**Data**

Surname data used to calculate the probability of belonging to a specific race/ethnicity ~~comes from data~~are derived from Census 2000 and were released by the US Census Bureau in 2007. This release lists each name held by at least 100 enumerated individuals along with a breakdown of the proportion of individuals with that name belonging to each of the six OMB defined race and ethnicity categories, defined as: White, Black, Asian/Pacific Islander, American Indian/Alaskan Native, Hispanic, and Multiracial/Some Other Race. This classification holds Hispanic as mutually exclusive from the other categories, with any individuals identified as Hispanic belonging only to that category, regardless of racial background. Similarly to the process for racial identification used in the HMDA loan application registry, the Census relies on self-identification of both race and ethnicity when determining identity for these individuals, with an exception made for classification to the "Multiracial/Some Other Race" category. In many cases individuals identifying as "Some Other Race" then specified a Hispanic nationality (e.g., Salvadoran, Puerto Rican); in these cases the Census identified the respondent as Hispanic. Additionally, the US Census did not provide exact counts/percentages for surname/race or ethnicity combinations with nonzero values less than five. In those instances we apportioned the remaining non-assigned total population for the surname evenly across the omitted categories. In total, the list provides 151, 671 surnames, covering approximately 90 percent of the population. Word (2008) provides a more detailed description of how the Bureau cleaned and developed the final list from initial responses. Information on racial/ethnic identity by geography uses census tract-level data from the more recent 2010 US Census Summary File 1.

To compare the constructed race/ethnicity proxies against a reported truth, we use data from the HMDA Loan/Application Register of one of the HMDA+ reporters, with additional information on credit, loan characteristics, name, and address provided. Under rules associated with collection of this data, the provider of the loan application should offer the applicant an opportunity to voluntarily self-identify their racial or ethnic background; if the applicant does not do so the lender must provide information on the applicant to the best of ~~their~~ its ability, with the exception of applications completed by mail, over the phone, or ~~online~~the internet[1]. As the proportion of online applications provided increases, the proportion of applications with race

[1] [Insert Reg C requirement]

omitted will increase as well; as a result imputation of race will become more important in this area when evaluating fair lending outcomes, even though current reporting rules allow for accurate measurement.

Comment [BES1]: Let's think about this word as it can mean something rather specific.

**Construction of Proxy**

Formatted: Indent: First line: 0"

For race and ethnicity, the name and geography information can be combined to form a joint probability[2] using the methodology described in Elliott, et al. [3] This methodology requires the following steps:

- Generate probabilities $p(r|s)$, the probability of belonging to race or ethnicity r given surname s, and $q(g|r)$, the proportion of the population of individuals in race or ethnicity r who live in geographic area g.
- Through application of Bayes' Theorem and the law of total probability, the likelihood that an individual with surname $s$ living in geographic area $g$ belongs to race or ethnicity $r$ is described by

$$\Pr(r|g,s) = \frac{p(r|s)q(g|r)}{\sum_{r\in R} p * q}$$

- For applicants with compound surnames or a co-applicant with a different surname, the probability generated for this paper uses the first surname found on the application. While this sidesteps issues surrounding the presence of multiple individuals on the application, including relationships between applicants, using only one surname results in a simpler implementation with little, if any, loss of accuracy.

Comment [BES2]: What does this mean and is it substantiated?

- For applicants with multiple addresses, only the address of the primary applicant is used for purposes of geocoding, for similar reasons.

To maintain statistical validity of this process, only one assumption is required: That the direct relationship between race and geography does not depend on surname. At a practical level,

[2] [Insert clarification in footnote: this is the posterior]

[3] Elliott et. al., "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," *Health Services and Outcomes Research Methodology*, Sept. 2009.

violation of this assumption requires that individuals with specific surname/race combinations tend to live in certain geographic areas that have a different racial/ethnic distribution than other individuals of the same race. For example, if African Americans with the last name Jones preferred to live in a certain neighborhood more than both African Americans in general and all people with the last name Jones, this would skew the results. The difficulty in coming up with an example that would meet both conditions demonstrates the specificity of this assumption.

### ———Refinements to Proxy

In the ~~census~~ surname data, the Census B~~b~~ureau suppressed exact counts for racial/ethnic categories with 2-5 occurrences for a given name.[4] Similarly to Elliott et. al., in these cases we distribute the sum of the suppressed counts for each surname evenly across all categories with missing nonzero counts. For applicants with multiple surnames, only one name was used to calculate a probability, sacrificing precision and paying the cost of measurement error in order to avoid the potentially contentious issue of how to infer the relationship across last names. In those multiple surname cases, the first surname to appear was matched against the surname list; if this name appeared in the list its distribution was used; otherwise if it appeared the second surname provided the race/ethnicity distribution used in estimation.

**Comment [BES3]:** This needs to be toned down.

**Comment [BES4]:** Need to explore sensitivities to this.

### Measurement Concerns

When evaluating the impact of use of proxies in place of reported race/ethnicity for purposes of estimating impacts on outcomes, among other considerations the roles of four general concerns arise: classification error, omitted variable bias, non-mean-preserving distributions of proxies, and lack of a one-to-one relation between reported race and the proxy.

**Comment [BES5]:** Should clarify this. I think the issues are more nuanced. There are proxy vs. truth issues; and there are proxy vs. reported value issues.

### Classification Error

Given general concerns across economics and statistics~~al~~ literature regarding the role of classical, mean-zero measurement error, one might worry that the use of a constructed proxy

[4] [Insert a footnote describing how often this occurs within the name list both unweighted and weighted by the counts]

would result in error that would mismeasure the true disparity. However, under the basic assumptions used to construct the ~~Bayesian~~ Integrated Surname and Geography proxy $\hat{r}_{isg}$, reconsider it as the mean probability of belonging to race $r$ for individuals with last name $s$ in geography $g$, $\hat{r}_{isg} = \bar{r}_{sg}$ . In expectation, we should expect the reported race to have the form

$$r_{isg} = \bar{r}_{sg} + \nu_{isg} \quad (1)$$

with $v(\nu, \bar{r}) = 0$ . This assumption states that the measurement error in estimating race has no systematic relationship with the proxy itself.

In terms of using the proxy to evaluate outcomes, we are interested in estimating

$$y_{isg} = \beta_r \beta r_{isg} + \varepsilon_{isg} \quad (2)$$

but since we do not know the race, must instead evaluate

$$y_{isg} = \beta_r \tilde{\beta} \bar{r}_{sg} + u_{isg} \quad (3)$$

This will yield the coefficient estimate

$$\beta_r \tilde{\beta} = \frac{cov(y_{isg}, \bar{r}_{sg})}{var(\bar{r}_{sg})} = \frac{cov(\beta_r \beta(\bar{r}_{sg} + \nu_{isg}) + \varepsilon_{isg}, \bar{r}_{sg})}{var(\bar{r}_{sg})} = \frac{\beta_r \beta \sigma_{\bar{r}_{sg}}^2}{\sigma_{\bar{r}_{sg}}^2} = \beta_r \beta$$

under the additional assumption that $cov(\varepsilon, \bar{r}) = 0$.

**Comment [BES6]:** So the point here is that the coefficient estimate is unbiased/consistent.

**Omitted Variables**

The consistent estimate shown above relies on a lack of omitted variables across three estimating equations; relaxation of these assumptions will have consequences. As an example, consider equation (2) in the presence of omitted variables; e.g., $cov(r, \varepsilon) \neq 0$. In this case, regressing $y$ on $r$ results in a parameter estimate of

$$\hat{\beta}_r \hat{\beta} = \beta_r \beta + \frac{\sigma_{r\varepsilon}}{\sigma_r^2} \quad (4)$$

Here, $\beta_r \beta$ represents the true direct impact of race on outcomes, what would be thought of in the legal literature as disparate treatment. $\hat{\beta}_r \hat{\beta}$, meanwhile, represents the true direct impact of race on outcomes, as well as the indirect relationship race has with outcomes based on its correlation with other factors that also influence the outcome. The additional amount beyond the direct relationship is also known as disparate impact. The impact term in the above equation can also be expressed in terms of the race proxy, as

**Comment [BES7]:** Is this a formulation that is common? I like the idea of this framing.

$$\hat{\beta}_r = \beta_r\beta + \frac{\sigma_{r\varepsilon}}{\sigma_r^2} = \beta_r\beta + \frac{cov(\bar{r}+v,\varepsilon)}{var(\bar{r}+v,\bar{r}+v)} = \beta_r\beta_{\bar{r}} + \frac{\sigma_{\bar{r}\varepsilon}+\sigma_{v\varepsilon}}{\sigma_{\bar{r}}^2+2\sigma_{\bar{r}v}+\sigma_v^2} \quad (5)$$

When $cov(r,\varepsilon) \neq 0$ running the regression in (3) results in an estimate of

$$\hat{\beta}_{\bar{r}}\tilde{\beta}_{\bar{r}} = \frac{cov\left(\beta_r\beta\left(\bar{r}_{sg}+v_{isg}\right)+\varepsilon_{isg},\bar{r}_{sg}\right)}{var\left(\bar{r}_{sg}\right)} = \frac{\beta_r\beta\sigma_{\bar{r}}^2+\beta_r\beta\sigma_{\bar{r}v}+\sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2} = \beta_r + \frac{\beta_r\sigma_{\bar{r}v}+\sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2} \quad (6)$$

Comparing equations (4) through (6), while not straightforward, allows us to draw conclusions about the consequences of violating the initial assumptions about the presence of omitted variables both in the true estimation and in the proxy equation. First, examination of the denominators of (5) and (6) and the dichotomous nature of $r$ versus the continuous nature of its proxy $\bar{r}$ provides evidence that the denominator of the bias for the reported race should be larger than that of the proxy, increasing the size of the proxy's estimated bias/disparate impact. Second, differences between covariances of the omitted variables and the reported race when estimating outcomes and those same variables and the proxy (i.e., $\sigma_{r\varepsilon} \neq \sigma_{\bar{r}\varepsilon}$) will impact the difference between the estimates of outcomes generated by regressions using these variables. Third, the amount of estimated bias in outcomes from the proxy relative to the use of reported race will be impacted by the amount of omitted variable bias present in the estimation of actual race. For example, if income predicts race and is correlated with the proxy, but the information it provides regarding race is not included in the proxy, then this will bias the estimate relative to the result in equation (4). With that said, the size of that bias is mitigated by the variance of the proxy, as well as any disparate treatment. If no disparate treatment exists, then this term would cancel out and not create any difference in disparate treatment estimates between the reported race and the proxy. With that said, as the precision of the proxy estimate increases, this will decrease the covariance terms associated with $v$ and the variance of the proxy, minimizing the issues listed above and resulting in outcome estimates generated from the proxy approaching those generated directly using reported race.

Comment [BES8]: What about covariance term

Comment [BES9]: This is not clear to me

Comment [BES10]: Need to clarify this discussion.

**Lack of One-to-One Relation between Reported Race and Proxy**

The above two sections, via equation (1), made the implicit assumption of a coefficient of 1 on the term for the proxy when estimating reported race. If this assumption does not hold, estimation bias results. Specifically, if we reframe equation (1) as

$$r_{isg} = \gamma \bar{r}_{sg} + \nu_{isg} \; (1a)$$

equation (6) becomes

$$\beta_{\bar{r}} \tilde{\beta}_{\bar{r}} = \gamma \beta_r + \frac{\beta_r \sigma_{\bar{r}\nu} + \sigma_{\bar{r}\varepsilon}}{\sigma_{\bar{r}}^2} \; (6a)$$

Here we see the impact of the coefficient from (1a) on measuring outcomes depends on its magnitude. While in the estimation results section we will attempt to infer the impact of omitted variables, we can directly estimate $\hat{\gamma}$ and show how this source of bias will impact the results.

**Results**

**Comparison of Proxy in Estimating Reported Race**

When discussing the performance of any instrument regarding both ability to proxy for some truth as well as that true value's impact on some outcome of interest, while a variety of test statistics and summary information can provide inference of the performance of that instrument relative to other options, no absolute quantitative threshold short of perfection exists for which one can say it serves as an acceptable substitute for the true or reported value. With that said, this section attempts to provide some information on absolute performance of the Bayesian proxy as well as provide clarity on the its performance relative to other alternatives in both the ability to measure the reported race as well as race's relationship with other outcomes.

In terms of measuring race, we would like for the distribution of instrumented races across the population to match that reported by data. Table 1 displays the reported race/ethnicity distribution of our sample data, along with the comparable distributions generated by the joint, name, and geography proxy, respectively. Due to the large sample size of our data set, all differences in percentages of race/ethnicity across distributions are statistically significant at the 1 percent level. In the context of the notation in the previous section, this means that error in our measurement of race is currently nonrandom, and that other covariates could help identify race with further precision. Ignoring the potential for selection, however, overestimating the size of the treatment groups in general should lead to underestimates of the true disparity size, since some individuals in the control group will instead count toward the estimated treatment coefficient. Additionally, a noteworthy pattern emerges from examination of the table; namely,

for five of the six classifications the joint proxy comes closest to matching the reported data, with the only exception being the catch-all "Multiracial/Other" category.

Beyond matching the general population, the Bayesian probability should perform better than existing methods, such as geography or surname used alone, in ordering individuals correctly. Specifically, we should believe that an individual with a high estimated probability of belonging to a specific race/ethnicity is actually more likely to belong to that group than an individual with a lower estimated probability. A Receiver Operating Characteristic (ROC) curve captures this sentiment by graphing how the True Positive rate $\left(\frac{True\ Positives}{True+False\ Positives}\right)$ and False Positive rate change as a threshold rule is applied, with the points on the curve representing movement of that threshold level from zero to one. The slope of the curve at a given point represents the tradeoff between accurately identifying members of the group of interest and the consequence of inaccurately counting members of other groups as part of that group. The ROC curves for each reported race/ethnicity in the HMDA+ data appear in Figures X-Y. Additionally, we can compare whether one estimator better identifies members of a race/ethnicity by comparing the areas under each of the ROC curves, and whether the areas of these curves are statistically significant. In addition to these values appearing in the Figures, Table X also represents these values, along with the test statistic the p-value for the test that the Bayesian proxy more accurately sorts individuals into the correct reported race/ethnicity classifications. Beyond the test for statistical significance, though, examining the curves demonstrated in the figures along with the differences in the areas underneath each curve shows two races with notable improvements in efficiency of sorting individuals. The joint estimates for African Americans, a group known for difficulty in identifying through indirect methods, show a 7% change in the probability of accurately providing a random African American individual a higher probability of being so than a random individual of another race/ethnicity. Additionally, the joint estimates for Non-Hispanic Whites also improve markedly, with a similar change of approximately 3.5%.

While the above section demonstrates the overall superiority of the Bayesian estimator to existing alternatives in sorting, current practice in fair lending compliance work typically relies on the use of a threshold rule to firmly identify individuals as either belonging to some protected class or control group; to demonstrate the relative performance of the Bayesian estimator in that context we also generated contingency tables using the HMDA+ data for each race/ethnicity,

then compared those tables to similar ones generated using surname or geography only. Finally, we calculated the Pearson chi-squared statistic of the null hypotheses that the distribution of the combined proxy with the 80 percent threshold categories matches those for surname and geography, respectively. These results appear as Table XX, with the main takeaway being that the distributions differ significantly across all race/ethnicities, with higher true positive rates and lower false negative rates for the Bayesian proxy among Non-Hispanic Whites and African Americans relative to their next closest alternatives. In particular, the false negative rate for Non-Hispanic Whites drops a sizeable 16%. Among Hispanics the Bayesian estimator provides a 2.8% increase in the true positive rate and .3% increase in the false negative rate relative to use of name only, while the same comparison for Asian/Pacific Islander finds a .8% decline in the true positive rate and a .7% decline in the false negative rate. Overall, the magnitude of gains presented by use of the Bayesian proxy in accurately identifying individuals appears to outweigh the losses.

For measuring the ability of a proxy to co-move with the reported true value, minimizing the residual error and its potential impact on biasing estimates using the proxy away from those that would be generated using the reported truth, we can also analyze the correlations between the proxy value and reported race/ethnicity. The square of the correlation between these two variables, shown in Table YY, is similar to an R-squared value generated by a regression of the reported race on a proxy measure that included a constant term. Examination of Table YY shows that, for each reported race/ethnicity category, the Bayesian proxy explains more of the variance in observed race within our data than use of either surname or geography alone, meaning that regressions of outcomes using the Bayesian estimate will feature lower susceptibility to bias generated by omitted characteristics.

**Comparison of Proxy in Estimating Relationship between Reported Race and Outcomes**

While examining the direct relationship between any proxy estimate and reported race plays a role in determining the utility of the given proxy, our primary interest is in how well use of that proxy matches that of reported race/ethnicity when attempting to compare outcomes across groups. As shown by equation (6a), measuring the magnitude of the relationship between the two variables in equation (1a), along with the amount of variation in the reported truth that the proxy

can explain, provides clues to the ability of the proxy to match use of reported race. Table XX provides the results of a seemingly unrelated regression (SUR) on the test data for the system of six equations that define the race/ethnicity combinations, across all three proxy estimators of geography, surname, and the Bayesian combination of both. The regression framework does not include a constant term, to emulate how we will use the proxy directly in place of reported truth when measuring outcomes. R-squared results shown at the bottom of the table reflect the portion of the variation across all six equations explained by the model, expressed as $R^2 = 1 - \frac{\sum_{r=1}^{R} SSR_r}{\sum_{r=1}^{R} SSR_{tot}}$. Examination of the table provides two main takeaways. First, while the Bayesian proxy does not always provide the race coefficient closest to one in estimation, the differences between the Bayesian proxy and the truth, especially among the largest race/ethnicity classifications, are negligible. Additionally, the four largest race/ethnic groups see coefficients practically identical to one in magnitude, muting concerns about dealing with the specification shown in 6(a). Finally, the R-squared provides the strongest argument for use of the Bayesian proxy, as it explains almost 10 percent more of the variation found across races and ethnicities while providing estimates of race equal to or more accurate than the alternatives. This reduces the likely potential for omitted variables and measurement error to result in differences between outcomes measured with reported race/ethnicity versus the constructed estimate.

**Evaluating the Role of Omitted Variables in Estimating Outcomes**

Given the demonstrated ability of the proxy to accurately capture the magnitude of variation in the probability of belonging to a given race/ethnicity for individuals, omitted variables remains as the largest concern when evaluating the ability of the proxy to estimate relationships between race and outcomes. In order to test this, however, we need to find two types of comparable outcomes in our dataset: one in which we should readily believe in the existence of omitted variables correlated with our constructed estimates of race and ethnicity when running a regression, and another where any omitted variables can be reasonably assumed as orthogonal our race and ethnicity proxies. When using HMDA data, finding an example of the former is relatively easy: a wide variety of outcome variables will be correlated with wealth, which is correlated with race. Additionally, due to the complexity of mortgage lending a number of features should impact loan characteristics that do not exist in available data. Finding the latter is

much more difficult; however, among a subset of loans the condition should hold under reasonable assumptions. Among those loans that qualify as GSE-conforming, whether the loan is sent to Fannie Mae or Freddie Mac should depend neither on race nor on any covariates correlated with race. Though any significant relationship between which GSE purchases the loan and race should be spurious and non-causal, we would expect to see that same relationship show up with the equivalent race proxy measure.

First, to show the potential pitfalls of using the proxy to estimate a process in the presence of additional admitted variables, using our HMDA dataset we regressed the requested loan amount (in thousands of dollars) separately against the race proxy and the self-reported race/ethnicity, then repeated the exercise while adding the applicant's debt-to-income ratio, loan-to-value ratio, credit score, and income as covariates. These results were then compared against similar results generated using the geographic and surname proxies alone. Results appear as Tables XX and YY. In the baseline model of Table XX we see the coefficient for the joint proxy diverge greatly from those estimated using the reported race values, with the greatest difference among major categories occurring for African Americans. More surprisingly, although the name proxy reasonably approximates the reported race result for Hispanics and Asian/Pacific Islanders, the estimated coefficient is almost double the magnitude for African Americans. Upon further reflection, this implies that individuals with more heavily African American surnames, regardless of actual race, applied for lower loan amounts than others. Adding additional covariates changes the magnitudes of the coefficients, but neither their general direction nor the relationship between the reported race coefficients and each of the proxies.

Next, we attempted to contrast the previous result with one that demonstrated the ability of results generated using the proxy to accurately mirror those generated using reported race and ethnicity. To do this we proposed restricting the data set exclusively to loans that were GSE-conforming and sold to either Fannie Mae or Freddie Mac, and then estimating whether a given loan among that set was purchased by Freddie Mac. After discussions with industry experts on mortgage securitization and purchasing, we make the informed discussion that, conditional on generally uniform underwriting quality and documentation, no omitted variables should remain that would generate differences between the reported race and our constructed proxies. An analysis of the overall bank portfolio, presented as Table ZZ, presents a different story. The table shows results of a regression of the dichotomous variable representing the purchaser of the loan

on each respective measure of race/ethnicity, along with a large set of covariates potentially relevant to an investor, including product type, LTV, FICO, DTI, loan amount, purchase/refi status, and month of origination. Here, while the coefficients for the joint proxy generally match the signs of those generated using the reported race, the magnitudes and statistical significance can diverge greatly, in a way that is not obviously systematic. Upon further research on the origins of our dataset we learned that the data contained loans generated by two distinct underwriters, one with a relatively well-regarded reputation for quality and another known for more "thin files" where underwriters failed to follow-up on potential red flags in loan applications. This creates two sets of applications in the data, and violates the assumptions made after consultation with mortgage industry experts. Given correlations between geography, income, race, and credit characteristics, it is unsurprising that our proxy measure might not produce accurate results for a set of data that displays selection on those characteristics.

After learning of this aspect of the data, we performed the same analysis as before, but restricted the sample further to only include loans made by the underwriter perceived by industry to have higher quality standards. These results appear as Table ZZ. Here we see that, as before, the sign of the coefficients for the proxy measures match those of reported race; now however the magnitude of these numbers also resemble those of the reported race. In particular, we see only a difference of 0.0004 (3 percent, in relative terms) in the coefficient for Hispanic, and a 0.0013 (7 percent) for Asian/Pacific Islanders. The coefficient for African American is also much closer to the reported race value, and remains statistically insignificant in both cases. Relative to the other proxy measures, the joint proxy also features smaller standard errors as a result of its better ability to accurately reflect the true probability of reporting as the same race/ethnicity.

Overall, the results estimated using the HMDA dataset point to both the potential and pitfalls of using a proxy measure in place of reported race or ethnicity. In particular, if a process features few relevant omitted variables (e.g., the process is easily modeled using available variables or no relevant omitted variables exist), and the given then the proxy should do a fairly accurate job of estimating a true disparity, should any exist. As the number of relevant omitted variables increases, the ability of our constructed proxy to match the performance of reported race deteriorates.