Demography (2019) 56:391–404
https://doi.org/10.1007/s13524-018-0735-y

A Research Note on the Prevalence of Housing Eviction Among Children Born in U.S. Cities

Ian Lundberg[1] · Louis Donnelly[2]

Published online: 27 November 2018

Abstract

A growing body of research suggests that housing eviction is more common than previously recognized and may play an important role in the reproduction of poverty. The proportion of children affected by housing eviction, however, remains largely unknown. We estimate that one in seven children born in large U.S. cities in 1998–2000 experienced at least one eviction for nonpayment of rent or mortgage between birth and age 15. Rates of eviction were substantial across all cities and demographic groups studied, but children from disadvantaged backgrounds were most likely to experience eviction. Among those born into deep poverty, we estimate that approximately one in four were evicted by age 15. Given prior evidence that forced moves have negative consequences for children, we conclude that the high prevalence and social stratification of housing eviction are sufficient to play an important role in the reproduction of poverty and warrant greater policy attention.

Keywords Eviction · Housing · Material hardship · Poverty · Children

Introduction

Rising rents and stagnant or declining wages have contributed to increasingly unaffordable housing options for many U.S. households, especially poor urban families with children (Desmond 2015). The majority of low-income households now devote more than one-half of their income to housing (Joint Center for Housing Studies of Harvard University 2017),

The replication code is available on the Harvard Dataverse: https://doi.org/10.7910/DVN/BVWFG1.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13524-018-0735-y) contains supplementary material, which is available to authorized users.

✉ Ian Lundberg
ilundberg@princeton.edu

[1] Department of Sociology and Office of Population Research, Princeton University, 227 Wallace Hall, Princeton, NJ 08544, USA

[2] Center for Research on Child Wellbeing and Office of Population Research, Princeton University, 286 Wallace Hall, Princeton, NJ 08544, USA

and housing eviction is a key aspect of America's affordable housing crisis. Drawing on ethnographic evidence, administrative records, and survey data collected in Milwaukee, Desmond (2016:98) claimed, "If incarceration had come to define the lives of men from impoverished black neighborhoods, eviction was shaping the lives of women. Poor black men were locked up. Poor black women were locked out."

Recent studies have bolstered the claim that eviction plays a previously unrecognized role in the reproduction of urban poverty. Forced moves are prevalent among the poorest Milwaukee households (Desmond et al. 2015), and observational studies have suggested that eviction negatively affects mothers' health (Desmond and Kimbro 2015) and neighborhood quality (Desmond and Shollenberger 2015). An emerging consensus indicates that eviction has negative consequences for disadvantaged urban children. What is less clear is the extent to which housing eviction is a common and socially stratified phenomenon for children across U.S. cities. This study seeks to answer these questions.

Table 1 summarizes prior estimates of eviction prevalence, outlining the data source, target population, period covered, and likely biases. In general, eviction research has relied on one of two sources: administrative records or survey data. A major limitation of administrative records is that non-court-ordered or "informal" evictions, frequently the cause of involuntary displacement among low-income renters (Desmond and Shollenberger 2015), are excluded from the estimand. Research using survey data can overcome this limitation but introduces other forms of potential bias, such as misreporting, and often relies on household or telephone sampling frames that exclude those most likely to have experienced residential instability, possibly as a result of eviction.

The most comprehensive analysis of administrative records, conducted by the Eviction Lab at Princeton University, indicated that between 2000 and 2016, the proportion of U.S. rental households experiencing a court-ordered eviction in a calendar year ranged from 2.3 % in 2016 to 3.1 % in 2006 (Desmond et al. 2018). Rates are substantially higher in large U.S. cities, ranging from 2.9 % to 16.5 % (Desmond et al. 2018). Other population estimates that relied on cross-sectional survey data spanned short periods and gave the impression that eviction is a more rare phenomenon. Analyses of the Survey of Income and Program Participation (SIPP), for example, indicated that only 0.3 % of U.S. households experienced eviction in 1998 (Bauman 2003:8). Even in the Detroit metropolitan area during 12 months of the Great Recession, only 2.4 % of households reported experiencing eviction (Gould-Werth and Seefeldt 2012).

We advocate a new focus of study: the proportion of children who were ever evicted during childhood. This estimand is important for those concerned with the well-being of children and the intergenerational transmission of poverty. Prior estimates of eviction prevalence, typically the proportion of households ever evicted over a given year, are likely to be much lower than the proportion of children ever evicted during childhood for two reasons. First, households with children are evicted at higher rates than those without children (Desmond et al. 2013). Second, if a nonzero risk of eviction is typical among disadvantaged families, then a large proportion of children may experience eviction at some point during childhood even if only a small proportion are evicted in any given year.[1] Prior studies have focused on the proportion evicted in a given year, but none have

[1] The misleading nature of annual reports has long been known in poverty research. Duncan and Rodgers (1988) demonstrated that approximately one-third of children under age 4 in 1968 were ever poor by age 15, despite annual estimates representing 1967, 1973, and 1979 showing that the proportion of families with children below the poverty line was only 11 % to 13 % in any given year (Danziger and Gottschalk 1985).

Table 1 Previous estimates of eviction prevalence

	Administrative Records	Survey Data				
Study	Desmond et al. (2018)	Desmond and Shollenberger (2015)	Gould-Werth and Seefeldt (2012)	Bauman (2003)	Mayer and Jencks (1989)	Present study
Data	Court records	Milwaukee Area Renters Study	Michigan Recession and Recovery Study	Survey of Income and Program Participation	Telephone surveys of Chicago residents	Fragile Families Study
Period Covered	2016 calendar year[a]	Two years preceding interviews conducted from 2009 to 2011	12 months preceding interview conducted between October 2009 and April 2010	12 months preceding interview in 1998	12 months preceding interview in 1983 or 1985	Birth in 1998–2000 to age 15
Target Population	Renter-occupied U.S. households	Private rental households in Milwaukee County	Households in Detroit Metropolitan Area	U.S. households	Non-Hispanic black and white Chicago households	Births in hospitals in U.S. cities with populations over 200,000
Limitations to Estimand (theoretical target quantity)	Omits informal evictions	Household sampling frame excludes evictees who become unhoused or move out of rental market.	Household sampling frame excludes evictees who become unhoused or move out of area.	Household sampling frame excludes evictees who become unhoused.	Telephone sampling frame excludes evictees without phone service or who move out of Chicago.	Prospective design avoids common limitations.
Expected Bias of Estimator	Unbiased for court-ordered evictions	Unbiased	Downwardly biased due to underreporting of eviction	Downwardly biased due to underreporting of eviction	Downwardly biased due to underreporting of eviction	Downwardly biased due to underreporting of eviction
Estimate	2.3 %	3 % formal eviction 6 % informal eviction 4 % other forced move[b]	2.4 %	0.3 %	0.9 %	14.8 %

[a] We highlight the most recent estimate by Desmond et al. (2018), but estimates are available for 2000–2016 and range from a low of 2.3 % in 2016 to a high of 3.1 % in 2006.

[b] The authors considered formal evictions, informal evictions, landlord foreclosure, and building condemnations to be forced moves.

estimated the cumulative probability that a child is ever evicted over the span of childhood. Using panel data on a probability sample, we estimate the proportion of U.S. children born in large cities between 1998 and 2000 who were ever evicted from their home (for nonpayment of rent or mortgage) by age 15 (a parameter we call τ).

In addition to its substantive contribution, this study presents one approach to a persistent problem of missing data common to demographic research about the proportion of people to ever experience an event over the life course (e.g., Amorim et al. 2017; Duncan and Rodgers 1988; Wildeman 2009). Those seeking to estimate such a quantity face two fundamental challenges: (1) survey nonresponse and (2) periods of interest about which no respondents were surveyed. We propose a combination of multiple imputation to address the former and model-based interpolation to address the latter. To translate limited data that do not cover the entire period of interest into transparent estimates with reasonable precision, we assume a parametric model. In other settings with more plentiful data that more completely cover the period of interest, one might prefer a nonparametric machine learning approach; we assess the robustness of our results to one such approach in the online appendix, part 5. Given that no approach yields precise, assumption-free inference when data are incomplete, we state our assumption that missingness is independent of eviction given observed covariates, speculate about likely violations, and place a lower bound on the estimand in the event that assumptions are violated. The approaches used in this study may be useful to demographers who aim to estimate the life course prevalence of a phenomenon with incomplete longitudinal data.

Data

The Fragile Families and Child Wellbeing Study (hereafter, the Fragile Families Study) is a population-based birth cohort study of 4,898 children born in 20 large U.S. cities between 1998 and 2000. To our knowledge, the Fragile Families Study is the only national panel study to record housing eviction at all follow-up survey waves covering a period from birth through adolescence. Earlier waves of the study have been used to study the consequences of eviction (Desmond and Kimbro 2015). The study includes a stratified, multistage probability sample that oversamples children born to unmarried parents (approximately 3 to 1), resulting in a disproportionately large number of children from low-income families. Because low-income families with children are among those most at risk for eviction (Desmond et al. 2013), the oversample enables greater precision in the estimation of eviction prevalence.

Weighted estimators on the subsample of children born in the 16 probabilistically selected sample cities are unbiased for the true population parameters for all births in 1998–2000 in the 77 U.S. cities with populations greater than 200,000. Weighting is important because of the unequal sample selection probabilities; although 76 % of the unweighted sample of children were born to unmarried parents, weighted estimates suggest that only 40 % of births in the sampling frame were to unmarried parents. The sampling frame is children *born in* large U.S. cities; the target population is births in city hospitals, not births to city residents.[2] For a comparison between our sample and vital records on

[2] Based on address records, we estimate that approximately 20 % of sampled families in the Fragile Families Study resided outside the sample city's municipal boundaries at the time of the focal child's birth.

births to residents of large cities, see the online appendix, part 7. For further details on the sampling design, see Reichman et al. (2001). An advantage of the Fragile Families Study design is that the sampling frame is defined prior to eviction, and attempts are made to follow respondents even if they become unhoused or leave the metropolitan area. This prospective design avoids a common problem with retrospective designs that restrict the sample by potential consequences of eviction, such as whether one lives in a rental unit.

The key outcome variable is based on a question from the SIPP material hardship scale: "In the past twelve months, were you ever evicted from your home or apartment for not paying the rent or mortgage?"[3] (originally adapted from Mayer and Jencks 1989). Parents[4] answered this question when children were approximately ages 1, 3, 5, 9, and 15. Figure 1 presents the prevalence of housing eviction in each report. In most years, approximately 1.5 % of children born in large cities were evicted from their homes, but the prevalence spiked to 3.0 % during the Great Recession, suggesting a role for macroeconomic factors. Unweighted estimates and estimates without multiple imputation are provided in the online appendix, Table A1.

Overall Prevalence Estimates

What proportion of children born in large cities at the turn of the twenty-first century were *ever* evicted[5] between birth and age 15? The answer to this question is not obvious because some periods have no data available. Figure 2 presents this challenge graphically and summarizes our three approaches to yield a cumulative estimate (for details, see Table A2 in the online appendix).

Absolute Lower Bound: Observed Evictions

Before introducing more complicated approaches, we begin with a simple assumption: no evictions occurred in years with missing responses or when no data were collected. Because of missing data and attrition, annual reports cover only four of five years for the average child. A long-term recall report covers five additional years (ages 9 to 14) for 74 % of the sample. Using this strategy, which is a lower bound[6] given that some

[3] We estimate that only 14.9 % (unweighted) of parents who reported an eviction lived in an owned home at the previous wave, suggesting that the large majority of evictions (85.1 %) were due to nonpayment of rent. This estimate excludes 26.2 % of eviction reports for which housing status in the prior wave is not known.

[4] For children who live at least one-half of the time with their mothers, we use the mother's report. Otherwise, for those who live at least one-half the time with their father, we use the father's report. All others are missing. At age 15, eviction is reported for all children by their primary caregiver.

[5] The proportion of children who have experienced multiple evictions is also an important research question. An absolute lower bound on this proportion is the 1.1 % (CI: 0.1–1.4 %) of the sample that reported eviction in two or more waves. Multiply imputing to fill missing survey reports, we estimate that 1.8 % (CI: 1.0–2.7 %) were evicted in multiple waves. Our parametric model implies that 4.9 % (CI: 3.9–6.1 %) of children born in large U.S. cities in 1998–2000 were evicted in multiple years between birth and age 15. However, these estimates may substantially understate the prevalence of serial evictions because questions asked whether respondents were ever evicted in each reporting period, not the number of evictions. Because serial evictions may occur in quick succession within a single reporting period, we leave more definitive claims about the prevalence of serial eviction to future research.

[6] The point estimate is a lower bound on the weighted sample proportion evicted. The bootstrapped 95 % confidence interval captures estimation uncertainty about the population proportion.

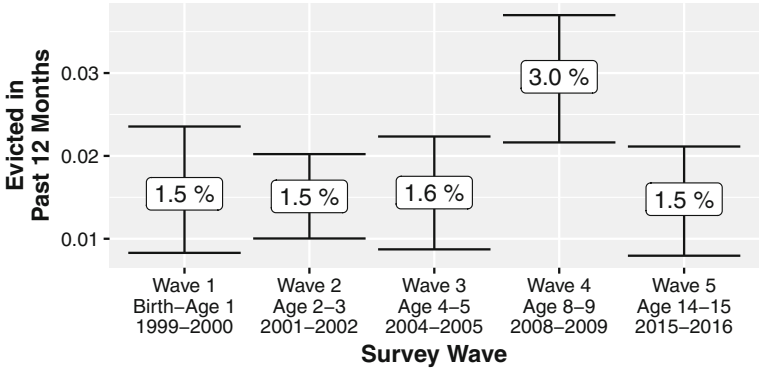


Fig. 1 Evicted in the past 12 months: Five cross-sectional reports. N = 3,442. Estimates represent the probability that a child born in 1998–2000 in a U.S. city with population over 200,000 was evicted in the 12 months preceding each interview wave. Missing data are multiply imputed (11 % to 29 % of observations). Error bars represent 95 % bootstrapped confidence intervals. Table A1 in the online appendix summarizes attrition across waves and how estimates change with multiple imputation and weighting. In general, unweighted estimates are slightly higher because the Fragile Families Study oversampled children born to unmarried parents, who are more likely to be evicted.

evictions likely occurred in the unobserved years, we find that 7.9 % (CI:[7] 7.1–8.9 %) of children born in large U.S. cities in 1998–2000 experienced an eviction by age 15.

Adjusting for Nonresponse: Multiple Imputation

The lower bound estimate understates eviction prevalence because it ignores evictions that are not reported as a result of (1) survey nonresponse (see Table A1 in the online appendix for response rates) and (2) periods in which the study did not ask about eviction (5 of 15 years). To address the first limitation, we determine whether each child with complete data (57 %) was ever evicted given all available reports and then multiply impute missing reports for the remaining children (see the online appendix, part 2, for details). Assuming that eviction is missing at random conditional on covariates, this approach captures the prevalence of eviction over 10 years: the four years preceding ages 1, 3, 5, and 9, and the six years between ages 9 and 15. Using this strategy, we estimate that 9.2 % (CI: 7.3–11.1 %) of children born in large U.S. cities were evicted during this period.

Preferred Estimate: Multilevel Logistic Regression

As shown in Fig. 2, there were five years in which the study did not ask about eviction (between ages 1–2, 3–4, and 5–8). Both our lower bound and multiple imputation estimators ignore evictions in these periods.

To estimate eviction prevalence over the entire period from birth to age 15, we apply a parametric[8] model: multilevel logistic regression with random intercepts to capture

[7] Throughout, *CI* refers to a 95 % quantile-based bootstrapped confidence interval for frequentist estimates or a 95 % quantile-based posterior credible interval for Bayesian estimates.

[8] Because our primary goal is prediction, we also considered random forests as a nonparametric alternative. For details, see the online appendix, part 5.

unobserved heterogeneity constant within cities of birth or within children.[9] We use bold characters topped by arrows to indicate vectors.

Random intercepts for each city of birth c:

$$\{\delta_c\} \sim^{\text{iid}} \text{Normal}\left(0, \sigma_\delta^2\right).$$

Random intercepts for each child i in each city c:

$$\left\{\epsilon_{c[i]}\right\} \sim^{\text{iid}} \text{Normal}\left(0, \sigma_\epsilon^2\right).$$

Linear predictor:

$$\eta_{c[i[t]]} = \alpha + \underbrace{\vec{\mathbf{X}}_{c[i]}\vec{\boldsymbol{\beta}}}_{\substack{\text{Child-level}\\ \text{predictors}}} + \underbrace{Age_{c[i[t]]}\gamma + Recession_{c[i[t]]}\lambda}_{\text{Time-varying predictors}} + \underbrace{\delta_c + \epsilon_{c[i]}}_{\substack{\text{Random}\\ \text{intercepts}}}.$$

Link function:

$$\underbrace{\pi_{c[i[t]]}}_{\substack{\text{Probability}\left(\text{Eviction in year } t \text{ for}\\ \text{child } i \text{ born in city } c\right)}} = \text{logit}^{-1}\left(\eta_{c[i[t]]}\right).$$

Stochastic component:

$$\underbrace{Y_{c[i[t]]}}_{\substack{\text{Eviction in year } t \text{ for}\\ \text{child } i \text{ born in city } c}} \sim \text{Bernoulli}\left(\pi_{c[i[t]]}\right).$$

We fit the model on data from the five annual reports. We exclude the six-year retrospective report from the age 15 survey out of concern that the long recall period may understate eviction prevalence. Child-level predictors $\vec{\mathbf{X}}_{c[i]}$ include race, mother's characteristics, family income relative to needs, type of housing, housing costs, neighborhood context, and city of birth. Time-varying predictors include child age and an indicator for observations in the Great Recession (2008–2009). Table 2 summarizes these variables. We selected these predictors with the goal of choosing variables likely to predict eviction while maintaining a parsimonious model to produce estimates with a reasonable degree of precision. If data were more plentiful, one could use a data-driven approach to select the most relevant predictors, but the limited data in our setting yielded such an approach of limited use (see Fig. A2 in the online appendix). We instead chose the variables based on

[9] Logistic regression estimates are consistent but biased in finite samples, often underpredicting rare events, such as eviction (King and Zeng 2001). Cross-validation suggests that the bias is small in our case (see the online appendix, part 4).

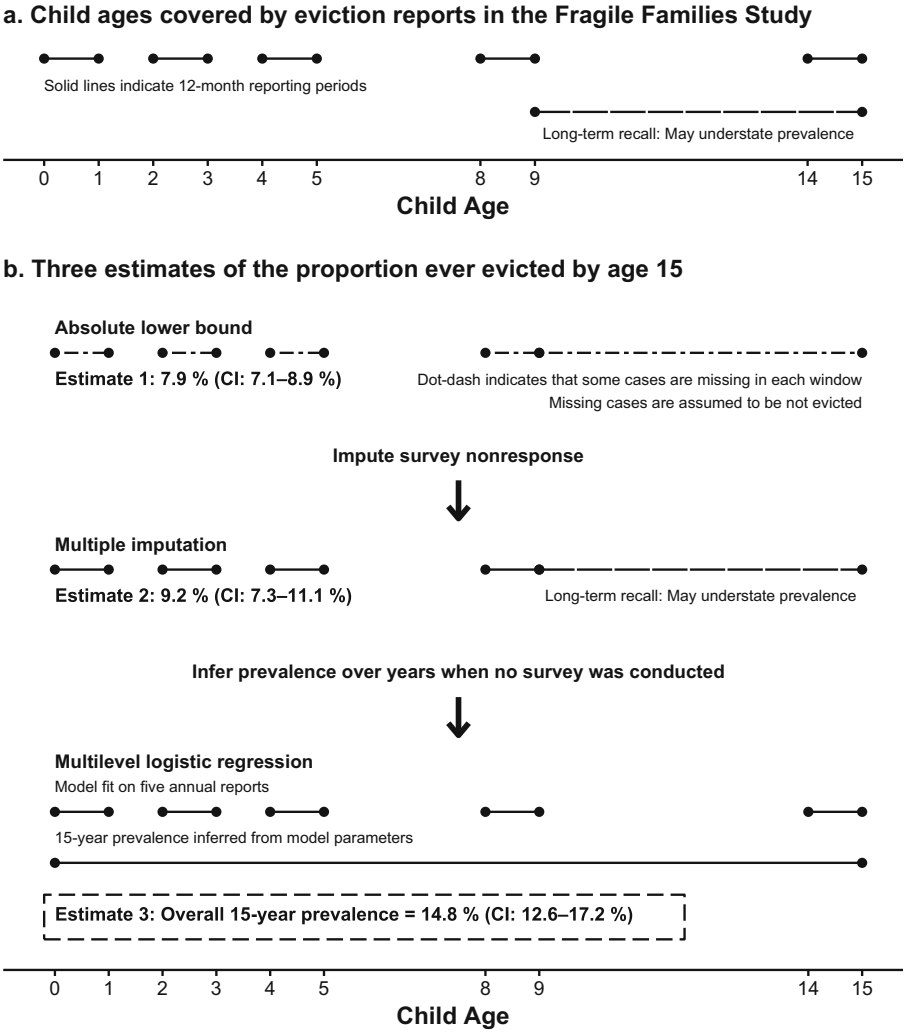


Fig. 2 Ever evicted by age 15: Data and estimation approaches. Estimates represent the probability that a child born in 1998–2000 in a U.S. city with population over 200,000 was ever evicted between birth and age 15. All estimates are weighted. Child ages are approximate because all children were not interviewed at precisely age 1, 3, 5, 9, and 15. Table A2 in the online appendix provides more details about the assumptions that produce each estimate.

theory, noting that the final conclusions are reasonably robust to the set of variables chosen (online appendix, Fig. A1).

To facilitate construction of uncertainty estimates, we adopt a Bayesian framework with Cauchy priors on α, $\vec{\beta}$, γ, and λ, and half-Cauchy priors on σ_u and σ_v. The Cauchy distribution is weakly informative; the prior density is greatest near 0 but has heavier tails than the normal distribution, thereby allowing the possibility of large parameter values (Gelman et al. 2008). We sample from the posterior by Hamiltonian Monte Carlo implemented in Stan using the R package *rstan* (Carpenter et al. 2017;

Stan Development Team 2017). Variable specifications, modeling details, and coefficient estimates are provided in the online appendix, parts 1–3.

To translate the model output into an estimate of eviction prevalence, we first calculate the predicted probability of eviction for each child i at every age t from birth to age 15. In the following notation, characters with hats indicate a single draw from the posterior distribution.

$$\hat{\eta}_{c[i[t]]} = \hat{\alpha} + \overrightarrow{X}_{c[i]}\widehat{\overrightarrow{\beta}} + Age_{c[i[t]]}\hat{\gamma} + Recession_{c[i[t]]}\hat{\lambda} + \hat{u}_c + \hat{v}_{c[i]}.$$
$$\hat{\pi}_{c[i[t]]} = \text{logit}^{-1}\left(\hat{\eta}_{c[i[t]]}\right).$$

Because we assume conditional independence of eviction across years given the probability of eviction $\pi_{c[i[t]]}$, we can collapse the age-specific probability estimates to a single cumulative probability estimate $\hat{\phi}_{c[i]}$ of any eviction between birth and age 15 for each child i born in city c.

$$\hat{P}\left([Ever\ evicted]_{c[i]}\right) = \hat{\phi}_{c[i]} = 1 - \prod_{t=1}^{15}\left(1 - \hat{\pi}_{c[i[t]]}\right).$$

A weighted average of the child-specific estimates $\hat{\phi}_{c[i]}$ over the sample yields an estimate of our target parameter τ: the population prevalence of any eviction from birth to age 15.

$$\text{Overall prevalence estimate} = \hat{\tau} = \frac{\sum_{c=1}^{16}\sum_{i=1}^{n_c} w_{c[i]}\hat{\phi}_{c[i]}}{\sum_{c=1}^{16}\sum_{i=1}^{n_c} w_{c[i]}}.$$

By repeating this process for each posterior draw of the parameters, we obtain 10,020 samples of $\hat{\tau}$ from its posterior distribution. The posterior mean indicates that 14.8 % (CI: 12.6–17.2 %) of U.S. children born in large cities in 1998–2000 were evicted by age 15.

Eviction Is Socially Stratified

Does eviction exacerbate existing inequalities? Prior work suggests that eviction has negative effects on children and families (Desmond and Kimbro 2015). This section demonstrates that children who are already disadvantaged are more likely to be evicted. To estimate subgroup-specific prevalence, we estimate the weighted average of the posterior cumulative probabilities ($\hat{\phi}_{c[i]}$) over all children in the subgroup. Figure 3 presents subgroup estimates by race/ethnicity and by family income at birth.[10] Eviction

[10] We also examined how eviction varied by city of birth. Prevalence is substantial in all sampled cities, but we found suggestive evidence of some variation from 12.1 % (born in Chicago, CI: 7.0–18.1 %) to 23.5 % (born in Detroit, CI: 18.9–28.7 %). Because estimates are imprecise and because sample cities indicate children's city of birth, not city of eviction, we treat this evidence as suggestive and call for future research on geographic heterogeneity in eviction prevalence. For details, see the online appendix, part 6.

Table 2 Descriptive statistics for predictors of eviction

	Weighted		Unweighted		
	Mean	SD	Mean	SD	Proportion Missing[a]
Household Characteristics					
Parents married at birth	0.24		0.60		
Permanent (income/poverty line)					
Below 50 %	0.06		0.10		.11–.30
50 % to 100 %	0.15		0.22		
100 % to 200 %	0.29		0.33		
200 % to 300 %	0.20		0.17		
Higher than 300 %	0.30		0.18		
Housing cost/income	0.30	0.18	0.34	0.18	.27–.39
Proportion of years living in an owned home	0.41	0.40	0.26	0.35	.10–.30
Mother's Characteristics					
Race/ethnicity					.00
Black	0.23		0.48		
Hispanic	0.31		0.27		
White/other	0.46		0.25		
Foreign-born	0.25		0.17		.00
Education (child's birth)					
< High school	0.29		0.35		.00
High school	0.30		0.30		
Some college	0.19		0.24		
College	0.22		0.11		
Age at birth	27.10	6.28	25.30	6.05	.00
Impulsivity (child age 3)	0.05	0.99	0.00	1.00	.14
Cognitive skills (WAIS-R, at child age 3)	7.04	2.79	6.70	2.67	.14
Father's Characteristics					
Ever in jail/prison by child age 1	0.20		0.34		.11
Neighborhood Context (census tract characteristics in 1999, residence at birth)					
Racial composition					
Proportion white	0.46	0.34	0.32	0.31	.04
Proportion black	0.25	0.30	0.40	0.37	
Proportion Hispanic	0.24	0.30	0.22	0.26	
Proportion all other	0.06	0.09	0.07	0.10	
Proportion of households below poverty line	0.16	0.15	0.19	0.14	.04
Median rent/household income	0.26	0.05	0.27	0.05	.04

Notes: $N = 3{,}442$ children for weighted estimates. $N = 4{,}898$ children for unweighted estimates. Missing data are imputed. Details on variable specification are provided in the online appendix, part 1.

[a] When a range is given, the estimate is averaged over five survey waves, with missing data imputed at the wave level before aggregation. The range indicates the least and greatest proportion missing for the variable across waves.

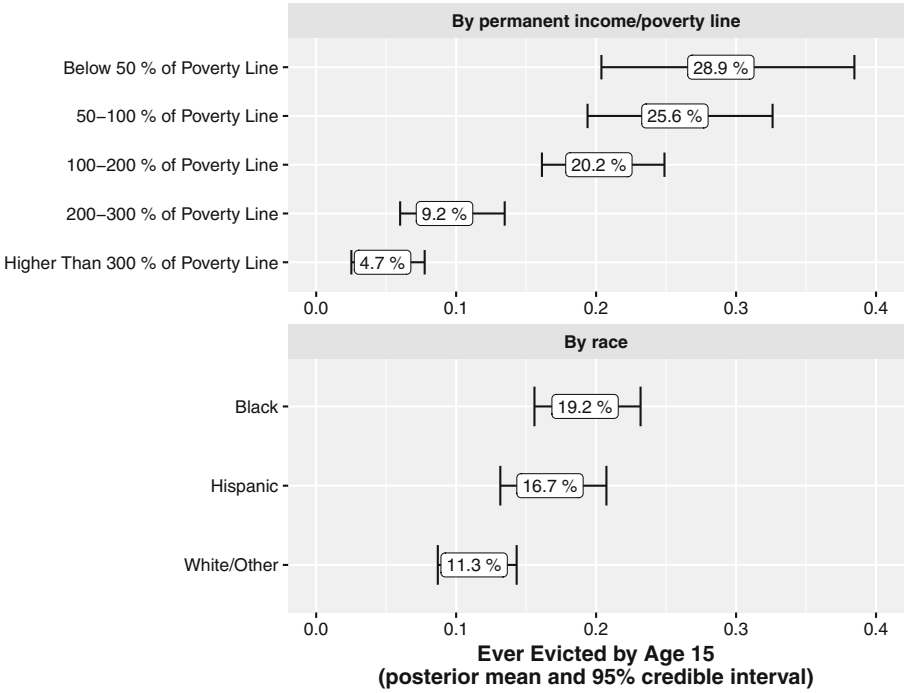


Fig. 3 Ever evicted by age 15, by race and family income**.** Estimates represent the probability that a child born in 1998–2000 in a U.S. city with population over 200,000 was ever evicted between birth and age 15. All estimates are weighted. Estimates are based on predicted probabilities from the multilevel logistic regression model. Error bars represent 95 % credible intervals.

was most common among children whose mothers were black (19.2 %, CI: 15.6–23.2 %) and Hispanic (16.7 %, CI: 13.2–20.7 %), compared with children born to mothers of white and other racial and ethnic backgrounds (11.3 %, CI: 8.7–14.3 %). This is consistent with prior evidence from Milwaukee that eviction is most prevalent among black renters (Desmond 2012). Children's probability of eviction, however, diverges most strikingly by household income. Approximately 28.9 % (CI: 20.4–38.5 %) of children living in deep poverty (permanent income relative to the poverty threshold below 50 %) were evicted. The prevalence of eviction declines monotonically as income rises, yet even the most advantaged children (higher than 300 % of the poverty threshold) faced a 4.7 % (CI: 2.5–7.8 %) probability of eviction.

Limitations

Our estimators may be biased by three noteworthy limitations: selection into the observed sample, survey question wording, and model specification errors. The first two are likely to lead to an underestimate of eviction prevalence, whereas the direction of the bias is less clear in the third case.

First, our estimates are vulnerable to nonignorable selection into the sample. For example, those at the greatest risk of eviction may have been least likely to consent to

the survey at birth, which would downwardly bias estimates of eviction prevalence. A greater potential source of bias is the assumption that attrition is conditionally ignorable. If eviction disrupts respondents' lives and leads to sample attrition, our estimates would be biased downward.

Second, the survey questions reference eviction due to rent or mortgage nonpayment. Persons evicted from their homes for other reasons, or informally without a court order, frequently do not report having been evicted (Desmond 2016). Fielded in 2009–2011, the Milwaukee Area Renters Study included questions about a variety of moves and found that fewer than one-half of forced moves were court-ordered evictions: the majority were informal evictions (no court order), landlord foreclosures, and building condemnations (Desmond and Shollenberger 2015). Given these findings, our data likely understate eviction prevalence and surely understate the prevalence of forced moves.

Finally, to infer eviction prevalence over 15 years using five annual reports, we assumed a parametric model and interpolated the prevalence of eviction in regions of the covariate space with no data (e.g., child age 7). We concluded that this model-based interpolation is reasonable given the relatively steady prevalence of eviction across childhood (with the exception of the Great Recession). In addition, this modeling approach produced estimates suggesting that eviction is alarmingly prevalent across a range of model specifications (see Fig. A1 in the online appendix). Nonetheless, different modeling assumptions could produce estimates as low as our lower bound or higher than our preferred estimate (see the online appendix, part 3).

The results presented in this article reflect our obligation as social scientists to make the best use of the data available. To provide more rigorous estimates in the future, social scientists will need access to better data. In particular, we recommend improvements on two fronts: surveys and administrative records. Cross-sectional and panel surveys of probability samples should incorporate detailed questions about the timing and nature of eviction (or other forced moves), following examples like the Milwaukee Area Renters Study (Desmond and Shollenberger 2015). Likewise, administrative records already show alarming eviction prevalence and represent a promising source for future research (Desmond et al. 2018). Improvements in these two areas would enable researchers to provide new insights into the prevalence and social stratification of eviction.

Conclusion

Drawing on a population-based panel study of a birth cohort, we estimate that 14.8 % (CI: 12.6–17.2 %) of children born between 1998 and 2000 in large U.S. cities were evicted from their homes by age 15. Although eviction is widespread across demographic groups and cities, it is most prevalent among children who already face other disadvantages: black children and children raised in poverty. The high prevalence of eviction and its unequal contours support recent ethnographic evidence suggesting that eviction plays an important role in the reproduction of poverty and spatial inequality.

Acknowledgments We thank Sara S. McLanahan, Brandon M. Stewart, Matthew J. Salganik, three anonymous reviewers, and members of the Stewart Lab and the Fragile Families Working Group for comments on early drafts. We thank the *Demography* copyeditors for helping to improve our prose. All errors are our own. Research reported in this publication was supported by the Robert Wood Johnson Foundation and by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879. Funding for the Fragile Families Study was provided through Award Numbers R01HD36916, R01HD39135, and R01HD40421, and by a consortium of private foundations. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Amorim, M., Dunifon, R., & Pilkauskas, N. (2017). The magnitude and timing of grandparental coresidence during childhood in the United States. *Demographic Research, 37*(article 52), 1695–1706. https://doi.org/10.4054/DemRes.2017.37.52

Bauman, K. J. (2003). *Extended measures of well-being: Living conditions in the United States: 1998* (Current Population Reports: Household Economic Studies Series P70–87). Washington, DC: U.S. Census Bureau.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1). https://doi.org/10.18637/jss.v076.i01

Danziger, S., & Gottschalk, P. (1985). *How have families with children been faring?* (Report prepared for the Joint Economic Committee of the Congress). Madison: University of Wisconsin Institute for Research on Poverty. Retrieved from http://files.eric.ed.gov/fulltext/ED268189.pdf

Desmond, M. (2012). Eviction and the reproduction of urban poverty. *American Journal of Sociology, 118,* 88–133.

Desmond, M. (2015). *Unaffordable America: Poverty, housing, and eviction* (Fast Focus Report No. 22). Madison: University of Wisconsin Institute for Research on Poverty.

Desmond, M. (2016). *Evicted: Poverty and profit in the American city.* New York, NY: Crown.

Desmond, M., An, W., Winkler, R., & Ferriss, T. (2013). Evicting children. *Social Forces, 92,* 303–327.

Desmond, M., Gershenson, C., & Kiviat, B. (2015). Forced relocation and residential instability among urban renters. *Social Service Review, 89,* 227–262.

Desmond, M., Gromis, A., Edmonds, L., Hendrickson, J., Krywokulski, K., Leung, L., & Porton, A. (2018). *Eviction lab national database: Version 1.0*. Princeton, NJ: Princeton University. Available from www.evictionlab.org

Desmond, M., & Kimbro, R. T. (2015). Eviction's fallout: Housing, hardship, and health. *Social Forces, 94,* 295–324.

Desmond, M., & Shollenberger, T. (2015). Forced displacement from rental housing: Prevalence and neighborhood consequences. *Demography, 52,* 1751–1772.

Duncan, G. J., & Rodgers, W. L. (1988). Longitudinal aspects of childhood poverty. *Journal of Marriage and the Family, 50,* 1007–1021.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics, 2,* 1360–1383.

Gould-Werth, A., & Seefeldt, K. S. (2012). *Material hardships during the Great Recession: Findings from the Michigan Recession and Recovery Study* (National Poverty Center Policy Brief No. 35). Ann Arbor: University of Michigan Poverty Solutions. Retrieved from http://www.npc.umich.edu/publications/policy_briefs/brief35

Joint Center for Housing Studies of Harvard University. (2017). *The state of the nation's housing* (Report). Cambridge, MA: Joint Center for Housing Studies. Retrieved from http://www.jchs.harvard.edu/sites/jchs.harvard.edu/files/harvard_jchs_state_of_the_nations_housing_2017.pdf

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9,* 137–163.

Mayer, S. E., & Jencks, C. (1989). Poverty and the distribution of material hardship. *Journal of Human Resources, 24,* 88–114.

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (2001). Fragile Families: Sample and design. *Children and Youth Services Review, 23,* 303–326.

Stan Development Team. (2017). *RStan: The R interface to Stan* (R package version 2.16.2) [Software]. Available from: http://mc-stan.org

Wildeman, C. (2009). Parental imprisonment, the prison boom, and the concentration of childhood disadvantage. *Demography, 46,* 265–280.

ONLINE APPENDIX

A research note on the prevalence of housing eviction among children born in U.S. cities

Ian Lundberg and Louis Donnelly

Parts 1–3 discuss the details of our preferred model, focusing on variable construction (Part 1), multiple imputation (Part 2), and model specification (Part 3). Part 4 assesses the bias and predictive performance of the preferred model. Part 5 compares the preferred parametric model to a random forest as a nonparametric alternative. Part 6 summarizes city differences in eviction prevalence. Part 7 compares the Fragile Families Study sample characteristics to the characteristics of a similar population from vital records.

Part 1: Variable construction

We construct a set of predictor variables that we expect may be related to a child's risk of eviction. Characteristics of the child's household include an indicator for whether the mother and father were married at the birth, the proportion of waves in which the child lived in an owned home, annual rent or mortgage cost as a proportion of household income[1] (averaged over waves), and permanent household income relative to

[1] In interviews at child age 5 and 9, only parents who had moved provided updated information about whether they lived in an owned home and about their housing costs. In

the poverty threshold for that household (averaged over waves). Because income may be nonlinearly associated with the log odds of eviction, we include it as a set of indicator variables: below 50 % of the poverty line, 50–100 % of the poverty line, 100–200 % of the poverty line, 200–300 % of the poverty line, or higher than 300 % of the poverty line. In the model, income is interacted with parents' marital status at the birth.

Because three quarters of the sample are children born to unmarried parents, the mother is the primary caregiver much more often than the father. To capture factors that affect a child's family's ability to afford rent, we include a set of mother's characteristics. Race and ethnicity is measured as black, non-Hispanic; Hispanic, any race; or white/other, non-Hispanic. Education is coded as less than high school, high school, some college, or college. Because younger mothers may be less financially secure and thus face a greater risk of eviction, we include a linear term for mother's age at the child's birth. Because eviction may be related to other caregiver characteristics, we include a scale designed to measure the mother's impulsivity (Dickman 1990) and the mother's score on the Wechsler Adult Intelligence Scale – Revised (Wechsler 1981).

Incarceration of the child's father is extremely common in these data: 34% (unweighted) of fathers have been in jail or prison by the interview when children were one year old. We include this indicator in the model since incarceration may disrupt family life and hinder a family's ability to pay rent.

---

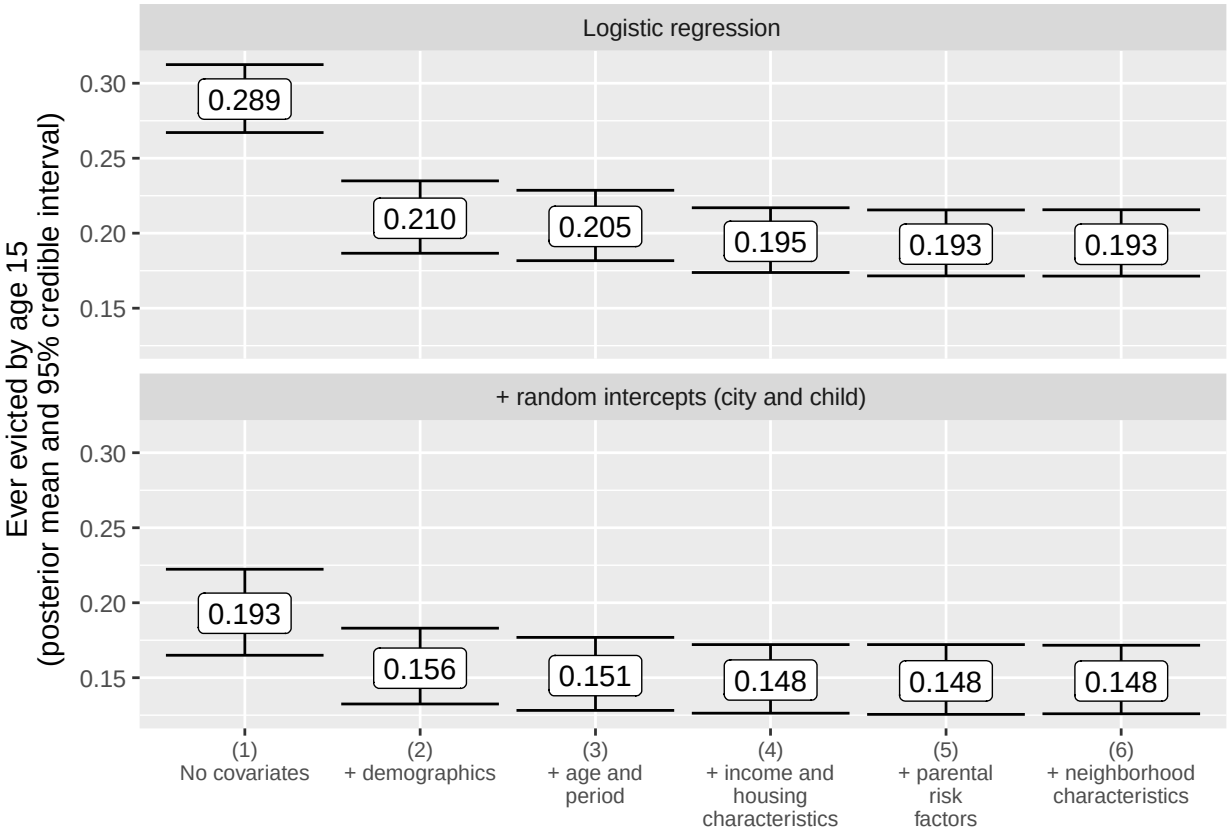
these waves, we assume for those who had not moved that these values were the same as in the previous wave.

We include a set of neighborhood context variables representing the Census tract in which the child's family lived when the child was born. These variables were measured in 2000 Census data and merged with the Fragile Families Study by Census tract. We measure racial composition as the proportion white, black, Hispanic, or other. We include the proportion of households below the poverty threshold. Finally, to capture rent burden in the neighborhood, we include the Census tract's median rent divided by household income.

Time-varying predictors include child age and whether the observation is during the Great Recession (2008-2009).

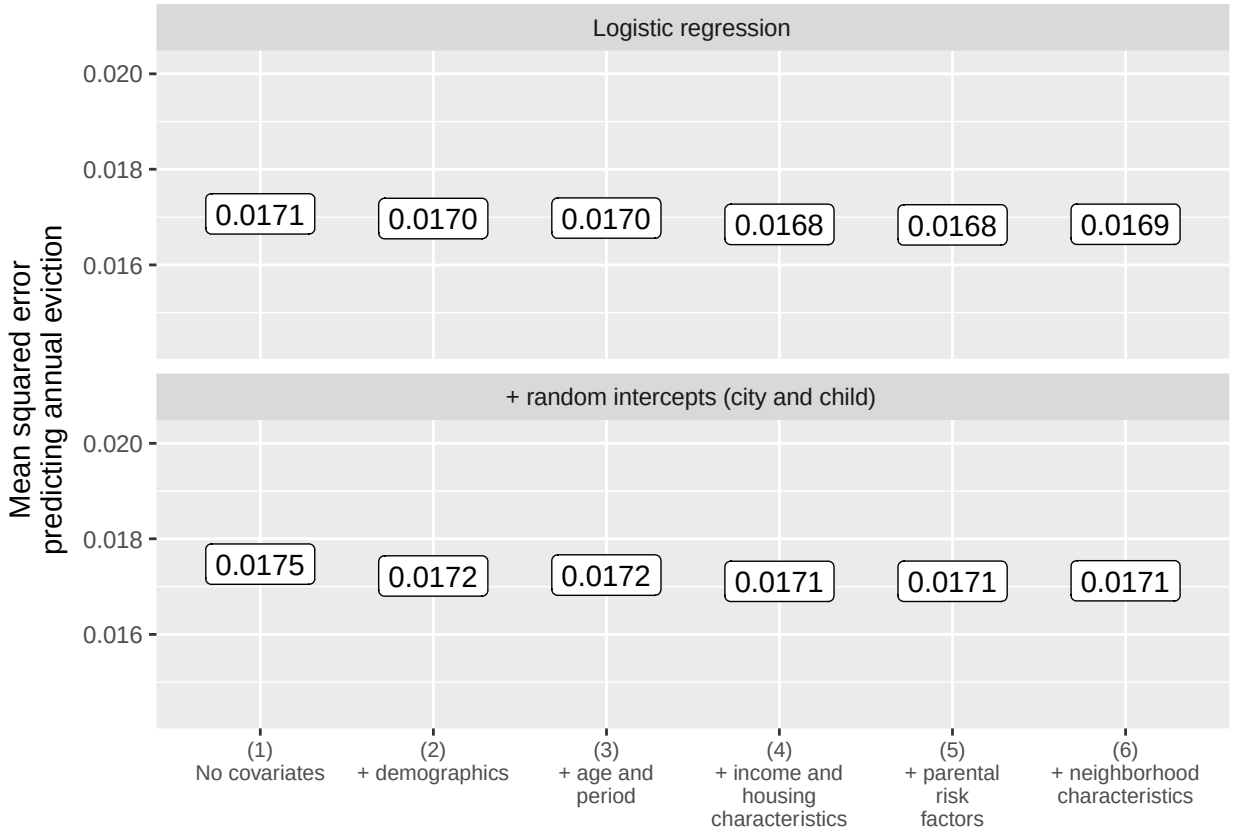
To assess the sensitivity of the results to the predictor set included, we re-estimated the model on various subsets of the predictors with and without the child- and city-specific random intercepts. Figure A1 shows that the implied eviction prevalence is greatest in the simplest model that includes only an intercept. Intuitively, more children will be evicted between birth and age 15 if all children have the same risk of eviction in all years, meaning that no child is safe from eviction. More complex models generally yield slightly lower estimates. Once a reasonably detailed set of variables is included, the addition of more predictors changes the implied proportion evicted only slightly. In addition, predictive performance varies only slightly across the models (Figure A2). Overall, results suggest little reason beyond prior beliefs to prefer one model over the others, but results appear relatively robust to any model that includes a reasonable set of predictors.

Figure A1. Ever evicted by age 15: Prevalence under various models



Note: The implied proportion of children born in large U.S. cities to be evicted by age 15 is reasonably robust to a range of model specifications. The estimate is slightly higher in models with fewer covariates or without random intercepts but is roughly constant across models with several covariates and random intercepts. Our preferred model, represented at the lower right, includes all of the covariates and random intercepts because we believe a priori that these variables are all relevant predictors and that there is unobserved child- and city-specific heterogeneity.

Figure A2. Performance of various models at predicting eviction in the past 12 months



Note: One might be tempted to select a model based on its ability to accurately predict held-out observations of eviction in the past year. Figure shows 5-fold cross-validated estimates of mean squared prediction error (Brier score). Predictive performance is relatively similar across all models, suggesting that a data-driven approach to model selection may not be useful. Our preferred model is the logistic regression with all of the predictors represented in the lower right of this figure; we choose this model not because of its predictive performance but because it captures the posterior distribution of our beliefs about eviction prevalence given prior beliefs that all of the variables and random effects are relevant predictors.

Part 2: Multiple imputation

This section discusses the independence assumptions required for multiple imputation and the particular estimation approach we chose.

Multiple imputation: Independence assumptions

Multiple imputation involves several independence assumptions, which we discuss in this section.

The key assumption required for multiple imputation is that data are missing at random: whether a child is missing a given variable is independent of the value that variable actually takes, conditional on all observed variables. As discussed in the limitations section, this assumption may not hold in our data, as the most disadvantaged children who experience the most housing instability may be most likely to attrite from the survey. This would cause our multiple imputation estimator (and all of our estimators) to underestimate eviction prevalence.

A second concern is whether the values taken for each variable (column) represent independent draws conditional on the other variables. A violation of this assumption, for example, would occur if repeated observations on a single respondent were represented by multiple rows. To the extent to which unobserved child-specific risk affects eviction prevalence (as assumed in our main model), multiple rows on a single respondent would be correlated in a way not captured by the variance estimation used by most algorithms for multiple imputation, potentially yielding misleading estimates of our uncertainty about imputed values.

To address potential concerns about repeated observations on individual respondents, we store data in wide format with one row per child for imputation. Time-varying variables (i.e. eviction) are stored in multiple columns, with one column per survey report. Thus, our imputation model assumes that eviction at age 5 for two children are independent, conditional on all other eviction reports for those children and on all the predictor variables.

One may be concerned about unmodeled dependence at levels higher than the respondent. The Fragile Families Study sampled cities, then hospitals within cities, then births within hospitals, with an oversample of non-marital births in each hospital. This nested design may induce dependence on several levels. Family incomes, for example, may be dependent if families in certain cities tend to have (a) higher levels or (b) greater variance of family incomes than families in other cities.

To address concern (a) about levels, we include city of birth as a predictor in the multiple imputation model. Likewise, we include mother's marital status as a predictor. We cannot address any concerns about dependence within hospitals because hospital identifiers are highly restricted and are not provided under our restricted use data agreement with the Fragile Families Study.

It is difficult to address concern (b): that family incomes and other variables in the imputation model may be heteroskedastic by city of birth, which could make uncertainty estimates unreliable. The use of multilevel models and other methods to account for heteroskedasticity is an area of active research (i.e. Goldstein et al. 2009, Quartagno and Carpenter 2017), but in our application existing approaches proved computationally

intractable with a dataset of our size. Our imputation model therefore maintains the assumption of homoskedasticity.

Multiple imputation: Estimation approach

We multiply impute 30 datasets using the Amelia package in R (Honaker, King, and Blackwell 2011). Amelia is attractive because it is computationally efficient. The algorithm draws 30 bootstrapped samples of the original data and imputes a value for each missing case based on the posterior mode of the parameters, with estimation uncertainty captured by bootstrapping rather than by computationally-intensive sampling from the full posterior. This makes Amelia more efficient in large datasets such as ours, which includes 4,898 observations on 28 continuous, 1 ordinal, 3 categorical (with 3-20 levels each), and 13 binary variables.

Race/ethnicity and city of birth are specified as categorical variables and education is specified as ordinal because this is how they will enter the model. To avoid losing information and thereby to maximize efficiency, we impute family income relative to the poverty line as a continuous variable within waves before aggregating across waves and binning for the analytic results. We likewise multiply impute eviction as a continuous variable because the mean eviction prevalence in each wave reliably falls between 0 and 1 (Figure 1) and because the package authors advise imputing variables as continuous when possible to maximize efficiency by including the maximum amount of information in the imputed value (Honaker, King, and Blackwell 2011). This does not create problems for the analytic results, which require a dichotomous outcome, because missing eviction reports are excluded from the fitting of the main model.

Our multiply imputed *cumulative* prevalence estimate requires a slightly different procedure. For this estimate, we first aggregate the indicators for eviction at each wave into a single indicator for whether an eviction was ever reported for the child, treating those with any missing reports as missing. This indicator is available for 57% of the sample. We then repeat the multiple imputation procedure above with this as the only measure of eviction included. Doing this is important because including both the wave-specific indicators and a cumulative indicator in a single imputation model leads to numeric estimation problems since all respondents who report an eviction in any given wave are either coded missing or 1 on the cumulative report.

Multiple imputation: Diagnostics

The assumption of missingness at random in multiple imputation is untestable without collecting additional data because it involves quantities that are unobserved: the unknown values that are missing. It is possible, however, to provide evidence in support of the estimation assumptions for multiple imputation. To do so, we use a technique called overimputation in which observed values are treated as though they were missing and imputed, thereby allowing us to assess the ability of the model to produce a distribution of imputed values that approximates the distribution of observed values for the same observations (Honaker, King, and Blackwell 2011). Figures A3 and A4 depict this comparison for binary and continuous outcomes, respectively. For the binary outcomes, the model imputes values with means very close to those observed (Figure A3). The distribution of imputed values appears similar to the observed values for many continuous outcomes, but the two are distinct in cases when the observed values are

bimodal (Figure A4). The observed values of rent or mortgage costs as a proportion of income include many values below 0.5 and a few values near 1, indicating families devoting an enormous share of their incomes to the cost of housing. The multivariate normal imputation model assumed by Amelia fails to capture this bimodal distribution, instead placing many of the imputed values near the middle of the distribution. This suggests that the multivariate normal estimation assumption in the imputation model may be a poor approximation to reality for these variables. On the other hand, it might be that the peak of observed values devoting nearly all their income to rent simply reflect measurement error in which income is systematically under-reported. Aside from the bimodal variables, the comparison in Figures A3 and A4 suggests that the imputation model yields a reasonable set of imputed values.

We estimate that about 2.8 % of the variance is between imputations, a quantity sometimes referred to as the fraction of missing information (Little and Rubin 1987:257). The fact that multiple imputation contributes a relatively small share of the variance compared with other sources of uncertainty suggests that our uncertainty estimates are likely to be reasonably robust to misspecification of the multiple imputation model.[2]

[2] As a rule of thumb, some authors argue that the fraction of missing information should be compared to the fraction of missing observations. In the univariate case, these two are equal in expectation. In the multivariate case, the relationship has no mathematical guarantees, yet Rubin (1987:114) writes that the fraction of missing information "is commonly less than the fraction of observations missing when there are covariates that predict" the missing values. Our case agrees with this intuition, with the 2.8 % fraction of

For readers concerned about selective attrition and the sensitivity of the results to the imputation of missing values, Table A1 summarizes the proportion evicted in each wave among complete cases with valid reports in all waves, available cases with valid reports in each individual wave, and all cases including imputed values for missing reports.

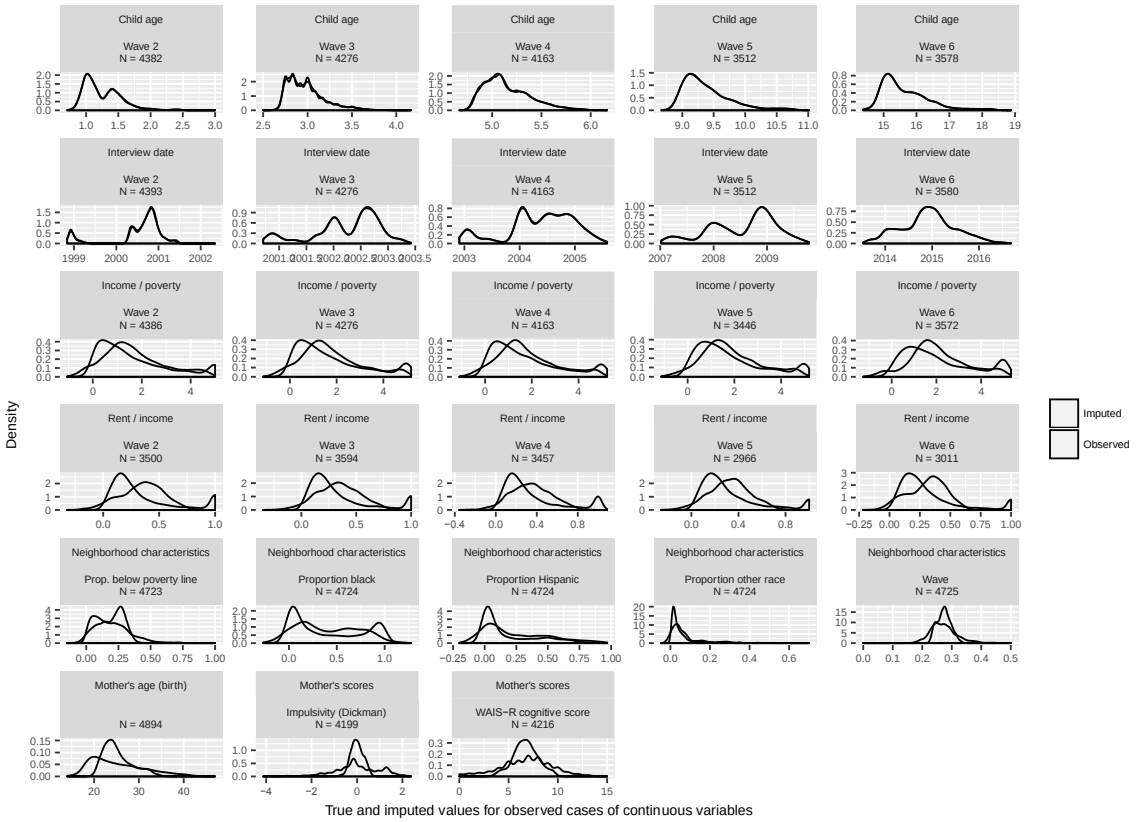
Figure A3. Diagnostics for multiple imputation of binary variables

Note: Figure reports the mean of the observed cases of binary variables as well as the mean that would be imputed if we treated each observed value as missing. Similarity between the predicted and observed means lends credibility to the estimation assumptions of the imputation model by demonstrating that model is well-calibrated.

---

missing information less than the fraction of missing observations on several predictor variables (see Table 2).

Figure A4. Diagnostics for multiple imputation of binary variables



Note: Figure reports the distribution of observed cases of continuous variables as well as the distribution of values that we impute by treating each observed case as missing. Similarity between the predicted and observed distributions lends credibility to the estimation assumptions of the imputation model by demonstrating that the model is able to approximately reproduce the distribution of observed values.

Table A1. Prevalence of reported eviction by survey wave: Attrition, imputation, and weighting

	Proportion with valid responses	Unweighted			Weighted
		Complete case	Available	Imputed	Imputed
Report by wave, in past year					
Wave 1 (Birth-Age 1; 1999-2000)	0.888	0.025	0.028	0.028	0.015
Wave 2 (Age 2-3; 2001-2002)	0.878	0.019	0.019	0.019	0.015
Wave 3 (Age 4-5; 2004-2005)	0.857	0.021	0.022	0.023	0.016
Wave 4 (Age 8-9; 2008-2009)	0.715	0.023	0.025	0.026	0.030
Wave 5 (Age 14-15; 2015-2016)	0.739	0.019	0.019	0.018	0.015
Report, since past wave					
Wave 5 (Ages 9-15; 2009-2016)	0.738	0.056	0.058	0.057	0.046
N		2,768	Varies	4,898	3,442

Note: The range of the median child age and median calendar year within wave are shown in parentheses. Complete case estimates are limited to those with valid responses to all eviction questions in all waves. Available estimates are restricted to those with valid responses in a given wave. The imputed columns are estimated using 30 multiply imputed datasets. The weighted imputed column restricts to the subsample that is a probability sample of children born in 1998–2000 in U.S. cities with populations over 200,000 and corresponds to the estimates presented in Figure 1 of the main text.

Table A2. Ever evicted by age 15: Estimation approaches

	Observed evictions	Multiple imputation	Parametric model
Prevalence of any eviction by age 15	0.079	0.092	0.148
95 % uncertainty interval	(0.071–0.089)	(0.073–0.111)	(0.126–0.172)
Interpretation	Lower bound	Underestimate	Preferred: Slight underestimate
Assumption about missing interviews	Assumes 0 evictions	Assumes interviews are missing at random given covariates and imputation model	Assumes interviews are missing at random given covariates, imputation model, and parametric outcome model
Credibility	Certainly violated	Violated if eviction is associated with attrition	Violated if eviction is associated with attrition
Implication	Yields a lower bound	Downward bias	Downward bias
Assumption about non-survey years	Assumes 0 evictions	Assumes 0 evictions	Assumes non-survey years follow the same data generating process as the parametric model fit to survey years
Credibility	Certainly violated	Certainly violated	Credible if parametric model is believable
Implication	Yields a lower bound	Yields a lower bound	No expected bias
Assumption about reporting accuracy	Assumes accurate reporting in 12-month recall and 6-year recall	Assumes accurate reporting in 12-month recall and 6-year recall	Assumes accurate reporting in 12-month recall
Credibility	Likely violated: - Respondents may underreport eviction - Respondents may not accurately recall 6 years	Likely violated: - Respondents may underreport eviction - Respondents may not accurately recall 6 years	Likely violated: - Respondents may underreport eviction
Implication	Downward bias	Downward bias	Downward bias

Part 3: Detailed modeling procedures

As discussed in the text, we present three estimates of eviction prevalence: a lower bound based on observed evictions, a multiply imputed estimate, and our preferred estimate based on a multilevel logistic regression model. Table A2 provides more detail on each of these approaches, which are summarized graphically in Figure 2 in the main text. The rest of this section presents details related to the multilevel logistic regression model with random intercepts for children and cities (our preferred estimate).

Frequentist approaches to generalized linear mixed models require computationally intensive numerical integration methods, especially when standard errors of the random effects are desired (Skrondal and Rabe-Hesketh 2009). In order to easily calculate uncertainty around our weighted cumulative prevalence estimates, we estimate a Bayesian multilevel logistic regression with random intercepts for children and cities of birth. This section gives details of the priors used in the model. Following the advice of Gelman et al. (2008), we first recenter all variables to have mean 0 and scale all continuous variables to have standard deviation 0.5, so that the standard deviation of continuous variables matches that of a binary variable with mean 0.5. We then assume the following model.

Grand intercept:	$\alpha \sim \text{Cauchy}(-4.5, 1)$
Coefficients:	$\{\beta_1, \ldots, \beta_k, \gamma, \lambda\} \sim^{\text{iid}} \text{Cauchy}(0,1)$
Random intercept standard deviations:	$\{\sigma_\delta, \sigma_\epsilon\} \sim^{\text{iid}} \text{HalfCauchy}(0,1)$
City random intercepts:	$\{\delta_c\} \sim^{\text{iid}} \text{Normal}(0, \sigma_u^2)$
Child random intercepts:	$\{\epsilon_{c[i]}\} \sim^{\text{iid}} \text{Normal}(0, \sigma_v^2)$

Linear predictor:

$$\eta_{c[i[t]]} = \alpha + \underbrace{\vec{\mathbf{X}}_{c[i]}\vec{\boldsymbol{\beta}}}_{\substack{\text{Child-level}\\ \text{predictors}}} + \underbrace{\text{Age}_{c[i[t]]}\gamma + \text{Recession}_{c[i[t]]}\lambda}_{\text{Time-varying predictors}} + \underbrace{\delta_c + \epsilon_{c[i]}}_{\substack{\text{Random}\\ \text{intercepts}}}$$

Link function:

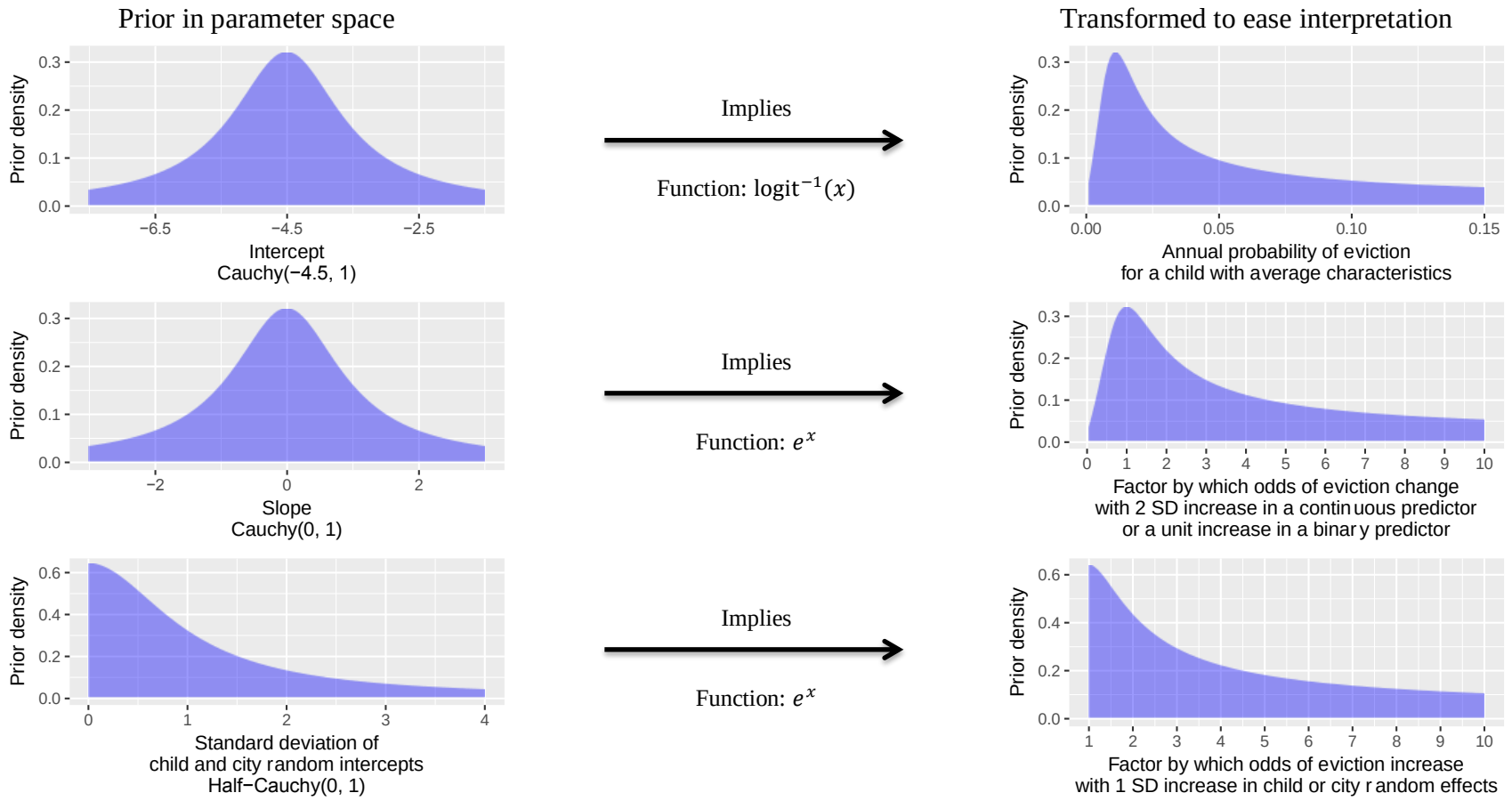
$$\underbrace{\pi_{c[i[t]]}}_{\mathbb{P}\left(\substack{\text{Eviction in year } t \text{ for}\\ \text{child } i \text{ born in city } c}\right)} = \text{logit}^{-1}\left(\eta_{c[i[t]]}\right)$$

Outcome:

$$\underbrace{Y_{c[i[t]]}}_{\substack{\text{Eviction in year } t \text{ for}\\ \text{child } i \text{ born in city } c}} \sim \text{Bernoulli}\left(\pi_{c[i[t]]}\right)$$

These priors are depicted graphically in Figure A5. While Gelman et al. (2008) recommend Cauchy(0,2.5) priors in general, we use narrower Cauchy(0,1) priors because we believe these are sufficiently diffuse in this empirical setting. The prior on the intercept α places the middle 50% of the density on annual eviction probabilities between 0.4% and 2.9% for a child with average characteristics. The middle 90% of the prior density extends to probabilities as extreme as 0.002% to 86.0%. The prior on the other coefficients implies odds ratios that typically fall between 0.37 and 2.7 (middle 50% of prior density) for a 2 standard deviation increase in a continuous predictor or a change from 0 to 1 for a binary predictor. However, the heavy tails of the Cauchy distribution allow odds ratios as extreme as 0.002 to 552.1 (middle 90% of prior density). The prior on the standard deviation of the random intercepts implies a 50% prior probability that a standard deviation increase in the child- or city-of-birth-specific random intercept multiplies the odds of eviction by an odds ratio between 1 and 2.7 and a 90% prior probability that this odds ratio lies between 1 and 552.1. These priors are weakly informative in the sense that they prefer values that seem reasonable (the middle 50%) but do not rule out extreme values (the middle 90% and beyond).

Figure A5. Prior distributions



Note: The prior distributions on the parameters on the left side imply the transformed values on the right. All plots are truncated; the support is all real numbers for the Cauchy priors and all non-negative numbers for the Half-Cauchy priors. For the motivation to choose Cauchy priors, see Gelman et al. (2008).

The prevalence of eviction implied by our model depends on the particular way the model shares information across children, as specified in the priors above. At one extreme, one could assume that the eviction experiences of each child are entirely uninformative about the experiences of other children. Under this assumption, one would have to estimate each child's probability of eviction from birth to age 15 using only reports from that child. One approach to do so is our lower bound estimator: treat each child as evicted if an eviction is ever reported, and otherwise assume the probability of eviction was zero. This shares no information across children and produces an estimate that is likely too low. At another extreme, one could assume that the probability of eviction in each year is constant across the entire population, with evictions representing independent and identically distributed draws across child-years. Observing that 2.3 % of observed child-years involve an eviction report,[3] one might infer that the probability of eviction from birth to age 15 is 29.0 % because $1 - (1 - .0226)^{15} = 0.290$. This shares all information across children and produces an estimate that is likely too high. Our preferred estimator involves partial pooling: the normal prior that we assume over the child-specific random intercepts allows some unobserved child-specific heterogeneity but shrinks all child-specific estimates toward the grand mean. In addition, the covariate-based portion of our model pools information to yield similar fitted probabilities for those with similar covariates. This partial-pooling approach produces an estimate that falls between the no pooling and full pooling estimators described above. While we believe

[3] We do not use weights in this calculation because they are unnecessary given the assumption that the probability of eviction is constant across all child-years.

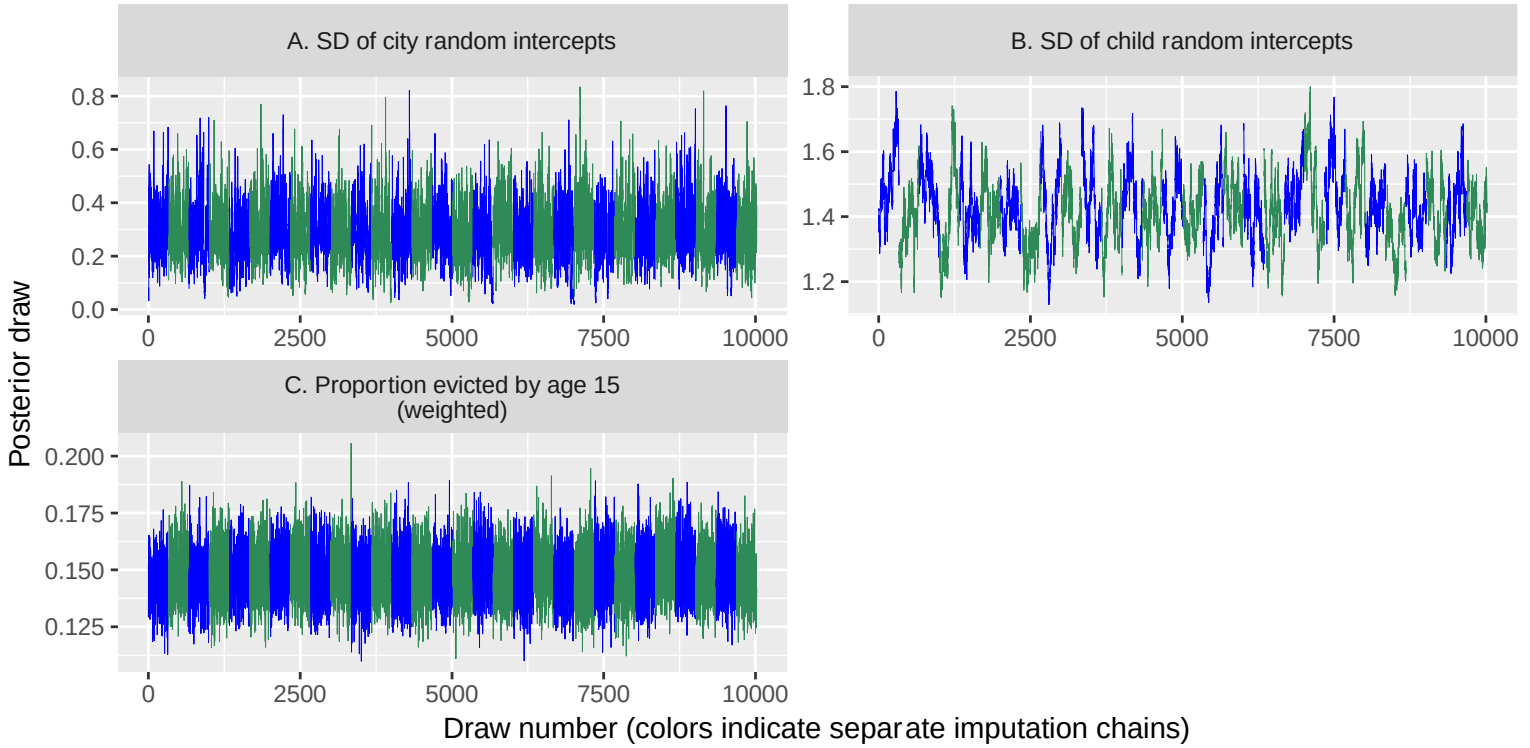
that our model represents a reasonable way to pool information across children, we acknowledge that different modeling assumptions could produce estimates as low as our lower bound or higher than our preferred estimator.

For each of the $m = 30$ multiply imputed datasets, we draw 334 posterior samples from the model above using the Hamiltonian Monte Carlo implemented in Stan through the RStan package in R (Carpenter et al. 2017, Stan Development Team 2017), employing the No-U-Turn Sampler (Hoffman and Gelman 2014). We mix the posterior samples from all 30 imputed datasets to produce 10,020 draws from a posterior distribution that accounts for imputation uncertainty (Gelman et al. 2013:452).

We assessed convergence of the Hamiltonian Monte Carlo sampling by examining trace plots for the coefficients. The trace plots for the standard deviation of the random intercepts and for the overall eviction prevalence estimate are presented in Figure A6. The fuzzy appearance of the trace plots indicates that posterior sampling did not stick in any particular parts of the parameter space but rather explored the full range of the posterior many times over the iterations.

Posterior mean estimates of the coefficients are presented in Table A3. However, the coefficients are only predictive and have no causal interpretation without additional assumptions.

Figure A6. Trace plots showing posterior sampling of key parameters



Note: Each panel shows 10,020 samples from the posterior for a given parameter, drawn in 30 sequential chains of 334 samples each. Each color change represents the start of a new chain drawn based on one of 30 multiply imputed dataset. Prior to each chain, we sample 1,000 burn-in draws. Panels A and B plot draws of the quantities one might imagine are most difficult to estimate from the data: the standard deviation of the random intercepts at the city and child levels. Panel C plots samples of the key quantity of interest: the overall weighted prevalence estimate.

Table A3. Multilevel logistic regression model of eviction

	Posterior mean of odds ratio	95 % quantile-based posterior credible interval
Household characteristics		
Permanent income		
(over 300% of poverty line omitted)		
Below 50% of poverty line	0.80	0.63–0.99
50-100% of poverty line	1.45	1.09–1.87
100-200% of poverty line	2.97	1.09–7.05
200-300% of poverty line	3.26	1.44–6.99
Parents married at birth	1.04	0.78–1.36
× below 50% of poverty line	2.44	0.34–9.31
× 50-100% of poverty line	1.09	0.29–2.86
× 100-200% of poverty line	1.40	0.48–3.26
× 200-300% of poverty line	0.59	0.13–1.54
Annual housing costs / income	6.44	0.57–30.81
Prop. of years living in an owned home	4.57	1.57–11.81
Mother's characteristics		
Race/ethnicity (white/other omitted)		
Black	0.73	0.48–1.04
Hispanic	0.88	0.56–1.31
Education (high school omitted)		
Less than high school	1.12	0.83–1.47
Some college	0.96	0.67–1.33
College	1.02	0.46–1.92
Age at birth	0.96	0.93–0.98
Foreign-born	0.74	0.40–1.25
Impulsivity (Dickman 1990, standardized scale)	0.61	0.48–0.76
Cognitive skills (WAIS-R, standardized scale)	1.11	1.05–1.19
Father's characteristics		
Ever in jail/prison by child age 9	0.71	0.55–0.91
Neighborhood context		
(Census tract characteristics in 1999)		
Racial composition (proportion white omitted)		
Proportion black	1.37	0.80–2.22
Proportion Hispanic	1.51	1.05–2.12
Proportion all other	0.50	0.04–2.27
Proportion of households below poverty line	0.58	0.09–2.02
Median rent / household income	2.74	0.02–17.04
Time-varying predictors		
Child age	0.99	0.96–1.02
Recession year	1.18	0.90–1.52
Unobserved heterogeneity		
Odds ratio associated with 1 standard deviation increase in city of birth random intercept	1.34	1.09–1.68
Odds ratio associated with 1 standard deviation increase in child random intercept	4.17	3.39–5.20

Note: Overall N = 4,898. Model is unweighted. Odds ratios are exponentiated coefficients. Credible intervals represent the 2.5% and 97.5% quantiles of the posterior distribution of each odds ratio.

Part 4: Downward bias in logistic regression

Logistic regression is a consistent but biased estimator for fitted probabilities (McCullagh and Nelder 1989). In a finite sample with a rare outcome, this bias tends to make the results under-predict the probability of an event (King and Zeng 2001). To test this possibility, we evaluated the predictive performance of our model using 5-fold cross-validation. We first sorted the data by eviction, city of birth, sample weight, and respondent. Then, we randomly assigned every block of 5 observations to 5 folds (subsets of the data), thereby splitting the sample randomly into 5 folds but forcing the folds to be similar on these variables. For each of the 5 folds, we held out the given fold, trained the model on the remaining 4 / 5 of the observations, and then evaluated predictive performance on the held-out fold. Finally, we averaged prediction errors over the 5 folds. Because the primary goal of this exercise is to evaluate bias and accuracy in the point estimates rather than to estimate standard errors, and because the procedure is computationally intensive, we only conducted cross-validation on the first imputed dataset.

If the bias in logistic regression were substantial, we would expect the predicted prevalence of eviction in the held-out cases to be less than the observed prevalence. In fact, the two were almost the same: the sample average posterior mean predicted probability of eviction was 2.20 %, compared with an observed proportion evicted of 2.26 %, suggesting a miniscule downward bias. Mean squared prediction error (Brier score) was 0.0227. Likewise, if we weight the observed person-years by the mother's baseline interview sampling weight, the weighted sample average posterior mean predicted probability of eviction was 1.59 %, compared with a true weighted proportion

evicted of 1.73 %, again suggesting a slight downward bias. Weighted mean squared prediction error (Brier score) was 0.0171. In either case, the bias in the logistic regression predictions is very small. Our logistic regression models may therefore understate eviction prevalence, but we believe only slightly.

Part 5: Nonparametric alternative

The task of this paper is fundamentally one of prediction: our goal is to predict the prevalence of eviction in unobserved years of childhood. Because nonparametric machine learning models typically outperform parametric models at prediction tasks, we considered one such model: random forests (Breiman 2001). Results were very similar to the main parametric specification, suggesting that 16.8 % of children born in large U.S. cities in 1998–2000 were evicted by age 15, slightly higher than the main parametric estimate of 14.8 %. Cross-validated mean squared prediction error (Brier score) was 0.020 unweighted and 0.016 weighted, representing only slightly better predictive performance than our preferred multilevel logistic regression model (unweighted MSE = 0.023, weighted MSE = 0.017).

We prefer the estimate in the main text for four reasons. (1) The random forest estimate is sensitive to the choice of tuning parameters, producing a point estimate ranging from 11.3–22.0 % across the tuning parameters we considered. Because the amount of data is relatively small, we have limited confidence in the reliability of cross-validation to select a preferred set of tuning parameters for the estimates reported above. We prefer Bayesian logistic regression because this strategy allows us to state explicitly our prior beliefs about the distribution of eviction risk across the population (model complexity) in a transparent way (see Figure A5). (2) The aim of the paper involves interpolation over regions of the covariate space with no data (i.e. child ages 7-8). Parametric assumptions make this interpolation transparent. (3) We expect that logistic regression will be more familiar to our target audience. (4) When in doubt, we choose to present the lower of two estimates (logistic regression instead of random forest). As a

result, our findings, if anything, are conservative. The substantive claim that eviction is alarmingly prevalent holds under either estimator.

The remainder of this section introduces random forests and outlines how we applied them in this scenario.

Random forests: An introduction

This introduction is necessarily brief; interested readers can find more thorough introductions in Hastie, Tibshirani, and Friedman (2009:305–317, 587–604) and Efron and Hastie (2016:124–128, 325–332).

A random forest is an ensemble of regression trees. A regression tree is a predictive algorithm that learns a partition of the data such that observations in each cell have similar predicted outcomes. The algorithm begins with a bootstrapped sample of the data. The sample is split into observations above or below a certain value on a chosen variable, with the split chosen to maximize the difference in the predicted outcome between the two branches. At the end of each branch, a new split is chosen along a (possibly different) variable. After many such splits, the sample is divided into a large number of terminal nodes, called leaves, which each contain a small number of observations that are relatively homogenous on the outcome variable. Given a new observation, the regression tree prediction rule drops that observation down all the branches it would follow defined by its covariates and then predicts the probability of eviction as the proportion evicted in the terminal node to which the observation would be assigned. Regression trees are appealing because the repeated splits allow deep interactions between predictors and because they do not assume linear associations

between the predictors and the outcome. Trees can theoretically approximate any functional form associating the predictors with the outcome.

Regression trees tend to be noisy predictors, often referred to as "weak learners." Random forests (Breiman 2001) address this problem by averaging over many trees. Each tree is grown on a bootstrapped sample of the observations. At each split, only a randomly selected subset of variables is considered for splitting. These two sources of randomness produce trees that are only weakly correlated with each other, so that the resulting average of many trees is a better predictor than any individual tree.

Random forests: Two challenges

We fit a random forest regression to the eviction data using the ranger package in R (Wright and Ziegler 2017). Our application involves two elements not common among uses of random forests: one predictor (respondent ID) has many categories, and the sample needs to be weighted.

There are 3,442 respondents in the nationally-weighted subsample of the Fragile Families Study sample, each of whom is observed five or fewer times. Hastie, Tibshirani, and Friedman (2009:310) warn that categorical variables with many values are often poor predictors in a random forest; splitting on any individual category is unlikely to improve predictive performance because each category contains only a few observations and therefore yields noisy predictions. If one must include a categorical predictor with many values, Hastie, Tibshirani, and Friedman (2009:310) recommend ordering the categories by the proportion evicted. Splitting on the 10^{th} respondent, for instance, might be useful in this encoding because the first 10 respondents all have higher observed eviction

prevalence than the remaining 3,432 respondents. We considered both approaches, excluding respondent identifiers or including them as an ordered categorical predictor, using cross-validation to select a model (described in the next section).

The second challenge particular to this setting is that random forests typically assume that the data come from a simple random sample, so that the bootstrapped sample used to grow each tree reasonably approximates the original sample. The Fragile Families Study is not a simple random sample: because children were sampled with unequal probabilities, the data only yield unbiased population estimates with survey weights. While one can ignore weights in a correctly-specified parametric model, one cannot ignore weights in a random forest because the forest predicts the unweighted proportion evicted in terminal nodes, averaged over trees, which is likely biased to the extent to which nodes involve observations with unequal weights. To address this problem, we specify the case.weights option in the ranger package to fit trees on weighted bootstrap samples.

Random forests: Tuning by cross-validation and results

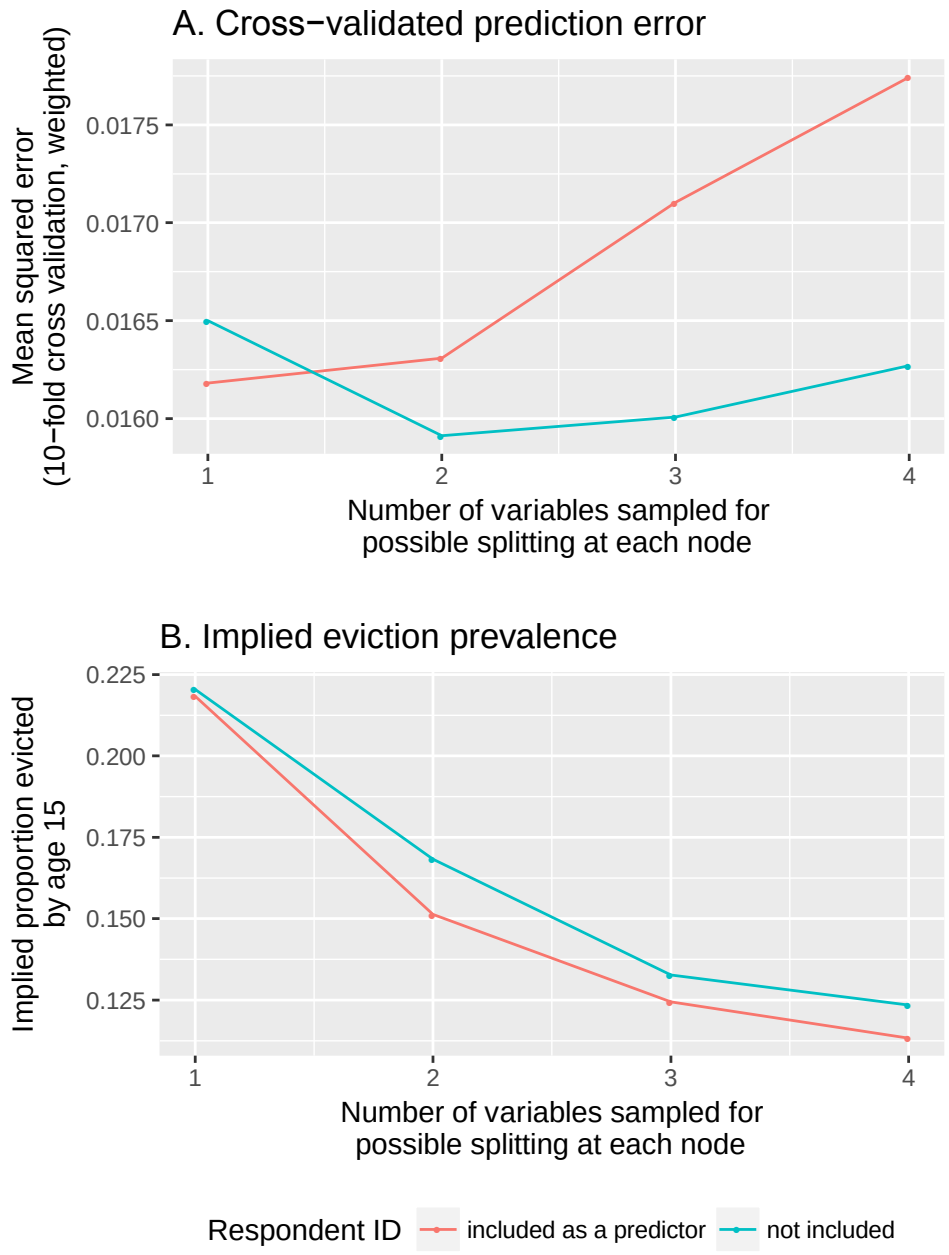
The use of weights makes estimation of generalization error more difficult. In a typical random forest, each observation will be omitted by chance from bootstrap samples used to train some of the trees. By comparing the observed outcomes with the predictions from the trees excluding a given observation, one can estimate "out-of-bag" prediction error. In our weighted case, however, some observations are almost never out-of-bag because their weights are large enough to be always included in every bootstrapped sample. Meanwhile, observations with very small weights are often out-of-

bag. For this reason, the out-of-bag error metric does not work well in our case. Instead, we rely on 5-fold cross-validation. As in Part 4, we order observations by eviction, city of birth, sample weight, and respondent ID, and then randomly assign every block of 5 observations to 5 folds for cross-validation. This strategy ensures that these key variables will be evenly distributed over the 5 folds. We hold out each fold in turn, building the random forest on the other nine folds and comparing predicted eviction with the truth in the held-out fold. The measures of predictive performance reported in Figure A7 represent mean squared error averaged over the folds.

The primary tuning parameter for a random forest is the number of variables sampled to consider at each split. Sampling more variables tends to yield better predictions from any individual tree, but also induces greater correlation among all the trees. The goal of tuning is to find the optimal parameter that yields sufficiently strong individual trees that are not too correlated with each other, thereby producing a forest with minimal out-of-sample prediction error. In our case, the predictor matrix includes 43 columns: a set of 28 columns corresponding to the model matrix from the main specification and a set of 15 columns representing a dummy encoding of the 16 national-sample cities. When respondent identifiers are included, this adds one additional column. We evaluated the fit of the model with 1 through 4 variables sampled at each split, and with or without respondent identifiers included as a possible variable on which to split. As shown in Panel A of Figure A7, the best predictive performance is achieved when two variables are sampled at each split and respondent identifiers are not considered for splitting. Under this model, the implied eviction prevalence is 16.8 % (Figure A7, Panel B). The random forest estimate is thus slightly higher than the 14.8 % estimate reported

in the main text, but the difference is well within the range of estimation uncertainty. One drawback of the random forest estimate is that it depends heavily on the chosen tuning parameters; over the 8 model specifications considered, point estimates ranged from 11.3–22.0 % (Figure A7, Panel B). The parametric specification reported in the main text, in contrast, allows us to argue for a particular model rather than depending on potentially imprecise estimates of mean squared error to choose automatically between a range of models with very different implications. For this reason, we report the parametric model as the main specification and treat the random forest as a robustness check.

Figure A7. Random forest tuning and implied eviction prevalence



Note: Figure shows the tuning of a random forest with 500 trees. Panel A shows that mean squared error estimated by 5-fold cross-validation is minimized when two columns are selected at random to consider in each split and respondent identifiers are excluded from the model. Panel B shows that estimated 15-year eviction prevalence generally declines as the number of variables considered at each split increases. The model chosen by the the lowest mean squared error in Panel A implies in Panel B that 16.8 % of children are evicted between birth and age 15. This is slightly higher than the overall estimate of 14.8 % from the parametric model presented in the main text. The predictive performance of the parametric model from the main text was competitive with the performance of the random forest estimators. The cross-validated weighted mean squared error of the parametric model was 0.0171 (see Part 4), slightly worse than most of the random forest estimators.

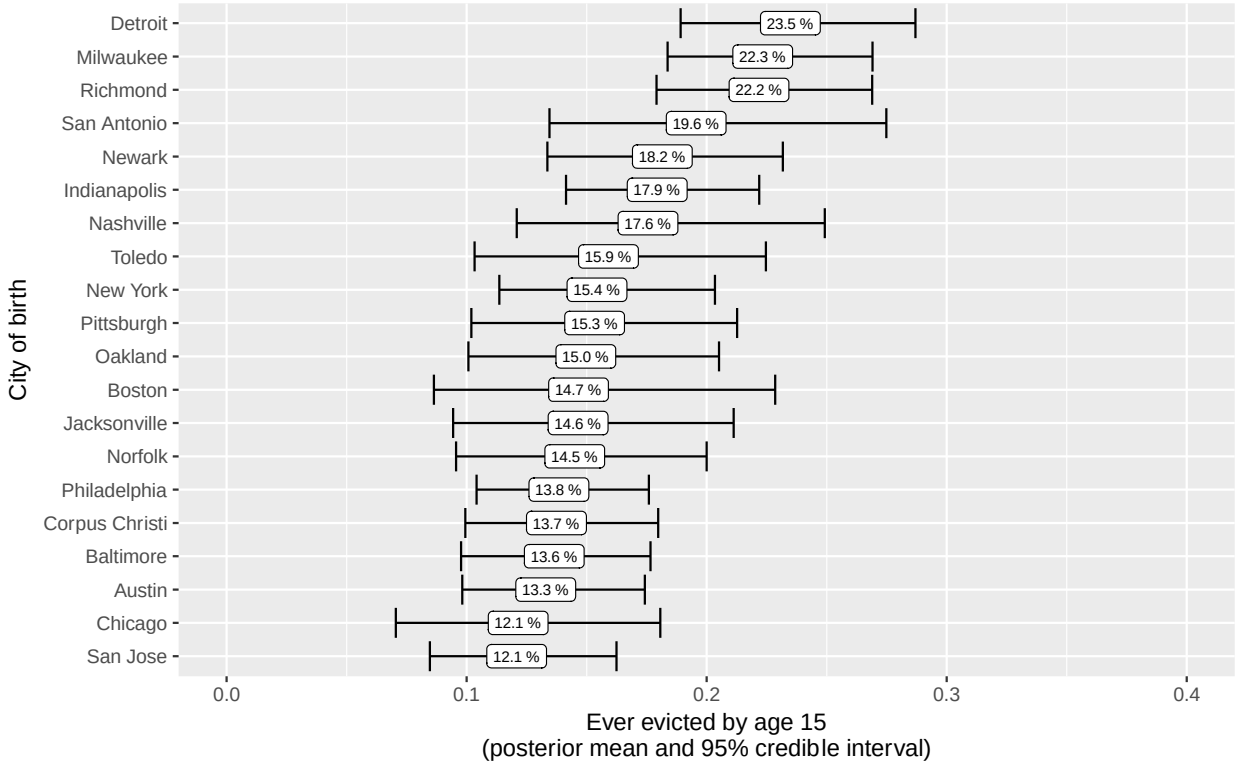
Part 6: Variation by city of birth

Given that the sample is nested in 20 cities of birth, we included city-level random intercepts in the preferred model to allow the probability of eviction to vary by city of birth net of covariates. This decision was methodologically necessary because, without accounting for city-level variation, one could easily imagine that observations clustered in cities would not be independent. This section reports the resulting city-level estimates and outlines why they should be interpreted cautiously, as suggestive of an area for future research rather than as definitive conclusions.

City-specific prevalence estimates are presented in Figure A8, ranging from a low of 12.1 % among children born in Chicago (CI: 7.0–18.1 %) to a high of 23.5 % among children born in Detroit (CI: 18.9–28.7 %). While the estimates suggest some geographic variation, they are very imprecise. As shown in Figure A9, differences between city-specific estimates and the national estimate are not statistically significant at the 0.05 level in most cases. In three cities of birth (Detroit, Richmond, and Milwaukee), prevalence is significantly higher than the national estimate. These differences remain significant even under a procedure to hold the false discovery rate at 0.05 (Benjamini and Hochberg 1995).

While it is tempting to infer that local policies may shape eviction prevalence, this interpretation should be discouraged until further research is conducted for two reasons. First, the city estimates presented here reflect heterogeneity by city of birth, which is not necessarily city of residence at the time of the eviction. The latter is likely more related to local policies. Second, the estimated city differences are descriptive rather than causally identified. The estimates in Figures A8 and A9 allow the other covariates to take the

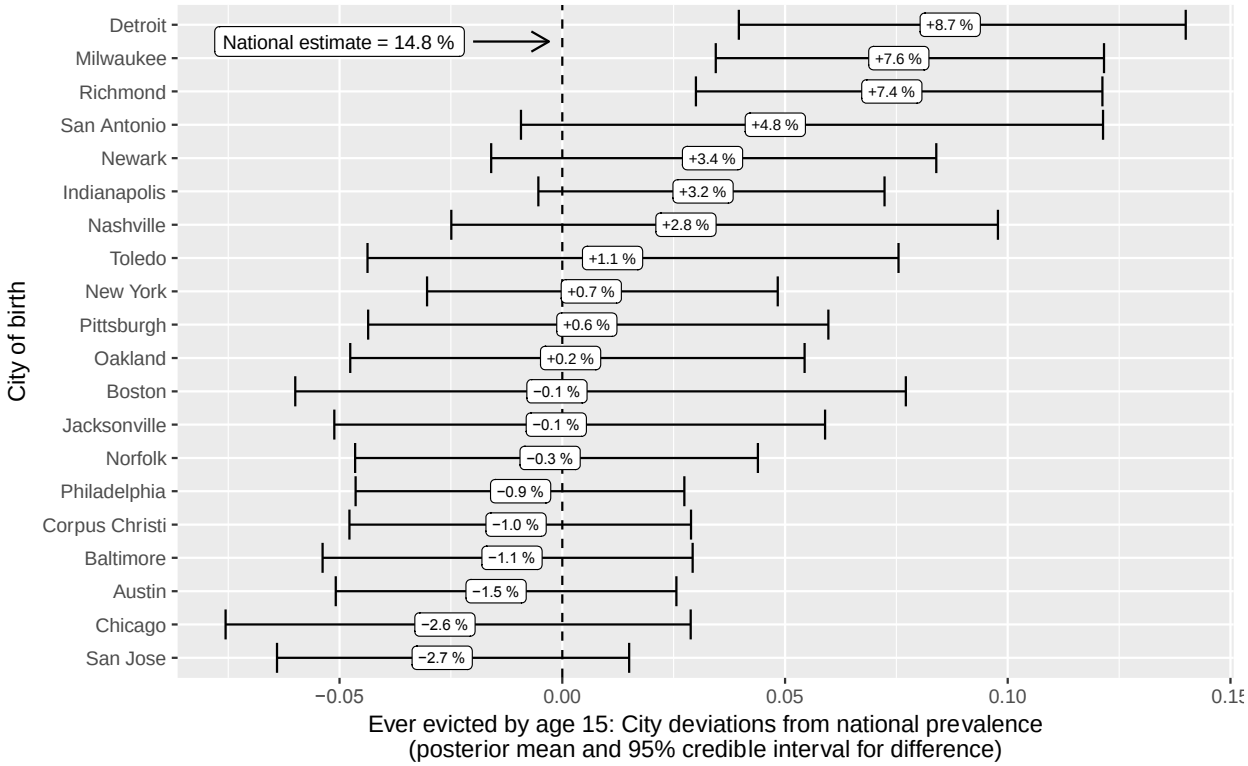
Figure A8. Ever evicted by age 15, by city of birth



Note: Estimates represent the probability that a child born in 1998–2000 in each city was ever evicted between birth and age 15. All estimates are weighted. Estimates are based on predicted probabilities from the multilevel logistic regression model. Error bars represent 95 % credible intervals. This figure shows eviction levels in each city; to more clearly see how eviction prevalence differs between cities, see Figure A9.

values observed in that city; nothing is held constant across cities. The benefit of this approach is that it estimates an interpretable descriptive estimand: the probability that a randomly chosen child born in a given city was evicted by age 15. Policymakers, however, would likely want a different estimand: the effect of living in a certain city on the risk of eviction, holding all confounding variables constant. These are not the same quantities, and our estimate is imprecise regardless. We therefore leave the question of geographic heterogeneity in eviction to future research.

Figure A9. Difference in the estimated proportion evicted by age 15 among those (a) born in each city, and (b) born in any city with population over 200,000. Point estimates and uncertainty capture the difference: (a) - (b).



Note: Estimates represent the probability that a child born in 1998–2000 in each city was ever evicted between birth and age 15. All estimates are weighted. Estimates are based on predicted probabilities from the multilevel logistic regression model. Error bars represent 95 % credible intervals. As the figure suggests, the three cities with the highest prevalence (Detroit, Richmond, and Milwaukee) have significantly higher eviction prevalence than the national estimate. This result is robust to a Benjamini-Hochberg correction to control the false discover rate at 5 %. However, the wide credible intervals in the figure suggest that the present study is underpowered to draw definitive conclusions about geographic differences in eviction. Future research is needed for more definitive claims.

Part 7: Comparing the sample to vital records

The Fragile Families Study sampling frame includes births in 1998–2000 in U.S. cities with populations over 200,000. Vital records made public for research make it possible to compare the weighted descriptive statistics for the sample to known quantities for a similar population: births in 1998–2000 to residents of cities with populations over 250,000. Although these two populations are distinct in terms of city size and whether the inclusion criterion is residency or location of the birth, one would expect the weighted Fragile Families Study sample to be descriptively similar to the known values for this similar population.

Table A4 presents the comparison for several key demographic variables between (1) the unweighted Fragile Families Study sample, (2) the weighted Fragile Families Study sample designed to produce unbiased estimates for births in large cities, (3) vital records on births to residents of large cities, and (4) vital records on all U.S. births. The most relevant comparison is between (2) and (3). The Fragile Families Study sampling design includes an oversample of children born to unmarried parents. Although marital births represent only 24 % of the Fragile Families Study sample (1), the weighted proportion born to married parents is 60 % (2), more comparable to the vital statistics estimate of 56 % (3). Compared to vital records for mothers who reside in large cities (3), Fragile Families Study estimates for mothers giving birth in large cities (2) suggest a slightly lower proportion black (23 % vs. 27 %) and Hispanic (31 % vs. 34 %), and slightly higher levels of education (i.e. 22 % with a college degree vs. 19 % in vital records). The tendency for the weighted Fragile Families Study sample to be slightly more advantaged may reflect differences in the definition of a large city (200,000 in the

Fragile Families Study vs. 250,000 in vital records) or may reflect the inclusion of some more advantaged families who reside outside the city limits but give birth in city hospitals in the Fragile Families Study sampling frame. These two explanations are consistent with the differences reflecting differences in the target populations rather than sampling errors. Overall, the discrepancies between the two are small, which lends credibility to the design of the Fragile Families Study.

Table A4. Comparison of mother characteristics in the Fragile Families Study and in vital statistics

	Fragile Families Study		Vital statistics	
	(1)	(2)	(3)	(4)
Target population	Sampled births (unweighted)	Births in 1998-2000 in U.S. hospitals in cities with populations over 200,000 (weighted)	Births in 1998-2000 to residents of U.S. cities with populations over 250,000	U.S. births in 1998-2000
Married	0.24	0.60	0.56	0.67
Age	25.28	27.08	26.73	27.11
Race / ethnicity				
Black	0.48	0.23	0.27	0.15
Hispanic	0.27	0.31	0.34	0.20
White/other	0.25	0.46	0.39	0.65
Education				
< High school	0.35	0.29	0.31	0.22
High school	0.30	0.30	0.31	0.32
Some college	0.24	0.19	0.19	0.22
College	0.11	0.22	0.19	0.24

Note: The justification for population estimates from the Fragile Families Study is its probability sampling design, which would produce unbiased estimates in the absence of nonresponse. Similarities between descriptive statistics of the weighted Fragile Families Study sample (2) and vital records for a similar population (3) nonetheless lend credibility to the finite-sample performance of the Fragile Families Study. Column (4) is included to clarify how the characteristics of mothers giving birth who are residents of large cities compare to characteristics of mothers giving birth in the nation as a whole.

Online Appendix References

Benjamini, Y., & Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57* (1), 289–300.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76* (1).

Dickman, S. J. (1990). Functional and dysfunctional impulsivity: Personality and cognitive correlates. *Journal of Personality and Social Psychology, 58*, 95–102.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. New York: Cambridge University Press.

Gelman, A, Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis,* Ed. 3. Boca Raton, FL: Taylor and Francis.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2* (4), 1360–1383.

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling, 9* (3), 173–197.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Ed. 2. New York: Springer.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15,* 1593–1623.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45* (7), 1–47.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9* (2), 137–163.

Little, R. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

McCullagh, P., and Nelder, J. A. (1989). *Generalized linear models*. Ed. 2. London: Chapman and Hall.

Quartagno, M., & Carpenter, J. (2017). *Jomo: A package for multilevel joint modelling multiple imputation*. R package version 2.5-2. <https://CRAN.R-project.org/package=jomo>.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society, 172*, (3), 659–687.

Stan Development Team. (2017). *RStan: The R interface to stan.* R package version 2.16.2. <http://mc-stan.org>.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale – Revised (WAIS-R manual).* New York: Psychological Corporation.

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77* (1).