Supplementary Materials

Mentored research experiences bridge the wealth gap in PhD admissions

Authors: Nicholas David[1], Misha Teplitskiy[2], Sidney Xiang[2], Daniel Romero[2,3,4], Dallas Card[2], Wenhao Sun[1*]

Corresponding author: Wenhao Sun (whsun@umich.edu)

The file includes:

Materials and Methods

Figs. S1 to S5

Tables S1 to S10

References

Materials and Methods

S1 Large-scale construction of a doctoral admissions database

To enable the large-scale quantitative analysis of doctoral admissions presented in the main text, we collected and processed data from internal university services combined with external linkage to other datasets. We outline the various acquisition methods and variables in the following sections. Figure **S1** illustrates the connection between various data sources.

S1.1 Prior existing digitized application features

Several application features are already available from the university's data warehouse which are pulled directly for applications submitted between 2013-2023. These include GPA (out of 4.0), GRE-V and GRE-Q (130-170), prior degree information, previous work experiences, citizenship status, and prior Pell Grant recipient status. Table **S1** outlines these variables and their types.

Faculty evaluations for applications submitted to the engineering college were also available in tabular format. The structure of these evaluations varies between departments, but all consist of topic specific ratings (e.g., academic record, recommendation letters, experience, statement of purpose), an overall rating (e.g., 3, 2, 1) and some brief textual comments. Table **S2** describes these evaluations in detail.

S1.2 PDF extraction of application features

Several application features were unavailable in a computationally analyzable format and therefore had to be text-mined from PDF versions of the application. Features that were extracted from PDFs include reported financial hardship (1–lowest; 5–highest), highest level of parental education, applicant's permanent address (for real-estate data linking), academic statement of purpose, personal statement, curriculum vitae, and recommendation information (letters and scores). Table **S3** outlines the variables extracted from applications in PDF format.

At a high level, the process for extracting this information is as follows:

(1) Segment the PDF according to a series of deterministic rules based on the structure of the compiled PDF. This allows us to construct bookmarks with the start and end of application sections, for example, the start and end of a recommendation. Page text is extracted using *PyMuPDF*[1], a PDF extraction software for Python.
(2) Extract the information from the page body or defined bounding box area for targeted extraction.
(3) Place the extracted information in a database of key-value pairs of the extracted type and the extracted value.

More specific details on this process are in Section **S1.2.1-S1.2.2**

S1.2.1 PDF segmentation

Compiled applications are stored in PDF format with preexisting bookmarks for the application, recommendations, and transcript sections. Within the application and recommendation sections exist subsections that do not come bookmarked. Our goal is to identify the span or range of these subsections using patterns and substructures of the PDF. The subsections include the application forms, academic statement of purpose, personal statement, curriculum vitae, and recommendation forms and letters. The application asks students to include a personal statement—different than the typical academic statement of purpose—which prompts applicants to explain how their background, life experiences, cultural, geographical, financial, educational, and other opportunities motivated their decision to pursue a graduate degree at the university. Below is the algorithm used to perform this segmentation.

(1) The generic form start page is the first page of the document. Since each generic form page shares the same header (applicant name, program name, term), we advance the end page to the generic form section until this is no longer the case.
(2) The academic statement of purpose always immediately follows the generic form section; therefore, its start is one more than the end of the generic form section.
(3) Define the start page of the personal statement by examining the first 20 lines of each successive page from the start of the academic statement of purpose for the terms "personal" and "statement" occurring in the same line. The end page of the academic statement of purpose is therefore one less than the start of the personal statement page. When matching these headers, we use the Levenshtein distance up to 1 edit distance away due to small misspellings.
(4) Determination of the start and end pages for the curriculum vitae and personal statement is decided using the following logic:
    (a) If the first 20 lines of a page after the start of the personal statement contains the words "curriculum vitae", "CV", or "resume", define this page to be the start. Then we may define the end of the personal statement as one less than the start of the curriculum vitae.
    (b) Otherwise, if the combined word count of the start page of the personal statement and the following page exceeds a threshold (600 words) and the bottom quarter of the following page is empty, we treat the personal statement as two pages; otherwise the personal statement is one page. The application requirements impose a 'soft self-enforced' 500 word limit on the personal statement.
(5) Occasionally, there exists a supplemental writing piece for some degree programs that directly follows the curriculum vitae. To avoid absorbing this information we do the following:
    (a) Check for a page-dimension change across each successive curriculum vitae page from the start until the start of the recommendations bookmark.
    (b) If the distance between the start of the curriculum vitae is larger than 5 pages, truncate the span to 5 pages.
(6) A recommendation consists of an easily recognized recommendation form (1 page) and the recommender's letter which follows directly after. This sequence repeats itself until the bookmarked transcript section. Therefore, by leveraging the consistent headers of the recommendation form pages we're able to segment each individual recommendation (typically three, sometimes more) into their form and letter.

The resulting deterministic segmentation results in a detailed table of contents for which we can perform targeted information extraction of application features. The only failed segmentations arise from either incomplete or malformed sections between the generic form sections and recommendations. A total of 11,873 of the 175,702 (6.8%) Fall matriculating PhD applications failed to be segmented into these sections. For these applications, the academic statement of purpose, personal statement, and curriculum vitae are extracted into a list of strings, in order of pages extracted. In addition, PDF segmentation and extraction failed on two application PDFs: one due the PDF originating as an image making text extraction unavailable and the other due to the inclusion of an exceedingly large image, resulting in a read error.

S1.2.2 Targeted extraction of form fields

Some identifying information, such as the recommender's credentials, applicant addresses, and other short entries appear as form field entries. Because this information appears in a consistent location across applications within the same cycle, we extract it using a predefined template that specifies the bounding box corresponding to the relevant content. The information extracted from these boxes is stored along with the name of the field (e.g., Address Line 1). In some cases, such as zip code, typing constraints are placed on the output. Additionally, we leverage ordering constraints from the form field to deduce the field labels associated with extracted information. For example, text appearing on the same line directly to the right of "Address Line 1" is consistently the applicant's supplied address line 1 information.

S1.2.3 Targeted extraction of checked boxes

Multi-choice selections solicit categorical information from applications in the form of check boxes. These boxes are either empty or contain an 'X' indicating a selection. These selections appear in a consistent format for each application term. Therefore, we're able to discern which selections are made using an image masking template which relates grayscale pixel values to selections, black pixel rations greater than 0.1, or no selections black pixel ratios less than or equal to 0.1. Using the position of the mask, the selection item (e.g., 'yes', 'no') is determined.

This extraction technique is used to gather some mentored research experiences, namely students' participation the Ronald E. McNair Post-Baccalaureate Achievement Program, Research Initiative for Scientific Enhancement (RISE), Post-baccalaureate Research Education Program (PREP), Minority Access to Research Careers (MARC), Michigan Humanities Emerging Research Scholars Program (MICHHERS), Mellon Mays Undergraduate Fellowship, Louis Stokes Alliance for Minority Participation (LSAMP), Graduate Engineering and Science Fellowships for Minorities (GEM), Bridges to the Baccalaureate, and the Big Ten Academic Alliance - Summer Research Opportunity Programs. Participation in these mentored research experiences is only solicited for U.S. citizens and permanent residents.

S1.3 Text-derived application features

A major component of a PhD application is the statement of purpose–a chance for a student to describe their past experiences and future ambitions–and the letters of recommendations. Within these bodies of text, along with the applicant's curriculum vitae, we search for possible mentions of publications and/or Research Experiences for Undergraduates (REUs) participation. To do this we use two separate regular expressions to flag an applicant as either having or not having each of these features. The regular expression for publications matches for "publication(s)", "publish(ing)", or "published". The expression for REU matches the National Science Foundation's REU program as well as other common acronyms for organized, university or program sponsored mentored research experiences. We use the following regular expressions to identify these REU programs.:

- "\b((?-i:REU)|(?:NSF\s+)?Research\s+Experience(s)?\s+for\s+Undergrad(uate)?s?)\b"
- "\b((?-i:SRIP)|Summer\s+Research\s+Intern(ship)?\s+Program(me)?)\b"
- "\b((?-i:SURE)|Summer\s+Undergrad(uate)?\s+Research\s+Experience(s)?)\b"

- "\b((?-i:SROP)|Summer\s+Research\s+Opportunit(y|ies)\s+Program(me)?)\b"
- "\b((?-i:UROP)|Undergrad(uate)?\s+Research\s+Opportunit(y|ies)\s+Program(me)?)\b"
- "\b((?-i:SURP)|Summer\s+Undergrad(uate)?\s+Research\s+Program(me)?)\b"
- "\b((?-i:SURF)|Summer\s+Undergrad(uate)?\s+Research\s+Fellow(ship)?)\b"

To validate the measurement of these features we manually analyze a total of 200 PhD applications, 100 for the REU objective and 100 for the publication objective, drawn randomly from our dataset enforcing a 50:50 split of labels. We find that the REU and publication features have an F1-score of 0.95 and 0.87 respectively.

S1.4 Engineering college faculty evaluations

The engineering college has a centralized system that collects faculty evaluations on applications. These evaluations consist of scores and comments, which are detailed in Table **S4**. Faculty evaluations are linked to applications using the department ID, department name, term, and employee ID fields provided. Department names and IDs are slightly different from those used by the university under the program name and program code fields. Because of this and the fact that a single person can apply to multiple programs in the same term, a total of 2,968 out of 76,048 (3.9%) faculty evaluations remain unlinked. Not every application has a faculty evaluation and some applications have multiple. Of the fall matriculating applicants, 22,968 have a faculty evaluation.

S1.5 Recommender features and their linked bibliometrics

Besides the intrinsic grades and experiences applicants bring to their applications, recommenders write letters and leave scores attesting to the applicant's readiness for PhD studies. This advocacy on behalf of the applicant is backed by the recommender's reputation in their field. Each complete application contains 3 or more recommendations which we process from PDFs into cumulative statistics for each application.

Recommenders rate applicants according to the six criteria. The categories they score applicants include (1) knowledge in chosen field, (2) motivation and perseverance toward goals, (3) ability to work independently, (4) ability to express thoughts in speech and writing, (5) ability/potential for college teaching, (6) ability to plan and conduct research. For each category, recommenders select from one of the following ratings: exceptional upper 5%, outstanding next 15%, very good next 15%, good next 15%, next 50%, and lastly "no basis for judgement"—which is excluded from any subsequent aggregation. Recommenders who do not mark a score for a particular category (leaving all boxes blank) are similarly excluded from aggregation for that category. Median scores for each category are computed across recommendations and averaged to produce an average "skill score" used in Figures 2 and 4 of the main text. In computing the average skill score, we first convert the categorical scores into midpoints of their respective ranges (i.e., 97.5, 87.5, 72.5, 57.5, 25). The skill scores seen in Figure 4 of the main text treat the scores as integers with 5 being the highest and 1 as the lowest.

In addition to the skill scores, recommenders leave an overall endorsement on the application, marked as "highly recommended", "recommended", "recommended with some reservations", "not recommended". These are converted to an integer 4, 3, 2, or 1 and are averaged across an applicant's recommendations to produce a single endorsement score.

To measure the prestige of a recommender, bibliometric data from Elsevier's Scopus[2] and SciVal[3] for recommenders that are probable researchers (i.e., with "professor", "research", "researcher", or "lecturer" in their title). The recommender's first name, last name, and affiliation (typically a university's name) is used to query Elsevier for the author's profile. 514,572 recommendations result in 243,857 unique queries and 123,127 author profiles. *H*-index data is gathered as of October 2024. Table **S5** outlines the variables gathered from Elsevier's Scopus and SciVal which were used in this study.

The last feature gathered from the recommendation "letter length" used in the regression and mediation analysis is a measure of the number of words used in an application's recommendations. A longer recommendation typically corresponds to a more detailed, meaningful recommendation than a shorter, brief one. Word counts were computed by joining text such that lines ending in a hyphen were concatenated to the subsequent line without a space, while all other line breaks were replaced with a space. The resulting text was then tokenized on whitespace. These recommendation letter lengths are then averaged across all recommendations in each application to produce a single estimate of the average letter length for each application. Missing letters were excluded from this averaging scheme.

S1.6 Linking QS World University Rankings

To compare the effects of undergraduate institution prestige on admission for both domestic and international applicants, we linked applicants' undergraduate institutions to the QS World University Rankings[4]. Table **S6** describes the variables and types gathered from the QS rankings. Our dataset contains unique undergraduate institution names that we manually matched to the names found in historical 2013-2023 QS World University Rankings to obtain a year normalized ranking for an applicant's undergraduate institution in their applied term. Because institution names differ slightly between the QS rankings and the ones listed in the application records, we compute the Levenshtein distance between each unique QS institution name and all undergraduate institution names in the applications database. The five most similar institution names, along with their country, state, and city, are presented to the annotator (the research team) for selection of the correct match. We performed this matching routine on all 1,270 QS institution names to application institution names. In some cases, none of these five are correct, which results in no match. In total, 115,321 (65.6%) applications have a linked QS World University Rankings rank.

S2 Linking applicant parcel-level real estate data

To measure the economic status of an applicant, we link their address to parcel-level real estate data to calculate the amount of equity their family (or themselves) has in their primary residence. To do this we link the applicants' permanent address (as opposed to current address) to historical basic property, deed transfer, and mortgage acquisition data, provided by CoreLogic[5–7]. Details of the variables contained and extracted from these data sources are in Tables **S7-S9**.

Since applicant addresses originate from a free-form text field, we standardize the address according to the United States Postal Service addressing standards along with CoreLogic specific formats. Addresses are then directly matched to CoreLogic entries by address fields. To increase yield, offline geocoding software *Nominatim*[8]—the backend for *OpenStreetMap*[9]—is used to convert addresses to geospatial boundaries, providing a secondary matching criterion to CoreLogic entries. Any geospatially matched parcel undergoes additional checks for certain matched address information, such as house number and zip code.

S2.1 Address linking algorithm

An applicant's permanent address contains an address 1 (addr1), address 2 (addr2), city, state, zip code, and country field. Because the address fields are free form, numerous standardizations and normalizations are applied to allow for querying CoreLogic's database.

(1) First, ensure the permanent address is within the United States. This is done by checking if "united" and "state"(s) is in the country field, or if its abbreviation US, USA, U.S.A., U.S.A, U.S., or U.S is present.
(2) Using only the addresses within the United States, remove punctuation ('.' and ','), uppercase all fields, strip whitespace, and truncate zip codes to 5 digits.
(3) Format addresses to align with CoreLogic formatting. That is split and convert street suffixes (e.g., STREET → ST), cardinal abbreviations (e.g., NORTH →N), state abbreviations (e.g., MICHIGAN → MI), and unit designators (e.g., APARTMENT → APT). Apply these substitutions in a forwards and backwards, tree-like manner creating every possible edit permutation according to these substitutions. In any unit designator substitution, replace the designator, and any trailing whitespace with a '#'. For example, "APT 5" becomes "#5".
(4) Transform each new standardized address into a query key. If address line 2 is non-empty, the key becomes:

    "{addr1} {addr2}_{city}_{state}_{zip}", otherwise "{addr1}_{city}_{state}_{zip}".

With this format established, we use the CoreLogic address to fields to construct one unique address. Since the SITUS STREET ADDRESS field contains the full address not including the city, state, or zip code, the CoreLogic key is constructed in the same manner as before with this field replacing the addr1 (or addr1 addr2) placement in the key string.

A local copy of *OpenStreetMap*, with *Nominatim* as the search tool, is used to process applicant addresses into geospatial coordinates. These coordinates are assembled into an array of spatial coordinates linked for query.

Using these key dictionaries we iterate over each parcel entry in CoreLogic, collecting parcel information that matches the address directly or is within the geospatial bounds of an applicant's parcel using the LATITUDE and LONGITUDE data fields if available.

Thereafter, we have a dataset of CoreLogic linked applicant addresses which we further refine in section **S3**.

S3 Validating housing wealth

The matched parcel data then undergoes a series of filters to ensure that the real-estate data gathered can be used to reliably measure an applicant's primary residence home equity:

(1) Familial ownership – The surname of the buyers (or owners) listed on the deed transfer must match the surname of the applicant.

(2) Family occupied – The property must be listed as owner occupied on the deed transfer (e.g., not absentee owners)
(3) Family debt – The surname of the borrowers on the mortgage must match the surname of the applicant.

If multiple matching deed transfers or mortgages are found for an applicant, the one closest to their applied date is chosen. For example, if an applicant applies in Fall of 2016, the parcel data occurring closest to December 1st, 2015 (the most likely application submission date) is chosen. Additionally, a homeowner may refinance their principal mortgage after purchasing their home, which in that case is used to calculate home equity if the refinancing occurs before the application date. Parcels that are indicated as not having a mortgage by CoreLogic assume no debt.

All real estate dates are limited to transactions on or after January 1st, 2000, and to parcels categorized as either a single-family residence, condominium, duplex, or apartment. Parcel sales not conducted as an arms-length transaction are filtered, as well as parcels selling for less than or equal to $1,000 due to the misspecification, or lack thereof, transaction types[10].

S3.1 Home equity calculation

Home equity is a strong indicator of socioeconomic status, often representing an individual's largest asset. To calculate home equity at the time of application, mortgage balances are amortized under a 30-year fixed rate model using historical average national mortgage rates from Freddie Mac[11].

The outstanding balance B on a loan after k months is:

$$B = L \times \frac{(1+r)^n - (1+r)^k}{(1+r)^n - 1}$$

Where L is the loan amount, r is the monthly rate, and n is the total length of the loan in months ($n = 360$ for a 30-year fixed loan). k is taken to be the number of months elapsed between the mortgage financing date and the application date.

Using the estimated debt on the parcel, we can subtract that from the estimated market value of the parcel to get the home equity of the applicant at the time of applying. The sale price of the parcel is adjusted to the application date using the *Annual Five-Digit ZIP Code FHFA House Price Index*®[12]. This accounts for yearly fluctuations in local home values. Now that both the parcel value and parcel debt are referenced to the application date, we can subtract the two to compute home equity.

home equity = estimated market value - estimated debt

The *Consumer Price Index for All Urban Consumers*[13] is then used to compare home equity values across application dates.

S4 Modeling admissions

To understand how different application features contribute to the admissions decision, we model admissions using a multivariate logistic regression. Application features included in the regression model are below:

(1) GPA – Referenced to a 4.0 scale using the university provided GPA scale attribute.
(2) GRE-Q – The score for the Quantitative Reasoning Graduate Record Examination. A score of 130 to 170. GRE related analyses are restricted to years prior to 2022, before some programs removed GRE as an application requirement.
(3) GRE-V – The score for the Verbal Reasoning Graduate Record Examination. A score of 130 to 170. GRE related analyses are restricted to years prior to 2022, before some programs removed GRE as an application requirement.
(4) Undergraduate institution rank – The QS World University Rankings rank for the applicant's undergraduate institution.
(5) Publication – A flag derived from the textual components of the application indicating whether the applicant has published a manuscript (1) or not (0).
(6) REU experience – A flag indicating whether the applicant has participated in a mentored research experience. This flag is 1 if the applicant participated in one of the mentored research experiences extracted in **S1.2.3** or in one of the textually derived experiences in **S1.3**.
(7) Master's degree – A flag indicating whether the applicant has completed a master's degree prior to applying (1 = yes; 0 = no).
(8) Number of work experiences – A number counting the number of reported work experiences the applicant has had. Each application contains a list of previous work experiences that is totaled to produce this feature.
(9) *H*-index – The *h*-indices of an applicant's recommenders, averaged.
(10) Skill score – The average skill score given to applicants from recommenders. For each category, the median skill score across recommenders is found, then across categories the mean skill rating is found to produce the skill score. Applications which receive no single recorded score, sometimes due to recommender elections not to, result in a null skill score.
(11) Letter word length – The averaged number words from the cleaned recommendation letters given to an applicant on an application. Letter word counts were computed by joining hyphen-broken lines, replacing other line breaks with spaces, tokenizing on whitespace, and averaging across an application's letters.
(12) Endorsement – The average endorsement score, computed across recommendations, given to an applicant on an application.
(13) Citizenship – The citizenship status of the applicant separated by U.S. citizens and permanent residents (0) and other (1).

In addition to these variables, we include the applicant's term and program as an interaction to account for within term and program differences in admission. Reference groups for programs and terms are chosen as the Computer Science and Engineering PhD and Fall 2013, respectively. This is expressed in the model formulation through the multiplication operator '*' (i.e., program_code * term) in *patsy*[14]. Two-sided 95% confidence intervals were computed using an asymptotic normal (z-based) distribution to estimate p-values. Specific to REU experiences, we filter out programs that are unlikely to provide such experiences to applicants (i.e., programs with less than 20 applicants or a computed REU rate less than 2.5%).

Beyond assessing the relationship between regression coefficients seen in Figure 2 and Table 1 of the main text, we evaluate the model using a 5-fold cross validation. Our model exhibits an F1 score of 0.50 ± 0.01, illustrating the difficulty of the classification task. The confusion matrix for one of these folds had 1,062 true positives, 1,179 false positives, 7,453 true negatives, and 883 false negatives.

S4.1 Visualizing the applicant-provided recommender-provided tradeoff

In Figure 2A, we visualize the modeled admission probabilities, using groupings of applicant-provided and recommender-provided features. To do this, we first take each feature and compute that feature's within term, within program, percentile. For example, if an applicant has a 4.0 GPA, they will score in the 100th percentile for GPA. We take the arithmetic mean of these percentiles for the two feature sets, which determines the application's location on the plot. These percentiles are binned into 0.02 × 0.02 boxes. Using the regression model, each application gets assigned an admissions probability, which is classified into 4 probabilities classes determined by quartiles. These quartiles are $0 < P < 0.05$ (pink), $0.05 < P < 0.12$ (yellow), $0.12 < P < 0.25$ (blue), $0.25 < P < 1$ (green). Each bin is colored according to the quartile that contains the largest number of applications within that bin. The opacity of each bin is normalized to the bin with the largest number of applications. Therefore, the most opaque bin contains the most applications and the most transparent bin contains the least applications. In Figures **S2-S4**, we show that the trend visualized is robust to different weightings of features. We allow for a random ±10% perturbation of each feature weight around the prior equal weighting scheme.

To fit a smoothed decision boundary across a discrete jagged transition, we first apply a Gaussian filter which smears the boundaries into a continuous transition. Scanning across the horizontal direction, transitions between the first quartile (0) and the second quartile (1) are determined by vertical locations where the smeared boundary is halfway (0.5) between each quartile. If the location doesn't fall exactly on a single bin, we use a linear interpolation to determine the crossing point. If multiple transition locations exist, the median vertical location is taken. Using this series of boundary points, we fit a univariate spline of degree 3 with a smoothing factor of 0.3 × the number of boundary points. These splines are plotted, three in total, truncating the spline at the start of bins with less than or equal to 25% of the maximum number of applications found in a single bin.

In Figure 2B, we take the subset of applications that were admitted from each quadrant in Figure 2A, and sample faculty evaluations from comments containing the most frequent textual bigrams and trigrams. These comments are shortened, only displaying the portion of the comment relevant to the theme of the bigram or trigram. Some identifying or extraneous information is summarized in brackets to preserve anonymity and improve readability.

S5 Field-weighted *h*-index as a proxy recommender prestige

Many argue that *h*-index is a poor indicator for measuring the productivity of a researcher, mainly because it does not account for field differences in citation numbers (e.g., medicine versus math). Though our use of *h*-index is primarily used for establishing the prestige of a recommender, we show that across all bibliometric linked recommendations shown in Figure 3 of the main text, the field-weighted *h*-index of a recommender is associated with higher probabilities of admission for every summed recommendation score (1.137, P = 0.015). We compute this field-weighted h-index as a percentile relative to other h-indices gathered, weighted by the number of citations that author has in different subject areas. This can be interpreted as the expected field-specific percentile of an author's h-index, when fields are sampled according to the author's own publication distribution by field. Mathematically,

$$FWH_i = \mathbb{E}_{f\sim w_i}\left[F_f(h_i)\right]$$

where, $F_f(h_i)$ is the cumulative distribution function (CDF) of the h-index within field f . The expected value of this would be weighting the CDF by the publication weight of field f for author i as $w_{if} = \dfrac{n_{if}}{n_i}$, where n_i is the total number of publications and n_{if} is the total number of publications in that field. Figure **S5** shows this relationship with field-weighted h-index percentiles equivalent to the h-indices shown in Figure 3 of the main text.

S6 Mediation analysis

Each applicant- and recommender- provided feature is placed in a mediation between each socioeconomic variable and whether the applicant is admitted. Specifically, the outcome Y is admission, the exposure X is a socioeconomic variable, and the mediator M is an application feature. The software used to perform this mediation is the package *mediation* in R[15].

Before performing the mediation, we make a few transformations to the socioeconomic status and application features for compatibility. The maximum parental education status of an applicant is converted into an integer—1 for partial high school, 2 for high school, 3 for associates degree, 4 for bachelor's degree, and 5 for graduate or professional degree. The applicant's level of reported financial hardship is also converted to an integer with 1 representing the least amount of hardship and 5 representing the most amount of hardship. Home equity and prior Pell Grant recipient status—shown in Table 2 of the main text—are left unaltered.

The mediation effect sizes reported in Figure 4D of the main text are estimates of the average causal mediation effect (ACME). Inputs into this mediation are the mediator model, outcome model, control value, treatment value T, and the number of simulations (sims = 5000) for uncertainty estimates. The mediator model is specified as a logistic regression for binary mediators or a linear regression for continuous mediators. That is,

$$P(M = T \mid X) = \text{logit}^{-1}(\alpha_0 + \alpha_1 X)$$

or for continuous mediators,

$$P(M = T \mid X) = \alpha_0 + \alpha_1 X$$

The outcome model is always modeled with a logistic regression including X, M, and their interaction,

$$P(Y = 1 \mid X, M) = \text{logit}^{-1}(\beta_0 + \beta_1 X + \beta_2 M + \beta_3 (X \times M))$$

The treatment and control values are defined as follows. For home equity, they correspond to the 25th and 75th percentiles of the sample, approximately \$70,000 and \$370,000, respectively. For self-reported financial hardship, the contrast is a one-unit difference between the lowest (1) and second lowest (2) hardship categories. For parental education, treatment contrasts applicants whose parents hold a bachelor's degree with those whose parents hold a graduate or professional degree. For prior Pell Grant status, treatment is defined as receipt of Pell Grant (1) versus non-receipt (0). Full details on these values and the values obtained from the mediation analysis are in Table **S10**.

S6.1 Visualizing mediation effects

Figure 4D of the main text presents the mediation effects calculated from the approach in section **S6**. They are organized into a grid of triangles oriented upwards indicating a positive mediation effect and downwards indicating a negative mediation effect. The size of each triangle represents the mediation effect size with an absolute minimum of 0.012×10^{-3} and absolute maximum of 72.504×10^{-3}. These triangles are color coded based on the significance of the mediation—red being most significant ($P < .001$) and gray being not significant ($P > .05$) with levels of orange ($P < .01$) and yellow ($P < .05$). The opacity of these triangles is altered to draw emphasis to larger mediation effects—effect sizes larger than 0.005 or 0.5% change in admission probability. The mediators are organized vertically according to the maximum significant effect size, with tiebreaking for the summed levels of significance (e.g., nine for three instances $P < .001$). School rank is inverted to indicate that positive associations correspond to smaller (better) ranks and negative associations correspond to larger (worse) ranks.

Figures 4A-C of the main text examines the fitted regressions from select mediations in Figure 4D. Within these subfigures, we report the coefficient of the mediator model, and the coefficients of the outcome model. The mediator is color-coded according to significance level on the mediation effect in Figure 4D.

Figures S1-S5

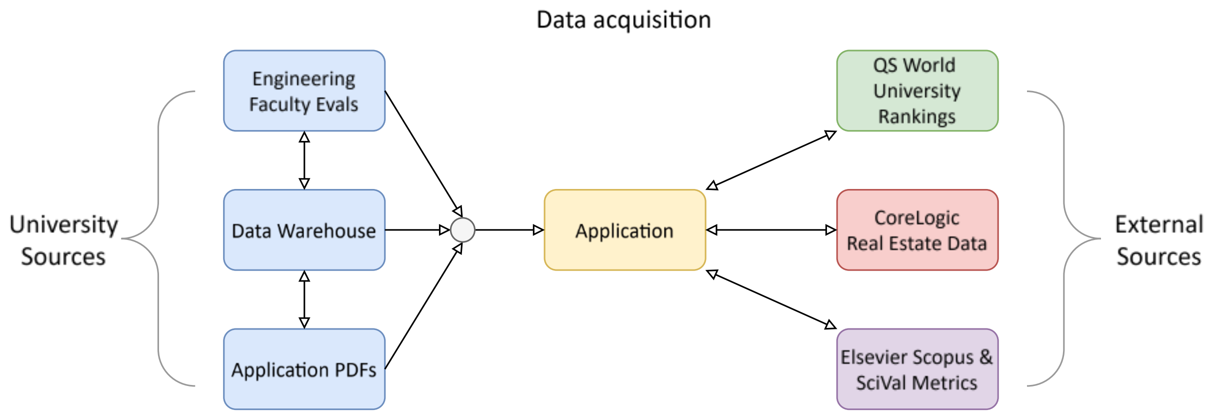

Fig. S1. Diagram illustrating the various data sources combined to construct an application.

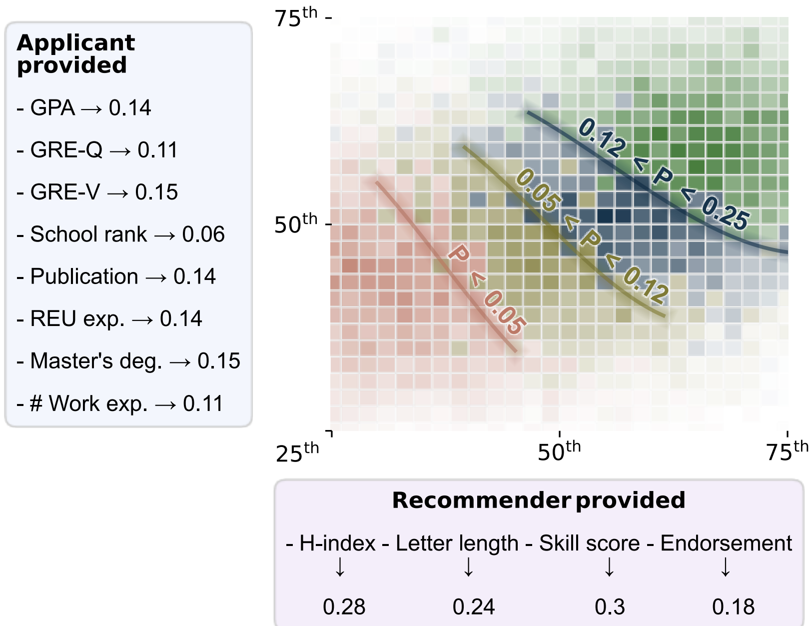

Fig. S2. Perturbation 1 of applicant- and recommender-provided weighting schemes within ±10% of an equal weighting scheme. Perturbed weights for each feature are shown next to or below their attached arrow.

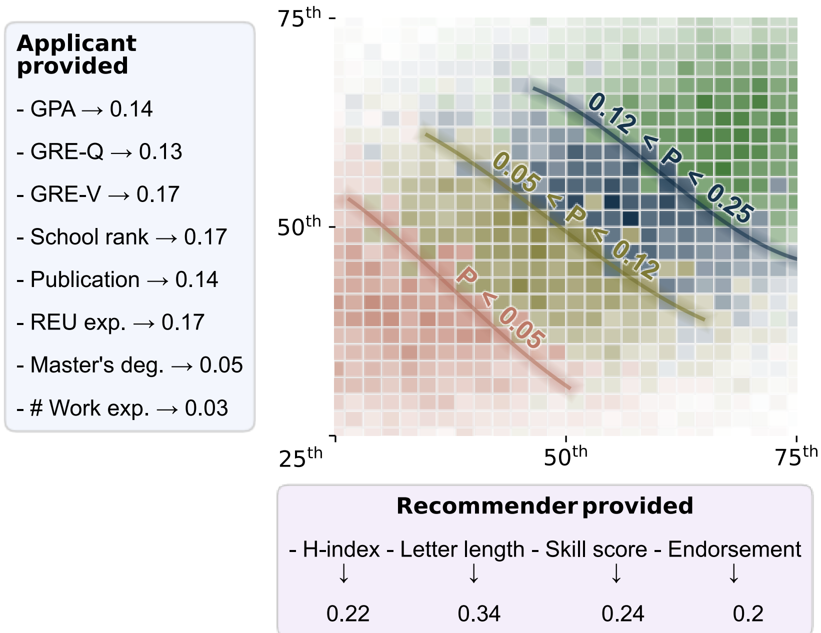

Fig S3. Perturbation 2 of applicant- and recommender-provided weighting schemes within ±10% of an equal weighting scheme. Perturbed weights for each feature are shown next to or below their attached arrow.

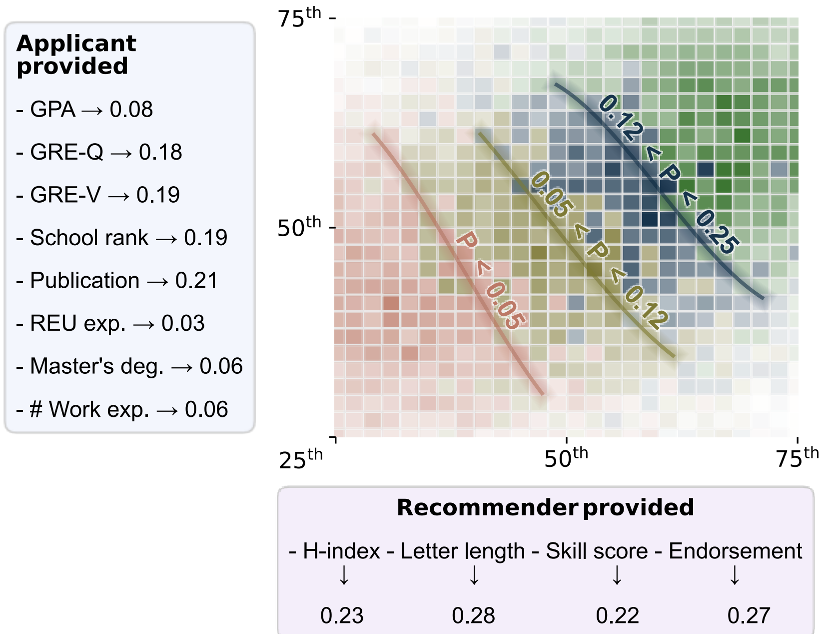

Fig. S4. Perturbation 3 of applicant- and recommender-provided weighting schemes within ±10% of an equal weighting scheme. Perturbed weights for each feature are shown next to or below their attached arrow.

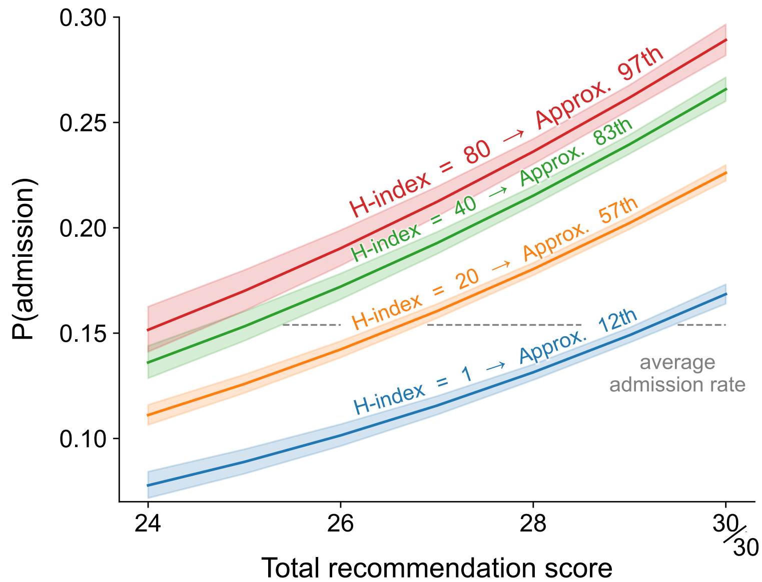

Fig S5. Interaction plots for the modeled probability of admission for levels of total recommendation score, with the second factor being a field-weighted *h*-index percentile. Approximate mappings are created between what that *h*-index might be.

Tables S1 to S10

Physical Element Name	Logical Element Name	Format Type	Element Definition	Examples of Valid Values
gre_quant2_score	Graduate Record Exam Quantitative 2 Score	Number	A number representing the score which a person achieved in the Quantitative component of the Graduate Record Exam. Self reported scores are not included. Note: Data reflects the new scaling implemented by ETS for those tests taken after August 1, 2011. Data for tests taken prior to August 1, 2011 will be blank unless the test was resubmitted.	Example of valid values: Range 130-170; 145; 162; 0
gre_verb2_score	Graduate Record Exam Verbal 2 Score	Number	A number representing the score which a person achieved in the Verbal component of the Graduate Record Exam. Self reported scores are not included. Note: Data reflects the new scaling implemented by ETS for those tests taken after August 1, 2011. Data for tests taken prior to August 1, 2011 will be blank unless the test was resubmitted.	Example of valid values: Range 130-170; 145; 162; 0
undergrad_gpa	Undergraduate Grade Point Average	Number	A number representing an applicant's undergraduate Grade Point Average	Example of valid values: 93.000; 2.860; 3.820
gpa_scale	Grade Point Average Scale	Number	A number representing the scale used in reporting the person's Grade Point Average	Example of valid values: 4; 5
undergrad_inst	Undergraduate Institution Name	String	A string representing the name of the undergraduate institution	Example of valid values: Wayne State University; Western Michigan University
undergrad_state	Undergraduate Institution State Abbreviation	String	A string abbreviation for the state in which the person's undergraduate institution resides in.	Example of valid values: OH; MI: PA
undergrad_country	Undergraduate Institution Country Abbreviation	String	A string abbreviation for the country in which the person's undergraduate institution resides in.	Example of valid values: USA; CHN; IND
EXT_CAREER	External Career Code	String	A string representing the level of study at an external organization	Example of valid values: UGRD; MAST

PRIORWORK_EXPER	Prior Work Experiences	Table	Prior job experiences include title, employer, location and start and end dates	Example of values Contained: PRIORWORK_CITY; PRIORWORK_TITLE_LONG
emplid	Employee Identification Code	String	A system assigned number to uniquely identify a person.	Example of valid values: 25582542
application_id	Application Identification Code	String	A system assigned number to uniquely identify an application.	Example of valid values: 00981506
term	Admit Term Short Description	String	An abbreviated textual description of the code representing the term for which a prospect anticipates applying to a program, or for which an applicant is applying to a program, or for which an admitted/matriculated applicant or a student has been admitted to a particular program. Note: FA = Fall, WN = Winter, SP = Spring, SU = Summer, SPSU = SS = Spring/Summer (combined session).	Example of valid values: WN 1998; SP 1998; FA 1998
field_of_study_code	Field of Study Code	String	A code representing an area of study within an academic program; usually consisting of a field/concentration and degree.	Example of valid values: 7880; 4390; 2760
field_of_study	Field of Study Description	String	A textual description of the field of study code.	Example of valid values: Electrical & Computer Engineering; Industrial & Operations Engin.
program_code	Academic Program Code	String	A code which uniquely identifies the area within the University to which an applicant is admitted and from which a student graduates.	Example of valid values: 00397; 00179; 00051
program_name	Academic Program Name	String	A textual description of the code which uniquely identifies the area within the University to which an applicant applies.	Example of valid values: Chemical Biology Doc; Electrical Engineering Mas
level	Education Level Code	String	A code representing the educational level of the applicant's intended degree	Example of valid values: MA; PHD
first_name	First Name	String	The person's first name	Example of valid values: John; Joe
middle_name	Middle Name	String	The person's middle name	Example of valid values: John; Joe

last_name	Last Name	String	The person’s last name	Example of valid values: Doe, Smith
citizenship	Citizenship Status	String	A textual description of the citizenship status of the person	Example of valid values: Non. Res.; U.S. Citzn; Perm. Res.
flag_pell	Pell Grant Recipient	String	A flag indicating whether a person has received a Pell Grant.	Example of valid values: Y; N
parent1_edu_level	Parent 1 Education Level	String	A textual description code indicating the education level of the parent number 1 of the person.	Example of valid values: Bachelor's degree or equivalent; Graduate or professional degree (ex.: M.A., Ph.D., M.B.A. or J.D.)
parent2_edu_level	Parent 2 Education Level	String	A textual description code indicating the education level of the parent number 2 of the person.	Example of valid values: Bachelor's degree or equivalent; Graduate or professional degree (ex.: M.A., Ph.D., M.B.A. or J.D.)

Table S1. Description of university application attributes.
Physical element name, logical element name, format type, element definition, and examples of valid values. Some details are taken directly from the university’s data warehouse data dictionary.

Physical Element Name	Logical Element Name	Format Type	Element Definition	Examples of Valid Values
num_work_experiences	Number of Work Experiences	Number	A number representing the number of work experience entries for a particular application.	Example of valid values: 5; 3; 0
applied_date	Application Due Date	Date	A date represents the typical due date of the application. Note this is only valid for Fall matriculating applicants and is December 1st of the year prior to the matriculating term.	Example of valid values: December 1st 2022; December 1st 2018
flag_MA	Master's Degree Flag	Number	A number representing whether an applicant has obtained a Master's degree prior to applying for their doctorate. Calculated as the presence of a Master's degree level in a person's prior degree history.	Example of valid values: 0; 1
max_parent_edu_level	Maximum Parental Education Level	String	The maximum parental education level as indicated by the highest education level between parent 1 and parent 2 education levels	Example of valid values: Bachelor's degree or equivalent; Graduate or professional degree (ex.: M.A., Ph.D., M.B.A. or J.D.)
appkey	Application Key	String	A unique application, person, year key defined as the person's "{emplid}_{program_code}_{year}". Used to link applications across different datasets.	Example of valid values: 12345678_00176_2014

Table S2. Derived application features from the data warehouse and other university sources.
Physical element name, logical element name, format type, element definition, and examples of valid values.

Physical Element Name	Logical Element Name	Format Type	Element Definition	Examples of Valid Values
addr1	Address Line 1	String	A string extracted from the Address Line 1 field of an applicant's permanent address information.	Example of valid values: 123 Main St.
addr2	Address Line 2	String	A string extracted from the Address Line 2 field of an applicant's permanent address information.	Example of valid values: APT 5; Unit 10
city	Address City	String	A string extracted from the City field of an applicant's permanent address information	Example of valid values: Ann Arbor; Lansing
state	Address State	String	A string extracted from the State field of an applicant's permanent address information	Example of valid values: Michigan; PA; OH
zip_code	Address Zip Code	String	A string extracted from the Zip Code field of an applicant's permanent address information	Example of valid values: 48105; 15213; 48109-1245
country	Address Country	String	A string extracted from the Country field of an applicant's permanent address information	Example of valid values: US; USA; CHN; UK; United Kingdom
hardship_short	Reported Financial Hardship	String	A string extracted from the response to the question "How much financial hardship did you or your family experience to pay for the total cost of your undergraduate education (tuition, housing books)?"	Example of valid values: 3 = some; 1 = (little/none)
hardship_long	Explanation of Reported Financial Hardship	String	A textual description of an optional explanation of the reported financial hardship response, limited to 500 characters.	Example of valid values: …
rec_first_name	Recommender First Name	String	A string extracted from the First Name field of a recommender form.	Example of valid values: Joe; Jim; Bob
rec_last_name	Recommender Last Name	String	A string extracted from the Last Name field of recommender form.	Example of valid values: Doe; Smith
rec_title	Recommender Title	String	A string extracted from the Title field of the recommender form	Example of valid values: Professor; Assistant Professor; Researcher

rec_institution_or_company	Recommender Institution or Company	String	A string extracted from the Institution or Company field of the recommender form	Example of valid values: University of Michigan; Harvard; Yale
asp	Academic Statement of Purpose	String	The textual contents of the Academic Statement of Purpose – as identified through document segmentation	Example of valid values: …; None
ps	Personal Statement	String	The textual contents of the personal statement – as identified through document segmentation.	Example of valid values: …; None
cv	Curriculum Vitae	String	The textual contents of the curriculum vitae—as identified through document segmentation	Example of valid values: …; None
flag_REU	Applicant Research Experience for Undergraduates Flag	Number	A flag used to indicate whether a student participated in a mentored research experience. Derived from the academic statement of purpose, curriculum vitae, personal statement, and recommendation letters.	Example of valid values: 0; 1
flag_publication	Applicant Publication Flag	Number	A flag used to indicate whether a student has published work. Derived from the academic statement of purpose and recommendation letters.	Example of valid values: 0; 1
mean_hindex	Average Recommender *H*-index	Number	The average *h*-index of an applicant's recommenders.	Example of valid values: 4.5; 10.2; 8.0
mean_skill_score	Average Skill Score	Number	The mean skill score, computed by averaging the median score for each category across recommenders.	Example of valid values: 5.0; 4.2; 4.7
mean_endorsement	Average Endorsement Score	Number	The average endorsement score, computed across recommendations, given to an applicant on an application.	Example of valid values: 4.0; 3.2; 3.7
avg_rec_letter_word_count	Average Recommendation Length Word Count	Number	The average number words from the recommendation letters given to an applicant on an application.	Example of valid values: 633.26; 840.71

Table S3. PDF derived application features.
Application features derived from PDF versions of the application or computed from other data sources. Many of these features are used to query external databases for other relevant features. Physical element name, logical element name, format type, element definition, and examples of valid values are included.

Physical Element Name	Logical Element Name	Format Type	Element Definition	Examples of Valid Values
Eval ID	Evaluation Identification	Number	A unique number assigned to the evaluation.	Example of valid values: 817; 818
Applicant ID	Applicant ID	Number	A system assigned number to uniquely identify an application.	Example of valid values: 12345678
Faculty Uniqname	Faculty Uniqname	String	The uniqname of the faculty or rater evaluating the application.	Example of valid values: jsmith
Dept ID	Department Identification	Number	A code used to represent different departments.	Example of valid values: 222500; 216002
deptName	Department Name	String	A text description of the department	Example of valid values: Robotics; Mechanical Engineering
comments	Faculty Comments	String	Comments left by the rater when reviewing applications	Example of valid values: ….
eval1	Evaluation 1	String/Number	The evaluation option or score for category 1 left by a faculty rater.	Example of valid values: Excellent; 3; Satisfactory
eval2	Evaluation 2	String/Number	The evaluation option or score for category 2 left by a faculty rater.	Example of valid values: Excellent; 3; Satisfactory
eval3	Evaluation 3	String/Number	The evaluation option or score for category 3 left by a faculty rater.	Example of valid values: Excellent; 3; Satisfactory
eval4	Evaluation 4	String/Number	The evaluation option or score for category 4 left by a faculty rater.	Example of valid values: Excellent; 3; Satisfactory
eval5	Evaluation 5	String/Number	The evaluation option or score for category 5 left by a faculty rater.	Example of valid values: Excellent; 3; Satisfactory
eval6	Evaluation 6	String/Number	The evaluation option or score for category 6 left by a faculty rater.	Example of valid values: Excellent; 3; Satisfactory
grade	Overall Evaluation Grade	Number	A number indicating the overall evaluation score given by the rater.	Example of valid values: 1; 2; 3
Eval Title 1	Evaluation Title 1	Number	A number indicating the title or description of evaluation 1.	Example of valid values: Academic Record; Experience

Eval Title 2	Evaluation Title 2	Number	A number indicating the title or description of evaluation 2.	Example of valid values: Academic Record; Experience
Eval Title 3	Evaluation Title 3	Number	A number indicating the title or description of evaluation 3.	Example of valid values: Academic Record; Experience
Eval Title 4	Evaluation Title 4	Number	A number indicating the title or description of evaluation 4.	Example of valid values: Academic Record; Experience
Eval Title 5	Evaluation Title 5	Number	A number indicating the title or description of evaluation 5.	Example of valid values: Academic Record; Experience
Eval Title 6	Evaluation Title 6	Number	A number indicating the title or description of evaluation 6.	Example of valid values: Academic Record; Experience

Table S4. Engineering faculty evaluation system attributes.
Physical element name, logical element name, format type, element definition, and examples of valid values.

Physical Element Name	Logical Element Name	Format Type	Element Definition	Examples of Valid Values
AuthorID	Author Identification	Numeric	A unique identifier assigned to an author.	Example of valid values: 6701858763; 20433296900
HIndices	*H*-index	Numeric	A measure of both the productivity and citation impact of an entity, based on the number of publications as well as the number of citations they have received. The highest number *h* where the author has *h* papers each with a least *h* citations.	Example of valid values: 2.0; 3.0
SubjectAreaCount	Subject Area Count	Array	The subject area abbreviations and the frequency (count) of publications an author has made in the subject area.	Example of valid values: [{ "@abbrev": "SOCI", "@frequency": "5"}, { "@abbrev": "PSYC", "@frequency": "4"}]

Table S5. Bibliometric data for recommenders.
Variables from Elsevier's Scopus[2] and SciVal[3] dated October 2024. Author IDs are found using the Author Search endpoint to then query the SciVal Author Metrics endpoint. Metrics which are "by year" were collected for the 10 years prior from the pull date. Query specifications include "includeSelfCitations": 'true', "includedDocs": "AllPublicationTypes", "journalImpactType": "CiteScore", "showAsFieldWeighted": 'false', "byYear": 'true', "indexType": "hIndex", "includeSelfCitations": 'true', "yearRange": "10yrs"}.

Physical Element Name	Logical Element Name	Format Type	Element Definition	Examples of Valid Values
# World Rank	World Rank Number	Numeric	The number an institution places in the world rankings.	Example of valid values: 1; 2;
Institution	Institution Name	String	The name of the institution.	Example of valid values: University of Michigan—Ann Arbor; Harvard
Country	Country	String	The country the institution resides in.	Example of valid values: USA; France
Year	Year	Numeric	The release year of the world ranking.	Example of valid values: 2015; 2016

Table S6. QS World University Ranking data.
Variables from QS World University Rankings obtained from University Rankings (.ch) (https://www.universityrankings.ch/results/QS/).

Field Number	Field Name	Format Type	Field Definition	Examples of Valid Values
1	CLIP	AlphaNumeric	Unique identification number assigned to each property.	Example of valid values: 1118609200
40	PROPERTY INDICATOR CODE	AlphaNumeric	A CoreLogic general code used to easily recognize specific property types (e.g., Single Family Residence, Condominium, Commercial).	Example of valid values: 10 = SINGLE FAMILY RESIDENCE; 11 = CONDOMINUM
46	BLOCK LEVEL LATITUDE	Numeric	CoreLogic derived geographic coordinate that specifies the north–south position of a point based on United States Postal Service address data for the parcel. The latitude contains numeric data values only conforming to the mask XX.ZZZZZZ, where the XX are the digits to the left of the decimal point, and ZZZZZZ are the digits to the right of the decimal point. Latitude is always a positive value north on the North American continent.	Example of valid values: 26.328367; 26.719467
47	BLOCK LEVEL LONGITUDE	Numeric	CoreLogic derived geographic coordinate that specifies the east–west position of a point based on United States Postal Service address data for the parcel. The longitude contains numeric data values only conforming to the mask -XXX.ZZZZZZ, where the XXX are the numeric digits to the left of the decimal point, and ZZZZZZ are the 6 numeric digits to the right of the decimal point. Longitude is negative on the North American continent.	Example of valid values: -80.133197; -80.188389
46	BLOCK LEVEL LATITUDE	Numeric	CoreLogic proprietary parcel-centroid coordinate that specifies the north–south position of the center point of a parcel. The latitude contains numeric data values only conforming to the mask XX.ZZZZZZ, where the XX are the digits to the left of the decimal point, and ZZZZZZ are the digits to the right of the decimal point. Latitude is always a positive value north on the North American continent.	Example of valid values: 26.328367; 26.719467
47	BLOCK LEVEL LONGITUDE	Numeric	CoreLogic proprietary parcel-centroid coordinate that specifies the east–west position of the center point of a parcel. The longitude contains numeric data values only conforming to the mask -XXX.ZZZZZZ, where the XXX are the numeric digits to the left of the decimal point, and ZZZZZZ are the 6 numeric digits to the right of the decimal point. Longitude is negative on the North American continent.	Example of valid values: -80.133197; -80.188389
50	SITUS HOUSE NUMBER	AlphaNumeric	The digits found to the left of a street name.	Example of valid values: 10; 105; 20A

51	SITUS HOUSE NUMBER SUFFIX	AlphaNumeric	The characters found to the immediate right of a house number (e.g., 202B Jones Rd).	Example of valid values: B; HH; C
52	SITUS HOUSE NUMBER 2	AlphaNumeric	The second set of numbers in an address range.	Example of valid values: 205; 47
53	SITUS DIRECTION	AlphaNumeric	This field represents the direction found to the left of the street name (e.g., 9340 N Dunhill Dr). AKA Directional Abbreviations, Pre-Directional.	Example of valid values: W; N
54	SITUS STREET NAME	AlphaNumeric	The name or number of the street where a parcel is located.	Example of valid values: MAIN; MARKET
55	SITUS MODE	AlphaNumeric	The Mode or Type of street found to the right of the street name (e.g., 9340 N Dunhill Dr). AKA Street Designators / Street Suffixes.	Example of valid values: ST; DR; LN
56	SITUS QUADRANT	AlphaNumeric	The quadrant is found to the right of Situs Mode (e.g., 2040 NW 100 ST SW). AKA Post-Directional.	Example of valid values: NW; SE
57	SITUS UNIT NUMBER	AlphaNumeric	The unit or suite number of the property address (e.g., 649 Lake Shore Dr #1400).	Example of valid values: 27; 102
58	SITUS CITY	AlphaNumeric	The city associated with the property address.	Example of valid values: ANN ARBOR, LANSING
59	SITUS STATE	AlphaNumeric	The two-letter USPS postal abbreviation associated with the state / protectorates / commonwealth (e.g., CA, VI).	Example of valid values AK; TX; PA
60	SITUS ZIP CODE	Numeric	Code assigned by the USPS. This is populated by various source files and other proprietary and non-proprietary processes. Data may be the 5-digit zip or 9-digit Zip+4 (e.g., 00501 or 954630042).	Example of valid values: 00501; 954630042
63	SITUS STREET ADDRESS	AlphaNumeric	Full Situs address (not including City/St/Zip) (e.g. 123 N Main St)	Example of valid values: 123 N MAIN ST; 456 S
76	OWNER 1 FULL NAME	AlphaNumeric	Full name of property owner - 1st occurrence.	Example of valid values: JOHN SMITH W
77	OWNER 1 LAST NAME	AlphaNumeric	Last name of property owner - 1st occurrence.	Example of valid values: SMITH; MATTHEWS
78	OWNER 1 FIRST NAME & MIDDLE INITIAL	AlphaNumeric	First name and middle initial of property owner - 1st occurrence.	Example of valid values: JOHN M; MATT P

79	OWNER 2 FULL NAME	AlphaNumeric	Additional owner name if more than one entity owns the property.	Example of valid values: JOHN SMITH W
80	OWNER 2 LAST NAME	AlphaNumeric	Additional owner name data if more than one entity owns the property.	Example of valid values: SMITH; MATTHEWS
81	OWNER 2 FIRST NAME & MIDDLE INITIAL	AlphaNumeric	Additional owner name data if more than one entity owns the property.	Example of valid values: JOHN M; MATT P
84	OWNER 3 FULL NAME	AlphaNumeric	Additional owner name if more than one entity owns the property.	Example of valid values: JOHN SMITH W
85	OWNER 3 LAST NAME	AlphaNumeric	Additional owner name data if more than one entity owns the property.	Example of valid values: SMITH; MATTHEWS
86	OWNER 3 FIRST NAME & MIDDLE INITIAL	AlphaNumeric	Additional owner name data if more than one entity owns the property.	Example of valid values: JOHN M; MATT P
88	OWNER 4 FULL NAME	AlphaNumeric	Additional owner name if more than one entity owns the property.	Example of valid values: JOHN SMITH W
89	OWNER 4 LAST NAME	AlphaNumeric	Additional owner name data if more than one entity owns the property.	Example of valid values: SMITH; MATTHEWS
90	OWNER 4 FIRST NAME & MIDDLE INITIAL	AlphaNumeric	Additional owner name data if more than one entity owns the property.	Example of valid values: JOHN M; MATT P
95	OWNER OCCUPANCY CODE	AlphaNumeric	CoreLogic derived code that indicates if the property owner resides at the situs (property site). Values of M, O, and S indicate owner occupied (A and T reflect absentee owners).	Example of valid values: M; O; S
149	SALE RECORDING DATE	Numeric	The date the sales transaction was recorded at the county. Format: YYYYMMDD.	Example of valid values: 20000101; 20040501
150	SALE DATE	Numeric	The date the sales transaction was legally completed (i.e., contract signed). Format: YYYYMMDD.	Example of valid values: 20000101; 20040501
151	SALE AMOUNT	Numeric	Dollar amount of the most recent sale of the property.	Example of valid values: 100500; 56000210

Table S7. Basic property data.
Variables from CoreLogic's Property Basic v2 dated November 5th, 2020[5]. Attributes are centered around basic historical information about the parcel.

Field Number	Field Name	Format Type	Field Definition	Examples of Valid Values
1	CLIP	AlphaNumeric	Unique identification number assigned to each property.	Example of valid values: 1118609200
15	PROPERTY INDICATOR CODE - STATIC	AlphaNumeric	A CoreLogic general code used to easily recognize specific property types (e.g. Single Family Residence, Condominium, Commercial). This value was determined from property information upon receipt of the transaction and reflects the Property Indicator as a static value at that point in time.	Example of valid values: 10 = SINGLE FAMILY RESIDENCE; 11 = CONDOMINUM
19	DEED SITUS HOUSE NUMBER - STATIC	AlphaNumeric	The digits found to the left of a street name.	Example of valid values: 8653; 2109; 6201
20	DEED SITUS HOUSE NUMBER SUFFIX - STATIC	AlphaNumeric	The characters found to the immediate right of a house number (e.g., 202B Jones Rd).	Example of valid values: B; HH; C
21	DEED SITUS HOUSE NUMBER 2 - STATIC	AlphaNumeric	The second set of numbers in an address range.	Example of valid values: 205; 47
22	DEED SITUS DIRECTION - STATIC	AlphaNumeric	This field represents the direction found to the left of the street name (e.g., 9340 N Dunhill Dr). AKA Directional Abbreviations, Pre-Directional.	Example of valid values: W; N
23	DEED SITUS STREET NAME - STATIC	AlphaNumeric	The name or number of the street where a parcel is located.	Example of valid values: MAIN; MARKET
24	DEED SITUS MODE - STATIC	AlphaNumeric	The Mode or Type of street found to the right of the street name (e.g., 9340 N Dunhill Dr). AKA Street Designators / Street Suffixes.	Example of valid values: ST; DR; LN
25	DEED SITUS QUADRANT - STATIC	AlphaNumeric	The quadrant is found to the right of Situs Mode (e.g., 2040 NW 100 ST SW). AKA Post-Directional.	Example of valid values: NW; SE
26	DEED SITUS UNIT NUMBER - STATIC	AlphaNumeric	The unit or suite number of the property address (e.g., 649 Lake Shore Dr #1400).	Example of valid values: 27; 102

27	DEED SITUS CITY - STATIC	AlphaNumeric	The city associated with the property address.	Example of valid values: ANN ARBOR, LANSING
28	DEED SITUS STATE - STATIC	AlphaNumeric	The two-letter USPS postal abbreviation associated with the state / protectorates / commonwealth (e.g., CA, VI).	Example of valid values: AK; TX; PA
29	DEED SITUS ZIP CODE - STATIC	AlphaNumeric	Code assigned by the USPS. This is populated by various source files and other proprietary and non-proprietary processes. Data may be the 5-digit zip or 9-digit Zip+4 (e.g., 00501 or 954630042).	Example of valid values: 00501; 954630042
30	DEED SITUS COUNTY - STATIC	AlphaNumeric	Situs county name.	Example of valid values: ANCHORAGE; EL PASO
32	DEED SITUS STREET ADDRESS - STATIC	AlphaNumeric	Full Situs address (not including City/St/Zip) (e.g. 123 N Main St)	Example of valid values: 123 N MAIN ST; 456 S
40	PRIMARY CATEGORY CODE	AlphaNumeric	CoreLogic proprietary transaction classification code that identifies the type of transaction.	Example of valid values: A = ARMS LENGTH; B = NON-ARMS LENGTH/PURCHASE
43	SALE AMOUNT	Numeric	Dollar amount of the most recent sale of the property.	Example of valid values: 100500; 56000210
44	SALE DERIVED DATE	Numeric	CoreLogic proprietary sale date derived from the original public record filing, representing the date that the Deed was prepared in advance of being notarized and prior to being recorded. Format: YYYYMMDD.	Example of valid values: 20000101; 20040501
45	SALE DERIVED RECORDING DATE	Numeric	CoreLogic proprietary sale recording date derived from the original public record filing, representing the date the sales transaction was recorded at the county. This date is generally considered the transfer of ownership date. Format: YYYYMMDD.	Example of valid values: 20000101; 20040501
54	MORTGAGE PURCHASE INDICATOR	Numeric	CoreLogic proprietary value that identifies if the property has a mortgage. 0 = not applicable, 1= applicable.	Example of valid values: 1; 0
63	BUYER 1 FULL NAME	AlphaNumeric	Full name of the BUYER at the time of transfer - 1st occurrence.	Example of valid values: JOHN SMITH W
64	BUYER 1 LAST NAME	AlphaNumeric	Last name of the BUYER - 1st occurrence.	Example of valid values: SMITH; MATTHEWS

65	BUYER 1 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BUYER - 1st occurrence.	Example of valid values: JOHN M; MATT P
66	BUYER 2 FULL NAME	AlphaNumeric	Full name of the BUYER - 2nd occurrence.	Example of valid values: JOHN SMITH W
67	BUYER 2 LAST NAME	AlphaNumeric	Last name of the BUYER - 2nd occurrence.	Example of valid values: SMITH; MATTHEWS
68	BUYER 2 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BUYER - 2nd occurrence.	Example of valid values: JOHN M; MATT P
71	BUYER 3 FULL NAME	AlphaNumeric	Full name of the BUYER - 3rd occurrence.	Example of valid values: JOHN SMITH W
72	BUYER 3 LAST NAME	AlphaNumeric	Last name of the BUYER - 3rd occurrence.	Example of valid values: SMITH; MATTHEWS
73	BUYER 3 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BUYER - 3rd occurrence.	Example of valid values: JOHN M; MATT P
75	BUYER 4 FULL NAME	AlphaNumeric	Full name of the BUYER - 4th occurrence.	Example of valid values: JOHN SMITH W
76	BUYER 4 LAST NAME	AlphaNumeric	Last name of the BUYER - 4th occurrence.	Example of valid values: SMITH; MATTHEWS
77	BUYER 4 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BUYER - 4th occurrence.	Example of valid values: JOHN M; MATT P
83	BUYER OCCUPANCY CODE	AlphaNumeric	CoreLogic derived code that indicates if the property buyer intends to reside at the situs (property site). Values of M, O, and S indicate buyer occupied (A and T reflect absentee buyers).	Example of valid values: M; O; S

Table S8. Owner transfer property data.
Variables from CoreLogic Owner Transfer v3 dated February 22nd, 2021[6]. Owner transfer data contains specific details about the sale of a parcel.

Field Number	Field Name	Format Type	Field Definition	Examples of Valid Values
1	CLIP	AlphaNumeric	Unique identification number assigned to each property.	Example of valid values: 1118609200
15	PROPERTY INDICATOR CODE - STATIC	AlphaNumeric	A CoreLogic general code used to easily recognize specific property types (e.g. Single Family Residence, Condominium, Commercial). This value was determined from property information upon receipt of the transaction and reflects the Property Indicator as a static value at that point in time.	Example of valid values: 10 = SINGLE FAMILY RESIDENCE; 11 = CONDOMINUM
19	DEED SITUS HOUSE NUMBER - STATIC	AlphaNumeric	The digits found to the left of a street name.	Example of valid values: 8653; 2109; 6201
20	DEED SITUS HOUSE NUMBER SUFFIX - STATIC	AlphaNumeric	The characters found to the immediate right of a house number (e.g., 202B Jones Rd).	Example of valid values: B; HH; C
21	DEED SITUS HOUSE NUMBER 2 - STATIC	AlphaNumeric	The second set of numbers in an address range.	Example of valid values: 205; 47
22	DEED SITUS DIRECTION - STATIC	AlphaNumeric	This field represents the direction found to the left of the street name (e.g., 9340 N Dunhill Dr). AKA Directional Abbreviations, Pre-Directional.	Example of valid values: W; N
23	DEED SITUS STREET NAME - STATIC	AlphaNumeric	The name or number of the street where a parcel is located.	Example of valid values: MAIN; MARKET
24	DEED SITUS MODE - STATIC	AlphaNumeric	The Mode or Type of street found to the right of the street name (e.g., 9340 N Dunhill Dr). AKA Street Designators / Street Suffixes.	Example of valid values: ST; DR; LN
25	DEED SITUS QUADRANT - STATIC	AlphaNumeric	The quadrant is found to the right of Situs Mode (e.g., 2040 NW 100 ST SW). AKA Post-Directional.	Example of valid values: NW; SE
26	DEED SITUS UNIT NUMBER - STATIC	AlphaNumeric	The unit or suite number of the property address (e.g., 649 Lake Shore Dr #1400).	Example of valid values: 27; 102

27	DEED SITUS CITY - STATIC	AlphaNumeric	The city associated with the property address.	Example of valid values: ANN ARBOR, LANSING
28	DEED SITUS STATE - STATIC	AlphaNumeric	The two-letter USPS postal abbreviation associated with the state / protectorates / commonwealth (e.g., CA, VI).	Example of valid values: AK; TX; PA
29	DEED SITUS ZIP CODE - STATIC	AlphaNumeric	Code assigned by the USPS. This is populated by various source files and other proprietary and non-proprietary processes. Data may be the 5-digit zip or 9-digit Zip+4 (e.g., 00501 or 954630042).	Example of valid values: 00501; 954630042
30	DEED SITUS COUNTY - STATIC	AlphaNumeric	Situs county name.	Example of valid values: ANCHORAGE; EL PASO
32	DEED SITUS STREET ADDRESS - STATIC	AlphaNumeric	Full Situs address (not including City/St/Zip) (e.g. 123 N Main St)	Example of valid values: 123 N MAIN ST; 456 S
41	PRIMARY CATEGORY CODE	AlphaNumeric	CoreLogic proprietary transaction classification code that identifies the type of transaction.	Example of valid values: A = ARMS LENGTH; B = NON-ARMS LENGTH/PURCHASE
42	MORTGAGE AMOUNT	Numeric	The amount of the mortgage at the time of the transaction.	Example of valid values: 100500; 50000
43	MORTGAGE DATE	Numeric	Borrower's signature date on mortgage/deed of trust. This represents the date the borrower’s obligation to the lender begins, i.e. loan open date. This date is the same or prior to the recording date. If the day and/or month is missing, CoreLogic populates these values with 00. Format: YYYYMMDD.	Example of valid values: 20000101; 20040501
44	MORTGAGE RECORDING DATE	Numeric	CoreLogic proprietary sale recording date derived from the original public record filing, representing the date the sales transaction was recorded at the county. This date is generally considered the transfer of ownership date. Format: YYYYMMDD.	Example of valid values: 20000101; 20040501
56	MORTGAGE TYPE CODE	AlphaNumeric	CoreLogic standardized mortgage classification for principal mortgages ('P'urchase, 'R'efinance) or subordinate mortgages ('J'unior). Won't be populated on pending record.	Example of valid values: P = Purchase; R = Refinance
81	BORROWER 1 FULL NAME	AlphaNumeric	Full name of the BORROWER at the time of transaction - 1st occurrence.	Example of valid values: JOHN SMITH W

82	BORROWER 1 LAST NAME	AlphaNumeric	Last name of the BORROWER - 1st occurrence.	Example of valid values: SMITH; MATTHEWS
83	BORROWER 1 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BORROWER - 1st occurrence.	Example of valid values: JOHN M; MATT P
84	BORROWER 2 FULL NAME	AlphaNumeric	Full name of the BORROWER - 2nd occurrence.	Example of valid values: JOHN SMITH W
85	BORROWER 2 LAST NAME	AlphaNumeric	Last name of the BORROWER - 2nd occurrence.	Example of valid values: SMITH; MATTHEWS
86	BORROWER 2 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BORROWER - 2nd occurrence.	Example of valid values: JOHN M; MATT P
89	BORROWER 3 FULL NAME	AlphaNumeric	Full name of the BORROWER - 3rd occurrence.	Example of valid values: JOHN SMITH W
90	BORROWER 3 LAST NAME	AlphaNumeric	Last name of the BORROWER - 3rd occurrence.	Example of valid values: SMITH; MATTHEWS
91	BORROWER 3 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BORROWER - 3rd occurrence.	Example of valid values: JOHN M; MATT P
93	BORROWER 4 FULL NAME	AlphaNumeric	Full name of the BORROWER - 4th occurrence.	Example of valid values: JOHN SMITH W
94	BORROWER 4 LAST NAME	AlphaNumeric	Last name of the BORROWER - 4th occurrence.	Example of valid values: SMITH; MATTHEWS
95	BORROWER 4 FIRST NAME AND MIDDLE INITIAL	AlphaNumeric	First name of the BORROWER - 4th occurrence.	Example of valid values: JOHN M; MATT P

Table S9. Property mortgage data.
Variables from CoreLogic Mortgage v3 dated May 1st, 2022[7]. Mortgage data details the loans taken out on parcels typically at the time of purchase.

Exposure	Mediator	Treatment value	Control value	Observations	ACME average	ACME upper CI	ACME lower CI	ACME P-value
Parental education	GRE-Q	5	4	35663	2.964E-02	2.717E-02	3.203E-02	0.000
Parental education	GRE-V	5	4	35663	1.986E-02	1.747E-02	2.229E-02	0.000
Parental education	GPA	5	4	62646	1.582E-02	1.453E-02	1.715E-02	0.000
Parental education	School rank	5	4	39286	5.079E-03	4.194E-03	5.977E-03	0.000
Parental education	Skill score	5	4	61502	5.703E-03	4.786E-03	6.689E-03	0.000
Parental education	*H*-index	5	4	46412	2.865E-03	2.392E-03	3.364E-03	0.000
Parental education	Master's	5	4	63520	1.659E-03	1.285E-03	2.364E-03	0.000
Parental education	Endorsement	5	4	61542	3.671E-03	2.830E-03	4.597E-03	0.000
Parental education	Publication	5	4	63520	7.887E-04	6.514E-04	1.580E-03	0.000
Parental education	Letter length	5	4	60886	2.370E-04	-1.430E-04	6.057E-04	0.227
Parental education	REU exp.	5	4	60773	-4.060E-03	-4.993E-03	-3.416E-03	0.000
Parental education	# Work exp.	5	4	63520	-1.229E-05	-6.322E-05	2.480E-05	0.526
Financial hardship	GRE-Q	5	1	55054	-7.250E-02	-7.674E-02	-6.854E-02	0.000
Financial hardship	GRE-V	5	1	55054	-3.488E-02	-3.785E-02	-3.198E-02	0.000
Financial hardship	GPA	5	1	83379	-5.024E-02	-5.298E-02	-4.747E-02	0.000
Financial hardship	School rank	5	1	51814	-4.648E-03	-6.011E-03	-3.299E-03	0.000
Financial hardship	Skill score	5	1	81924	-1.580E-02	-1.810E-02	-1.370E-02	0.000
Financial hardship	*H*-index	5	1	61283	-1.211E-02	-1.366E-02	-1.060E-02	0.000
Financial hardship	Master's	5	1	84503	-9.776E-03	-1.098E-02	-8.717E-03	0.000
Financial hardship	Endorsement	5	1	81962	-1.557E-02	-1.795E-02	-1.337E-02	0.000
Financial hardship	Publication	5	1	84503	-3.104E-03	-3.797E-03	-2.201E-03	0.000
Financial hardship	Letter length	5	1	80949	-3.090E-03	-4.236E-03	-1.955E-03	0.000
Financial hardship	REU exp.	5	1	80547	2.956E-03	1.932E-03	4.516E-03	0.000
Financial hardship	# Work exp.	5	1	84503	7.586E-04	2.222E-04	1.304E-03	0.010
Home equity	GRE-Q	352338.18	66782.69	7285	1.285E-02	9.954E-03	1.637E-02	0.000
Home equity	GRE-V	352338.18	66782.69	7285	6.210E-03	4.579E-03	8.277E-03	0.000

Home equity	GPA	402695.65	85115.63	12546	4.230E-03	2.704E-03	5.942E-03	0.000
Home equity	School rank	447059.00	102029.38	8589	5.905E-03	4.399E-03	7.728E-03	0.000
Home equity	Skill score	402632.98	84714.41	12335	1.880E-03	3.808E-04	3.367E-03	0.014
Home equity	*H*-index	425650.01	90739.10	9447	3.391E-03	2.331E-03	4.640E-03	0.000
Home equity	Master's	403753.71	84679.76	12679	1.471E-03	6.665E-04	2.895E-03	0.002
Home equity	Endorsement	402670.66	84724.68	12343	-1.225E-04	-2.122E-03	1.726E-03	0.892
Home equity	Publication	403753.71	84679.76	12679	1.813E-03	8.591E-04	3.175E-03	0.000
Home equity	Letter length	405606.66	85813.03	12207	1.709E-03	9.600E-04	2.541E-03	0.000
Home equity	REU exp.	403526.99	85730.00	12231	-7.783E-04	-2.461E-03	9.945E-04	0.382
Home equity	# Work exp.	403753.71	84679.76	12679	-1.280E-04	-3.066E-04	-9.041E-06	0.035

Table S10: Mediation effect values.
Effect values and statistics from the mediation analysis shown in Figure 4D of the main text.

References

1. PyMuPDF. Artifex (2024). (https://pymupdf.readthedocs.io/en/latest/)
2. Elsevier. Scopus. (https://www.elsevier.com/products/scopus)
3. Elsevier. Scival. (https://www.elsevier.com/products/scival)
4. QS World University Rankings: Methodology. *Top Universities* https://www.topuniversities.com/world-university-rankings/methodology (2026).
5. CoreLogic, Inc. Property Basic. (2020).
6. CoreLogic, Inc. Onwer Transfer. (2021).
7. CoreLogic, Inc. Mortgage. (2022).
8. Nominatim. (https://nominatim.org/)
9. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. (2025).
10. Lambie-Hanson, L., Li, W. & Slonkosky, M. Real estate investors and the U.S. housing recovery. *Real Estate Economics* **50**, 1425–1461 (2022).
11. Freddie Mac. Primary Mortgage Market Survey. (2025).
12. FHFA House Price Index® Datasets | FHFA. https://www.fhfa.gov/data/hpi/datasets.
13. U.S. Bureau of Labor Statistics. All items in U.S. city average, all urban consumers, not seasonally adjusted.
14. Wardrop, M., Capretto, T. & Smith, N. Patsy. (2025).
15. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software* **59**, 1–38 (2014).