conditional marginal treatment effect curves in the binary treatment case; here both the extrapolation and number of instruments needed for identification are guided by a multiple-treatment Roy model and do not depend on the distribution of quasi-experimental controls except through the set of non-parametric instrument propensity scores.[9]

The parametric extrapolation of reduced-form moments is most clearly seen in the case of $J = 2$

[9]Brinch et al.'s approach requires estimating the functions $E[Y_i|D_{ij} = 1, Z_{i\ell} = 1, X_i = x]$ for each j, ℓ, and value in x in the support of the quasi-experimental controls X_i. In practice this can be infeasible when the controls are continuous or take on many discrete values, as in my setting. Standard asymptotic theory may also provide only poor approximations for the sampling distribution of estimators based on many stratified conditional means, an issue discussed in Robins and Ritov (1997) and Angrist and Hahn (2004) and that motivates Hirano, Imbens, and Ridder's (2003) inverse propensity score weighting approach for efficiently estimating average treatment effects. This chapter's approach follows in the same spirit.

institutions, for which the model given by Assumptions 3.2 and 3.3 is a bivariate probit and the conditions for identification in Propositions 2 and 3 coincide. Without loss of generality, we may then normalize $\pi_{2\ell} = \eta_{i2} = 0$ and drop j subscripts from the latent utility parameters for institution 1 to write

$$Y_i = \mathbf{1}[h_{i1}D_{i1} + h_{i2}D_{i2} \geq 0] \tag{3.14}$$

$$D_{i1} = \mathbf{1}[\pi' Z_i + \eta_i \geq 0] = 1 - D_{i2}, \tag{3.15}$$

where, under Assumption 3.3, $(h_{i1}, h_{i2}, \eta_i)' \sim N\left((\beta_1, \beta_2, 0)', \Sigma\right)$. Here the covariance matrix Σ has two health-utility correlations, ρ_1 and ρ_2, which along with β_1, β_2, and π yield $L + 4$ parameters in θ_0. Under Assumption 3.1 we observe L sets of linearly-dependent choice probabilities and $2L$ mean selected outcomes by the formulas in Lemma 3.1, and $L \geq 2$ ambulance companies satisfies the order condition.

Bivariate probit mean selected outcomes, $E[Y_{ij}|\pi_\ell + \eta_i \geq 0]$, are monotone in the first-stage parameters π_ℓ.[10] Thus any two instrument values ℓ and m for which $\pi_\ell > \pi_m$ inform the sign of selection bias at each institution. If, for example, we learn by Lemma 3.1 that $Pr(\pi_\ell + \eta_i \geq 0) > Pr(\pi_m + \eta_i \geq 0)$ and $E[Y_{i1}|\pi_\ell + \eta_i \geq 0] < E[Y_{i1}|\pi_m + \eta_i \geq 0]$, we would know that patients with *lower* utility admissions utility η_i, who only select hospital 1 when assigned to ambulance company ℓ (that is, in the language of Imbens and Angrist (1994), the ambulance company "compliers"), have *worse* health outcomes at hospital 1 than those who would be admitted by either ambulance company (the quasi-experiment's "always-takers"). By normality, hospital 1's average potential outcome in the population of patients (i.e., its quality $E[Y_{i1}]$) is therefore lower than that of patients who actually choose hospital 1: $E[Y_{i1}|D_{i1} = 1] - E[Y_{i1}] > 0$, so that hospital 1 is positively selected.

Along with the direction of selection bias, joint-normality prescribes a particular translation of admitted patient health to the population. In the bivariate model, the quality index $\beta_j = \Phi^{-1}(q_j)$ can be written as a linear combination of the health of patients who would be admitted by the two

[10] Namely, $E[Y_{ij}|x + \eta_i \geq 0] = Pr(h_{ij} \geq 0|x + \eta_i \geq 0) = \int_{-\infty}^{x} \Phi\left((\beta_j - \rho_j t)/\sqrt{1 - \rho_j^2}\right) \frac{\phi(t)}{\Phi(x)} dt$ when h_{ij} and η_i are normally distributed. The derivative of this function with respect to x is proportional to $\Phi\left((\beta_j - \rho_j x)/\sqrt{1 - \rho_j^2}\right) - \int_{-\infty}^{x} \Phi\left((\beta_j - \rho_j t)/\sqrt{1 - \rho_j^2}\right) \frac{\phi(t)}{\Phi(x)} dt \gtrless 0 \iff \rho_j \lesseqgtr 0$.

ambulance companies: e.g.,

$$\beta_1 = E[h_{i1}|\pi_\ell + \eta_i \geq 0]\omega + E[h_{i1}|\pi_m + \eta_i \geq 0](1-\omega) \tag{3.16}$$

for

$$\omega = 1/\left(1 - \frac{\phi(\pi_\ell)}{\Phi(\pi_\ell)}/\frac{\phi(\pi_m)}{\Phi(\pi_m)}\right), \tag{3.17}$$

where $\phi(\cdot)$ denotes the standard normal probability density function. The inverse Mills ratio $-\phi(\pi_\ell)/\Phi(\pi_\ell)$ is increasing in the first-stage parameters, so with $\pi_\ell > \pi_m$ we have $\omega > 1$, and the non-convex weighting scheme given by equation (3.16) extrapolates in the direction of the larger patient subpopulation. This is illustrated in panel A of Figure 3.1, in the case of positive selection bias ($\rho_1 > 0$). The two vertical dashed lines show the inverse Mills ratio for two ambulances' first-stage parameters, while the two horizontal dashed lines show the associated average health of patients who would be admitted by each company. The downward-sloping line that intercepts the maximum inverse Mills ratio of zero at β_1 (with a slope of $-\rho_1$) gives the extrapolation from these two patient subpopulations to population health.

When $L > 2$ in the bivariate probit model, any two ambulance companies with different referral preferences identify hospital quality in this way, and the minimum distance quality estimator given by equation (3.7) efficiently aggregates all pairwise comparisons. If the relative preference parameters π were known, this would amount to solving a variance-weighted nonlinear least squares problem of fitting estimated mean selected outcomes to a particular parametric curve. An example of this is plotted in panel B of Figure 3.1, using data simulated from the same probit specification used in panel A. The nonlinear curve of best fit is parameterized by an intercept, hospital quality q_j, along with a shape parameter ρ_j that determines the sign and extent of selection bias. The R-squared for the curve's fit informs the overidentification test statistic $\hat{T}$ from equation (3.9).

The same extrapolative logic applies to estimation of multi-institution models, when $J > 2$. Each new institution adds a shape parameter to the multivariate probit curve, thereby necessitating an additional mean selected outcome point. Other parameterizations of the selection model would yield other curves with different quasi-experimental data requirements. Note that, unlike with linear IV (see, e.g., Angrist (1991) and the references therein), the least-squares interpretation of these nonlinear estimators no longer holds when the π_j are estimated, as the IV moment vector is

not linear in the first-stage parameters. Nevertheless, a procedure wherein the first stage is initially obtained from the set of choice probabilities and then used to fit appropriate parametric curves to mean selected outcomes will yield consistent semi-parametric estimates.

3.3 Estimating Hospital Quality

3.3.1 Data and RAMs

I use the preceding framework to estimate the quality of U.S. hospitals according to their effects on short-term patient mortality. Policymakers currently base observational hospital RAMs on three-year windows of emergency Medicare claims (YNHHSC/CORE, 2013); correspondingly, I draw a sample of 405,173 Medicare fee-for-service beneficiaries brought to an acute-care hospital by an ambulance for one of 29 emergency conditions in 2010-2012.[11] Observations come from a nationally-representative 20% sample of administrative inpatient claims from the Centers of Medicare and Medicaid Services (CMS) and include information on basic patient demographics (such as age, sex, race, and home ZIP code); diagnoses and procedures from previous inpatient and outpatient claims ("comorbidities"); the identity of, ZIP code location of, and procedures performed by a patient's assigned ambulance company; the identity and location of the hospital; and subsequent mortality. As in Card et al. (2009), I restrict the sample to patients admitted for a "nondeferrable" primary condition, i.e. those with a weekend admissions rate close to 2/7ths. These are the same conditions used by Doyle et al. (2015) and are listed in the notes to Table 3.1. I also follow standard CMS risk-adjustment methodology in attributing outcomes to a patient's first hospital admission in 2010-2012, ignoring all subsequent transfers or readmissions. Finally, I divide the national sample of patients, ambulances, and hospitals into hospital service areas (HSAs), which are sets of ZIP codes defined by the Dartmouth Atlas of Health Care as narrow regions where patients receive most of their emergency care. I use HSAs to delineate local emergency care markets, within which it is plausible that ambulance company propensity scores have full support. As Appendix Table 3.A3 illustrates, I obtain similar findings throughout with hospital referral regions (HRRs). A data appendix describes the sample construction in detail.

Table 3.1 summarizes the distribution of diagnoses, ambulances, hospitals, HSAs, and 30-day survival probabilities. Hospital RAMs were first developed to measure quality by the mortality of

[11] Unlike in some RAMs, I am not able to include Veterans Affairs facilities in this analysis.

Medicare patients with circulatory and respiratory conditions, such as acute myocardial infarction, heart failure, and pneumonia, though often with the stated goal of extending the methods to a broader patient population (Krumholz et al., 2006).[12] Panel A of Table 3.1 shows that circulatory and respiratory diagnoses make up 42% of nondeferrable admissions in my sample, with the remainder split between digestive (7%), injury (18%), and all other conditions (34%).

Each patient in the analysis sample was assigned to one of 9,590 ambulance companies and admitted to one of 4,821 hospitals.[13] Panel B of Table 3.1 reports that the distribution of within-HSA hospital counts is highly skewed, with around half (2,464) of all hospitals operating in their own single-hospital market. Since the ambulance design leverages within-market admissions variation, my analysis focuses on local comparisons for the other 2,357 hospitals in 695 multi-hospital HSAs. Column 5 of Table 3.1 summarizes average 30-day patient survival, which is the usual outcome of mortality RAMs. Around 83% of patients survive the first 30 days following their emergency admission, with survival rates as low as 78% for patients with respiratory conditions and as high as 93% for those with injuries. Panel B shows that average survival does not seem to vary much by the number of available hospitals.

I first use this sample to obtain a set of observational RAM quality predictions, following standard CMS risk-adjustment methodology. These specify an additively-separable latent index model for 30-day survival:

$$Y_{ij} = \mathbf{1}[\alpha_j + \epsilon_i \geq 0], \tag{3.18}$$

where

$$\epsilon_i = \gamma' W_i - \nu_i \tag{3.19}$$

for a set of observed risk-adjusters W_i. Thus in a conventional RAM

$$Y_i = \mathbf{1}[\alpha' D_i + \gamma' W_i \geq \nu_i], \tag{3.20}$$

[12] A related quality measurement effort models patient readmissions. Since a patient who dies at a low-quality hospital cannot be readmitted, more involved assumptions are required to causally attribute variation in these outcomes to hospital performance; I leave this issue for future work.

[13] 41% of Medicare patients hospitalized for a nondeferrable condition in 2010-2012 were admitted by an ambulance company; these and other comparisons are reported in columns 1 and 2 of Appendix Table 3.A1 and discussed in the data appendix.

where α collects the quality indices α_j. Identification of the RAM parameters α and γ follows from a selection-on-observables assumption that hospital choice is independent of latent health conditional on the included controls, $\nu_i \perp\!\!\!\perp D_i \mid W_i$. Following YNHHSC/CORE (2013), I parameterize η_i by an independent logit distribution and obtain quality predictions $\hat{\alpha}_j$ by estimating logit regressions of 30-day survival on hospital random effects and patient age, sex, and diagnosis and comorbidity indicators; the data appendix details the RAM estimation procedure.

Observational RAMs in my sample leave unexplained most of the national variation in survival outcomes. This is illustrated in Figure 3.2, which plots the ratio of residual to total 30-day survival variance in five diagnosis-specific RAMs. Only around 7% of circulatory and respiratory survival variance is due to a patient's hospital, admitting diagnosis, and year of admission. The reduction is smaller for digestive conditions and injuries, and larger, around 14%, for other diagnoses in the analysis sample. Patient demographics and comorbidities account for an additional 4% of circulatory and respiratory survival variance, with similarly modest declines for the other diagnosis categories.

If the significant residual survival determinants are exogenous to the hospital selection process, predictions from these RAMs may still provide unbiased measures of hospital quality. However, to the extent survival variance may be further reduced by observable admission determinants, such as a patient's assigned ambulance company, observational RAMs are likely to be biased. The econometric appendix formalizes this argument and develops instrument-based tests for nonlinear RAM unbiasedness that extend earlier methods for validating linear education VAMs (Kane and Staiger, 2008; Chetty et al., 2014a; Deming, 2014; Angrist et al., 2016b). These tests, summarized in Appendix Table 3.A2, decisively reject the null of selection-on-observables ($p < 0.001$), suggesting scope for bias in the observational RAMs. Motivated by these findings, I next describe the implementation of the semi-parametric IV techniques that I use to quantify and characterize hospital selection bias and quality.

3.3.2 Estimation

I use the identification result in Proposition 3.3 to semi-parametrically estimate the quality of 1,041 hospitals operating in one of 626 multi-hospital HSAs with at least 25 patients in the analysis sample and sufficient quasi-experimental admissions variation. Doyle et al. (2015) first propose that in regions served by multiple ambulance companies, centralized policies of rotational and simultaneous

911 dispatch generate plausibly-exogenous company assignment, while the subsequent expression of non-random ambulance preferences can systematically affect the admissions of otherwise identical patients. Table 3.2 explores both of these claims by comparing individuals in the same ZIP code who are assigned to different ambulance companies likely to refer to hospitals with high and low RAM predictions. Specifically, I compute the distance between each ambulance company's office and each nearby hospital using the provider ZIP codes contained in Medicare claims, and label companies as likely to deliver patients to a low- or high-ranked provider if their closest hospital is in the first or fourth quartile of RAM quality predictions in the HSA. I then regress patient characteristics on either these group indicators (with group means reported in columns 1 and 2) or the ambulance company's closest hospital's predicted RAM itself (with the coefficient reported in column 4), along with a full set of ZIP code fixed effects in the subsample of 254,101 admissions in multi-hospital HSAs.

Table 3.2 shows that patients assigned to ambulance companies based close to a high-ranked hospital see significantly increased RAM-predicted hospital quality, despite appearing identical to other patients in terms of their demographics, the location of their emergency, and their admitting diagnosis (panel A), as well as a host of comorbidity indicators describing their medical history (panel B). This balance of observable characteristics validates the quasi-random assignment of ambulance company indicators $Z_{i\ell}$, conditional on patient location X_i (Assumption 3.1). Ambulance assignment also appears balanced across a set of ambulance services performed pre-hospitalization (such as distance traveled in excess of the hospital ZIP code distance, whether the patient was assigned paramedics, or whether intravenous medication was delivered en route), a fact documented in panel C of Table 3.2. This supports the exclusion of ambulance-based instruments from potential survival outcomes Y_{ij}, allowing for interpretation of reduced-form ambulance effects on mortality outcomes by way of first-stage admission effects (a weaker restriction than in Doyle et al. (2015), where ambulances can only affect outcomes by changing the treatment intensity of a patient's provider). The p-value for a joint test of balance on assignment to ambulances based close to high- vs. low-RAM hospitals, across all 32 covariates in panels A, B, and C, is 0.89.[14]

[14]Similarly, Doyle et al. (2015) find no relationship between their ambulance-based instrument and a patient's probability of emergency room admission conditional on ZIP code; see their Figure A1. They likewise validate instrument balance in their analysis sample (see their Tables 1 and A3) and report anecdotal evidence for Assumption 3.1 from a 30-city survey of dispatch policies. My interviews with ambulance technicians in Connecticut, Massachusetts, Nevada, Philadelphia, Washington, and Wyoming further corroborate the assumption of quasi-random assignment. Note that the findings in Sanghavi et al. (2015) that advanced life support (ALS) services lead to higher cardiac arrest mortality are not at odds with my framework, since most ambulance companies provide both ALS and basic

As in Doyle et al. (2015) I leverage a first-stage monotonicity restriction, namely that differences in ambulance referral patterns do not systematically vary by patient characteristics (Assumption 3.2). Although not directly testable, Doyle et al. (2015) provide anecdotal support for monotone referral from their interviews with emergency care technicians – differences in referral patterns across ambulance companies appear to be driven by institutional and personal relationships with hospitals, rather than by patient heterogeneity. This is especially plausible in the relatively homogenous sample of emergency Medicare patients studied here. Differential treatment of uninsured patients by profit-driven ambulance companies, for example, is not a concern for this population.

My own interviews with current and former emergency medical staff across the U.S. support the alternative-specific model used in Proposition 3.3 as appropriate for ambulance assignment instruments: when differentially redirecting patients, ambulance companies seem to prefer returning to the hospital based closest to their offices in order to minimize excess travel time and maximize local availability.[15] The estimation strategy given by Proposition 3.3 is also attractive in practice as the analysis sample contains many hospital-ambulance combinations with relatively few non-zero observations of $D_{ij}Z_{i\ell}$, which may lead to unreliable choice probability and quality estimates from Proposition 3.2. I thus use the closest-hospital mapping from Table 3.2 to partition instrument vectors to alternative-specific subvectors and use only the largest ambulance company in each Z_{ij} to estimate π_{jk} for $k \neq j$. Table 3.A3 shows qualitatively similar results when Z_{ij} instead comprises the ambulance companies that most-often refer patients to hospital j in the universe of 2010-2012 Medicare claims (excluding observations in the analysis sample).[16]

My estimates of hospital choice probabilities and mean selected survival outcomes are based on a flexible probit specification for ambulance company propensity scores $p_\ell(X_i)$ that model the latent risk of assignment by a cubic polynomial in company-patient distance:

$$E[Z_{i\ell}|X_i] = \Phi\left(\delta_{0\ell} + \delta_{1\ell}d_\ell(X_i) + \delta_{2\ell}d_\ell(X_i)^2 + \delta_{3\ell}d_\ell(X_i)^3\right), \tag{3.21}$$

where $d_\ell(x)$ denotes the distance between ambulance company ℓ's institutional address and a patient located in ZIP code x. Minimum distance quality estimates correct for first-step error in

life support services and preference variation across companies is unlikely to be correlated with ALS availability.

[15]This appears especially true for ambulances owned by municipal and local fire departments, which are often the only local emergency transport provider and thus have a strong preference to return when dispatched outside of their home ZIP code.

[16]Judgments based on ambulance company size are also made on the basis of this larger disjoint sample.

approximating these conditional expectations. For robustness I also include the vector of RAM controls W_i in the propensity scores of my benchmark specification, though, consistent with Assumption 3.1, Table 3.A3 demonstrates that all results are essentially unchanged when these are excluded from the probit model.[17] This table also illustrates robustness to the health and utility probit specification (Assumption 3.3), with similar conclusions drawn from a fatter-tailed multivariate Student's $t(2)$ distribution that yields quality identification under the same assumptions as in the normal case. Quality is only identified by Proposition 3.3 for hospitals with $L_j \geq J(h(j))$ ambulance companies in their instrument subvectors Z_{ij}, where $J(h)$ is the hospital count of HSA h and $h(j)$ indexes hospital j's HSA; for these I use only the $J(h(j))$ largest companies in order to keep the model just-identified and reduce the scope for finite sample bias from many-weak IV identification.[18]

Figure 3.3 summarizes the available quasi-experimental data by plotting the joint distribution of differences in estimated hospital choice probabilities and mean selected outcomes for each of the 1,041 hospitals with enough ambulance company instruments to identify their quality. These differences are taken over the two ambulance companies generating the highest choice probability gap for each hospital; the marginal x-axis distribution thus summarizes the maximal variation in institutional choice generated by the instruments. The average choice probability difference is 0.4, with 43% of hospitals seeing a higher estimated choice probability difference. The average associated mean selected outcome difference is negative, and increasingly so as the first stage gap grows. As in the bivariate probit example in section 3.2.3, this suggests most hospitals in the sample see positive selection bias, which the generalized Roy model later confirms.

The solid curve in Figure 3.4 plots the distribution of the 1,041 minimum distance estimates of hospital quality indices, $\beta_j = \Phi^{-1}(q_j)$. Due to the HSA-stratified estimation procedure, the wide dispersion in these estimates reflects both causal (within-HSA) differences in potential survival outcomes for the same patient population and variation in average patient health across different HSAs, along with estimation error. I next outline an empirical Bayes procedure to account for these different variance components and produce more accurate posterior predictions of hospital quality.

[17] In some small samples where maximum likelihood estimates of equation (3.21) fail to converge, RAM controls and higher-order distance terms are sequentially dropped until convergence is achieved.

[18] See Cattaneo et al. (2016) for discussion of many-weak bias in estimating generalized Roy models. Appendix Figure 3.A1 plots the distribution of minimum distance first stage F-statistics that test equality of choice probabilities for each hospital against quality estimate standard errors. As expected, the hospitals with lower first stage F-statistics tend to have higher quality standard errors; less weight will be placed on these estimates in the empirical Bayes procedure.

3.3.3 Posteriors

Under Assumptions 3.1-3.3 we obtain, for a subset of hospitals j with sufficient quasi-experimental data, minimum distance estimates $\hat{\beta}_j$ that are noisy but consistent measures of the true hospital quality indices β_j. At the same time, we observe a full set of observational RAM predictions $\hat{\alpha}_j$ from equation (3.20), which are likely positively, but not perfectly, correlated with quality due to the selection bias detected in section 3.3.1. Following Morris (1983) and Raudenbush and Byrk (1986), I next estimate a hierarchical linear model (HLM) to link these two quality measures.[19] This is

$$\hat{\beta}_j = \kappa + \lambda\hat{\alpha}_j + \mu_{h(j)} + \upsilon_j + \iota_j, \tag{3.22}$$

where $\kappa + \lambda E[\hat{\alpha}_j] = E[\beta_j]$ is the average hospital quality index, $\mu_{h(j)}$ is a random effect for the HSA of hospital j, υ_j is the residual true quality index of hospital j, and ι_j is a mean-zero estimation error term. The HSA random effects, assumed to be identically normally-distributed with mean zero and variance σ^2, capture between-HSA variation in unmeasured quality, while within-HSA variation in residual quality indices $\upsilon_j \sim N(0, \phi^2)$ reflect causal differences not accounted for by observational RAMs. Subject to the usual first-order asymptotic approximation, the estimation error term ι_j can also be modeled as normally-distributed, with a known covariance structure. Consistent estimation of the HLM's hyperparameters κ, λ, σ, and ϕ comes from an ordinary least squares (OLS) regression of quality index estimates $\hat{\beta}_j$ on RAM predictions $\hat{\alpha}_j$, while efficient estimates come from a feasible generalized least squares (FGLS) procedure that uses first-step estimates of σ and ϕ and the covariance of ι_j to iteratively solve for the hyperparameters by weighted least squares. The econometric appendix describes these procedures in more detail.

Table 3.3 reports OLS and FGLS hyperparameter estimates of equation (3.22), where for ease of interpretation the standard deviation of $\hat{\alpha}_j$ has been normalized to one. Column 1 shows that minimum distance quality estimates are indeed correlated with observational RAM predictions, though the OLS estimate of $\hat{\lambda} = 0.11$ is far from statistically significant due to the relative imprecision of the equal-weighted regression. Using the OLS residual variance estimates of $\hat{\sigma} = 0.88$ and $\hat{\phi} = 0.23$ to compute inverse-variance weighted FGLS estimates in column 3 dramatically increases precision: the standard error of $\hat{\lambda}$ falls from 0.16 to 0.04 without much change in the coefficient estimate.

[19] McClellan and Staiger (1999) also use a HLM to combine multiple hospital quality measures.

Iterating this procedure to convergence yields modest additional precision gains in column 4, and a Hausman (1978) test of the random-effects specification relative to a model with HSA fixed effects (reported in column 2) returns a p-value of 0.79. The HLM's decomposition suggests that 90% of the national variation in quality indices β_j is found between HSAs, with only 20% of the remaining within-HSA variation explained by observational RAM predictions and 80% left unexplained.

I use these estimates to generate empirical Bayes posterior predictions of hospital quality that, as in chapter 2 and Chetty and Hendren (2016), shrink noisy quasi-experimental estimates of institutional quality towards precise, but likely biased, observational predictions. The random-effects structure of equation (3.22) further allows the vector of estimates for each HSA to be jointly shrunk towards a HSA-specific mean, thereby accounting for the high local correlation in hospital quality found in Table 3 by $\hat{\sigma} > 0$. In particular, the posterior mean and variance of a HSA's quality indices given vectors of its RAM predictions $\hat{\alpha}_h$ and minimum distance estimates $\hat{\beta}_h$ are

$$E[\beta_h|\hat{\alpha}_h, \hat{\beta}_h] = \Omega_h\hat{\beta}_h + (I_{J(h)} - \Omega_h)(\kappa + \lambda\hat{\alpha}_h) \tag{3.23}$$

$$Var(\beta_h|\hat{\alpha}_h, \hat{\beta}_h) = (I_{J(h)} - \Omega_h)(\phi^2 I_{J(h)} + \sigma^2), \tag{3.24}$$

where Ω_h is a weighting matrix given by the variance hyperparameters and Ξ_h, the variance-covariance matrix of estimation error:

$$\Omega_h = (\phi^2 I_{J(h)} + \sigma^2)(\phi^2 I_{J(h)} + \sigma^2 + \Xi_h)^{-1}. \tag{3.25}$$

Without HSA-level random effects ($\sigma = 0$) and correlated estimation error across hospitals serving the same HSA population (so that Ξ_h is diagonal), these formulas yield the usual empirical Bayes procedure seen in Morris (1983), applied hospital-by-hospital. When additionally $\lambda = 0$, so that observational RAM predictions do not reveal anything about true hospital quality, the minimum distance estimates are shrunk towards the grand mean κ in proportion to one-minus the quality signal-to-noise ratio, as with the simplest empirical Bayes procedures. Given the posterior mean and variance of hospital j's quality index β_j, posterior mean hospital quality is given by

$$E[q_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] = E[\Phi(\beta_j)|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}] = \Phi\left(\frac{E[\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)}]}{\sqrt{1 + Var(\beta_j|\hat{\alpha}_{h(j)}, \hat{\beta}_{h(j)})}}\right) \tag{3.26}$$

since β_j is normally-distributed conditional on $\hat{\alpha}_{h(j)}$ and $\hat{\beta}_{h(j)}$.[20]

I construct hospital quality posteriors using these formulas and the iterated FGLS estimates of the hyperparameters κ, λ, σ, and ϕ.[21] The dashed line in Figure 3.4 shows the distribution of quality index posteriors for the 1,041 hospitals with first-step estimates (Appendix Figure 3.A2 instead plots the full distribution of quality posteriors). As expected, the posterior mean distribution is tighter than the estimate distribution, reflecting empirical Bayes shrinkage and theoretically-improved mean squared prediction error. The posterior mean distribution is also more symmetric, as equation (3.23) downweights the heteroskedastic distribution of estimation error ι_j. The dotted green line in Figure 3.4 shows the distribution of posterior within-HSA quality indices $\kappa + \lambda\hat{\alpha}_j + \upsilon_j$, which is narrower still.

Importantly, equation (3.22) also produces posterior quality predictions for hospitals without a first-step quality estimate due to insufficient quasi-experimental data. In the 69 HSAs without any minimum distance estimates (mostly two hospital HSAs with fewer than 25 admissions), the posterior quality index is simply the HLM fitted values $\hat{\kappa} + \hat{\lambda}\hat{\alpha}_h$, which uses the population relationship between observational RAM and hospital quality to extrapolate to underidentified regions. In the other 626 HSAs these predictions are then shrunk toward the HSA-average quality estimate due to the HLM's random-effects structure. This extrapolation is valid when equation (3.22) describes the relationship between quality indices and observational RAM across all hospitals, whether or not they have enough quasi-experimental variation. Appendix Tables 3.A1 and 3.A4 show that the average characteristics of patients and hospitals across these two groups are quite similar, while Table 3.A3 shows that all main results continue to hold or are strengthened when the HLM includes interactions with the HSA's hospital count, which is the main driver of minimum distance estimate availability and the only observable characteristic that meaningfully varies across the columns of Table 3.A4. I next discuss these findings in detail.

[20] If $x \sim N(m, v)$, $E[\Phi(x)] = Pr(y - x < 0)$ for independent $y \sim N(0, 1)$. Thus $E[\Phi(x)] = \Phi(-E[y - x]/\sqrt{Var(y - x)}) = \Phi(m/\sqrt{1 + v})$.

[21] As usual with empirical Bayes procedures, I treat hyperparameter estimates as known when constructing posteriors. The high degree of precision in Table 3.3's iterated FGLS estimates justifies this simplification in my setting.

3.4 Results

The hyperparameter estimates in Table 3.3 indicate significant within-HSA variation in true hospital quality that is positively, but only partially, correlated with observational RAM predictions. I next use the 2,357 empirical Bayes posterior mean predictions of hospital quality from 695 multi-hospital HSAs to characterize this variation as well as the non-random patient sorting that causes observational and quasi-experimental quality estimates to diverge. I then quantify the significance of this selection bias in two quality-based policies currently in place in U.S. healthcare markets.

3.4.1 Hospital Quality and Patient Sorting

Within-HSA comparisons of quality $E[Y_{ij}] = q_j$ reflect average causal effects of moving a representative patient across different local hospital types. I quantify these effects by regressing various hospital characteristics on a quality measure and HSA fixed effects in the set of multi-hospital HSAs. The characteristics include indicators for a hospital's ownership structure (either private non-profit, private for-profit, or government owned); an indicator for whether it is a teaching hospital; log average hospital spending on emergency Medicare patients; log emergency Medicare patient volume; and log bed capacity. Correlations with posterior quality are reported in the first row of Table 3.4, while the second row regresses hospital characteristics on posterior quality indices $\beta_j = \Phi^{-1}(q_j)$. For comparison purposes, the last two rows report coefficients from regressions on two existing quality measures, conventional RAM predictions and observed hospital survival, and all regressors are normalized to standard deviation units.

The first two rows of Table 3.4 show that moving patients to providers with higher posterior quality and quality indices tends to place them in hospitals that spend more on emergency Medicare patients, have a larger HSA market share, and are less likely to be government-run. I do not find a statistically-significant difference in the probability of admission to for-profit vs. non-profit hospitals, nor any significant correlation with teaching status or bed capacity, though the associated standard errors are sometimes large. With a quality posterior standard deviation of around 12 percentage points, the estimates in the first row of Table 3.4 imply that moving a random patient to a hospital with a 1 percentage point higher potential 30-day survival rate reduces the chances of admission in a government-run provider by 0.7 percentage points and places the patient in a hospital with 1.9% higher emergency Medicare spending and 4.3% higher volume, on average. A

supplementary results appendix section analyzes additional quality dimensions and finds significant within-hospital correlation in quality posteriors across time and by admitting conditions, positive correlations between quality posteriors and measurable inputs (in particular average staff salary), and increases in average quality following a hospital merger.

The findings in the first row of Table 3.4 are broadly consistent with previously documented correlates of observational quality measures, including in Sloan et al. (2001), Silber et al. (2010), Foster et al. (2013), Doyle et al. (2015), and Chandra et al. (2015).[22] Moreover, the third row of Table 3.4 shows similarly signed coefficients from each hospital characteristic regression on RAM predictions, though the strength of the relationship is attenuated with the more-biased quality measure. Hospitals with quality posteriors (RAM predictions) one standard deviation above the HSA mean are 8% (2%) less likely to be government owned, spend 23% (7%) more per Medicare patient, and have a 50% (16%) larger Medicare market share, on average.

This attenuation suggests a *negative* correlation between true hospital quality and the residual selection bias of observational RAMs: better hospitals appear to attract relatively sicker patients, thereby reducing the observed relationship between, say, average spending and mortality. Indeed, the fourth row of Table 3.4 shows no statistically-significant correlation between the most biased quality proxy, observed survival $E[Y_{ij}|D_{ij} = 1]$, and any of the spending, volume, or ownership structure measures found to correlate with the quality posteriors.[23] The negative quality-bias correlation is more broadly illustrated in Figure 3.5, which plots observed survival against quality posteriors net of their HSA means. Points above the dashed 45 degree line represent hospitals with relatively higher selection bias, $E[Y_{ij}|D_{ij} = 1] - E[Y_{ij}]$, while points below are less positively selected than average. The figure shows that hospitals with relatively higher quality posteriors – those to the right of the origin – tend to fall below the 45 degree line and thus be less positively selected. Overall, I find a within-HSA correlation of quality and bias posteriors of -0.83.

The generalized Roy (1951) framework underlying these estimates provides another way to characterize selection: the extent to which patient sorting exploits comparative advantage by admitting at more appropriate hospitals (i.e., selection-on-gains). To explore this, Figure 3.6 plots the distribution of volume-weighted average selection bias posteriors for all multi-hospital HSAs.

[22]The instrumented quality measures used by McClellan and Staiger (2000) and Geweke et al. (2003) also show small and rarely significant differences between for-profit and non-profit hospitals.

[23]For consistency I also shrink observed survival rates towards their grand mean in proportion to one minus the signal-to-noise ratio, though all results are virtually unchanged by this empirical Bayes procedure.

In a constant effects framework, HSA-average bias equals zero by construction; in contrast, the wide distribution in Figure 3.6 suggests a large degree of comparative advantage across emergency healthcare providers. Moreover, most HSAs (86%) appear to have positive average selection bias. In these markets, a typical patient is more likely to survive at the selected hospital than at a hospital picked at random from the market, thus implying that patients benefit from positive Roy selection. Only 15 HSAs (2%) have an average bias posterior of less than -10 percentage points, while the average bias posterior in 440 HSAs (63%) exceeds 10 percentage points.

This finding does not appear to be driven by hospitals specializing in treating different emergency conditions: the shares of positively-selected HSAs in models, described in the supplementary appendix, that estimate quality separately by diagnostic category all exceed 85%. Nor does the result appear driven by the normality assumption, as Table 3.A3 shows a similar 80% of HSAs have positive average selection bias when a Student's $t(2)$ distribution is used. Recall that the non-parametric estimates plotted in Figure 3.3 also suggest pervasive positive selection bias.

A more plausible driver of match-specific quality is differential hospital distance, since individuals suffering from an acute emergency may only survive if brought to the closest available emergency room. Table 3.5 examines the extent to which distance explains selection-on-gains by estimating the average selection bias that would be found if patients were not more likely to attend hospitals close to them. Virtually all HSAs have a negative volume-weighted "distance bias," $E[d_{ij}|D_{ij} = 1] - E[d_{ij}]$, where d_{ij} denotes the ZIP code distance between patient i and hospital j. The mean of this measure across the 695 multi-hospital HSAs is -0.91 miles. However, there is also considerable variation, with patients in some regions sorting to hospitals no more than 0.1 miles closer to them than a provider picked at random from the HSA.

Panel A of Table 3.5 regresses HSA-level survival bias on flexible polynomials in HSA-level distance bias and indeed finds a strong correlation. Nevertheless, the constant in even the most flexible cubic regression in column 3, representing average outcome selection bias in a HSA given zero selection-on-distance, remains significantly positive at 15 percentage points. Panel B reports non-parametric estimates of this quantity by directly computing mean selection bias in HSAs with relatively little distance bias. Even in the 39 regions where average distance bias is above -0.01 miles, patients are still around 9 percentage points more likely to survive at their chosen hospitals then via random admission (74% of these HSAs have positive average bias posteriors). Thus differential hospital distance appears to explain some, but not all, of the Roy selection shown in Figure

3.6.[24] Accommodating unobservable hospital comparative advantage and selection-on-gains with the heterogenous-effects multivariate probit specification – features ruled out by other models such as the linear IV specification from chapter 2 or the fixed-coefficient probits of conventional RAMs and Geweke et al. (2003) – is therefore empirically important in this setting.[25]

3.4.2 Policy Consequences of RAM Bias

Non-random patient sorting generates a sizable distribution of posterior selection bias, with a within-HSA standard deviation of 2.8 percentage points. Although conventional risk-adjustment appears to offset some of this bias, quality posteriors and RAM predictions often disagree, with a within-HSA correlation of 0.68.[26] Around 19% of hospitals (131) with the best quality posteriors in each multi-hospital HSA are ranked differently by RAM, while a similar 20% of HSAs (138) see disagreements on the worst local hospital. Nevertheless, it is difficult to gauge the economic importance of RAM bias from these statistics alone – as shown in chapter 2, policy decisions based on biased quality rankings may still generate large social gains. Furthermore, the distribution shown in Figure 3.5 means that policies that reward or punish hospitals according to observational RAM rankings are most likely to understate true quality differentials, as in Table 3.4. To better assess the economic implications of RAM bias, I next simulate these policies directly.

Medicare Reimbursement

I first consider how payments from Medicare's value-based purchasing (VBP) program would differ if hospital ranks were based on quality posteriors instead of RAMs. VBP was launched in 2013 with the goal of incentivizing hospitals with quality-linked Medicare reimbursement adjustments in a budget-neutral way (DHHS/CMS, 2015). Along with clinical process-of-care measures and patient surveys, risk-adjusted mortality became a part of a "total performance score" (TPS) assigned to each hospital receiving Medicare reimbursement payments in fiscal year 2014. CMS withheld 1.25% of

[24] I find similarly reduced average selection bias within diagnosis categories, with the largest for circulatory and injury conditions.

[25] The EMS staff I interviewed were very receptive to the possibility of comparative advantage and selection on salient unobserved local factors: many hospitals have specialized services such as trauma centers or advanced CT scanners, for example, that are essential for some but not all patients. Ambulance company EMTs and paramedics seem well-poised to exploit these gains; in some states like Massachusetts there are explicit "Point of Entry" guidelines formalizing this institutional knowledge.

[26] For comparison, chapter 2 reports a correlation between conventional middle school value-added predictions and quasi-experimental quality posteriors of 0.85-0.93 in Boston.

each participating hospital's FY2014 diagnosis-related group (DRG) payment, redistributing around $1.1 billion of total withholdings by a linear TPS schedule. Currently, VBP affects only a small share of a hospital's reimbursements; in FY2014, the average VBP penalty was a 0.26 percentage points and the average bonus was a 0.24 percentage points (Conway, 2013). Nevertheless, the program has proved quite controversial as the withholding rate has steadily increased, reaching to 2% in 2016 (Pear, 2014), and as CMS recently announced new plans to tie 90% of all traditional Medicare payments to quality programs like VBP by 2018 (DHHS, 2015). In recent work Norton et al. (2016) show that hospitals indeed respond to the program's seemingly modest incentives, with providers facing higher marginal VBP returns improving their TPS components in subsequent years, while Gupta (2016) finds large incentive effects from the hospital readmissions reduction program, another recently-introduced quality-based reimbursement policy.

I replicate the FY2014 VBP payment schedule to simulate payment adjustments under alternative hospital rankings. Total performance scores combine "achievement points," which are based on hospital quality estimates in the most recent period, and "improvement points," which are based on a hospital's gain relative to a previous period. In FY2014, CMS computed points from hospital risk-standardized mortality rates, defined with the notation of equation (3.20) as

$$RSMR_j = \frac{1 - \sum_{i:D_{ij}=1} F_\nu(\hat{\alpha}_j + \hat{\gamma}' W_i)}{1 - \sum_{i:D_{ij}=1} F_\nu(\bar{\alpha} + \hat{\gamma}' W_i)} (1 - \bar{Y}), \tag{3.27}$$

where F_ν is the distribution of the observational RAM error term ν_i, $\hat{\gamma}$ is an estimate of the RAM parameter γ, $\bar{\alpha}$ is the mean RAM prediction $\hat{\alpha}_j$, and $1 - \bar{Y}$ is the average mortality rate in the sample. In practice, risk-standardized survival rates, $1 - RSMR_j$, correlate strongly with observational RAM predictions ($\rho = 0.98$).

These rates are converted to points by a coarse schedule, with the greater of achievement and improvement points constituting a hospital's outcome domain score. In FY2014 outcome scores made up 25% of a hospital's TPS. Hospitals were refunded none of their DRG withholdings if they scored the minimum level across all three quality domains and linearly accrued payments with higher TPSs. In simulating the distribution of FY2014 payments I hold the non-outcome domains and FY2014 DRG totals fixed, generating benchmark outcome achievement points from the estimated 2010-2012 RAM and computing improvement points from the gain in a hospital's

risk-standardized mortality rate between 2007-2009 and 2010-2012. I then compare simulated VBP reimbursement adjustment rates with those that would be produced with posteriors of the within-HSA component of hospital quality, $\kappa + \lambda\hat{\alpha}_j + \nu_j$, rather than $1 - RSMR_j$. The data appendix describes the construction of simulated payments in more detail.

The results of this simulation are summarized in Table 3.6. Column 1 reports, for different hospital types, the percentage point change in the relative value-based purchasing adjustment from incorporating quasi-experimental data, compared with the prevailing RAM-based adjustment. Column 2 contains this benchmark adjustment, while column 3 reports the implied percentage change in relative VBP adjustments. The results indicate that when using quality posteriors, non-profit and teaching hospitals would see an average of 8.7% and 14.9% higher VBP adjustments, respectively, while government-run hospitals would have their relative VBP adjustment lowered by 8.5%. Table 3.6 also suggests that higher-volume and higher-capacity hospitals would see their VBP payments raised, though the coefficient on log average spending is not statistically significant. As in Table 3.4 and Figure 3.5, the estimates in Table 3.6 show the residual bias in conventional RAM rankings tends to attenuate quality-based VBP differentials rather than changing the types of hospitals that are generally rewarded by performance-linked subsidies.

The magnitudes of changes in column 3 of Table 3.6 are modest, reflecting both the low weight of the outcome domain (25%) and the coarseness of achievement and improvement point schedules. As columns 4-6 show, eliminating the contributions of process-of-care measures and patient surveys magnifies the average change in relative adjustment rates for non-profit, government-run, and teaching hospitals to 44.3%, -45.6%, and 70.2%, respectively. VBP adjustments for relatively higher-volume and higher-capacity hospitals similarly increase, and higher spending hospitals begin to see both higher benchmark reimbursement adjustments and increased payments for quality. Although the policies represented by these columns are far from current VBP practice, together the simulation results suggest bias in observational RAMs has significant capacity to affect performance-based hospital incentive schemes, especially as outcome-based measures become more important. Nevertheless, reducing bias in performance rankings primarily rewards benchmark-subsidized hospitals further and intensifies existing incentive margins, at least along observable dimensions.

Patient Guidance

Along with hospital incentives, supervisory quality rankings have begun to shape patient admission decisions. The federal Hospital Compare website, launched in 2005 to help consumers make informed decisions about their inpatient options, reports multiple hospital performance measures, including observational RAM predictions starting in 2008. At the same time a growing number of private organizations, including the U.S. News and World Report, Consumer Reports, and the Joint Commission, have developed competing hospital "report cards" with alternative risk-adjustment measures. Although patients increasingly consult such rankings (Rice, 2014), and research shows that higher-ranked hospitals tend to see increased future emergency patient market shares (Chandra et al., 2015), there is little evidence on how quality-based admissions may affect patient survival.

The hyperparameter estimates in Table 3.3 suggest that redirecting a typical patient from a random hospital to the provider with the highest RAM ranking likely increases her expected 30-day survival, and that decisions based on less-biased quality posteriors should generate even better average health outcomes. At the same time, the significant degree of positive selection bias shown in Figure 3.6 suggests these gains may be offset by the fact that a typical patient's admissions is better than random: on average, patients already see large survival gains from selecting more appropriate hospitals.

I quantify these effects by simulating 250 realizations of quality indices β_j from the iterated FGLS estimates of κ, λ, σ, and ϕ, holding the distribution of observational RAM predictions fixed. I then draw estimation error components ι_j and construct simulated quality estimates and posteriors. From these data, I compute the average 30-day survival rates for a typical patient admitted to a random hospital within her HSA, the local hospital with the highest survival rate, or the local hospital ranked best by either RAM predictions or quality posteriors. While abstracting away from various general equilibrium effects or capacity constraints, these estimates give a rough sense of the relative public health value of guiding patient admissions by various supervisory quality rankings.

Results of this exercise are plotted in Figure 3.7. Selection bias notwithstanding, an emergency patient sent to the lowest-mortality local hospital is on average 0.9 percentage points more likely to survive their first 30 days after admission, relative to the random admissions benchmark. Using a conventional RAM for admissions further increases the policy's health effect, to 2.8 percentage points. This reduction in 30-day emergency condition mortality is quite large in the historical

context: among Medicare patients admitted for pneumonia, for example, Ruhnke et al. (2011) estimate an average mortality decline due to technological advances of around 3.4 percentage points between 1987 and 2005.

Incorporating quasi-experimental data leads to larger survival gains from report card admission policies, though this improvement is limited by imprecision in minimum distance quality estimates. The last two bars in Figure 3.7 depict the range of possible improvements, from a feasible admission policy with the actual estimation error level found in my sample to an infeasible regime in which all choice probabilities and mean selected outcomes used to construct minimum distance quality estimates are assumed to be known without error. Sending patients to hospitals with the highest quality posteriors leads to incremental 30-day survival rate gains of between 0.5 and 1.7 percentage points, or 18-60% of the 2.8 percentage point gain from RAM-based admission policies. This suggests using less-biased hospital rankings to guide admissions would deliver meaningful partial-equilibrium health returns, particularly when rankings are estimated on larger administrative datasets or by more efficient semi-parametric methods.

At the same time, the simulation results in Figure 3.7 highlight the inherent limitation of supervisory quality-based admission policies applied to settings with significant institutional comparative advantage and positive Roy selection. Moving a patient from the selected (rather than a randomly-chosen) hospital to the local hospital with the highest average quality actually *decreases* expected survival by 11 percentage points. Consumer guidance policies that make average emergency care patients more likely to select high-ranked hospitals (in circumstances where their ambulance operator gives them the choice), as well as policies that close or limit the growth of low-ranked providers, may therefore undermine the prevailing health benefits of hospital selection-on-gains and have unintended negative consequences for average patient health.

3.5 Conclusions

Policymakers in many settings now rely on outcome-based quality measures to incentivize institutions and inform consumers, despite concerns that existing observational methods only partially offset bias from non-random institutional choice. This chapter develops a flexible framework for quantifying institutional performance and selection bias with quasi-experimental data. Quality in these models can be non-parametrically estimated from rich instrument variation, while distributional

restrictions may substitute for constant effects to extrapolate from narrower quasi-experimental designs. Unlike previous likelihood-based estimation methods, a tractable minimum distance procedure implements this semi-parametric approach. Moreover, the models estimated here allow for both institutional comparative advantage and Roy-style selection-on-gains, two important features previously lacking in both linear and nonlinear IV frameworks.

These features are highly relevant in emergency healthcare. I both find a large degree of match-specific hospital quality and that most markets exhibit positive Roy selection, with patients admitting to more appropriate hospitals on average. This non-random sorting generates pervasive selection bias, with a negative quality-bias correlation obscuring important relationships with hospital ownership structure, patient volume, and average spending. Observational risk-adjustment methods remove some of this bias, generating survival gains in simulations of ranking-based guidance policies, while quasi-experimental quality posteriors can further improve the targeting of both Medicare reimbursement and patient guidance programs.

Ultimately, more work is needed to characterize the ways in which these policies may shape long-run hospital quality supply and demand. As long as biased quality measures are used to structure the value-based purchasing program, providers may find ways to "game the system," boosting their payments without improving actual performance. While the simulations in section 3.4 show that most observable hospital characteristics currently rewarded by VBP are only further subsidized by policies based on less-biased quality posteriors, there may remain various hospital-controlled unobservables that correlate with RAM rankings but not true quality. Detecting VBP "gaming" may become easier as the scope of performance-linked healthcare reimbursement and the strength of incentives grow.

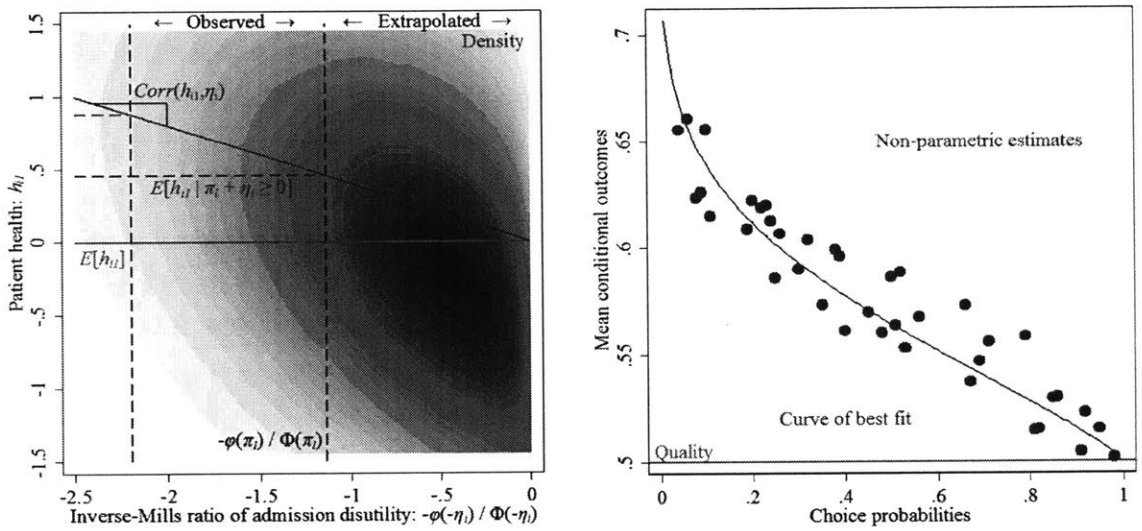
The simulations also raise new questions about the efficacy of demand-side interventions, including the large and growing set of hospital report cards currently consulted by patients. With constant causal effects, the finding that higher-ranked hospitals tend to attract more emergency patients in the future, as in Chandra et al. (2015), has unambiguously positive implications for public health. Accounting for the significant extent of selection on match-specific quality, however, requires a more nuanced analysis. On one hand, report cards may cause patients to update weak or incorrect priors on their most appropriate hospital and induce the selection of providers with high average quality, thus increasing patients' chances of survival. However, widely-known rankings may also disrupt prevailing beneficial selection patterns, to the extent they also influence patients

with better private information. Understanding the ways in which hospital performance measures actually affect admission decisions and characterizing the optimal design of public quality signals in settings with Roy selection are two important goals raised by the heterogeneous-effects framework.

3.6 Figures and Tables

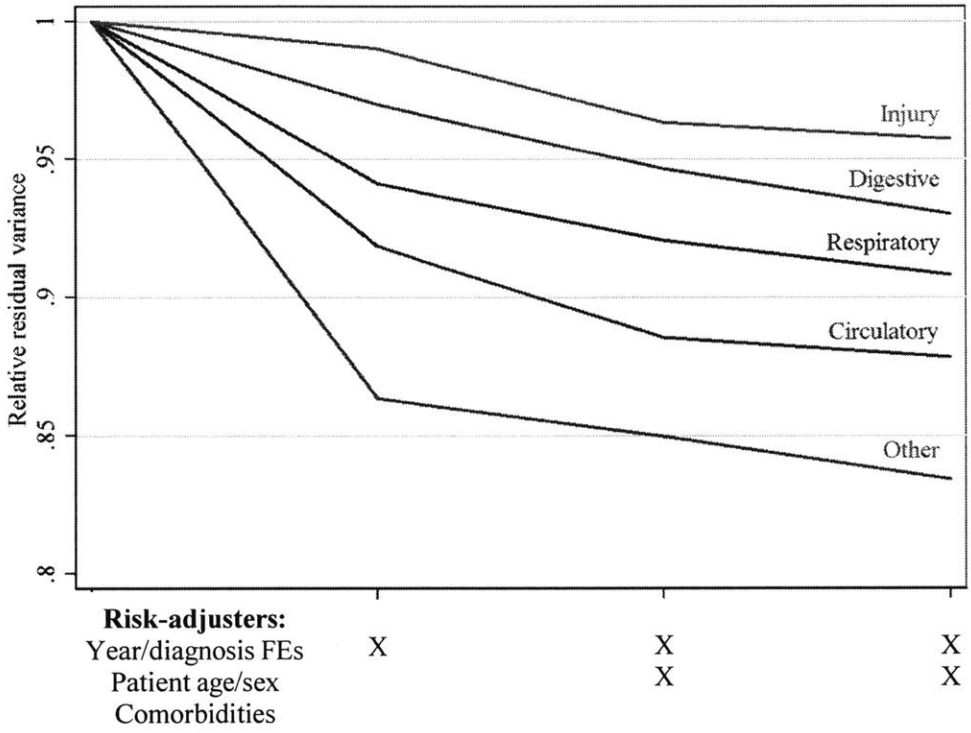
Figure 3.1: Quality Identification and Estimation in a Bivariate Probit Model

A. Identification ($L = 2$) B. Estimation ($L > 2$)



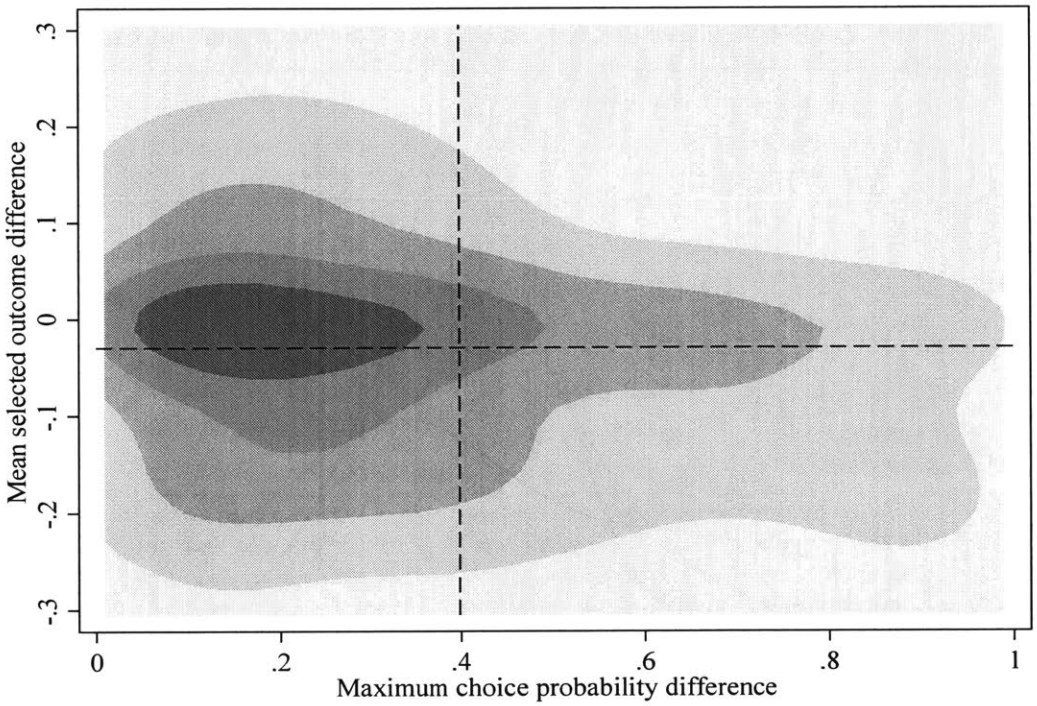
Notes: Panel A shows the probability density function of potential patient health and the inverse Mills ratio of latent admission disutility for a hospital with positive selection bias and joint-normal health and utility. The vertical dashed lines indicate inverse Mills transformations of the first-stage preference parameters for two different ambulance companies, while the horizontal dashed lines indicate the average health of patients that would be admitted to the hospital by each company. The downward-tilted line meets zero on the x-axis at the hospital's population average health ($\beta_1 = 0$) on the y-axis, and its slope ($-\rho_1 = -0.4$) represents the population correlation of health and disutility. Panel B shows estimated mean conditional outcomes (survival probabilities) for patients admitted by a set of ambulance companies against the associated choice probability from the same model. The curve of best fit equals the population survival probability ($\Phi(\beta_1) = 0.5$), that is, the hospital's quality, when the choice probability equals one.

Figure 3.2: Residual Survival Variance in Observational RAMs



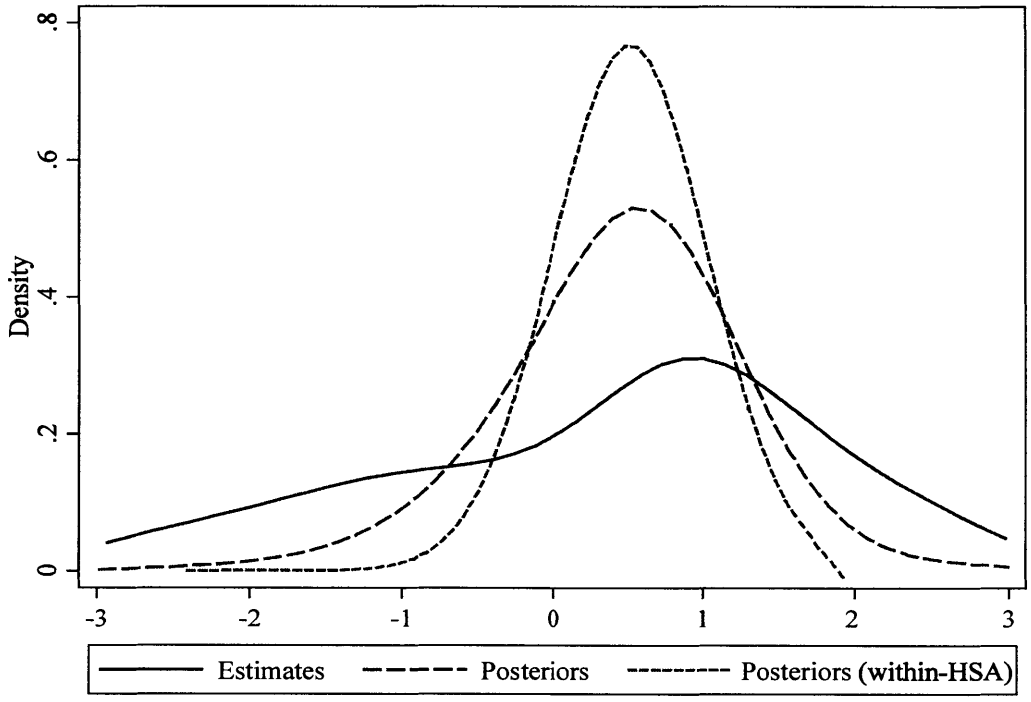
Notes: This figure plots the variance of risk-adjusted 30-day survival relative to the unadjusted survival variance for three risk-adjustment models, estimated separately by diagnosis category. See Table 3.1 for a description of each diagnosis category, Table 3.2 for a list of included comorbidities, and the data appendix for a description of the RAM estimation procedure.

Figure 3.3: The Joint Distribution of Ambulance Effects on Hospital Choice and Patient Survival



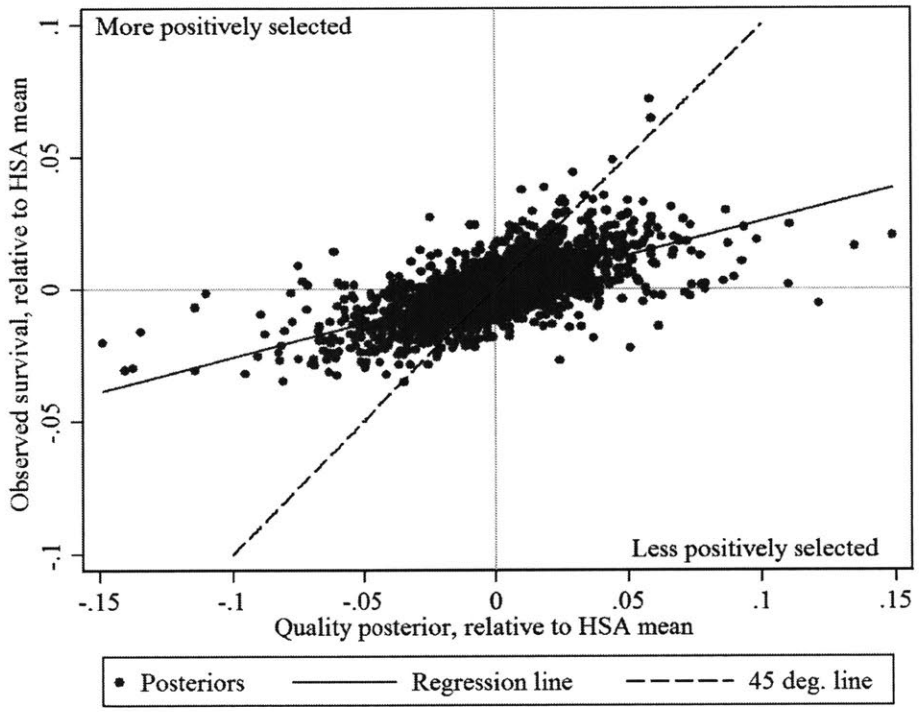
Notes: This figure plots a Gaussian kernel density estimate of the joint distribution of estimated mean selected outcome differences and estimated choice probability differences for 1,041 hospitals with minimum distance quality estimates. Differences are taken across the two ambulance companies with the maximal estimated choice probability difference for each hospital and estimate causal effects of differential ambulance company assignment on hospital choice and 30-day survival for admitted patients. The vertical and horizontal bandwidths used to estimate this distribution are 0.05 and 0.1. Dashed lines indicate sample means.

Figure 3.4: The Distribution of Hospital Quality Index Estimates and Posteriors



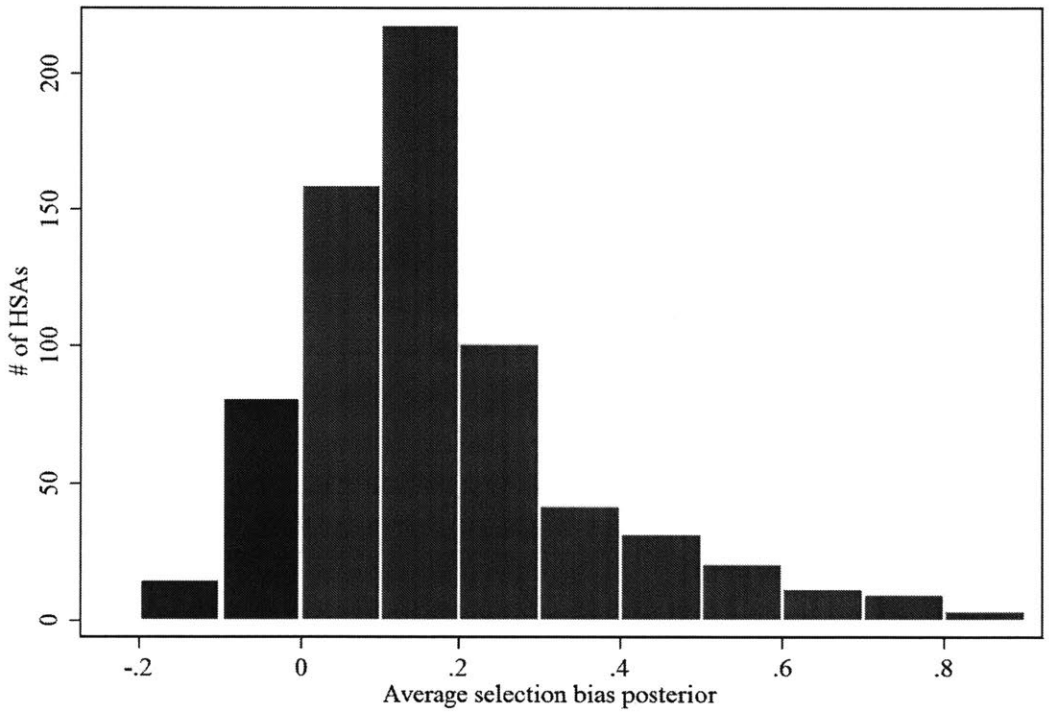
Notes: This figure plots Gaussian kernel density estimates of the distribution of minimum distance hospital quality index estimates and empirical Bayes posteriors of both the overall and within-HSA quality indices. The sample includes 1,041 hospitals operating in 626 multi-hospital HSAs with a first-step quality estimate. The bandwidth used to estimate each distribution is 0.5.

Figure 3.5: Within-HSA Variation in Hospital Quality and Selection Bias



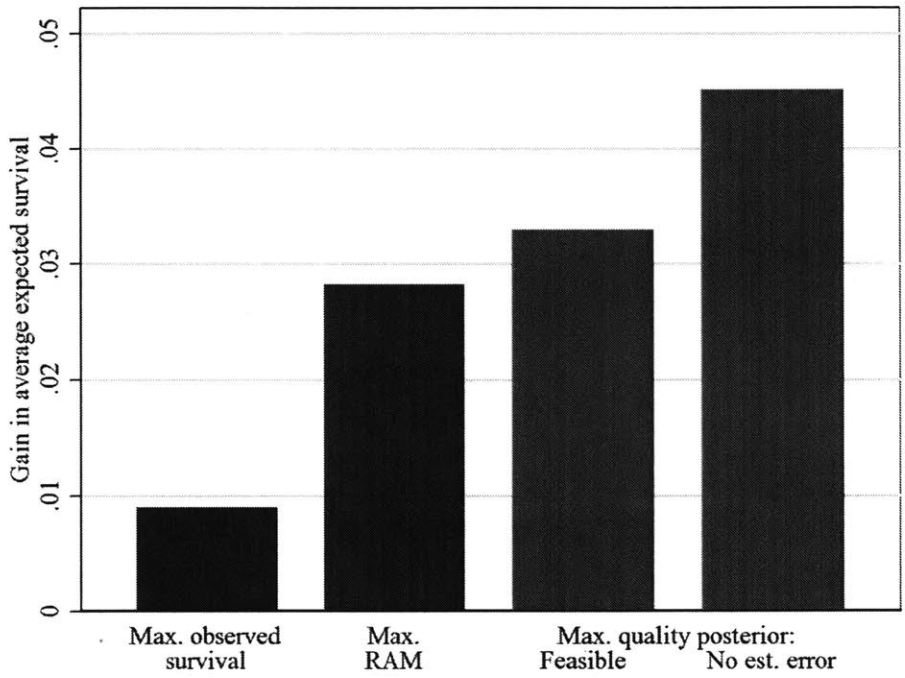
Notes: This figure plots posterior hospital survival rates against posterior quality, both net of their HSA means. The sample includes 2,357 hospitals operating in 695 multi-hospital HSAs. Points above the dashed 45-degree line represent hospitals that are relatively more positively selected within their HSA, while hospitals below the 45-degree line are relatively less positively selected.

Figure 3.6: The Distribution of HSA-average Selection Bias



Notes: This figure plots the distribution of volume-weighted average posterior selection bias across 695 multi-hospital HSAs. HSAs with negative selection bias would see higher average 30-day survival if patients were randomly allocated to hospitals, while a positively-selected HSA would have a lower survival rate under random admissions.

Figure 3.7: Survival Gains from Selecting a Top-ranked Hospital, Relative to Random Admissions



Notes: This figure plots simulated gains in average expected survival for a random patient sent to the highest-ranked hospital in her HSA, relative to a random admission, according to the hospital's 30-day survival rate, observational RAM prediction, or quality posterior (with and without estimation error). The sample consists of 2,357 hospitals operating in 695 multi-hospital HSAs. Estimates are from 250 draws of the hierarchical model described in the text.

Table 3.1: The Analysis Sample

	Diagnoses (1)	Patients (2)	Ambulances (3)	Hospitals (4)	HSAs (5)	30-day survival (6)
Full sample	29	405,173	9,590	4,821	3,159	0.833
			A. By diagnosis category			
Circulatory	5	89,077	7,578	3,879	2,777	0.807
Respiratory	4	81,021	7,432	4,224	2,980	0.781
Digestive	6	26,359	5,244	3,323	2,354	0.902
Injury	8	71,616	7,396	3,634	2,561	0.931
All other	6	137,100	8,064	4,441	2,997	0.815
			B. By HSA hospital count			
One	29	151,072	6,756	2,464	2,464	0.831
Two	29	84,634	3,578	800	400	0.837
Three	29	44,399	2,302	396	132	0.835
Four	29	24,398	1,227	212	53	0.829
Five or more	29	100,670	3,775	949	110	0.832

Notes: This table summarizes the distribution of diagnoses, ambulances, hospitals, and 30-day survival in the sample of Medicare FFS patients admitted for one of 29 nondeferrable diagnoses in 2010-2012. Circulatory diagnoses include acute myocardial infarction, intracerebral hemorrhage, occlusion and stenosis of the precerebral artery, occlusion of cerebral arteries, and transient cerebral ischemia. Respiratory diagnoses include pneumonia due to solids and liquids, pneumonia (organism unspecified), other bacterial pneumonia, and other diseases of the lung. Digestive diagnoses include diseases of the esophagus, gastric ulcer, duodenal ulcers, vascular insufficiency of the intestine, intestinal obstruction without mention of hernia, and other/unspecified noninfectious gastroenteritis and colitis. Injury diagnoses include fracture of the ribs, sternum, larynx, and trachea; fracture of the pelvis; fracture of the neck or femur; fracture of the tibia and fibula; fracture of the ankle; poisoning by angesicis; antipyretics, and antirheumatics; poisoning by psychotropic agents; and other/unspecified injury. All other diagnoses include septicemia; malignant neoplasm of the trachea, bronchus, and lung; secondary malignant neoplasm of respiratory and digestive systems; other disorders of the urethra and urinary tract; disorders of muscle, ligament, and fascia; and general symptoms.

Table 3.2: Ambulance Company Assignment Balance

	Assigned ambulance company's closest hospital		Equality p-value	Regressions on RAM of the ambulance's closest hospital
	Low RAM	High RAM		
	(1)	(2)	(3)	(4)
RAM prediction	-0.058	0.015	<0.001	0.101 (0.010)
	A. Demographics			
Age	81.56	81.62	0.790	0.148 (0.214)
Male	0.380	0.384	0.784	-0.002 (0.013)
White	0.859	0.849	0.259	-0.004 (0.009)
Black	0.092	0.099	0.395	0.004 (0.007)
Referred from home	0.639	0.607	0.008	-0.043 (0.013)
Referred from accident	0.130	0.125	0.546	-0.009 (0.009)
Circulatory diagnosis	0.236	0.229	0.531	-0.014 (0.011)
Respiratory diagnosis	0.187	0.185	0.852	0.009 (0.010)
Digestive diagnosis	0.065	0.067	0.800	0.003 (0.006)
Injury diagnosis	0.174	0.184	0.294	0.002 (0.010)
	B. Comorbidities			
Hypertension	0.262	0.271	0.419	0.009 (0.012)
Stroke	0.011	0.012	0.708	0.002 (0.003)
Cerebrovascular disease	0.032	0.034	0.614	0.003 (0.005)
Renal failure	0.117	0.121	0.643	0.009 (0.009)
Dialysis	0.012	0.012	0.943	0.001 (0.003)
Chronic obstructive pulmonary disease	0.107	0.107	0.992	0.005 (0.008)
Pneumonia	0.052	0.055	0.682	0.003 (0.006)
Diabetes	0.120	0.133	0.126	0.012 (0.009)
Protein-calorie malnutrition	0.035	0.037	0.680	0.003 (0.005)
Dementia	0.082	0.093	0.132	0.009 (0.007)
Paralysis	0.032	0.037	0.213	0.006 (0.005)
Peripheral vascular disease	0.073	0.077	0.498	0.001 (0.007)
Metastic cancer	0.020	0.020	0.896	0.001 (0.004)
Trauma	0.057	0.058	0.820	0.003 (0.006)
Substance abuse	0.039	0.037	0.727	0.001 (0.006)
Major psychological disorder	0.029	0.031	0.672	0.001 (0.005)
Chronic liver disease	0.007	0.007	0.796	0.002 (0.002)
	C. Ambulance services			
Excess miles transported	-0.044	0.048	0.985	0.072 (0.069)
Emergency transport	0.956	0.961	0.331	0.013 (0.005)
Advanced life support	0.727	0.728	0.994	0.001 (0.012)
Intravenous fluids administered	0.009	0.008	0.869	-0.004 (0.002)
Intubation performed	<0.001	<0.001	0.228	-0.001 (0.001)
Demographics, comorbidities, and ambulance services joint p-value: 0.887				

Notes: This table compares the characteristics of patients referred by ambulance companies located close to hospitals with high and low RAM predictions, within patient ZIP codes. The sample includes 254,101 patients admitted to 2,357 hospitals in 695 multi-hospital HSAs. Columns 1 and 2 report average characteristics of patients assigned to ambulances that are closest (in terms of ZIP code centroid distance) to hospitals in the first and fourth quartiles of RAM predictions in their HSA, controlling for ZIP code fixed effects. Column 3 reports robust p-values for tests of equality across the two groups. Column 4 reports coefficients and robust standard errors from regressions on the assigned ambulance's closest hospital's RAM, controlling for ZIP code fixed effects. Excess miles transported is computed as a patient's transported miles minus the ZIP code centroid distance to a patient's hospital.

Table 3.3: Hierarchical Linear Model Estimates

	OLS	HSA fixed effects	Random effects (FGLS)	
			Two-step	Iterated
	(1)	(2)	(3)	(4)
RAM coefficient (λ)	0.106	0.029	0.111	0.111
	(0.155)	(0.311)	(0.038)	(0.038)
Constant (κ)	0.222		0.502	0.502
	(0.155)		(0.043)	(0.042)
Variance components:				
Within-HSA (φ)	0.232	0.150	0.189	0.248
	(0.039)	(0.053)	(0.045)	(0.037)
Between-HSA (σ)	0.878		0.802	0.811
	(0.034)		(0.037)	(0.037)
Hausman test statistic (1 d.f.)				0.07
				[0.790]

Notes: This table reports estimated parameters of the hierarchical linear model outlined in the text. The sample consists of 1,041 minimum distance quality index estimates and RAM predictions from 626 multi-hospital HSAs. Column 1 reports OLS coefficients and variance estimates from a regression of quality index estimates on RAM predictions and a constant. Column 2 reports estimates an OLS regression with HSA fixed-effects, while columns 3 and 4 report two-step and iterated FGLS random-effect estimates. See the econometric appendix for details on this estimation procedure. The Hausman statistic tests equality of the estimates in columns 2 and 4. Standard errors, clustered by HSA, are reported in parentheses; the test's p-value is reported in brackets.

Table 3.4: Correlations of Quality Measures and Hospital Characteristics

Regressor:	Non-profit hospital (1)	For-profit hospital (2)	Government hospital (3)	Teaching hospital (4)	Log (avg. spending) (5)	Log (volume) (6)	Log (# of beds) (7)
Quality posterior	0.055	0.024	-0.079	-0.018	0.230	0.517	-0.057
	(0.040)	(0.033)	(0.029)	(0.041)	(0.111)	(0.223)	(0.084)
Quality index posterior	0.044	0.025	-0.069	-0.018	0.161	0.388	-0.015
	(0.034)	(0.029)	(0.022)	(0.033)	(0.106)	(0.219)	(0.088)
RAM prediction	0.022	-0.002	-0.020	-0.016	0.065	0.159	-0.013
	(0.012)	(0.010)	(0.009)	(0.015)	(0.036)	(0.068)	(0.031)
Observed survival	0.003	0.009	-0.012	-0.050	0.009	-0.089	-0.137
	(0.012)	(0.011)	(0.009)	(0.013)	(0.032)	(0.058)	(0.027)

Notes: This table reports coefficients from regressions of the hospital characteristic in each column on the quality measure in each row, controlling for HSA fixed effects. All regressors are normalized to standard deviation units. Observed survival posteriors shrink observed rates towards the grand mean in proportion to one minus the signal-to-noise ratio. The sample is 2,357 hospitals operating in 695 multi-hospital HSAs. Standard errors, clustered by HSA, are reported in parentheses.

Table 3.5: Estimates of Average Selection Bias Net of Hospital Distance Bias

Regressors:	(1)	(2)	(3)
	A. Parametric		
Constant	0.165	0.159	0.154
	(0.008)	(0.008)	(0.008)
Avg. distance bias (marginal effect)	-0.009	-0.018	-0.031
	(0.003)	(0.005)	(0.007)
Polynomial:	Linear	Quadratic	Cubic
HSAs:		695	
	B. Non-parametric		
Constant	0.160	0.117	0.089
	(0.015)	(0.019)	(0.022)
Bandwidth:	1 mile	0.1 miles	0.01 miles
HSAs:	142	66	39

Notes: This table summarizes regressions of average selection bias posteriors for 695 multi-hospital HSAs. Panel A regresses bias posteriors on polynomials in HSA-average distance bias. A hospital's distance bias is the difference between its average ZIP code centroid distance to its admitted patients and its average distance to all potential patients in the HSA. Panel B reports average selection bias posteriors for HSAs with an average distance bias that falls within narrow bandwidths of zero. The constant from both sets of regressions estimates average selection bias that is not explained by the relative distance between admitted patients and their hospitals. Robust standard errors are reported in parentheses.

Table 3.6: Correlates of Changes in Value-based Purchasing Adjustments

	25% outcome domain weight			100% outcome domain weight		
	Percentage point change	Benchmark adjustment	Change as % of | benchmark |	Percentage point change	Benchmark adjustment	Change as % of | benchmark |
Regressor:	(1)	(2)	(3)	(4)	(5)	(6)
Non-profit hospital	0.014	0.160	8.70	0.074	0.168	44.28
	(0.005)	(0.029)		(0.026)	(0.042)	
For-profit hospital	-0.003	-0.032	-10.04	-0.022	-0.055	-39.57
	(0.007)	(0.036)		(0.032)	(0.054)	
Government hospital	-0.018	-0.217	-8.50	-0.094	-0.206	-45.57
	(0.006)	(0.035)		(0.030)	(0.049)	
Teaching hospital	0.013	0.089	14.92	0.072	0.103	70.21
	(0.006)	(0.033)		(0.029)	(0.047)	
Log (avg. spending)	-0.001	-0.137	-0.99	0.100	0.348	28.62
	(0.008)	(0.062)		(0.036)	(0.069)	
Log (volume)	0.008	0.196	4.12	0.053	0.246	21.61
	(0.002)	(0.010)		(0.011)	(0.017)	
Log (# of beds)	0.008	0.132	6.09	0.059	0.208	28.18
	(0.003)	(0.018)		(0.015)	(0.024)	

Notes: Columns 1 and 4 report coefficients from regressions of changes in simulated value-based purchasing reimbursement adjustment percentages from using quality posteriors instead of risk-standardized survival rates. Each row represents a different regression. Columns 2 and 5 report coefficients from regressions of the original adjustment percentages, while columns 3 and 6 report the changes as a percentage of absolute benchmarks. Columns 1-3 use a 25% outcome domain weight while columns 4-6 use a 100% outcome domain weight. Both simulations use FY2014 balance sheet information, withholdings, and non-quality domain scores. See the data appendix for a detailed description of the reimbursement schemes. The sample is 2,565 hospitals with balance sheet information and quality posteriors from both the 2007-2009 and 2010-2012 periods. Robust standard errors, clustered by HSA, are reported in parentheses.

3.7 Data Appendix

I follow Doyle et al. (2015) in constructing an analysis sample from 2010-2012 CMS claims. I first link a 20% random sample of Medicare beneficiaries that originate an ambulance company claim in the CMS Carrier file to their inpatient claims, which indicate admitting hospitals and diagnoses. The claims data also include basic demographic information on beneficiaries, including birth date, sex, race, and the ZIP code where official correspondence is sent. The data are further linked to vital statistics that record when a patient dies, thereby generating the primary 30-day survival outcome. Ambulance company data, including the company's registered ZIP code, information on miles traveled, the mode and method of transport, and any pre-hospital interventions for each claim are retained from the Carrier file. Hospital ZIP codes provided by inpatient claims and linked to hospital service areas defined by the Dartmouth Atlas. Data on hospital ownership structure (non-profit private, for-profit private, and government owned) and the total number of hospital beds come from the CMS Provider of Service files, while teaching status and total diagnosis-related group payments for FY2014 come from hospital Cost Report data. Hospital volume is computed as the total number of admitted patients in the analysis sample, while average spending includes all Medicare reimbursement paid to the hospital from the first 30 days following a patient's admission, excluding those for drugs covered under Medicare Part D due to data limitations.

Following Card et al. (2009) and others, I limit the sample to patients who were admitted by ambulance through a hospital's emergency room and receiving a primary diagnosis of one of 29 "nondeferrable" conditions wherein selection into inpatient care is unlikely to be discretionary. These are the same conditions Doyle et al. (2015) identify as having weekend admissions rates close to the 2/7ths, which would be expected given no selection, and are listed in the footnote of Table 1. As with the CMS risk-adjustment methodology established by YNHHSC/CORE (2013), I keep only a patient's first hospital admission in 2010-2012. Unlike Doyle et al. (2015), I do not drop small ZIP codes, ambulances, or hospitals, nor do I limit the sample to hospitals within 50 miles of the patient's ZIP code centroid in order to minimize endogenous sample selection concerns.

Appendix Table 3.A1 summarizes patient demographics in the analysis sample. Around 41% of beneficiaries admitted for a nondeferrable condition in 2010-2012 (column 1) were done so via ambulance (column 2); this subsample is slightly older and more female, with somewhat higher average Medicare spending and 30-day mortality. Columns 3 and 4 of Table 3.A1 further report

demographics for patients admitted in multi-hospital HSAs and to hospitals with enough quasi-experimental data to estimate quality by the minimum distance procedure outlined in the text; these subsamples appear quite representative of the full analysis sample.

Observational RAMs are estimated as hierarchical logit regressions with normal random hospital effects, separately by each of the five diagnosis categories listed in Table 3.1. RAM predictions $\hat{\alpha}_j$ are the volume-weighted average posterior means of the hospital effects. The benchmark RAM specification includes diagnosis and year fixed effects, patient age and sex, and the 17 diagnosis comorbidities listed in Panel B of Table 3.2. Appendix Table 3.A2 also uses estimates from replicated CMS-RAM models. For these I follow YNHHSC/CORE (2013) as closely as possible in constructing a 20% sample from 2010-2012 inpatient claims and defining diagnosis and procedure comorbidities specific to each of their AMI, heart failure, and pneumonia risk-adjustment models. The AMI model is estimated using a sample of 107,916 patients and includes indicators for the comorbidities listed in Table 2 of YNHHSC/CORE (2013). The heart failure model uses a sample of 206,363 patients and includes the comorbidity controls listed in their Table 6. Lastly, the pneumonia specification is estimated using a sample of 205,980 patients and includes comorbidity indicators that YNHHSC/CORE (2013) list in their Table 12. Regressions of reported CMS hospital scores on those generated in my samples produce coefficients of 0.93 (AMI), 1.05 (heart failure), and 1.03 (pneumonia) with standard errors on the order of 0.04, which suggests a faithful reproduction.

To simulate payments from the CMS value-based purchasing program, I replicate as closely as possible the methodology outlined in DHHS/CMS (2015). I obtain FY2014 non-outcome domain scores from the VBP website and hold them fixed throughout. Achievement and improvement scores for the outcome domain are obtained from either conventional risk-standardized survival rates, as described in the text, or from $\hat{\kappa} + \hat{\lambda}\alpha_j + E[v_j|\hat{\alpha}, \hat{\beta}]$, the posterior mean of a hospital's within-HSA quality index. I generate achievement scores from the main 2010-2012 analysis sample, while improvement scores come from changes in these measures between 2007-2009 and 2010-2012. Achievement points are awarded on a linear 0-9 scale, with zero points given to hospitals that score below the median achievement score and 9 points awarded to those scoring above the mean of hospitals in the top tenth percentile. No improvement points are assigned to hospitals with negative improvement scores but are earned linearly from positive improvement with 8 points awarded to hospitals above the mean of the top tenth percentile of improvement. A hospital's overall score for the outcome domain is the maximum of achievement and improvement points multiplied by 10,

which is combined with the non-outcome domains with a weight of either 25%, 40%, or 100% to arrive at its Total Performance Score. Total VBP withholdings equal 1.25% of total hospital DRG payments in FY2014 and are fully redistributed to hospitals by a linear schedule, with hospitals scoring zero on their Total Performance Score earning back zero withholdings. VBP percentage adjustments are given by these payments divided by a hospital's withholdings.

3.8 Econometric Appendix

Proof of Lemma 3.1

Consider the choice probability for institution j and instrument value ℓ:

$$\begin{aligned} Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) &= E[E[\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]|X_i]] \\ &= E[E[\mathbf{1}[U_{ij}(Z_i) \geq U_{ik}(Z_i), \forall k]|Z_{i\ell} = 1, X_i]] \\ &= E[E[D_{ij}|Z_{i\ell} = 1, X_i]] \\ &= E\left[E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}|X_i\right]\right] \\ &= E\left[\frac{D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \end{aligned} \tag{3.28}$$

The first and fifth equalities follow from the Law of Iterated Expectations, the second holds under Assumption 3.1, and the third and fourth use the model for D_{ij} and definition of $p_\ell(X_i) = E[Z_{i\ell}|X_i]$. Similar logic yields equation (3.4), as each mean selected outcomes can be written

$$E[f(Y_{ij})|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k] = \frac{E[f(Y_i)\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]]}{Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)}, \tag{3.29}$$

and, following the same steps as above,

$$E[f(Y_i)\mathbf{1}[U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k]] = E\left[\frac{f(Y_i)D_{ij}Z_{i\ell}}{p_\ell(X_i)}\right]. \tag{3.30}$$

Combining equations (3.28)-(3.30) completes the proof. □

Multivariate Probit Choice Probabilities and Mean Selected Outcomes

Under the model given by equations (3.10) and (3.12), Assumption 3.3, and the necessary normalizations,

$$Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k) = Pr(\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k)$$
$$= \frac{1}{(2\pi)^{(J-1)/2}|\tilde{\Sigma}_\eta|^{1/2}} \int_{-\infty}^{\tilde{\pi}_{j\ell}-\tilde{\pi}_{1\ell}} \cdots \int_{-\infty}^{\tilde{\pi}_{j\ell}-\tilde{\pi}_{J'\ell}} \exp\left(-\frac{1}{2}t'\tilde{\Sigma}_\eta^{1/2}t\right) dt, \tag{3.31}$$

where $\tilde{\pi}_{j\ell} = \pi_{j\ell} - \pi_{\bar{j}\ell}$ for fixed $\bar{j}$, $J' = J$ if $j \neq J$ and otherwise $J' = J - 1$ (i.e. the integration is taken over all $k \neq j$), and $\tilde{\Sigma}_\eta = 1 + I_{J-1}$ is the variance matrix of the vector of $\eta_{ik} - \eta_{ij}$, for $k \neq j$. By the exchangeable correlation structure of $\tilde{\Sigma}$, we can furthermore use the result in Dunnett (1989) to rewrite this as a more computationally-tractable expression involving a single integral:

$$Pr(U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k)$$
$$= \frac{1}{\sqrt{\pi}} \int_0^\infty \left(\prod_{k\neq j} \Phi\left(-\sqrt{2}t - (\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell})\right) + \prod_{k\neq j} \Phi\left(\sqrt{2}t - (\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell})\right) \right) \exp\left(-t^2\right) dt. \tag{3.32}$$

Similarly, $E[Y_{ij}|U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k] = Pr(h_{ij} \geq 0, \tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k)/Pr(\tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k)$ with

$$Pr(h_{ij} \geq 0, \tilde{\pi}_{j\ell} - \tilde{\pi}_{k\ell} \geq \eta_{ik} - \eta_{ij}, \forall k)$$
$$= \frac{1}{(2\pi)^{J/2}|\tilde{\Sigma}(\rho_j)|^{1/2}} \int_{-\infty}^{\beta_j} \int_{-\infty}^{\tilde{\pi}_{j\ell}-\tilde{\pi}_{1\ell}} \cdots \int_{-\infty}^{\tilde{\pi}_{j\ell}-\tilde{\pi}_{J'\ell}} \exp\left(-\frac{1}{2}t'\tilde{\Sigma}(\rho_j)^{1/2}t\right) dt, \tag{3.33}$$

where $\rho_j = (\rho_{j1}, \ldots, \rho_{jJ'})$ for $\rho_{jk} = Corr(h_{ij}\eta_{ik} - \eta_{ij})$ and

$$\tilde{\Sigma}(\rho_j) = \begin{bmatrix} 1 & -\sqrt{2}\rho_j' \\ -\sqrt{2}\rho_j & \tilde{\Sigma}_\eta \end{bmatrix}. \tag{3.34}$$

Testing Hospital RAMs

For the general quality model given by equations (3.1)-(3.2), consider the null hypothesis

H_0 (*RAM Validity*): $Y_{ij} = \mathbf{1}[a_j + \gamma' W_i \geq \nu_i]$, where $\nu_i \mid \left((Z_{i\ell}, (U_{ij}(z_\ell))_{j=1,\ldots,J})_{\ell=1,\ldots,L}, W_i\right) \sim F_\nu$

for some distribution F_ν. In the health context, H_0 rules out hospital comparative advantage and says that the patient sorting mechanism is independent of latent health, conditional on the controls in W_i. In particular, H_0 implies $\nu_i \perp\!\!\!\perp D_i \mid W_i$, the usual basis for consistent estimation of equation (3.20), and is equivalent when ignoring knife-edge cases of perfectly-offsetting dependencies between health, utility, and ambulance company assignment.

By the Law of Iterated Expectations, we have under H_0 that

$$\begin{aligned} E[Y_i|Z_{i\ell} = 1, D_{ij} = 1, W_i = w] &= Pr(\alpha_j + \gamma' w \geq \nu_i | Z_{i\ell} = 1, U_{ij}(z_\ell) \geq U_{ik}(z_\ell), \forall k, W_i = w) \\ &= F_\nu(\alpha_j + \gamma' w) \\ &= E[Y_i|D_{ij} = 1, W_i = w], \end{aligned} \tag{3.35}$$

for any ambulance company ℓ, hospital j, and w in the support of W_i. Given a first-step estimate of the propensity score $p_\ell(D_i, W_i) = E[Z_{i\ell}|D_i, W_i]$, a non-parametric test statistic for H_0 based on equation (3.35) is therefore the sample analogue of

$$\begin{aligned} E\left[Y_i \left(\frac{Z_{i\ell} - p_\ell(D_i, W_i)}{p_\ell(D_i, W_i)(1 - p_\ell(D_i, W_i))}\right)\right] &= E\left[E\left[\frac{Y_i Z_{i\ell}}{p_\ell(D_i, W_i)} - \frac{Y_i(1 - Z_{i\ell})}{1 - p_\ell(D_i, W_i)} | D_i, W_i\right]\right] \\ &= E[E[Y_i|Z_{i\ell} = 1, D_i, X_i] - E[Y_i|Z_{i\ell} = 0, D_i, X_i] = 0 \end{aligned} \tag{3.36}$$

where the last equality follows by H_0.

An alternative test procedure leverages knowledge of the error distribution F_ν, noting that under H_0,

$$\begin{aligned} E[Y_i|Z_i] &= E[E[Y_i|Z_i, D_i, W_i]|Z_i] \\ &= E[F_\nu(\alpha' D_i + \gamma' W_i)|Z_i], \end{aligned} \tag{3.37}$$

so that $E[Y_i|Z_i] - E[F_\nu(\alpha' D_i + \gamma' W_i)|Z_i] = 0$. Given first-step coefficient estimates of the RAM parameters α and γ, this equality can be verified by a Lagrange Multiplier test statistic that checks orthogonality of the RAM's residuals $Y_i - F_\nu(\alpha' D_i + \gamma' W_i)$ with the instrument. As when validating linear VAMs (Angrist et al., 2016b), a first-order equivalent Wald test statistic uses the fact that

equation (3.37) implies vector-equality of the coefficients μ_Y and μ_F in the regressions:

$$Y_i = \mu_Y' Z_i + e_Y \tag{3.38}$$

$$F_\nu(\alpha' D_i + \gamma' W_i)) = \mu_F' Z_i + e_F. \tag{3.39}$$

A final approach notes that equations (3.38) and (3.39) are the reduced form and first stage equations of a two-stage least squares (2SLS) procedure that uses Z_i to instrument for RAM-predicted survival in a regression of realized survival Y_i. Since $\mu_Y = \mu_F$ under H_0, this procedure should produce a 2SLS of one when the RAM is valid. As in the education setting, testing the L restrictions of the Lagrange Multiplier and Wald statistic can be viewed as combining a single degree-of-freedom test for "forecast bias" or that the 2SLS "forecast coefficient" equals one (Kane and Staiger, 2008), with the 2SLS model's $L - 1$ overidentifying restrictions.

Panel A of Appendix Table 3.A2 reports chi-squared statistics and associated p-values for non-parametric propensity score tests, using 100 randomly-selected ambulance companies admitting at least 100 patients in the main analysis sample to simplify computation. For each observational RAM specification in Figure 3.2, I approximate the propensity scores $p_\ell(D_i, W_i)$ by a probit model and jointly test significance of the 100 sample analogues of equation (3.36), correcting inference for first-step estimation error. Adding patient demographics and comorbidity controls to W_i reduces the resulting chi-squared test statistic, with 100 degrees of freedom, from 295 in column 1 to 238 in column 3. Nevertheless, all three RAM specifications reject the null hypothesis of RAM validity with $p < 0.001$. This is similar to the rejection in column 4, which tests replicated AMI, heart failure, and pneumonia RAMs from the 2013 CMS risk-adjustment methodology (see the data appendix for details of this replication).

Panel B of Table 3.A2 reports chi-squared statistics and associated p-values for tests of forecast bias, overidentification, and the full set of parametric restrictions given by equation (3.37), for the same set of 100 randomly-chosen ambulance companies. Adding demographic and comorbidity controls to the RAM brings the forecast coefficient from 1.30 to 1.09, and the latter is not statistically distinguishable from one at conventional levels. Nevertheless, p-values for tests of the 2SLS model's overidentifying restrictions (with 99 degrees of freedom) are all less than 0.001. As with the non-parametric test in panel A, joint test statistics for all forecast restrictions (again with 100 degrees of freedom) are all around 200 and produce correspondingly small p-values. Although

the forecast coefficient is not statistically distinguishable from one in the CMS-RAM subsample of AMI, pneumonia, and heart attack patients, the model's overidentifying restrictions continue to drive rejections of RAM validity.

Estimating the Hierarchical Linear Model

Given minimum distance quality index estimates $\hat{\beta}_j$ for hospital j in HSA $h(j)$ and vectors H_j of conventional RAM predictions and a constant, an OLS procedure applied to equation (3.22) consistently estimates the HLM hyperparameters $\Gamma = (\kappa, \lambda)'$, σ, and ϕ:

$$\hat{\Gamma}_0 = (H'H)^{-1}H'\hat{\beta} \tag{3.40}$$

$$\hat{\sigma}_0^2 = \frac{\sum_h \overline{w}_{h0}\left(\left(\overline{\hat{\beta}}_h - \overline{H}_h'\hat{\Gamma}_0\right)^2 - \overline{\Xi}_h\right)}{\sum_g \overline{w}_{g0}} \tag{3.41}$$

$$\hat{\phi}_0^2 = \frac{\sum_j w_{j0}\left(\left(\left(\hat{\beta}_j - \overline{\hat{\beta}}_{h(j)}\right) - \left(H_j - \overline{H}_{h(j)}\right)'\hat{\Gamma}_0\right)^2 - \tilde{\Xi}_j\right)}{\sum_k w_{k0}}, \tag{3.42}$$

where $\hat{\beta}$ and H collect observations of $\hat{\beta}_j$ and H_j, $\overline{\hat{\beta}}_h$ and $\overline{H}_h$ denote HSA-level averages of $\hat{\beta}_j$ and H_j, $\overline{\Xi}_h$ denotes the variance of HSA-average estimation error, $\tilde{\Xi}_j$ is the variance of hospital estimation error net of this HSA average, and $\overline{w}_{h0}$ and w_{j0} are known weights. The step-s feasible generalized least squares estimates are

$$\hat{\Gamma}_s = (H'V_s^{-1}H)^{-1}H'V_s^{-1}\hat{\beta} \tag{3.43}$$

$$\hat{\sigma}_s^2 = \frac{\sum_h \overline{w}_{hs}\left(\left(\overline{\hat{\beta}}_h - \overline{H}_h'\hat{\Gamma}_s\right)^2 - \overline{\Xi}_h\right)}{\sum_g \overline{w}_{gs}} \tag{3.44}$$

$$\hat{\phi}_s^2 = \frac{\sum_j w_{js}\left(\left(\left(\hat{\beta}_j - \overline{\hat{\beta}}_{h(j)}\right) - \left(H_j - \overline{H}_{h(j)}\right)'\hat{\Gamma}_s\right)^2 - \tilde{\Xi}_j\right)}{\sum_k w_{ks}}, \tag{3.45}$$

where V_s is a block-diagonal matrix with HSA blocks $V_{hs} = \hat{\phi}_{s-1}^2 I_{J(h)} + \hat{\sigma}_{s-1}^2 + \Xi_h$, and this procedure may be iterated to convergence. Many sequences of weights $\overline{w}_{hs}$ and w_{js} will yield consistent estimates of the variance hyperparameters σ^2 and ϕ^2. In practice I follow Efron and

Morris (1973) in setting

$$\overline{w}_{hs} = 1/(\hat{\sigma}^2_{s-1} + \overline{\Xi}_h)^2 \tag{3.46}$$

$$w_{js} = 1/(\hat{\sigma}^2_{s-1} + \hat{\phi}^2_{s-1} + \Xi_j)^2, \tag{3.47}$$

where Ξ_j is the estimated asymptotic variance of hospital-level estimation error.

In Appendix Table 3.A5 I also use these weights to estimate covariances between two sets of minimum distance quality estimates $\hat{\beta}^A_{jh}$ and $\hat{\beta}^B_{jh}$. Since these are all generated from different datasets A and B, under independent random sampling $Cov(\hat{\beta}^A_{jh}, \hat{\beta}^B_{jh}) = Cov(\beta^A_{jh}, \beta^B_{jh})$, and we also obtain estimates of true quality covariance. Quality correlations are then given by dividing by iterated FGLS estimates of $\sqrt{Var(\beta^A_{jh})Var(\beta^B_{jh})}$.

3.9 Supplementary Results

Correlation Across Time and Conditions

Appendix Table 3.A5 reports additional correlations, computed as described in the econometric appendix, of hospital quality measures over time and across admitting diagnosis categories. For example column 1 of Panel A shows that the most naïve quality yardstick, a hospital's observed 30-day survival rate, appears to follow a white noise process over three-year windows, with signal-to-noise ratio of around 0.5. In contrast, the time-series correlation of observational RAM predictions (panel B) and hospital quality indices (panel C) tends to decline, suggesting selection bias underlies some of the permanent component in observed survival rates. The time-series correlation for quality is also somewhat lower than RAM, with both reduced relative to the raw survival rates in panel A.

Reducing selection bias also appears to increase many of the correlations of observational quality measures across different diagnoses categories, the full matrix of which is shown in columns 2-6 of Table 3.A5. These correlations are computed over a broader analysis sample of Medicare patients admitted in 2001-2012 to increase precision in the narrower patient groups. The correlation reduction from panels A to panel C is especially apparent for circulatory, respiratory, and other survival outcomes, though correlations within digestive and injury diagnoses are smaller in panel C. Encouragingly, all correlations in panel C are at or above 0.2, suggesting quality measures based on one subpopulation of patients may serve as rough proxies for overall hospital effectiveness.

Correlation with Measured Inputs

Appendix Table 3.A6 presents a more fine-grained analysis of causal hospital quality variation by correlating posteriors with observed inputs: the log average staff salary, the use of electronic records and case management systems, and the numbers of earned accreditations and available imaging technologies. These measures come from 2010-2012 Annual Surveys of the American Hospital Association and are only available for two-thirds of my national hospital sample. Usefully, column 1 of Table 3.A6 shows that quality posteriors are unrelated to this availability. To make efficient use of the limited coverage, I regress posteriors of within-HSA hospital quality indices, $\kappa + \lambda\hat{\alpha}_j + \upsilon_j$, rather than including HSA fixed effects as in Table 3.4, though point estimates are qualitatively similar across the two specifications.[27] For interpretability, all regressors and the quality measure in Table 3.A6 are normalized to have a standard deviation of one.

All five input measures positively correlate with quality index posteriors, with the salary and accreditation proxies remaining statistically significant in a multivariate "horse race" regression that includes all measures (column 7). The salary-quality association appears particularly robust, holding even when the strong quality predictor of hospital volume is also included in column 8. Together these results suggest these kinds of input measures may also serve as hospital quality proxies.

Hospital Mergers

The correlations in Tables 3.4 and 3.A5 need not be causal in the sense of, say, predicting the effect on hospital quality from an exogenous volume or salary increase, as omitted factors such as staff experience or doctor expertise may be correlated with both quality posteriors and observed characteristics. To explore one possible determinant of quality, I exploit the plausibly-exogenous timing of hospital mergers and estimate the average causal effect of hospital acquisition on quality posteriors. For this I obtain a list of 39 hospitals that acquired another provider between 2001 and 2012 from the American Hospital Association Summary of Changes database and recompute quality estimates, the HLM, and quality posteriors in rolling three-year lagged windows for each year in 2001-2012. I then match each acquiring hospital to a comparison group by ownership structure, teaching status, and terciles of total patient volume, average spending, and bed capacity in the

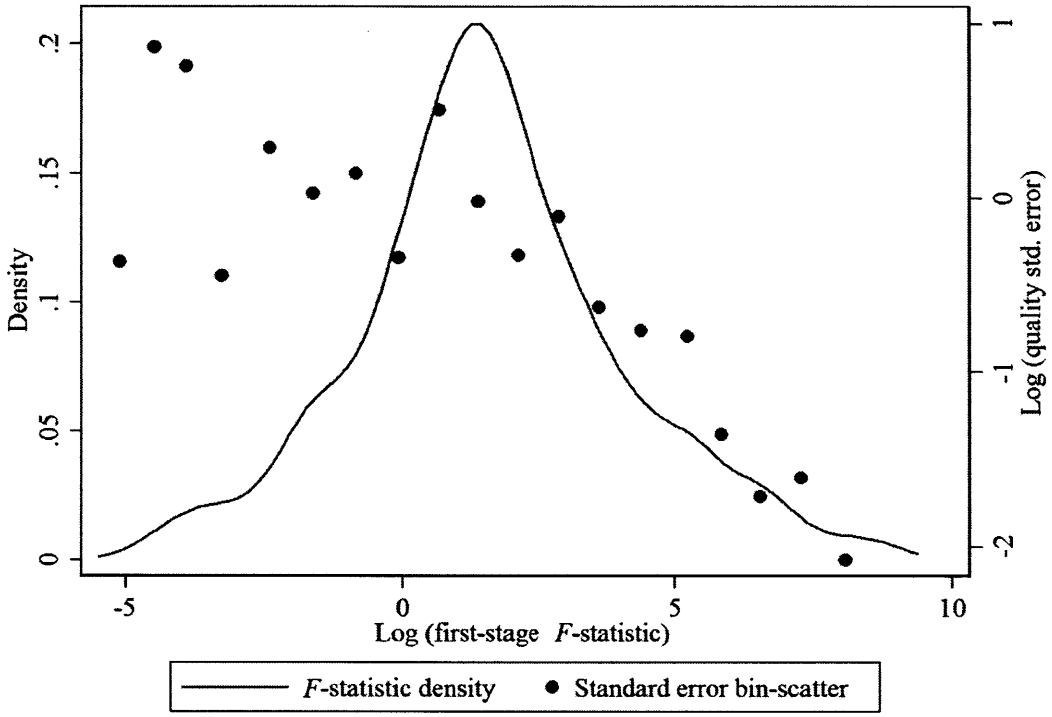
[27] For example, the coefficient on log average staff salary in column 8 of Table 3.A6 becomes 0.03 when HSA fixed effects are included, with a standard error of 0.11.

year preceding the merger. The solid line in Appendix Figure 3.A3, panel A, shows the trajectory of the average quality index posterior for acquiring hospitals, while the dashed line plots the same for the set of matched comparison hospitals. That these follow roughly parallel trends up to the merger year (year zero) suggests differences in quality growth post-merger may be interpreted as causal effects in a standard difference-in-differences (DD) framework.

DD estimates of acquiring hospital merger effects are plotted in panel B of Figure 3.A3 along with associated 95% confidence intervals. Pre-trend differences across the treatment and control groups are statistically indistinguishable from zero, while quality indices rise by an average of 0.16 in the first post-merger year and persist through year three. These pre-trend and effect estimates are quite robust and similar to those obtained using quality posteriors as the outcome, as shown in Appendix Table 3.A7, with an average quality gain of around 4 percentage points. I find no significant merger effect on either observed hospital survival or observational RAM predictions, however, consistent with earlier investigations by Ho and Hamilton (2000) and Capps (2005). While a detailed analysis of hospital merger effects is outside the scope of this chapter (see Gaynor and Town (2012) for a comprehensive overview of the existing literature), these results are consistent with a theory in which economies of scale in emergency healthcare production increase merging hospital quality despite offsetting anti-competitive forces. At the same time, changes in selection bias due to the new patient population post-merge appear to obscure these effects when gauged by observational measures.

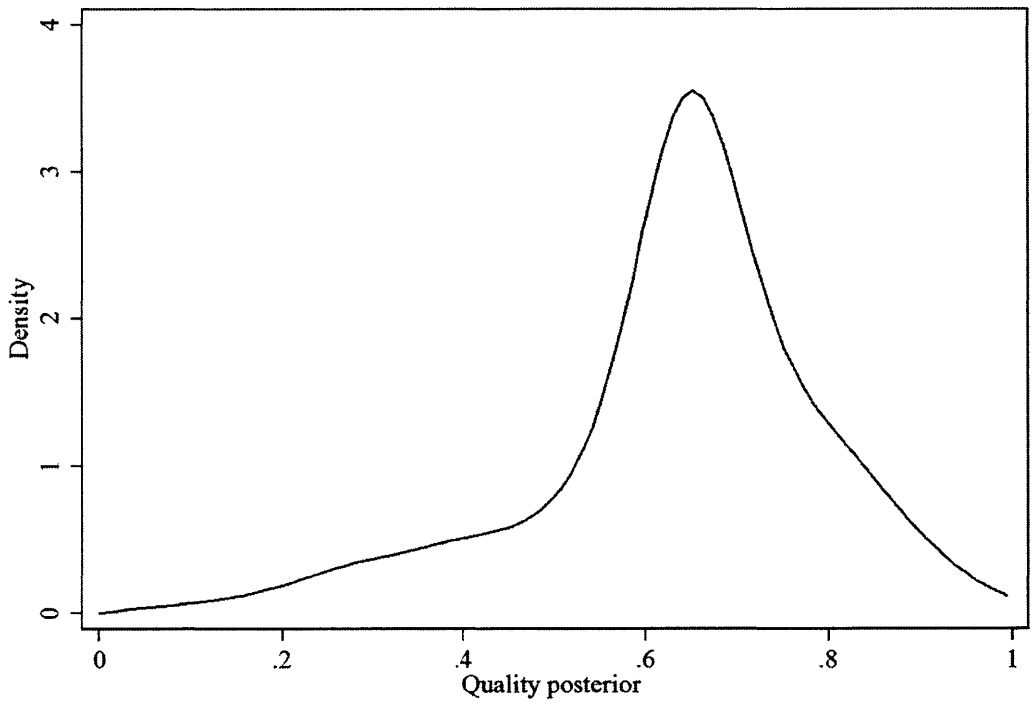
3.10 Appendix Figures and Tables

Figure 3.A1: The Distributions of First-stage F-statistics and Quality Estimate Standard Errors



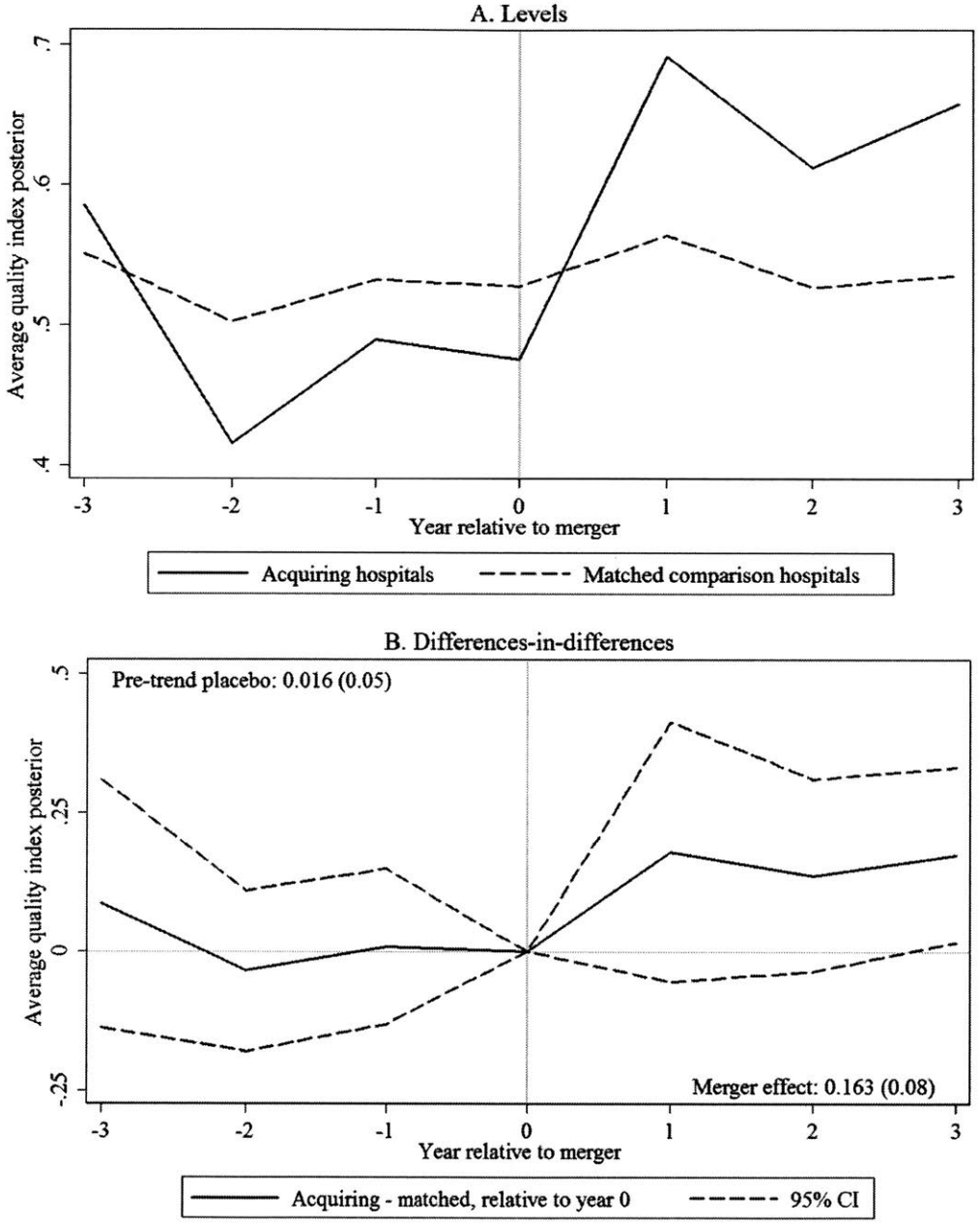
Notes: The solid line in this figure plots a Gaussian kernel density estimate of the distribution of the log of first-stage F-statistics for the 1,041 minimum distance quality estimates. F-statistics test the equality of estimated choice probabilities across all ambulance company instruments. The bandwidth used to estimate this distribution is 0.5. Points plot the median log first-stage F-statistic against the median log quality estimate standard error, in 20 equal-sized bins.

Figure 3.A2: The Distribution of Hospital Quality Posteriors in Multi-hospital HSAs



Notes: This figure plots Gaussian kernel density estimates of the distribution of empirical Bayes hospital quality posteriors. The sample includes 2,357 hospitals operating in 695 multi-hospital HSAs. The bandwidth used to estimate the distribution is 0.05.

Figure 3.A3: Estimated Effects of Hospital Mergers on Quality Indices



Notes: The solid line in panel A plots average quality index posteriors for a set of 39 hospitals acquiring another hospital in a merger between 2001-2012, relative to the merger year. The dashed line shows average quality indices for a set of matched comparison hospitals that were not merged in these years. Acquiring hospitals are matched by their ownership structure (non-profit, for-profit, or government owned), teaching status, and terciles of total patient volume, average spending, and bed capacity in the year preceding the merger. Quality posteriors for each calendar year are estimated in three-year windows using the preceding two years of data and the hierarchical model described in the text. Panel B plots difference-in-differences estimates of merger effects on quality posteriors. The dashed red lines indicate 95% confidence intervals, clustered by HSA. Average effects pre- and post-merger periods are displayed along with cluster-robust standard errors

Table 3.A1: Patient Characteristics

	All nondeferrable Medicare admissions	Analysis sample		
		All hospitals	Multi-hospital HSAs	With min. dist. quality ests.
	(1)	(2)	(3)	(4)
30-day survival	0.875	0.833	0.834	0.834
Log (spending)	8.821	9.284	9.323	9.304
Age	80.22	81.76	81.58	81.61
Male	0.410	0.379	0.382	0.383
White	0.873	0.875	0.857	0.891
Black	0.082	0.082	0.093	0.073
Circulatory diagnosis	0.233	0.220	0.234	0.238
Respiratory diagnosis	0.208	0.200	0.188	0.187
Digestive diagnosis	0.101	0.065	0.066	0.065
Injury diagnosis	0.118	0.177	0.177	0.181
Patients	998,489	405,172	254,101	179,589

Notes: Column 1 of this table reports average characteristics of patients, from a 20% random sample of Medicare inpatient claims, who were admitted to a hospital in 2010-2012 for one of the 29 nondeferrable conditions listed in the notes to Table 1. Column 2 summarizes patient characteristics in the ambulance company analysis sample, while columns 3 and 4 subset this sample to patients admitted in a multi-hospital HSA and to a hospital with enough quasi-experimental data to construct minimum distance quality estimates, respectively.

Table 3.A2: Hospital RAM Bias Tests

	Benchmark RAM			CMS-RAM
	(1)	(2)	(3)	(4)
		A. Propensity score test		
Test statistic (100 d.f.)	295.37	287.78	237.52	186.42
	[<0.001]	[<0.001]	[<0.001]	[<0.001]
		B. Forecast tests		
Forecast coefficient	1.301	1.187	1.086	1.294
	(0.123)	(0.106)	(0.095)	(0.262)
Test statistics (d.f.):				
Forecast bias (1)	6.04	3.12	0.82	1.26
	[0.014]	[0.077]	[0.365]	[0.262]
Over-identification (99)	189.98	184.71	183.67	149.94
	[<0.001]	[<0.001]	[<0.001]	[<0.001]
All restrictions (100)	201.56	192.80	189.43	171.02
	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Risk-adjusters:				
Diagnosis/year FEs	Y	Y	Y	Y
Patient age/sex		Y	Y	Y
Comorbidities			Y	Y
Patients:		405,173		82,815

Notes: This table summarizes tests for bias in hospital risk-adjustment models with ambulance company instruments. All RAMs are hierarchical logit models of 30-day survival, estimated separately for each diagnosis category in Table 3.1. Columns 1-3 estimate RAMs in the full analysis sample, while the model in column 4 uses a nationally-representative sample of AMI, heart failure, and pneumonia Medicare patients admitted in 2010-2012. The first column includes year and diagnosis fixed effects as RAM controls, while the second adds patient age and sex, and the third includes all comorbidity indicators listed in the notes to Table 3.2. The specification in column 4 replicates the 2013 CMS 30-day risk-standardized mortality models. Tests use 100 randomly-selected ambulance companies referring at least 100 patients. Panel A reports test statistics for the joint significance of each company in the propensity score weighting scheme outlined in the appendix. Panel B reports forecast coefficients from 2SLS regressions of realized survival on RAM-predicted survival, instrumented by ambulance company indicators. The forecast bias test statistic is for the null hypothesis that the forecast coefficient equals 1. The full test combines forecast bias and overidentifying restrictions and is implemented by regressing RAM residuals on ambulance indicators and testing their joint significance. Propensity scores for panel A are estimated by company-specific probit models. Test statistics are robust to heteroskedasticity and account for first-step propensity score estimation error. Robust standard errors are reported in parentheses; test p-values are reported in brackets.

Table 3.A3: Robustness of Main Results to Alternative Specifications

	Within-HSA rank corr. w/ preferred spec.	% positively-selected regions	Potential survival correlates			FY2014 VBP change (%)		Ranking survival gains (%)	
			Government hospital	Log (avg. spending)	Log (volume)	Non-profit hospital	Teaching hospital	Max. RAM prediction	Max. quality post.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Preferred specification	1.000	86.0	-0.079	0.230	0.517	8.70	14.92	3.39	3.97
			(0.029)	(0.111)	(0.223)	(3.40)	(6.78)		
Regions are HRRs (not HSAs)	0.926	99.3	-0.067	0.216	0.572	6.99	9.03	3.75	4.64
			(0.029)	(0.120)	(0.222)	(4.13)	(5.04)		
Z_{il} is in Z_{ij} if j is company l's most-referred (not closest) hospital	0.939	87.1	-0.079	0.254	0.571	8.43	9.20	3.07	3.55
			(0.034)	(0.119)	(0.232)	(5.45)	(4.95)		
No risk-adjusters in propensity scores $p_l(X_i)$	0.942	85.5	-0.082	0.298	0.699	11.05	18.64	3.44	4.13
			(0.038)	(0.153)	(0.295)	(2.80)	(5.10)		
Health/utility distributed by Student's *t(2)* (not normally)	0.959	79.6	-0.101	0.280	0.681	6.75	16.00	3.36	3.79
			(0.039)	(0.144)	(0.290)	(2.47)	(4.75)		
HLM includes J and J×RAM prediction interaction	0.681	85.6	-0.116	0.434	1.020	15.84	21.93	2.72	3.27
			(0.064)	(0.220)	(0.422)	(6.94)	(9.13)		

Notes: The first row in this table reports estimates from the hierarchical model described in the text. The second row modifies this model by using hospital referral regions rather than hospital service areas to define local emergency care markets. The third row uses referral rates rather than ambulance-hospital distance to partition ambulance company instruments, while the fourth excludes patient age, sex, and RAM comorbidities from ambulance company propensity scores. The fifth row assumes patient health and utility indices are distributed by a multivariate Student's t distribution with two degrees of freedom. The sixth row adds the total number of hospitals in a hospital's HSA and its interaction with RAM predictions to the benchmark hierarchical linear model. Column 1 reports the within-HSA rank correlation of quality posterior from each model with the preferred specification. Column 2 reports the percentage of hospital regions with positive average selection bias, and columns 3-5 report within-HSA regressions, as in Table 3.4. Columns 6 and 7 report regressions of the change in VBP reimbursement rates, as in Table 3.5, while columns 8 and 9 report simulated gains from rank-based admissions policies, as in Figure 3.6. Standard errors, clustered by region, are reported in parentheses.

Table 3.A4: Hospital Characteristics

	Multi-hospital HSAs		Single-hospital HSAs
	With min. dist. quality estimates	Without quality estimates	
	(1)	(2)	(3)
Observed survival	0.838	0.837	0.838
RAM prediction	0.011	0.008	0.006
Risk-adjustment index	1.977	1.972	1.992
Non-profit	0.763	0.696	0.681
For-profit	0.119	0.207	0.154
Government-owned	0.118	0.096	0.166
Teaching	0.584	0.578	0.573
Log (avg. spending)	9.600	9.693	9.496
Log (volume)	5.453	4.979	4.801
Log (# of beds)	5.916	5.802	5.159
HSA hospital count	3.654	11.174	1.000
Hospitals	1,041	1,316	2,464

Notes: This table reports average characteristics of hospitals in the analysis sample; column 1 summarizes the sample of hospitals with enough quasi-experimental data to construct minimum distance quality estimate, while columns 2-3 characterize the remaining sample. Observed survival posteriors shrink observed rates towards the grand mean in proportion to one minus the signal-to-noise ratio. The risk-adjustment index comes from the estimated fixed portion of the observational RAM.

Table 3.A5: Correlation Structure of 30-day Survival, RAM Predictions, and Quality Indices

	Over time	Across diagnosis categories					
		Circulatory	Respiratory	Digestive	Injury	All other	
	(1)	(2)	(3)	(4)	(5)	(6)	
			A. Observed survival rates				
2010-2012	1.000	1.000					Circulatory
2007-2009	0.529	0.179	1.000				Respiratory
2004-2006	0.517	0.394	0.204	1.000			Digestive
2001-2003	0.491	0.402	0.167	0.424	1.000		Injury
		0.259	0.168	0.334	0.471	1.000	All other
			B. RAM predictions				
2010-2012	1.000	1.000					Circulatory
2007-2009	0.323	0.298	1.000				Respiratory
2004-2006	0.286	0.209	0.193	1.000			Digestive
2001-2003	0.238	0.211	0.182	0.157	1.000		Injury
		0.255	0.316	0.178	0.154	1.000	All other
			C. Hospital quality indices				
2010-2012	1.000	1.000					Circulatory
2007-2009	0.292	0.385	1.000				Respiratory
2004-2006	0.204	0.413	0.279	1.000			Digestive
2001-2003	0.196	0.431	0.285	0.351	1.000		Injury
		0.319	0.430	0.191	0.222	1.000	All other

Notes: This table reports estimated correlation coefficients for a hospital's 30-day survival rate, RAM prediction, and quality index. Column 1 correlates data from the 2010-2012 analysis sample with corresponding data from 2007-2009, 2004-2006, and 2001-2003, while columns 2-6 report correlations across the five patient diagnosis categories over the entire 2001-2012 period. The sample in column 1 includes 2,357 hospitals operating in 695 multi-hospital HSAs; columns 2-6 report correlations for 5,805 hospitals in 2,021 multi-hospital HSAs. See the econometric appendix for a description of the estimation of correlations in panel C.

Table 3.A6: Standardized Regressions of Residual Hospital Quality Index Posteriors on Measured Inputs

Regressors:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Has inputs data	0.016							
	(0.023)							
Log (average staff salary)		0.075					0.060	0.053
		(0.017)					(0.018)	(0.018)
Uses electronic records			0.050				0.043	0.040
			(0.030)				(0.032)	(0.031)
Uses case management				0.071			-0.016	-0.047
				(0.031)			(0.035)	(0.035)
# of accreditations					0.072		0.056	0.030
					(0.018)		(0.022)	(0.025)
# of imaging technologies						0.049	-0.010	-0.044
						(0.016)	(0.022)	(0.029)
Log (volume)								0.072
								(0.032)
Hospitals:	4,821				3,199			

Notes: Each column of this table reports coefficients from regressions of hospital quality index posteriors, net of HSA random effects, on hospital characteristics. All variables are normalized to standard deviation units. Column 1 regresses quality on an indicator for the availability of input data from the American Hospital Associations Annual Survey in the full sample, while columns 2-8 report coefficients from regressions in the sample of hospitals with measured inputs. Inputs are measured in the first year in which data are available in 2010-2012. Average staff salary is computed by dividing total facility payroll by full-time equivalent total personnel. Accreditations include those by The Joint Commission, recognition for one or more Accreditation Council for Graduate Medical Education accredited programs, medical school affiliation with the American Medical Association, affiliation with the National League for Nursing, accreditation by the Commission on Accreditation of Rehabilitation Facilities, membership in the Council of Teaching Hospitals of the Association of American Medical Colleges, Blue Cross contracting or participating, Medicare certification by the U.S. Department of Health and Human Services, accreditation by the Healthcare Facilities Accreditation program of the American Osteopathic Association, approval of an internship by the American Osteopathic Association, approval of a residency by the American Osteopathic Association, and DNV Healthcare accreditation. Imaging technologies include CT scanners, diagnostic radioisotope facilities, EBCT systems, full-field digital mammography, MRI machines, IMRI machines, magnetoencephalography machines, multislice spiral computed tomography scanners, PET scanners, PET/CT scanners, SPECT scanners, and ultrasounds. Standard errors, clustered by HSA, are reported in parentheses.

Table 3.A7: Difference-in-differences Estimates of Hospital Merger Effects

	Pre-trend placebo (1)	Merger effect (2)	Hospitals (3)	Obs. (4)
	A. Alternative quality measures			
Observed survival	-0.003	-0.001		
	(0.002)	(0.002)		
RAM prediction	-0.003	-0.009		
	(0.006)	(0.006)	456	2,225
Quality index posterior	0.016	0.163		
	(0.050)	(0.080)		
Quality posterior	-0.002	0.039		
	(0.016)	(0.023)		
	B. Quality index estimate robustness checks			
Balanced panel	0.098	0.190	196	1,226
	(0.066)	(0.098)		
Volume weighting	0.001	0.192	456	2,225
	(0.085)	(0.113)		
Quintile matching	-0.010	0.176	195	934
	(0.055)	(0.085)		
Nearest-neighbor matching	0.005	0.119	138	867
	(0.065)	(0.063)		

Notes: This table reports difference-in-differences estimates of the effect of mergers on various hospital quality measures. The sample and estimation for panel A is described in the notes to Figure 3.4. Panel B reports alternative estimates of the effect of mergers on hospital quality index posteriors. The first row restricts the sample to hospitals observed for three years preceding and following the merger year. The second row weights the difference-in-differences specification by a hospital's Medicare patient volume, while the third uses quintiles (rather than terciles) of total patient volume, average spending, and bed totals to match hospitals. The final row in panel B matches each merging hospital to a comparison hospital with the same ownership structure and teaching status and the shortest Mahalanobis distance in volume, spending, and bed capacity. Standard errors, clustered by HSA, are reported in parentheses.

Bibliography

ABADIE, A. (2003): "Semiparametric instrumental variables estimation of treatment response models," *Journal of Econometrics*, 113(2), 231–263.

ABDULKADIROĞLU, A., J. D. ANGRIST, S. M. DYNARSKI, T. J. KANE, AND P. A. PATHAK (2011): "Accountability and flexibility in public schools: Evidence from Boston's charters and pilots," *Quarterly Journal of Economics*, 126(2), 699–748.

ABDULKADIROĞLU, A., J. D. ANGRIST, P. D. HULL, AND P. A. PATHAK (2016): "Charters without lotteries: Testing takeovers in New Orleans and Boston," *American Economic Review*, 106(7), 1878–1920.

ABDULKADIROĞLU, A., J. D. ANGRIST, Y. NARITA, AND P. A. PATHAK (forthcoming): "Market design meets research design: Using centralized assignment for impact evaluation," *Econometrica*.

ABDULKADIROĞLU, A., P. A. PATHAK, A. E. ROTH, AND T. SÖNMEZ (2006): "Changing the Boston school choice mechanism," NBER working paper no. 11965.

ALTONJI, J. G. AND L. M. SEGAL (1996): "Small-sample bias in GMM estimation of covariance structures." *Journal of Business and Economic Statistics*, 14(3), 353–366.

ANDREWS, D. W. K. (1991): "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59(2), 307–345.

ANGRIST, J. D. (1991): "Grouped-data estimation and testing in simple labor-supply models," *Journal of Econometrics*, 47(2), 243–266.

——— (1997): "Conditional independence in sample selection models," *Economics Letters*, 54(2), 103–112.

——— (1998): "Estimating the labor market impact of voluntary military service using Social Security data on military applicants," *Econometrica*, 66(2), 249–288.

ANGRIST, J. D., S. R. COHODES, S. M. DYNARSKI, P. A. PATHAK, AND C. R. WALTERS (2016a): "Stand and deliver: Effects of Boston's charter high schools on college preparation, entry and choice," *Journal of Labor Economics*, 34(2), 275–318.

ANGRIST, J. D. AND I. FERNANDEZ-VAL (2013): "ExtrapoLATE-ing: External validity and overidentification in the LATE framework," Cambridge University Press, vol. 3 of *Advances in Economics and Econometrics*.

ANGRIST, J. D. AND J. HAHN (2004): "When to control for covariates? Panel asymptotics for estimates of treatment effects," *Review of Economics and Statistics*, 86(1), 1–15.

ANGRIST, J. D., P. D. HULL, P. A. PATHAK, AND C. R. WALTERS (2016b): "Interpreting tests of school VAM validity," *American Economic Review: Papers & Proceedings*, 106(5), 388–392.

ANGRIST, J. D. AND G. W. IMBENS (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association*, 90(430), 431–442.

ANGRIST, J. D., G. W. IMBENS, AND D. B. RUBIN (1996): "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91(434), 444–455.

ANGRIST, J. D., P. A. PATHAK., AND C. R. WALTERS (2013): "Explaining charter school effectiveness," *American Economic Journal: Applied Economics*, 5(4), 1–27.

ANGRIST, J. D. AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

ANGRIST, J. D. AND M. ROKKANEN (2016): "Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff," *Journal of the American Statistical Association*, 110(512), 1331–1344.

BEHAGHEL, L., B. CRÉPON, AND M. GURGAND (2013): "Robustness of the encouragement design in a two-treatment randomized control trial," IZA discussion paper no. 7447.

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile prices in market equilibrium," *Econometrica*, 63(4), 841–890.

BLACKWELL, M. (forthcoming): "Instrumental variable methods for conditional effects and causal interaction in voter mobilization experiments," *Journal of the American Statistical Association.*

BLOOM, H. S. AND R. UNTERMAN (2014): "Can small high schools of choice improve educational prospects for disadvantaged students?" *Journal of Policy Analysis and Management*, 33(2), 290–319.

BRINCH, C. N., M. MOGSTAD, AND M. WISWALL (forthcoming): "Beyond LATE with a discrete instrument," *Journal of Political Economy*, statistics Norway Discussion Paper No. 703.

CAPPS, C. (2005): "The quality effects of hospital mergers," Department of Justice Economic Analysis Group discussion paper no. 05-6.

CARD, D., C. DOBKIN, AND N. MAESTAS (2009): "Does Medicare save lives?" *Quarterly Journal of Economics*, 124(2), 597–636.

CARD, D., J. HEINING, AND P. KLINE (2013): "Workplace heterogeneity and the rise of West German wage inequality," *Quarterly Journal of Economics*, 128(3), 967–1015.

CATTANEO, M., M. JANSSON, AND X. MA (2016): "Marginal treatment effects with many instruments," Working Paper.

CHAMBERLAIN, G. AND G. W. IMBENS (2004): "Random effects estimators with many instrumental variables," *Econometrica*, 72, 295–306.

CHANDRA, A., A. FINKELSTEIN, A. SACARNY, AND C. SYVERSON (2015): "Healthcare exceptionalism? Performance and allocation in the U.S. healthcare sector," *American Economic Review*, 106(8), 2110–2144.

CHANDRA, A. AND D. STAIGER (2007): "Productivity spillovers in health care: Evidence from the treatment of heart attacks," *Journal of Political Economy*, 115(1), 103–140.

CHETTY, R., J. N. FRIEDMAN, N. HILGER, E. SAEZ, D. W. SCHANZENBACH, AND D. YAGAN (2011): "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126(4), 1593–1660.

CHETTY, R., J. N. FRIEDMAN, AND J. ROCKOFF (2014a): "Measuring the impact of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 104(9), 2593–2632.

——— (2014b): "Measuring the impact of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review*, 104(9), 2633–2679.

——— (2016): "Using lagged outcomes to evaluate bias in value-added models," *American Economic Review: Papers & Proceedings*, 106(5), 393–399.

CHETTY, R. AND N. HENDREN (2016): "The impacts of neighborhoods on intergenerational mobility II: County-level estimates," NBER working paper no. 23002.

CHETTY, R., J. NFRIEDMAN, AND J. ROCKOFF (forthcoming): "Measuring the Impacts of Teachers: Reply to Rothstein (2016)," *American Economic Review*.

COHODES, S. R. (2015): "The long-run impacts of tracking high-achieving students: Evidence from Boston's Advanced Work Class," Working Paper.

CONWAY, P. (2013): "CMS releases latest value-based purchasing program scorecard," Available at `https://blog.cms.gov/2013/11/14/cms-releases-latest-value-based-purchasing-program-scorecard/`. Last accessed October 30, 2016.

CULLEN, J. B., B. A. JACOB, AND S. LEVITT (2006): "The effect of school choice on participants: evidence from randomized lotteries," *Econometrica*, 74(5), 1191–1230.

DE CHAISEMARTIN, C. (forthcoming): "Tolerating defiance? Identification of treatment effects without monotonicity," *Quantitative Economics*.

DEMING, D. (2014): "Using school choice lotteries to test measures of school effectiveness," *American Economic Review: Papers & Proceedings*, 104(5), 406–411.

DEUTSCH, J. (2013): "Proposing a test of the value-added model using school lotteries," Working paper.

D'HAULTFOEUILLE, X. AND A. MAUREL (2013): "Another look at the identification at infinity of sample selection models," *Econometric Theory*, 29(1), 213–224.

DHHS (2015): "Better, smarter, healthier: In historic announcement, HHS sets clear goals and timeline for shifting Medicare reimbursements from volume to value," Available at http://bit.ly/1QhLv5b. Last accessed October 26, 2016.

DHHS/CMS (2015): "Hospital value-based purchasing," Available at https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf. Last accessed March 20, 2016.

DiNardo, J., J. McCrary, and L. Sanbonmatsu (2006): "Constructive proposals for dealing with attrition: An empirical example," Working Paper.

Dobbie, W. and R. G. Fryer (2015): "The medium-term impacts of high-achieving charter schools," *Journal of Political Economy*, 123(5), 985–1037.

Doyle, J., J. Graves, J. Gruber, and S. Kleiner (2015): "Measuring returns to hospital care: Evidence from ambulance referral patterns," *Journal of Political Economy*, 123(1), 170–214.

Dranove, D. and A. Sfekas (2008): "Start spreading the news: A structural estimate of the effects of New York hospital report cards," *Journal of Health Economics*, 27(5), 1201–1207.

Duflo, E., R. Glennerster, and M. Kremer (2008): "Using randomization in development economics research: A toolkit," Elsevier, Handbook of Development Economics 4, 3895–3962.

Dunnett, C. (1989): "Algorithm AS 251: Multivariate normal probability integrals with product correlation structure," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 38(3), 564–579.

Efron, B. and C. Morris (1973): "Stein's estimation rule and its competitors: an empirical Bayes approach," *Journal of the American Statistical Association*, 68(341), 117 – 130.

Ferguson, T. (1958): "A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities," *Annals of Mathematical Statistics*, 29(4), 1046–1062.

Fletcher, J. M., L. I. Horwitz, and E. Bradley (2014): "Estimating the value added of attending physicians on patient outcomes," NBER working paper no. 20534.

Foster, D., L. Zrull, and J. Chenoweth (2013): "Hospital performance differences by ownership," Truven Health Analytics. Available at http://100tophospitals.com/portals/2/assets/HOSP_12678_0513_100TopHopPerfOwnershipPaper_RB_WEB.pdf. Last accessed May 31, 2016.

Frangakis, C. E. and D. B. Rubin (2002): "Principal stratification in causal inference," *Biometrics*, 58(1), 21–29.

Frölich, M. and M. Huber (2014): "Treatment evaluation with multiple outcome periods under endogeneity and attrition," *Journal of American Statistical Association*, 109(508), 1697–1711.

Fryer, R. G. (2014): "Injecting charter school best practices into traditional public schools: Evidence from field experiments," *Quarterly Journal of Economics*, 129(3), 1355–1407.

Gaynor, M. and R. Town (2012): "Competition in health care markets," in *Handbook of Health Economics*, ed. by M. McGuire, V. Pauly, and P. Barrow, Elsevier, vol. 2, chap. 9, 1 ed.

GEMAN, S. AND C.-R. HWANG (1982): "Nonparametric maximum likelihood estimation by the method of sieves," *Annals of Statistics*, 10(2), 401–414.

GEWEKE, J., G. GOWRISANKARAN, AND R. TOWN (2003): "Bayesian inference for hospital quality in a selection model," *Econometrica*, 171(4), 1215–1238.

GRONAU, R. (1974): "Wage comparisons – a selectivity bias," *Journal of Political Economy*, 82(6), 1119–1143.

GROSSMAN, M. (1972): "On the concept of health capital and the demand for health," *Journal of Political Economy*, 82(2), 223–255.

GUPTA, A. (2016): "Impacts of performance pay for hospitals: The Readmissions Reduction Program," Working Paper.

HADLEY, J. AND P. CUNNINGHAM (2004): "Availability of safety net providers and access to care of uninsured persons," *Health Services Research*, 39(5), 1527–1546.

HANSEN, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50(4), 1029–1054.

HASTINGS, J. S. 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. (1978): "Specification tests in econometrics," *Econometrica*, 46(6), 1251–1271.

HAUSMAN, J. AND D. WISE (1978): "A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences," *Econometrica*, 46(2), 403–426.

HECKMAN, J. J. (1976): "The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models," NBER, vol. 5 of *Annals of Economic and Social Measurement*, 475–492.

HECKMAN, J. J. AND B. HONORE (1990): "The empirical content of the Roy model," *Econometrica*, 58(5), 1121–1149.

HECKMAN, J. J. AND S. URZÚA (2010): "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics*, 156(1), 27–37.

HECKMAN, J. J., S. URZÚA, AND E. VYTLACIL (2006): "Understanding instrumental variables in models with essential heterogeneity," *The Review of Economics and Statistics*, 88(3), 389–432.

——— (2008): "Instrumental variables in models with multiple outcomes: The general unordered case," *Annals of Economics and Statistics*, 91(3), 151–174.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient estimation of average treatment effects using the estimated propensity score," 71(4), 1161–1189.

HO, V. AND B. HAMILTON (2000): "Hospital mergers and acquisitions: Does market consolidation harm patients?" *Journal of Health Economics*, 19(5), 767–791.

HOXBY, C. (2015): "Computing the value-added of American postsecondary institutions," Working paper.

HUDSON, S., P. D. HULL, AND J. LIEBERSOHN (2015): "Interpreting instrumented difference-in-differences," Working paper.

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and estimation of local average treatment effects," 62(2), 467–475.

IMBENS, G. W. AND D. B. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

JACOB, B. A. AND L. LEFGREN (2008): "Principals as agents: Subjective performance assesment in education," *Journal of Labor Economics*, 26(1), 101–136.

JUDGE, G. G. AND R. C. MITTLEHAMMER (2004): "A semiparametric basis for combining estimation problems under quadratic loss," *Journal of the American Statistical Association*, 99(466), 479–487.

——— (2007): "Estimation and inference in the case of competing sets of estimating equations," *Journal of Econometrics*, 138(2), 513–531.

KANE, T., J. ROCKOFF, AND D. STAIGER (2008): "What does certification tell us about teacher effectiveness? Evidence from New York City," *Economics of Education Review*, 27, 615–631.

KANE, T. AND D. STAIGER (2008): "Estimating teacher impacts on student achievement: An experimental evaluation," NBER working paper no. 14607.

KANE, T. J., D. F. MCCAFFREY, AND D. O. STAIGER (2013): "Have we identified effective teachers? Validating measures of effective teaching using random assignment," *Gates Foundation Report.*

KANE, T. J. AND D. O. STAIGER (2002): "The promise and pitfalls of using imprecise school accountability measures," *Journal of Economic Perspectives*, 16(4), 91–114.

KEANE, M. (1992): "A note on identification in the multinomial probit model," *Journal of Business and Economic Statistics*, 10(2), 193–200.

KINSLER, J. (2012): "Assessing Rothstein's critique of teacher value-added models," *Quantitative Economics*, 3(2), 333–362.

KIRKEBØEN, L., E. LEUVEN, AND M. MOGSTAD (2016): "Field of study, earnings, and self-selection," *Quarterly Journal of Economics*, 131(3), 1057–1111.

KLINE, P. AND C. R. WALTERS (2016): "Evaluating public programs with close substitutes: The case of Head Start," *Quarterly Journal of Economics*, 131(4), 1795–1848.

KLING, J. R., J. B. LIEBMAN, AND L. F. KATZ (2007): "Experimental analysis of neighborhood effects," *Econometrica*, 75(1), 83–119.

KOEDEL, C. AND J. R. BETTS (2011): "Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique," *Education Finance and Policy*, 6(1), 18–42.

KRUMHOLZ, H., Y. WANG, J. MATTERA, Y. WANG, L. F. HAN, M. INGBER, S. ROMAN, AND S.-L. NORMAND (2006): "An administrative claims model suitable for profiling hospital performance based on 30-Day mortality rates among patients with an acute myocardial infarction," *Circulation*, 113, 1693–1701.

LEE, D. S. (2009): "Training, wages, and sample selection: Estimating sharp bounds on treatment effects," *Review of Economic Studies*, 76(3), 1071–1102.

LEWBEL, A. (2007): "Endogenous selection or treatment model estimation," *Journal of Econometrics*, 141(2), 777–806.

LITTLE, R. J. A. AND D. B. RUBIN (1987): *Statistical Analysis with Missing Data*, Wiley.

MCCLELLAN, M. AND D. STAIGER (1999): "The quality of health care providers," NBER working paper no. 7327.

——— (2000): "Comparing hospital quality at for-profit and not-for-profit hospitals," in *The Changing Hospital Industry*, ed. by D. Cutler, University of Chicago Press.

MCCOLLOCH, R. AND P. ROSSI (1994): "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64(1), 207 – 240.

MCFADDEN, D. (1989): "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 57(5), 995 – 1026.

MITTLEHAMMER, R. C. AND G. G. JUDGE (2005): "Combing estimators to improve structural model estimation and inference under quadratic loss," *Journal of Econometrics*, 128(1), 1–29.

MORRIS, C. (1983): "Parametric empirical Bayes inference: Theory and applications," *Journal of the American Statistical Association*, 78(381), 47 – 55.

NEWEY, W. K. (1985): "Generalized method of moments specification testing," *Journal of Econometrics*, 29(3), 229–256.

NEWEY, W. K. AND K. D. WEST (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28(3), 777–787.

NORTON, E., J. LI, A. DAS, AND L. CHEN (2016): "Moneyball in Medicare," NBER working paper no. 22371.

PAKES, A. AND D. POLLARD (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57(5), 1027 – 1057.

PEAR, R. (2014): "Health law's pay policy is skewed, panel finds," The New York Times. Available at `http://nyti.ms/1rvCy8O`. Last accessed September 1, 2015.

PINTO, R. (2015): "Selection bias in a controlled experiment: The case of moving to opportunity," Working paper.

RAUDENBUSH, S. AND A. BYRK (1986): "A hierarchical model for studying school effects," *Sociology of Education*, 59(1), 1–17.

REARDON, S. F. AND S. W. RAUDENBUSH (2013): "Under what assumptions do site-by-treatments identify average causal effects?" *Sociological Methods Research*, 42(2), 143–163.

RICE, S. (2014): "Experts question hospital raters' methods," Modern Healthcare. Available at `http://www.modernhealthcare.com/article/20140531/MAGAZINE/305319980`. Last accessed June 1, 2016.

ROBINS, J. AND Y. RITOV (1997): "Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models," *Statistics in Medicine*, 16(1), 285–319.

ROTHSTEIN, J. (2009): "Student sorting and bias in value-added estimation: Selection on observables and unobservables," *Education Finance and Policy*, 4(4), 537–571.

——— (2010): "Teacher quality in educational production: Tracking, decay, and student achievement," *Quarterly Journal of Economics*, 125(1), 175–214.

——— (forthcoming): "Revisiting the impacts of teachers," *American Economic Review*.

ROY, A. (1951): "Some thoughts on the distribution of earnings," *Oxford Economic Papers*, 3(2), 135–146.

RUBIN, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66(5), 688–701.

——— (1981): "The Bayesian bootstrap," *Annals of Statistics*, 9(1), 130–134.

RUHNKE, G., M. COCA-PERRAILLON, B. KITCH, AND D. CUTLER (2011): "Marked reduction in 30-Day mortality among elderly patients with community-acquired pneumonia," *American Journal of Medicine*, 124(2), 171–178.

SANGHAVI, P., A. JENA, J. NEWHOUSE, AND A. ZASLAVSKY (2015): "Outcomes after out-of-hospital cardiac arrest treated by basic vs advanced life support," *Journal of the American Medical Assocation Internal Medicine*, 175(2), 196–204.

SARGAN, J. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica*, 26(3), 393–415.

SILBER, J., P. ROSENBAUM, T. BRACHET, R. ROSS, L. BRESSLER, O. EVAN-SHOSAHN, S. LORCH, AND K. VOLPP (2010): "The hospital compare mortality model and the volume-outcome relationship," *Health Services Research*, 45(5), 1148–1167.

SLOAN, F., G. PICCONE, D. TAYLOR, AND S.-Y. CHOU (2001): "Hospital ownership and cost and quality of care: Is there a dime's worth of difference?" *Journal of Health Economics*, 20(1), 1–21.

WALTERS, C. R. (2014): "The demand for effective charter schools," NBER working paper no. 20640.

WHITE, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48(4), 817–838.

——— (1982): "Instrumental variables regression with independent observations," *Econometrica*, 50(2), 483–499.

WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

YNHHSC/CORE (2013): "2013 measures epdates and specifications: Acute myocardial infarction, heart failure, and pneumonia 30-Day risk-standardized mortality measure (version 7.0)," Available at `http://www.qualitynet.org/dcs/ContentServer?cid=1228774398696&pagename=QnetPublic%2FPage%2FQnetTier4&c=Page`. Last accessed November 3, 2015.