Evaluation Tradecraft

Evaluation Tradecraft presents short articles about the art of evaluation in housing and urban research. Through this department of Cityscape*, the Office of Policy Development and Research presents developments in the art of evaluation that might not be described in detail in published evaluations. Researchers often describe what they did and what their results were, but they might not give readers a step-by-step guide for implementing their methods. This department pulls back the curtain and shows readers exactly how program evaluation is done. If you have an idea for an article of about 3,000 words on a particular evaluation method or an interesting development in the art of evaluation, please send a one-paragraph abstract to marina.l.myhre@hud.gov.*

Use of Genetic Matching in Program Evaluation: The Case of RAD

David Ruiz
Dennis Stout
Christine Herlihy
Econometrica

In this article, we describe the use of genetic matching in program evaluation, define cases in which this approach would be appropriate, and detail the value that this approach can provide. In particular, we focus on how the researchers used genetic matching in the ongoing evaluation of the Rental Assistance Demonstration (RAD) program, the results they obtained, and how they assessed its success. Clinical researchers and social scientists have developed genetic matching as a sampling technique for conducting nonrandomized observational studies in a quasi-experimental fashion. The method matches each member of the treatment group with one or more members of the control group. The match uses a set of key covariates, which the analyst selects based on prior expectations about possible treatment group participation factors. In the RAD evaluation, the research staff used stratified random sampling to select the RAD project sample (treatment group) from the participating RAD population. For the non-RAD sample (control group), researchers used a genetic matching algorithm to select a matched group of non-RAD public housing projects from the nonparticipating public housing population. Postsampling analysis confirmed that, on covariates likely to impact participation in RAD, the control group and the treatment group were similarly distributed. This matching technique can be a useful tool in program evaluation when membership in the treatment or control group is not random; for instance, if participation is voluntary, as is the case in the RAD program.

Overview of RAD

RAD was authorized in 2012[1] as a pilot program for converting public housing projects that are subsidized through public housing programs to assisted housing projects that are subsidized through project-based Section 8 Housing Assistance Payment (HAP) contracts. Participation in the program is voluntary for public housing authorities (PHAs). For a PHA to participate, RAD requires that it submit a project application with supporting documentation and analysis. Over a period of several months, the U.S. Department of Housing and Urban Development (HUD) reviews and approves the RAD application, grants a Comprehensive Housing Assistance Plan (CHAP), and issues a RAD Conditional Commitment (RCC). During this process, a project can be withdrawn by the PHA or have its CHAP revoked by HUD. At the end of this approval process, the PHA and HUD agree to convert the project to a project-based Section 8 HAP contract. After conversion, the former public housing project will receive its program funding from the project-based Section 8 program instead of from public housing programs. The primary intent of the RAD program is to preserve and improve the quality of subsidized housing by enabling PHAs to use their long-term Section 8 HAP contracts to leverage external capital for rehabilitation or new construction and financial stabilization.

Congress requires HUD to assess how the RAD program has been implemented and its impact on the physical and financial condition of converted housing and tenants. The core research questions revolve around whether RAD has produced better-quality housing and put that housing on a firmer financial foundation while continuing to serve low-income tenants. The evaluation began in 2014 and will continue through 2018. An interim report on the evaluation was released in September 2016.

Genetic Matching in Observational Studies

Few program evaluations can replicate the research design used in typical clinical experiments to test the efficacy of drugs and other medical treatments. In such experiments, the treatment group is administered the test drug, while the control group is given a placebo. Such studies are *double blind* in the sense that the assignment of each participant to the treatment or control group is random, and neither the research scientists nor the subjects of the experiment know to which group each subject has been assigned.

Random selection is the preferred approach for clinical research because any variation between the two groups after such assignment is random, rather than systematic. This allows researchers to more accurately attribute any difference in impact to the treatment alone (that is, receiving the drug as opposed to a placebo), rather than to potentially confounding variables.

Studies that use observational data rather than experimental data—as is the case with most program evaluations, including the evaluation of the RAD program—are more likely to produce biased results because assignment into the treatment and control groups has not been randomized. However, observational studies can be conducted in a quasi-experimental fashion by matching each member of the treatment group with one or more controls based on a set of key covariates that are postulated to have some effect on the propensity of a given individual to participate in the treatment. Using this method,

[1] Consolidated and Further Continuing Appropriations Act of 2012, Pub. L. 112–55.

the researcher selects a matched group of controls with a similar distribution of covariates to that of the treatment group. A high-quality match will minimize all observed sources of bias.[2] The quality of the match is measured by calculating the bias for each variable, as follows.

$$\boldsymbol{Bias} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{(\sigma_T^2 + \sigma_C^2)}{2}}} \text{ ,} \tag{1}$$

where $\bar{X}_T$ and $\bar{X}_C$ are the means of a covariate $\boldsymbol{X}$ for the treatment and control groups, respectively, and σ_T^2 and σ_C^2 are their variances.[3] This bias should not be statistically different from 0.

There are many ways to match treatment and control samples. No consensus has emerged in the research literature on which matching method is best, and empirical matching is as much art as science (Stuart, 2010). For the evaluation of the RAD program, research staff opted for a flexible matching method—the genetic matching algorithm GenMatch (GM) written in R (Sekhon and Mebane, 1998).[4] GM is "a multivariate matching method that uses an evolving search algorithm developed to maximize the balance of covariates across matched treated and control units" (Diamond and Sekhon, 2013: 2).[5] "Balance" means that the treatment and control groups have the same joint distribution of the covariates. GM minimizes a loss function[6] that combines two statistical tests: (1) a parametric *t*-test for the difference in means of each covariate and (2) a nonparametric Kolmogorov–Smirnov test (KS test) that minimizes the difference between the empirical cumulative distribution functions of each covariate (Diamond and Sekhon, 2013). It is helpful to think of genetic matching as a generalized matching algorithm that can incorporate both nonparametric (for example, Mahalanobis distance) and parametric (for example, propensity score matching [PSM]) matching algorithms, if appropriate.[7] GM, through its iterations, allows us to identify the most relevant characteristics to match treatment and comparison entities. GM outperforms PSM and other, more simplistic nonparametric methods because it returns an optimal set of matches even if such optimal balance is best achieved by a differential combination of weights across characteristics on which the treatment and comparison groups are matched.[8]

[2] It does not, of course, control for unobserved sources of bias.

[3] The bias is also known as the "standardized difference," because the difference in means is "standardized" by dividing it by the pooled standard deviation.

[4] R is a programming language and development environment for statistical computing and data visualization; see https://www.r-project.org/about.html for more information.

[5] Genetic algorithms (including but not limited to genetic matching) are tools that can be used in machine learning. These algorithms have their roots in and borrow concepts from evolutionary biology (for example, mutation, crossover, and selection).

[6] The loss function that is minimized is the maximum *p*-value from either the Kolmogorov–Smirnov (KS) test or the paired *t*-test over all variables that are matched over. One can write the loss function as—
$L = max\ [pt\text{-}test(var_1),\ pKS\text{-}test(var_1),\ pt\text{-}test(var_2),\ pKS\text{-}test(var_2), \ldots\ pt\text{-}test(var_k),\ pKS\text{-}test(var_k\})]$.
The best match is the one that minimizes this loss function. Rather than relying on the *t*-test alone, the GM algorithm combines the *t*-test with the KS test to get results that are as well-balanced as possible with respect to both tests. Using *p*-values enables one to compare results from both tests on the same (probability) scale.

[7] Mahalanobis matching, for example, minimizes a distance measure that does not rely on an econometric model or distributional assumptions—and the lack of distributional assumptions enables us to find matches among smaller sample sizes.

[8] As a practical example, if propensity score methods were indeed the best method for identifying comparison entities that were statistically similar, the genetic matching algorithm would place a weight of 1 on the propensity score "characteristic" and a weight of 0 on all other characteristics.

The *t*-test for the difference in means is familiar but subject to two crucial limitations. First, the *t*-test is based on a single parameter, the mean, even though distributions with identical means might have widely different underlying distributions. Second, the *t*-test depends on the assumption that both underlying distributions are normal, which may be untrue. The KS test, on the other hand, is nonparametric in that it does not depend on any assumptions about the underlying distributions. It compares the empirical distribution of each variable for a given RAD project with its potential matches in the non-RAD population and calculates the maximum difference between the two cumulative distribution functions. The GM algorithm minimizes the largest discrepancy based on *p*-values from KS tests and *t*-tests for all covariates.[9] For example, if one is trying to achieve balance on *A*, *B*, and *C* characteristics, the algorithm begins with the assumption that *A*, *B*, and *C* all matter equally in achieving balance. It assesses this by minimizing the *p*-values of the *t*-test and the KS test between the treatment and control groups. GM checks multiple weighting schemes—across *A*, *B*, and *C*—to identify which weighting scheme minimizes *p*-values, thereby identifying which variables are most important to minimize statistical bias.

Research Design for Evaluation of the RAD Program

The research design for the evaluation of the RAD program called for analyzing data for a small sample of RAD projects (the treatment group) and a small sample of non-RAD projects (the control group) to identify how RAD might impact the physical and financial condition of converted properties. The analysis focused on changes in the RAD cohort, before and after conversion, compared with changes in the non-RAD cohort over a comparable period of time for a range of variables, such as short- and long-term capital needs, reserves, and cashflow. Small sample sizes were a necessity, given the need to manage the high cost and burden on PHAs of collecting primary data. Data included a project's physical and financial condition, collected by professional engineers on site through a Physical Condition Assessment or similar format, and the PHA's views and experiences of the program, collected by online surveys and telephone interviews. The data collection burden on PHAs was managed by limiting the number of PHAs participating in the study to small sample groups. The research team decided on a minimum of 24 RAD projects in the treatment group and 48 non-RAD projects in the control group. The ratio of 2 non-RAD projects for every RAD project in the sample reflected the expectation that the control group would exhibit greater data variability and that PHAs with non-RAD projects would be less inclined to participate in the study, requiring substitutes.

The sample frame for the treatment group consisted of the universe of 132 RAD properties that had an approved CHAP as of December 31, 2013, and had either closed or reached the RCC stage by December 31, 2014. This sample frame meant that the resulting sample would be representative of

[9] The *p*-value represents the probability of getting a value of the test statistic greater than the one obtained, given that the null hypothesis is actually true. As applied in the RAD study, the null hypothesis would be that no difference exists between RAD and non-RAD developments. A low *p*-value (less than 5 percent) would result in rejection of the null hypothesis (with 95 percent confidence). A high *p*-value, however, would mean that the null hypothesis could not be rejected. Using *p*-values enables results for different test statistics to be compared on the same scale. That is, the GM algorithm can directly compare the *p*-value from a *t*-test with the *p*-value from a KS test because both are measured in terms of probabilities.

RAD projects that had applied earlier and had moved from CHAP to RCC or had closed in a timely manner.[10] To select the RAD sample, researchers used a stratified random selection methodology along two dimensions: (1) PHA size and (2) project performance rating. Each dimension was split into three subcategories, yielding nine (3 × 3 = 9) potential buckets. However, only eight of these buckets contained actual housing projects. Exhibit 1 shows the number and percentage of projects in each bucket for the RAD treatment group sample. The bucket of substandard-performing projects managed by small PHAs is empty, because no projects in the population of 132 RAD projects from which the RAD project sample was drawn were in that bucket.

The sample frame for the control group consisted of 5,993 public housing projects that had not applied to the RAD program based on the HUD's inventory of all public housing projects and RAD program data on applications. To select non-RAD projects to serve as the control group from the non-RAD population, the genetic matching algorithm was used to select without replacement the two non-RAD properties that were the best matches for each participating RAD property. Because the sample included 24 RAD properties, the result was 48 matching, non-RAD comparison projects from the HUD inventory of non-RAD projects.

Exhibit 2 shows the number and percentage of projects in each bucket for the non-RAD control group. The bucket of substandard-performing projects managed by small PHAs is empty again because it is empty for the RAD sample, and the non-RAD control group is intended to match the RAD sample. The distribution of projects across the other buckets for both samples is broadly similar. Although not identical in all buckets, the distribution is the same for the PHA size subcategories. The slight differences reflect the use of other variables during the iterative genetic matching process.

Exhibit 1

Distribution of Projects in RAD Treatment Group Sample by PHA Size and Project Performance Rating

PHA Size	Substandard		Standard		High-Standard		Total RAD Projects in Sample by PHA Size	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
Small	0	0	2	8	3	13	5	21
Medium	1	4	8	33	6	25	15	63
Large	1	4	2	8	1	4	4	17
Total RAD projects in sample by performance rating	2	8	12	50	10	42	24	100

PHA = public housing authority. RAD = Rental Assistance Demonstration.
Source: Public and Indian Housing Information Center and RAD program data as summarized by Econometrica, Inc.

[10] The RAD program was expanded at about the time of our sampling. Due to the time limitations of our study, we did not sample from the entire universe of 1,074 public housing projects that eventually applied to RAD. Projects that applied later, were progressing slowly, or dropped out were not in our sample frame because they would have offered little information on the impact of RAD. Researchers selected a supplemental sample of projects that withdrew from RAD or had their CHAPs revoked by HUD to analyze the factors contributing to that outcome.

Exhibit 2

Distribution of Projects in Non-RAD Control Group by PHA Size and Project Performance Rating

PHA Size	Substandard		Standard		High-Standard		Total Non-RAD Projects in Control Group by PHA Size	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
Small	0	0	2	4	8	17	10	21
Medium	2	4	17	35	11	23	30	63
Large	3	6	3	6	2	4	8	17
Total non-RAD projects in control group by performance rating	5	10	22	46	21	44	48	100

PHA = public housing authority. RAD = Rental Assistance Demonstration.
Source: Public and Indian Housing Information Center and RAD program data as summarized by Econometrica, Inc.

Accounting for Bias

Drawing a simple random sample of 48 projects from the population of 6,664 non-RAD public housing projects would introduce *self-selection bias*, because PHAs *choose* to participate (or not participate) in RAD; they are not randomly assigned to RAD. Failing to account for this choice could result in biased estimates that would reduce the accuracy and reliability of the findings. However, at the start of the study, the research team expected that RAD projects could differ systematically from non-RAD projects due to self-selection bias. For instance, PHAs might prefer to submit a well-managed project to the RAD program because such a project would be less risky.[11] In addition, the goal of RAD is to generate capital for rehabilitation, and PHAs might therefore select projects that need more capital improvements than other public housing developments to take advantage of that feature of the program. The size of a PHA could also affect its participation in RAD; if smaller PHAs have less mixed-finance experience and therefore less familiarity with the financing tools that RAD makes available, they may not understand the advantages of RAD or they may feel they cannot make RAD work for them.

To eliminate potential self-selection biases, such as those described above, and to give more confidence to the findings, statisticians matched the 24 projects in our RAD sample with non-RAD public housing projects based on observable characteristics that could account for differences in the likelihood that a given project would participate in RAD. Using RAD program data, HUD administrative data from the Public and Indian Housing Information Center, and data from the 2008–2012 American Community Survey 5-year estimates, the research team created a data set of non-RAD properties with usable information for 13 matching variables, or covariates. These variables were selected to capture key characteristics of PHAs, public housing projects, and the neighborhoods in which the projects are located. The covariates used for this matching are listed in exhibit 3, along with the rationale for each covariate and the source of the data. The only PHA-level variable was the size of the PHA based on the number of public housing units under

[11] Because RAD projects can assume project debt, which is repaid out of project cashflows, PHAs may consider better-managed projects to be less likely to default under RAD.

management. Property-related variables included information on the property's size (number of Annual Contributions Contract units), age (Date of Full Availability, construction date, or date of last modernization), structural type (building and development type), bedroom mix (percentage of zero-, one-, or two-bedroom units), physical condition (Real Estate Assessment Center inspection score), and vacancy rate. Neighborhood-level variables capture information on the strength or weakness of local affordable housing market conditions, such as rents that are high relative to average household income (cost-burden rate), overcrowded living conditions (overcrowding rate), degree of poverty in the community (poverty rate), extent to which households in the area rent rather than own their homes (percentage of renters), and the prevalence of vacant housing (vacancy rate) in the area.

Exhibit 3

Covariates Used To Match RAD Properties With Non-RAD Properties (1 of 3)

Variable	Description	Rationale	Data Source
ACC_Unit_Cnt	Number of Annual Contributions Contract units in a property	Indicator of the size of the development. Property maintenance and replacement costs are expected to be commensurate with the number of units in a property.	PIC database
Bldg_Type_Code	Building type of project = 1, 2, 3, 4, 5, where: 1 = ES, elevator structure 2 = RW, rowhouse or townhouse style 3 = SD, semidetached 4 = SF, single-family detached 5 = WU, walkup/multifamily apartment	Property maintenance and replacement costs are driven in part by building type, in that the cost of maintaining or replacing a physical asset such as an elevator will impact the level of capital needs.	PIC database
Dev_Type_Code	Development type of the project = 1, 2, 3, where: 1 = elderly 2 = mixed 3 = family	According to *Capital Needs in the Public Housing Program* (Abt Associates, 2010), average capital needs vary by type of housing. For example, the average amount of capital needs for an elderly unit is lower than that of a family unit.	PIC database
DOFA	Date of Full Availability for the project	Indicates the age of the building, which is important for determining replacement needs. DOFA establishes when a development can access the operating subsidy from a PHA's Operating Fund. In most cases, this date is the same as the construction date. We also considered the last modernization date, if available.	PIC database

Exhibit 3

Covariates Used To Match RAD Properties With Non-RAD Properties (2 of 3)

Variable	Description	Rationale	Data Source
Percent_1_2_Bed	Percentage of units in the project that have either zero, one, or two bedrooms	Indicator of the size of the unit. Costs associated with the unit size of individual units are not equally distributed.	PIC database
PHA_Size_Code	PHA size = 1, 2, 3, where: 1 = small, ≤ 250 units 2 = medium, 251–1,250 units 3 = large, > 1,250 units	Large PHAs differ from small PHAs. A PHA's planning process is unique to the PHA but related to the size of the PHA. The PHA plan includes policies, programs, operations, and strategies for meeting local housing needs and goals. Factors must be consistent with the housing and community development plans of the jurisdiction (as described in the Consolidated Plan); thus, PHA size matters.	PIC database
Rounded_Inspection_Score	Physical inspection score (rounded) for the project	REAC conducts approximately 20,000 physical inspections on housing properties annually to ensure that families living in public housing have decent, safe, and sanitary housing that is in good repair. Scores range from 0 to 100. Properties that receive a Public Housing Assessment System score greater than 90 are considered high performers; properties that score between 70 and 89 are standard; properties that score lower than 70 are substandard or troubled. High-scoring properties are inspected every 3 years, standard performers are inspected every 2 years, and troubled properties are inspected every year. The inspection score served as a proxy for estimating capital needs; properties with high scores are likely to have fewer capital needs than those with lower scores.	REAC file
Vacancy_Rate	Vacancy rate in the project	Calculated as the percentage of units occupied. Indicator of both the condition of the development and the quality of PHA management. One would expect that a well-managed development in good physical condition would be 100% occupied.	PIC database

Exhibit 3

Covariates Used To Match RAD Properties With Non-RAD Properties (3 of 3)

Variable	Description	Rationale	Data Source
Cost_Burden_Rate	Cost-burden rate in the census tract	Measures the percentage of renters with gross rent greater than or equal to 35 percent of their income. Indicator of both the cost of housing in the local market and of poverty in the neighborhood in which the development is located.	ACS data—U.S. Census Bureau; by census tract
Overcrowd_Rate	Overcrowding rate in the census tract	Calculated as number of persons/ number of rooms. A ratio greater than 1 is defined as overcrowded. Indicator of local housing market conditions and poverty in the neighborhood in which the development is located.	ACS data—U.S. Census Bureau; by census tract
Poverty_Rate	Poverty rate in the census tract	Percentage of neighborhood residents below the poverty level.	ACS data—U.S. Census Bureau; by census tract
Renter_Rate	Renter rate in the census tract	Percentage of neighborhood housing stock occupied by renters. Indicator of the type of housing available in the neighborhood in which the development is located.	ACS data—U.S. Census Bureau; by census tract
Vacant_Rate	Vacancy rate in the census tract	Percentage of vacant homes in the neighborhood in which the development is located. Indicator of demand and supply conditions in the local housing market. The vacancy rate determines the choices open to consumers in a market. As housing supply expands, housing vacancies rise, and demand will either remain the same or decrease as more residents find available units; as vacancies decrease, the housing supply either remains the same or contracts while demand grows.	ACS data—U.S. Census Bureau; by census tract

ACS = American Community Survey. PHA = public housing authority. PIC = Public and Indian Housing Information Center. RAD = Rental Assistance Demonstration. REAC = Real Estate Assessment Center.

One can see the amount of bias in the RAD sample directly by comparing the group of 24 RAD projects with the entire set of 6,644 non-RAD (NR) public housing projects. Exhibit 4 compares the number of projects in each group, the mean value for each variable for the RAD (Mean_{RAD}) and non-RAD (Mean_{NR}) groups, the standard deviation (StdDev) and standard error (StdErr) for difference in means, the *t*-value, and the bias. The bias is similar to a *t*-test for the difference in two means.[12] A high bias will often result in a *t*-test that rejects the null hypothesis that the two means are equal. For example, Vacant_Rate has a bias of -41.9 percent and a *t*-value of -2.05 (which would lead to a rejection of the null hypothesis that the two means are the same at the 95-percent confidence level). Exhibit 4 shows that the RAD sample and the population of non-RAD projects are fairly dissimilar. Even though the average bias of 4.2 percent is fairly low, the disaggregated results for each of the covariates show much higher levels of bias: Vacancy_Rate (within the project) has a 39.5-percent bias, and Poverty_Rate and Renter_Rate both have biases greater than 20 percent.

Due to the dissimilarity between RAD projects and the population of non-RAD projects, the research team gathered a matched sample of non-RAD projects for our control group. The goal was to establish a control group that has a lower overall bias and lower bias for individual covariates. The means to achieve this goal was the GM algorithm. The target was a control group of 48 non-RAD projects matched to the randomly selected sample of 24 RAD projects. However, to reach this target, statisticians implemented the GM algorithm using a 4-to-1 match (that is, 4 matches were selected for each of the 24 projects in the sample of RAD projects). This approach resulted in the selection of 96 unique non-RAD projects, which was twice the size of the desired sample of non-RAD projects in case some matches had to be rejected. Matches were selected without replacement, so each non-RAD project could be selected as a control only once. When possible, the two best matches were selected for the study. However, in several cases, matched non-RAD projects

Exhibit 4

Comparison of Means by Covariates for RAD Sample and Non-RAD Population

Variable	Mean_{RAD}	Mean_{NR}	StdDev	StdErr	*t*-Value	Bias (%)
ACC_Unit_Cnt	159.6	154.8	208.3	42.6037	0.11	2.3
Bldg_Type_Code	2.5833	2.4104	1.1134	0.2277	0.76	15.5
Dev_Type_Code	2.9167	2.8639	0.4234	0.0866	0.61	12.5
DOFA	1974.5	1976.6	16.9795	3.4722	– 0.60	– 12.4
Percent_1_2_Bed	0.6204	0.6610	0.3103	0.0635	– 0.64	– 13.1
PHA_Size_Code	1.9583	1.9738	0.8351	0.1708	– 0.09	– 1.9
Rounded_Inspection_Score	84.5417	84.8222	12.9316	2.6444	– 0.11	– 2.2
Vacancy_Rate	0.1658	0.0919	0.1868	0.0382	1.93	39.5
Cost_Burden_Rate	44.5833	42.1859	13.1244	2.6838	0.89	18.3
Overcrowd_Rate	0.0344	0.0375	0.0472	0.0097	– 0.32	– 6.4
Poverty_Rate	31.4750	27.8604	15.1196	3.0918	1.17	23.9
Renter_Rate	55.6292	51.0282	22.1308	4.5256	1.02	20.8
Vacant_Rate	10.9125	14.5284	8.6328	1.7653	– 2.05	– 41.9
Average bias						**4.2**

RAD = Rental Assistance Demonstration.

Source: Public and Indian Housing Information Center and American Community Survey data as analyzed by Econometrica, Inc.

[12] The *t*-value is calculated by dividing the difference in means by the standard error instead of the standard deviation. Loosely speaking, standard error = (standard deviation) × sqrt(1/n). The square root term causes the standard error to approach 0 as the sample size increases. The precise mathematical relationship between the bias and the *t*-value is Bias = tValue * sqrt($1/n_1$ +$1/n_2$).

had to be rejected for various reasons. For instance, one project was no longer a public housing project and was eliminated. Several projects turned out to be RAD projects (they had applied to the program after the cap was lifted) and thus were not appropriate as controls. One project was rejected because it was a single highrise building that had been matched against a project of scattered townhouse units. In other cases, the PHAs declined to participate and were dropped from the study. When a potential control was rejected, staff selected the next best match in the list. In a few cases, the controls ran out. Consequently, for some of the projects in the RAD sample, staff reran the genetic matching program to select 4 more controls to complete the final sample of 48 projects.

Assessing Genetic Matching Results

Using genetic matching reduced bias in the samples. The results of the GM algorithm are given in exhibit 5. On one variable, PHA_Size_Code, the samples were a perfect match under both the *t*-test for difference in means and the KS test. This means that the RAD sample and the non-RAD sample could be sorted into equal numbers of projects of the exact same size class. On Poverty_Rate, the samples were a near-perfect match—perfect in terms of the difference in means but not with respect to the KS test. The worst match in terms of *p*-values was on Bldg_Type_Code, at 29.6 percent under the KS test and 48.8 percent under the difference in means test. Another way of describing this result is to say that one can reject the null hypothesis that the distribution of building types in the two samples is the same with 70-percent confidence. Typically, analysts demand a higher degree of confidence—in the 90- or 95-percent range. Bias might still be present with respect to the Bldg_Type_Code and Vacant_Rate variables. For the latter, one can accept the null hypothesis that the means are the same with only 45-percent confidence.

Exhibit 6 compares the bias of the unmatched set of all non-RAD public housing projects against the final set of 48 matched non-RAD projects in the sample of comparison projects. For all but three covariates, bias decreased after the match. For two covariates—ACC_Unit_Cnt and

Exhibit 5

Results of the Genetic Matching Algorithm

Variable	*p*-Values (%)	
	***t*-Test for Difference in Means**	**KS Test**
ACC_Unit_Cnt	72.8	55.2
Bldg_Type_Code	48.8	29.6
Dev_Type_Code	81.2	92.2
DOFA	67.0	76.3
Percent_1_2_Bed	77.3	98.4
PHA_Size_Code	100.0	100.0
Rounded_Inspection_Score	76.6	79.0
Vacancy_Rate	84.3	98.8
Cost_Burden_Rate	87.4	81.6
Overcrowd_Rate	88.9	72.6
Poverty_Rate	100.0	99.8
Renter_Rate	77.3	82.4
Vacant_Rate	53.3	56.3

KS Test = Kolmogorov–Smirnov test.

Source: Based on analysis by Econometrica, Inc.

Exhibit 6

Comparison of the Degree of Bias Before and After Match

Variable	Mean_{RAD}	Mean_{NR}	Unmatched			Matched		
			N_{NR}	Mean_{NR}	Bias (%)	N_{NR}	Mean_{NR}	Bias (%)
ACC_Unit_Cnt	24	159.6	6,644	154.8	2.3	48	149.4	8.8
Bldg_Type_Code	24	2.5833	6,644	2.4104	15.5	48	2.3958	17.4
Dev_Type_Code	24	2.9167	6,644	2.8639	12.5	48	2.9375	– 5.9
DOFA	24	1974.5	6,644	1976.6	– 12.4	48	1976.1	– 10.8
Percent_1_2_Bed	24	0.6204	6,644	0.6610	– 13.1	48	0.6378	– 7.3
PHA_Size_Code	24	1.9583	6,644	1.9738	– 1.9	48	1.9583	0.0
Rounded_ Inspection_Score	24	84.5417	6,644	84.8222	– 2.2	48	85.5208	– 7.4
Vacancy_Rate	24	0.1658	6,644	0.0919	39.5	48	0.1518	5.0
Cost_Burden_Rate	24	44.5833	6,644	42.1859	18.3	48	44.1458	3.9
Overcrowd_Rate	24	0.0344	6,644	0.0375	– 6.4	48	0.0332	3.6
Poverty_Rate	24	31.4750	6,644	27.8604	23.9	48	31.4583	0.1
Renter_Rate	24	55.6292	6,644	51.0282	20.8	48	56.8750	– 7.2
Vacant_Rate	24	10.9125	6,644	14.5284	– 41.9	48	11.7083	– 15.7
Averages					**4.2**			**– 1.2**

Source: Based on analysis by Econometrica, Inc.

Rounded_Inspection_Score—the bias was higher after the match (in absolute value) but still less than 10 percent. Given the results in exhibits 5 and 6, some further caution might be due with respect to three covariates: Bldg_Type_Code, DOFA, and Vacant_Rate. Our overall conclusion is that the genetic matching reduced bias. After-match average bias is less than one-third its original size, falling from 4.2 to -1.2 percent.

Concluding Remarks

As evidenced by the results outlined in this article, genetic matching is a powerful tool. It can help researchers mitigate systemic bias in quasi-experimental situations in which assignment to the treatment group is nonrandom and a specific set of covariates is believed to influence the propensity to participate. Within the public policy space, genetic matching is particularly well-suited for program evaluation in which participation in the treatment or exposure to the policy reform has not been randomized (that is, it may be voluntary or the result of participation in previous reforms) and the researchers seek to draw conclusions about the impacts of the policy in question (Sekhon and Grieve, 2008).

In such instances, genetic matching is used to control for imbalances that are expected to impact the propensity of selection to, or participation in, the treatment group and *not* imbalances related to each group's values for the dependent variables of interest. Of course, in some situations, genetic matching would not be an appropriate methodological choice. These situations include research questions that do not involve the identification or selection of a comparison group (for example, longitudinal analysis using data for a single population) and research designs that employ random selection to choose the units that will receive treatment.

Acknowledgments

The authors thank Nathan Bossie of the U.S. Department of Housing and Urban Development (HUD) for his guidance in submitting this article and acknowledge the assistance of Bethany Hase of Econometrica, Inc., and anonymous HUD reviewers in the preparation of this article for publication.

Authors

David Ruiz is the Lead of the Data and Analytics Practice at Econometrica, Inc.

Dennis Stout is Director of Housing and Community Development at Econometrica, Inc.

Christine Herlihy is a data scientist at Econometrica, Inc.

References

Abt Associates. 2010. *Capital Needs in the Public Housing Program*. Washington, DC: U.S. Department of Housing and Urban Development.

Diamond, Alexis, and Jasjeet S. Sekhon. 2013. “Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies,” *Review of Economics and Statistics* 95 (3): 932–945.

Sekhon, Jasjeet S., and Richard Grieve. 2008. “A New Non-Parametric Matching Method for Bias Adjustment With Applications to Economic Evaluations,” *SSRN Electronic Journal*. DOI: 10.2139/ssrn.1138926.

Sekhon, Jasjeet S., and Walter R. Mebane, Jr. 1998. “Genetic Optimization Using Derivatives,” *Political Analysis* 7: 189–203.

Stuart, Elizabeth A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward,” *Statistical Sciences* 25 (1): 1–21.