Credit Invisibles and the Unscored

Kenneth P. Brevoort

Philipp Grimm

Michelle Kambara

Consumer Financial Protection Bureau

The views expressed in this article are those of the authors and do not necessarily reflect the official positions or policies of the Consumer Financial Protection Bureau or the U.S. government.

Abstract

Having a credit record and a credit score can be an important determinant of credit access. Surprisingly little is known, however, about people who lack credit records or scores. This article provides the first documented analysis of the characteristics of consumers without credit records, called "credit invisibles," and of consumers whose records are treated as "unscorable," by a widely used credit-scoring model. Our estimates suggest that 26 million adults, representing about 11 percent of the adult population, lack credit records. An additional 8.3 percent, or 19.6 million adults, have credit records that are unscored. We find that the incidence of having a credit record is not evenly distributed. Young, elderly, minority, and lower-income consumers are more likely to be credit invisible or have an unscored record. In addition, our analysis finds that observable credit performance is not widely available for such consumers, which may hinder the ability of alternative data to expand credit access for these consumers.

Introduction

In the United States, nationwide credit-reporting agencies (NCRAs) compile and sell records that detail the credit histories of millions of consumers.[1] Lenders use these records pervasively to assess creditworthiness when underwriting or pricing credit. They are widely used for purposes beyond credit granting as well. For example, they may be checked when setting auto and homeowner insurance premiums, establishing new utility accounts, renting housing, or hiring new employees. As a consequence, credit records affect the financial well-being of consumers in many ways.

[1] The three NCRAs are Equifax, Experian, and TransUnion.

The widespread use of credit records has drawn the attention of policymakers toward consumers with limited credit histories, meaning either that their credit record contains very little information or that they have no credit record at all. Much of this attention has focused on alternative sources of data that might supplement the information collected by the NCRAs and mitigate the problems that these consumers face. Examples of alternative data that have been suggested include utility payments (Experian, 2014; Schneider and Schutte, 2007; Turner et al., 2006), rental histories (Experian RentBureau, 2014), and remittance histories (CFPB, 2014).[2]

Despite this attention, very little is known about the scale of the problem or about the characteristics of consumers who are affected. Estimates of the number of people without credit records vary widely and the methodology used to produce these estimates has rarely been disclosed.[3] Moreover, the varying estimates of the number of consumers with limited credit histories provide little information about the populations themselves. Yet, such information is crucial for evaluating potential solutions. For example, utility payments may have a lot of value in predicting credit performance, but they can help only consumers with limited credit histories who have utility accounts in their own names.

Our analysis takes the first detailed look at consumers with limited credit histories. We focus on two groups of such consumers. The first group, "credit invisibles," includes consumers without NCRA credit records. These consumers likely face restricted access to credit because lenders cannot use NCRA records to assess their creditworthiness. The second group, the "unscored," consists of consumers whose NCRA credit records cannot be scored by conventional credit-scoring models. Generally speaking, a credit record may be treated as unscorable for two reasons: (1) it contains insufficient information to generate a reliable score, meaning that the record has too few accounts with sufficiently long payment histories; or (2) the information has become "stale," in that the record has no recently reported information.[4] Because many lenders rely on credit scores to assess creditworthiness, an unscored credit record can impair credit access in much the same way as not having a credit record. We present results for both types of unscored credit records, which we refer to as "insufficient-unscored" and "stale-unscored."

Reliable data on the population with limited credit histories are difficult to come by, particularly for the credit invisibles. Although samples of credit bureau data will generally contain information about the number of consumers with unscorable records, by definition, they contain no information on credit invisibles. Credit-record samples also do not contain any information about the demographic or other non-credit-related characteristics of the consumers, making profiling even those with unscored records difficult based on credit-record data alone. Other data sets, such as the Survey of Consumer Finances or the American Community Survey (ACS), that contain representative information about the adult population do not indicate which consumers have limited credit histories and, therefore, by themselves, are of limited use in profiling consumers with limited credit histories.

[2] These data sources have not been without their critics. For example, see Howat (2009).

[3] See, for example, VantageScore (2015), which provides estimates of the number of consumers with different types of limited credit history but does not describe how the estimates were derived. In particular, no information is provided about how fragment files were handled.

[4] Credit records will also generally be treated as unscorable when they indicate that the consumer is deceased. Because our focus is on living consumers with limited credit history, we ignore this cause of an unscored credit record.

Our approach combines multiple sources of data. We start with the Consumer Financial Protection Bureau's (CFPB's) Consumer Credit Panel (CCP), a 1-in-48 random sample of deidentified credit records from one of the NCRAs. These data include the census tract where each consumer resides and a commercially available credit score that indicates whether a credit record was unscorable and, if so, the reason. We compare these data from 2010 with the distribution of the U.S. population from the 2010 census. The difference between a census tract's population and our estimate of the number of credit records in that tract provides an estimate of the number of consumers who are credit invisible. The number of consumers in each tract with an unscorable record can be estimated directly from the CCP. We use these tract-level estimates, along with the demographic characteristics of each tract from the 2010 census and the 2008–2012 ACS, to estimate the demographic characteristics of consumers with limited credit histories.

Using these data, we conducted three related analyses. First, we estimate the number of consumers with limited credit histories and profile some of their demographic characteristics. In producing these estimates, we carefully detail how the estimates were calculated and provide detail on how the underlying assumptions affect the final estimates. Second, we use multivariate analysis to better understand the factors (such as income, education, and living conditions) that may affect the likelihood of having a limited credit history. These results are instructive in helping to identify the potential for different types of alternative data to reduce the problems caused by having a limited credit history. Finally, we use the data assembled in this study to investigate an often-ignored issue to expanding the universe of consumers with scorable credit records, the necessity of having observable performance. Expanding the coverage of credit-scoring models requires more than just alternative data that can serve as predictive factors (or right-hand-side variables) to forecast performance. It also requires observable performance on credit obligations. Because this performance information generally comes from credit-record information, we look at how often such information is available for consumers with limited credit histories.

Background and Data

This section provides background information about the analysis described in this article. We begin by describing the types of information contained in the credit records maintained by the three NCRAs. We then describe the specific sources of data that are used in this study to conduct our analysis.

Credit-Record Background

The credit records assembled by the NCRAs contain detailed information about the past and current credit usage of American consumers. These records include four types of information.[5] The first type of information is "tradelines"—credit accounts voluntarily reported by lenders or loan servicers. Each tradeline contains information about a single credit account that details the date the account was opened, the original amount on the loan, the credit limit (if a revolving account), the current balance, whether the account remains open, and up to 7 years of payment history. The second type is "collections"—accounts reported by third-party debt collectors. Although some

[5] For a more indepth discussion of the types of information included in credit records, see Avery et al. (2003).

collection accounts derive from credit accounts, most of the reported collections are for non-credit-related items, such as unpaid medical or cell phone bills. The third type of information is "public records," such as bankruptcy filings and tax liens. The final type is "inquiries"—records created by the NCRAs whenever a consumer's credit record is accessed in connection with an application for credit.[6]

The credit records maintained by the NCRAs contain nearly comprehensive information about many mainstream credit products, including auto loans, mortgages, and credit cards. Largely missing from this information, however, are accounts with nontraditional credit sources such as payday or auto-title lenders and pawnshops. Non-credit-related bills, like medical and utility bills, are sometimes reported to the NCRAs, although such reporting is rare and often limited to reporting by debt collectors.

Any one of the four types of information, by itself, is sufficient to trigger the creation of a credit record. So, a consumer can have a credit record with as little information as a single inquiry. The four information types, however, are not equally informative. Inquiries, although they can be treated as negative risk factors in credit-scoring models, contain little information about past credit experiences. Collections and public records both provide only information about negative experiences (although if these have been paid in full, they may be less negative than those that remain unpaid). Establishing a positive credit history requires having at least one tradeline that has been reported to the NCRAs with a long enough history to reflect either positive or negative payments.

An important challenge in working with credit-record data is dealing with so-called "fragment files"—credit records that contain a portion of a consumer's credit history that exists outside the consumer's primary file. For example, a consumer with a credit record opens a new credit card. When the lender reports that account, the NCRA attempts to match it with the correct credit record. If the NCRA is unable to find a match or finds multiple matches, perhaps reflecting erroneous or incomplete information reported with the new account, then the newly reported credit card will be placed in its own credit record. Most fragment files are temporary. Over time, as more information comes in, the NCRA may realize that the accounts in a fragment file belong to a consumer with an existing credit record. When this happens, the fragment file and all the information it contains will be subsumed into the consumer's primary credit record.

The existence of fragment files suggests that some consumers will have multiple credit records. Left unaddressed, the presence of fragment files will cause the number of credit invisibles to be understated. Moreover, because many of these fragment files will be unscorable, reflecting their limited contents, failing to exclude fragment files will overstate the number of consumers with unscorable credit records. In the next section, we describe in detail the steps taken to prune fragment files from our sample.

[6] The types of inquiries used in this article are also referred to as "hard inquiries." Inquiries can also be created for other reasons, such as when credit records are accessed to solicit new business, for account maintenance purposes, or for other reasons. Such "soft inquiries" are not included in the CCP and are not supplied to lenders who receive credit records.

Data

The analysis described in this paper proceeds by comparing the geographic distribution of records in the CFPB CCP with the distribution of demographic characteristics of the population of adults in the United States from multiple sources. In this section, we provide background information about the CCP and describe the sources of demographic information that we use.

CFPB Consumer Credit Panel

The primary source of data we use is the CFPB's CCP. The CCP is a nationally representative, 1-in-48 sample of deidentified consumer credit records from one of the NCRAs. We use archive data from December 2010, which provides a representative sample of credit records as they existed at that time.

Each credit record contains deidentified information about the consumer's credit history, including information on each account's type, the date it was opened, outstanding balance, payment history, and current status. The CCP includes, when available, the consumer's year of birth.[7] We calculate each consumer's age at the end of December 2010.[8]

As shown in Exhibit 1, the CCP data for 2010 include about 4.96 million records. From these, we exclude records that indicate the consumer was deceased or living outside the United States. These exclusions make the composition of sample credit records more comparable with the census population.

We also attempt to remove fragment files using three methods. First, we exclude credit records that were consolidated into other credit records during the next 4 years (through December 2014, which was the most recently available data at the time of this analysis). When two credit records are consolidated, the newer of the records, the fragment, is destroyed and the older record remains (with the additional information absorbed from the fragment). Dropping the newer files removes an additional 2.8 percent of sample records.

Exhibit 1

Count of Excluded Records by Reason

	Observation Count	Percent of Sample	Stale-Unscored (%)	Insufficient-Unscored (%)
Total observations	4,956,746	100	5.1	6.9
Exclusions				
Outside United States	43,828	0.9	5.0	2.4
Deceased	179,279	3.6	4.1	2.1
Consolidated	138,152	2.8	18.1	35.9
Disappeared	104,575	2.1	9.2	66.4
Age missing	153,308	3.1	4.9	6.6
Bad geography	2,804	0.1	6.5	8.6
Sample observations	4,334,800	87.5	4.6	4.8

Source: Consumer Financial Protection Bureau, Consumer Credit Panel

[7] Actual credit records include the date of birth. The CCP excludes the month and day of birth to enhance the anonymity of the data.

[8] Although the ages we calculated are not perfectly comparable with the age data from the 2010 census, which asks consumers their age as of April 2010, the differences should be negligible, particularly because our analysis primarily uses 5-year age buckets.

Second, we exclude credit records that were destroyed between 2010 and 2014, despite not being involved in a consolidation. Although these excluded records were not absorbed by another file, their disappearance suggests that the information they contained was removed, which resulted in the destruction of the file. Although this disappearance could reflect information that has become so old it has migrated off the credit record (such as information on delinquent accounts that is required by law to be removed after 7 years), most of these credit records were recently reported in December 2010. The recent reporting suggests that most of these records were not destroyed because the information aged.[9] Instead, we believe the disappearance of these records likely reflects erroneous information that was subsequently re-reported by the data furnisher and correctly associated with the proper file. This exclusion removed another 2.1 percent of credit records.

Third, we exclude credit records without a year of birth. To help ensure that the lack of a reported year of birth was not a temporary characteristic of these records, we supplement the year of birth information in 2010 with the information from 2014 and exclude only those records missing years of birth in both periods. The absence of this information suggests that these are fragment files created because of incomplete information that prevented successfully assigning the information in these records with the right consumer's primary credit record. Consistent with this theory, most of these credit records involve authorized user accounts. As described in detail by Brevoort, Avery, and Canner (2013), authorized users are people who are permitted to use a revolving account (normally a credit card), but who are not legally liable for any charges incurred. Because they are not liable for the charges, lenders may collect only partial information, which inhibits the ability of the NCRA to match the account information to the correct credit record. Excluding these records removes 3.1 percent of the sample.

After these exclusions, we are left with a sample of about 4.3 million credit records.[10] Although we think these data restrictions provide the best available measure of the number of consumers with credit records, we may be excluding a nontrivial number of primary credit records. If so, our estimate of the number of credit invisibles will be overstated and, because many of the excluded credit records are unscorable, our estimate of consumers with unscored records would be understated. It is also likely, however, that some of the credit records that remain in the sample are themselves fragment files. For example, we have opted not to exclude credit records containing only collection accounts or public records. Although some of these are likely fragments, we concluded that they were more likely primary files. Nevertheless, to the extent that these credit records include a material number of fragment files, our estimate of the number of credit invisibles will be understated and the number of unscorables overstated.

For each credit record in our sample, we determine whether the sample record contained a credit score. For records without a score, a code was provided indicating whether the record was insufficient-unscored or stale-unscored. The exact definition of what makes a credit record insufficient or stale

[9] Additional evidence that information aging was not a significant cause of the disappearance of these records is provided by the fact that two-thirds of the records that were excluded because they had disappeared were insufficient-unscored compared with less than 10 percent that were stale-unscored. If account-information aging was an important cause of the disappearance of these records, a much larger share should have been stale-unscored in 2010.

[10] We also exclude from the sample a small number of records that had either missing or invalid census tract information. Excluding these records removed 0.1 percent of the sample.

differs across credit-scoring models, because each model uses its own proprietary definition. Our analysis is based on the commercially available credit-scoring model that generated the scores included in the CCP. We think this credit-scoring model uses a relatively narrow definition of a scorable credit record, but a definition that is consistent with most credit scores in use today.

Using the CCP data, we estimate the number of consumers in each census tract whose credit record was insufficient-unscored, stale-unscored, or scored by multiplying the number of sample credit records in each tract by 48 to account for the sampling rate. We then estimate the number of credit invisibles in each tract as the difference between the adult population of the census tract from the 2010 census and our estimate of the number of consumers with credit records. We calculate these totals for each of 13 different age categories, discussed in more detail in the next section.

Demographic Data

The credit-record data contained in the CCP contain no demographic information other than age. To develop our profile of consumers with limited credit history, we supplement the CCP data with information from the 2010 census and the 2008–2012 ACS.

From the 2010 census, we use information about the racial and ethnic composition of each census tract. We calculate the share of the population in each tract that was in each of the following groups: Hispanic or Latino ("Hispanic"); non-Hispanic Black or African-American ("Black"); non-Hispanic Asian ("Asian"); non-Hispanic White ("White"); and other non-Hispanic ("Other"), which includes American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and multiracial individuals.

We also use data from the 2010 census on the share of the population that lives in group quarters. We calculate the share of the population that was living in college or university student housing, correctional facilities for adults, military quarters (nondisciplinary), and nursing facilities.

Additional demographic information was taken from the 2008–2012 ACS. To better understand the relationship between the likelihood of having a limited credit history and income level, we use ACS data to calculate the "relative income" of each tract. Relative income is the ratio of the median household income in the tract and the median household income of the surrounding area. The surrounding area is defined as the metropolitan statistical area (MSA) for tracts in MSAs or the tract's county otherwise. Following the definitions used by the Community Reinvestment Act, a tract is considered "low income" if its relative income is less than 0.5, "moderate income" if it is between 0.5 and 0.8, "middle income" if it is between 0.8 and 1.2, or "upper income" if it is 1.2 or higher.

We also use the ACS to calculate the share of adults in each tract by citizenship status (native, foreign-born citizen, and noncitizen) and for five levels of education (less than a high school diploma, high school, some college, bachelor's degree, or graduate degree). ACS data also provide the share of consumers living below the poverty level, the share who speak a language other than English at home, and the share who moved in the past year. Finally, we use the ACS to provide information about the share of households in the tract across four different types: married-couple families, other families, nonfamily households, and single-person households.

The analyses using these demographic data, which are described in more detail in the next section, are conducted separately for the 13 different age groups shown in the left column of exhibit 2. Most of the demographic variables included in this study are not available at the tract level for each age group. For example, the population living in group quarters is provided only at the tract level for two adult age groups: 18 to 64 and 65 and older. In other cases, such as income and household type, which are both calculated at a household level, tract-level information was not available broken out by age at all. In such cases, we use the narrowest age group available for each of 13 age categories. A list of demographic variables (some expressed as variable groups), along with the age group mappings, is provided in exhibit 2. A complete list of variables along with selected summary statistics are provided in exhibit 3.

Exhibit 2

Age Groups of Explanatory Census Variables

Age Group	Moved Last Year	Race/ Ethnicity	Percent Below Poverty	Education	Group Quarters	Non-English Speaking	Citizenship
18–19	18–19	18–19	18–24	18–24	18–64	18–64	18+
20–24	20–24	20–24					
25–29	25–29	25–29	25–34	25–34			
30–34	30–34	30–34					
35–39	35–39	35–39	35–44	35–44			
40–44	40–44	40–44					
45–49	45–49	45–49	45–54	45–64			
50–54	50–54	50–54					
55–59	55–59	55–59	55–64				
60–64	60–64	60–64					
65–69	65–69	65–69	65–74	65+	65+	65+	
70–74	70–74	70–74					
75+	75+	75+	75+				

Sources: 2010 census; 2008–2012 American Community Survey 5-year data

Exhibit 3

Sample Summary Statistics (1 of 2)

Variable	Mean	Median	Standard Deviation
Race/ethnicity			
Asian	4.8	1.6	9.3
Black	11.6	3.3	19.8
Hispanic	14.2	5.1	20.9
White	67.0	77.6	29.9
Other	2.3	1.5	4.7
Citizenship status			
Native citizen	84.3	91.0	17.3
Foreign-born citizen	7.2	4.1	8.6
Noncitizen	8.5	4.1	11.0
Moved in last year	16.0	10.3	17.6
Relative household income			
Lower income	0.1	0.0	0.2
Moderate income	0.2	0.0	0.4
Middle income	0.4	0.0	0.5
Upper income	0.3	0.0	0.5
Poverty rate	13.5	9.2	14.3

Exhibit 3

Sample Summary Statistics (2 of 2)

Variable	Mean	Median	Standard Deviation
Education			
Less than high school	14.5	10.6	13.7
High school diploma	28.4	28.2	13.6
Some college	31.1	30.0	13.6
Bachelor's degree	16.7	14.0	12.6
Graduate degree	9.3	6.2	10.3
Group quarters			
College	1.0	0.0	6.7
Correctional	0.8	0.0	5.8
Military	0.1	0.0	2.9
Nursing	0.6	0.0	3.1
Household type			
Married-couple family	49.5	51.0	16.0
Other family	17.7	15.7	9.6
Nonfamily	6.1	4.8	5.7
Living alone	26.6	25.4	11.2
Credit record type			
Scored	80.6	77.7	47.7
Limited credit history	19.4	22.3	47.7
Insufficient-unscored	4.2	0.0	11.6
Stale-unscored	4.1	0.0	9.6
Credit invisible	11.0	14.3	49.3

Notes: Summary statistics are calculated across census-tract/age-group pairs, with each observation weighted by the population in that census tract/age group. Variable values reflect the percentage of the population in that age group (for example, the percentage of the population that is Asian or is a native citizen), except for the relative household income variables, which are dummy variables reflecting the household income level of the tract, and the household type variables, which reflect the share of households in the tract.

Sources: 2010 census; 2008–2012 American Community Survey 5-year data

Who Has Limited Credit History?

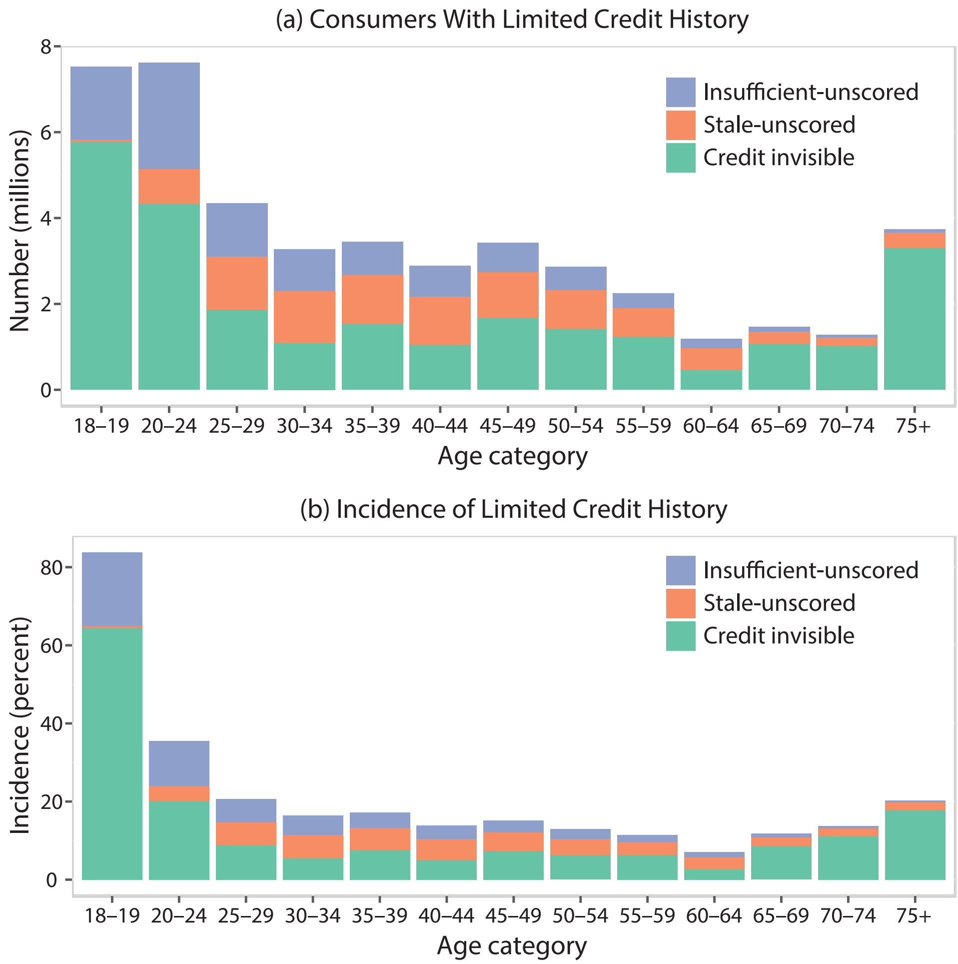
The data assembled for this study indicate that 45 million adults in the United States have a limited credit history.[11] This figure includes the 26 million credit-invisible adults in the United States who lack a credit record, representing about 11 percent of the adult population. It also includes 19.4 million people, or 8.3 percent of the adult population, who have unscored credit records, which are nearly evenly split between those whose records are insufficient-unscored (9.9 million) and stale-unscored (9.6 million). The remaining adult population of 188.6 million has scored credit records.

The likelihood of having a limited credit history varies significantly by age. As shown in panel (a) of exhibit 4, most consumers with limited credit histories are either younger than 30 or older than 74. This pattern is generally consistent with patterns of credit usage by age in the Survey of Consumer Finances (Bucks et al., 2009). Limited credit histories appear to be found among the young in particular. Consumers younger than 30 account for one-third of adults with limited credit histories. Moreover, as shown in panel (b), 18-to-19-year-olds are significantly more likely to have a limited credit history than any other age group.

[11] A preliminary version of the analysis in this section was originally released as a CFPB Data Point (Brevoort, Grimm, and Kambara, 2015).

Exhibit 4

Distribution and Incidence of Limited Credit History by Age

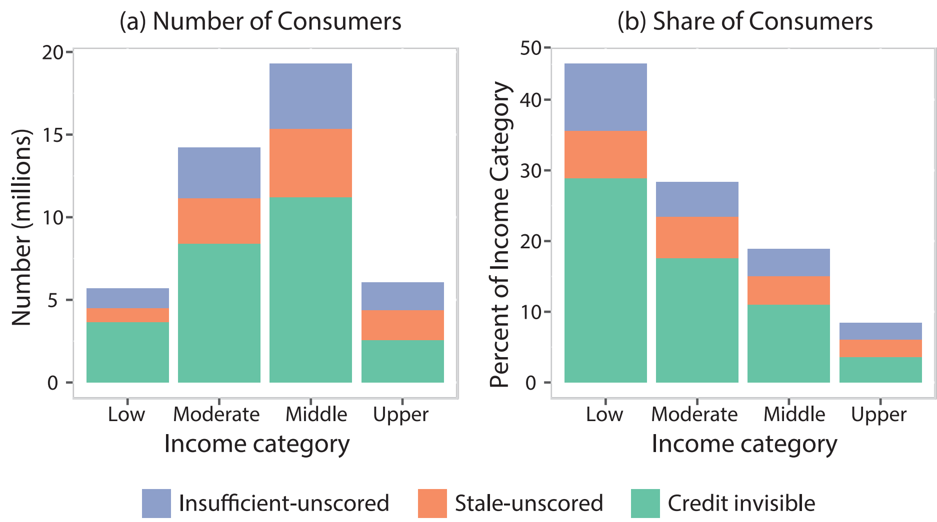


The age-related patterns vary by the type of limited credit history. The incidence of being credit invisible largely mirrors the general pattern for limited credit histories, with older and younger consumers being more likely to be credit invisible. For consumers with unscored credit records, the patterns are different. The incidence of having an insufficient-unscored credit record decreases with age, and the incidence of having a stale-unscored record is highest for middle-aged consumers (ages 30 to 49) and notably lower for younger or older consumers.

Income also appears to be highly related to the likelihood of having a limited credit history. Exhibit 5 shows both the distribution of consumers with limited credit histories by the relative income level of their tract in panel (a) and the incidence of having a limited credit history for each income level in panel (b). A little more than one-half of consumers with limited credit histories

Exhibit 5

Distribution and Incidence of Limited Credit History by Income



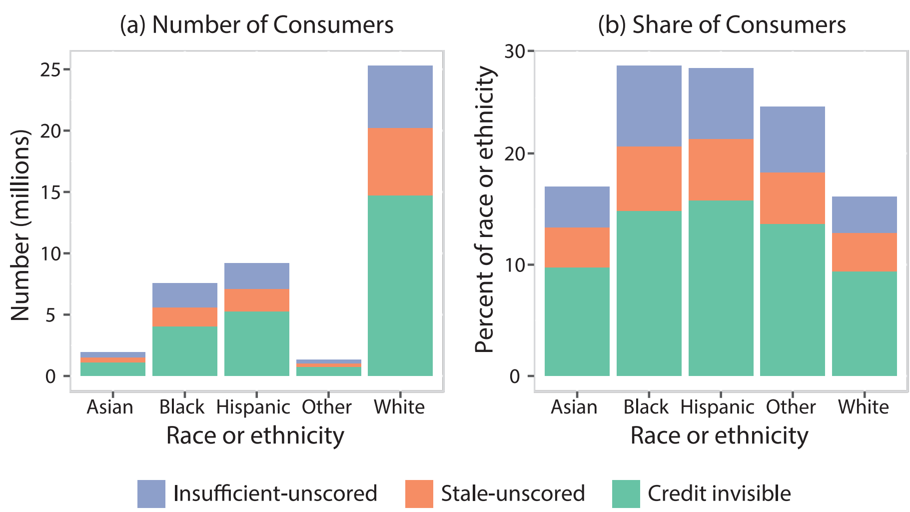
live in middle- or upper-income neighborhoods. This statistic is not surprising, given that these neighborhoods are home to most adults in the United States. The incidences, however, show that consumers in lower-income neighborhoods are much more likely to have a limited credit history. Indeed, 30 percent of consumers in low-income neighborhoods are credit invisible and an additional 15 percent have an unscored record.

There also appear to be significant differences in the likelihood of having a limited credit history by race or ethnicity. Assuming that within each tract, each racial or ethnic group has the same likelihood of being credit invisible or having an unscored credit record, we can estimate the number of consumers of each racial or ethnic group with a limited credit history. The results of these calculations are provided in exhibit 6, which shows the distribution and incidence of having a limited credit history by race or ethnicity. A higher proportion of Black and Hispanic consumers have limited credit histories than do either Asian or White consumers, who have similar incidences of being credit invisible or having an unscorable credit record. Although the incidence of having a limited credit history in general is similar for Black and Hispanic consumers, Hispanic consumers are more likely to be credit invisible (by about 1 percentage point) and less likely to be unscorable.

These univariate patterns in the incidence of having a limited credit history across age, income, and race or ethnicity suggest that the problems associated with limited credit histories are borne unevenly across these groups of consumers. In the next section, we employ multivariate analyses to better understand how these and other characteristics are associated with the likelihood of having a limited credit history.

Exhibit 6

Distribution and Incidence of Limited Credit History by Race or Ethnicity



Factors Associated With Limited Credit History

To better understand the factors that are associated with the likelihood of having a limited credit history, we conduct a multivariate analysis that exploits variation across tracts. We examine how the share of the population in a tract with limited credit history $\boldsymbol{h}$ varies with the demographic characteristics of consumers in the tract. Using $\boldsymbol{t}$ to index census tracts and $\boldsymbol{j}$ to index the 13 age categories, we estimate equations of the form

$$L_{tj}^{h} = \alpha_{jc}^{h} + R_{tj}\beta_{j}^{h} + I_{tj}\,\delta_{j}^{h} + X_{tj}\,\gamma_{j}^{h} + \epsilon_{tj}^{h}, \qquad (1)$$

where L_{tj}^{h} is the percentage of population in tract $\boldsymbol{t}$ in age group $\boldsymbol{j}$ that has the type of limited credit history indexed by $\boldsymbol{h}$. R_{tj}, I_{tj} and X_{tj} are row vectors with population characteristics described in more detail in the following paragraphs. β_{j}^{h}, δ_{j}^{h}, and γ_{j}^{h} are coefficient vectors to be estimated, ϵ_{tj}^{h} is an i.i.d. error term, and α_{jc}^{h} is a county-level fixed effect.

The first row vector of population characteristics, R_{tj}, contains variables related to the race, ethnicity, or national origin of the tract's population. This vector includes variables that reflect the percentage of the population that is in each of the five racial or ethnic groups described in the previous section (with the White group serving as the omitted group). We include these variables to better understand how limited credit histories are associated with race or ethnicity after controlling for other observable tract-level factors.

Avery, Brevoort, and Canner (2012) found that the credit scores of immigrants (in particular, recent immigrants) tend to understate their creditworthiness. This result derived from immigrants having shorter credit histories reflected in NCRA files than natural-born citizens have. This result suggests that tracts with relatively more immigrants should have a higher incidence of limited credit histories. To test for this pattern, we include in $\boldsymbol{R_{tj}}$ the percentage of the population in a tract that consists of foreign-born citizens or noncitizens. We also include the percentage of the population that speaks a language other than English at home.

The second vector, $\boldsymbol{I_{tj}}$, contains information about the income and education levels of consumers in the tract. We measure income using the four relative income levels discussed in the previous section and include in $\boldsymbol{I_{tj}}$ dummy variables that reflect whether the tract was low, moderate, or upper income (middle income is the omitted category). Because higher-income individuals tend to have greater access to credit, we would expect income to be negatively associated with limited credit histories.

We would also expect the education levels of the population to be similarly related to the incidence of limited credit histories. To test for this relationship, we calculate the percentage of the tract's adult population at each of five education levels: less than high school, high school diploma, associate's degree or some college, a bachelor's degree, or a graduate degree. In the estimations, the percentage of the population with a high school diploma is the omitted category. We would expect education to be negatively related to limited credit history.

The final vector, $\boldsymbol{X_{tj}}$, contains information about the living arrangements of consumers in the tract. This vector includes four variables that measure the percentage of the tract's population residing in four types of group quarters: college dormitories, correctional facilities, military quarters, and nursing homes. Consumers in these different living arrangements may differ from the rest of the population in their credit usage patterns, which could affect their likelihood of having a limited credit history.[12]

We also include in $\boldsymbol{X_{tj}}$ variables relating to the type of households in each tract. These variables include the percentage of households comprising a single person living alone, a married-couple family, a non-married-couple family, and nonfamily households (with the percentage of households comprising single persons serving as the omitted category). Including this set of explanatory variables serves two purposes. First, living conditions may affect credit usage patterns. For example, students who continue to live with their parents might have less cause to establish a credit history than consumers of the same age who are living independently.

The second purpose is to gain some insight about the potential for alternative data to enhance the credit records of consumers with limited credit histories. As discussed previously, two of the most commonly cited sources of alternative data are rental histories and utility payments. Although several studies have explored the predictive value of this information, no study that we are aware of has addressed how much of the population with limited credit histories might be helped. Even if rental histories or utility payments are highly predictive of future credit performance, unless a significant share of the population with limited credit histories has rent or utility payments in

[12] An alternative possibility is that the address on file at the NCRA may not be the actual address for consumers living in group quarters. For example, if the percentage of the population that lives in a correctional facility is positively correlated with the number of credit invisibles, this could indicate that prisoners are less likely to have credit records; however, it could also reflect a mismatch between the address in the credit record and the address information collected by the Census Bureau. Such a mismatch could result if prisoners in correctional facilities do not fill out change-of-address forms upon being incarcerated. If so, this would be expected to increase the number of credit invisibles in a tract, although it should not affect our estimate of the number of consumers with unscored credit records (which are directly observed in the CCP).

their own name, the potential of these data sources to help this population will be limited. We conjecture that people who live alone are more likely to be making rental payments and to have utility payments in their own names than are consumers in other household situations. To the extent that other household types are more prevalent in areas with higher incidences of limited credit histories, rental histories and utility payments may have less potential to provide information about consumers with limited credit histories.

We estimated equation 1 for each of the 13 age groups using the percentage of each tract's population with a limited credit history as the dependent variable. To facilitate the comparison of the estimated coefficients across age groups, we present the results graphically in exhibit 7. In appendix A, we also present the results obtained from estimating equation 1 for each type of limited credit history (that is, credit invisible, insufficient-unscored, and stale-unscored) separately.

Exhibit 7

Coefficient Estimates, Limited Credit History

The results of these estimations are largely consistent with our expectations. A positive correlation appears to exist between the percentage of the population that is Black or Hispanic and the percentage of the population with a limited credit history. We were somewhat surprised to find that census tracts with larger elderly Asian populations tend to have a higher incidence of elderly consumers with limited credit histories, although little relationship appears to exist between the Asian share of the population and limited credit history at younger ages.

Consistent with the results by Avery, Brevoort, and Canner (2012), we find that the percentage of noncitizens in a tract is associated with a higher incidence of having limited credit history for most age ranges. We were somewhat surprised to find that the percentage of the population composed of foreign-born citizens is negatively related to the percentage of the population with a limited credit history. We find little consistent relationship between the likelihood of having a limited credit history and either the percentage of the population that speaks a language other than English at home or the percentage of the population that moved in the past year, although moving in the past year appears to be associated with a higher incidence of having a limited credit history for young consumers.

As expected, both income and education appear to be important factors associated with having a limited credit history. The incidence of having a limited credit history is significantly higher in low- and moderate-income tracts for consumers age 30 or older. Consumers in upper income tracts appear to have a persistently lower likelihood of having a limited credit history. Moreover, tracts where a larger percentage of consumers have spent time in college tend to have lower incidences of limited credit history, and tracts with more consumers with less than a high school education have significantly higher incidences, particularly in the middle-age estimations.

The percentage of the population that lives in group quarters also appears to be strongly related to the incidence of having a limited credit history. Incidences are notably higher in tracts with more people living in correctional facilities, particularly among the young and middle aged. The percentage of the population in college dormitories or in military housing also appears to be related to having a limited credit history, although the direction of these effects changes across ages. Both are positively associated with having a limited credit history for younger consumers and negatively associated for older consumers. The percentage of consumers living in nursing homes is also positively associated with the limited credit history for older consumers. These results suggest that consumers in these environments (college, prison, military service, and nursing homes) are more likely to have a limited credit history. Because these populations tend to be small relative to the entire population, however, these populations likely account for only a small share of the total population with limited credit history.

Finally, a significant relationship appears to exist between living arrangements and the incidence of a limited credit history. Compared with the omitted group—the percentage of households composed of single adults—a larger percentage of households involving nonfamily members was associated with a higher incidence of limited credit history among younger consumers. A greater percentage of family households not including a married couple similarly was associated with a higher incidence of limited credit history among middle-aged consumers. By contrast, married-couple family households were associated with a lower incidence of limited credit history among young consumers.

By themselves, these results cannot establish that rental and utility histories will be insufficient to score the credit records of consumers with limited credit histories. Nevertheless, to the extent that consumers with limited credit histories do not have rental or utility payment information that might be used to supplement their credit records, we would expect them to live in non-married-family or nonfamily households. The fact that the incidence of limited credit histories is higher in areas with a larger percentage of these households suggests that there may be a significant portion of the population that would continue to fall through the cracks in the credit-reporting system even after rental and utility payment histories were incorporated. Additional research is necessary to establish the extent to which these forms of alternative data may help alleviate the problem of limited credit histories.

The Challenge of Assessing the Accuracy of Models Using Alternative Data

As discussed previously, credit records will not be scored when they have characteristics that the model's builders considered "unscorable." The term *unscorable*, which is widely used to refer to records that remain unscored, is somewhat misleading. Credit scores *could* be empirically derived for any credit record using the same model-building techniques that generate standard credit-scoring models. In fact, scores could be generated for consumers without credit records by, for example, estimating a scoring model that includes only an intercept using an estimation sample of credit records created after the start of the performance period used in model development.[13] The reasons these records remain unscored, therefore, go beyond a lack of explanatory variables, which is the problem that alternative data is meant to alleviate.

Among the most important reasons relates to the difficulty in assessing the credit performance of consumers with insufficient, stale, or nonexistent credit records. Consumers with such records tend not to have outstanding credit accounts on which performance can be evaluated. This lack of accounts with observable performance is almost tautological, because the records of these consumers would likely be considered scorable if they had such accounts. The lack of observable performance makes building and validating a credit-scoring model much more difficult.

For example, consider the case of stale-unscored records. These records have enough credit history to be scored (otherwise, they would have been insufficient-unscored), but their lack of recent updates suggests that they are unlikely to have active accounts on which to assess performance. Any attempt to build or validate a model for stale-unscored records would be limited to only those records with observable performance. But after it is estimated, the model would score all stale-unscored records.[14]

[13] Of course, this sample of credit records would have to contain some created early enough in the performance period to have credit accounts with observable performance. Such credit records could occur as a result of consumers opening their first accounts during the performance period or as a result of the reporting of accounts that had not been reported previously to the NCRAs.

[14] Model builders could segment the population of stale-unscored records based on observable characteristics; for example, by creating a scorecard for people with "moderately" instead of "severely" stale records. In this case, the model could be limited to that subset—in this case, moderately stale records. The underlying point, however, remains valid: models that are estimated on a portion of a subset of the population with observable performance will score the entire subset. A model built for the moderately stale will score all moderately stale records, even if only a small subset is expected to have observable performance.

Having observable performance for a small and possibly unrepresentative share of the sample leads to a well-known problem. Records with observable performance may misrepresent the performance that is observed when the model is deployed. Consumers with observable performance were able to find willing lenders, perhaps based on characteristics not observable in credit records or on the strength of co-applicants. By contrast, consumers who wanted credit but could not find willing lenders, perhaps because of weaker unobservable characteristics, will not have observable performance.[15] Putting the model into practice alters the ability of consumers to obtain credit, possibly enabling some consumers to borrow who otherwise would have been unable to find willing lenders. The result could be default rates that are higher than were expected based on the experience of consumers with observable performance before the model was implemented.

One way that lenders respond to this bias in performance is to employ "credit overlays,"[16] which are restrictions that lenders impose in addition to credit score cutoffs. For example, VantageScore (2015) suggests mortgage lenders typically consider only applicants whose credit records have at least three tradelines with updates during the past 6 months, regardless of their credit score. To the extent that lenders employ credit overlays, the effects of expanding the number of consumers with scored credit records will be reduced. Even if scores are generated for consumers with limited credit histories, lenders' credit overlays may prevent credit from being extended to these populations.

A credit score, therefore, is not necessarily a sufficient condition for improved credit access for consumers with limited credit histories (even for those consumers whose new score generally would be considered prime). One must also establish that the new scores can accurately reflect the creditworthiness of such consumers when the models are deployed; otherwise, lenders will likely use credit overlays. Being able to demonstrate that little bias exists in using observable performance is an important aspect of establishing that the new scores can accurately reflect the creditworthiness of such consumers when the models are deployed.

To shed some light on the extent to which performance is observable for consumers with limited credit histories, we use the December 2012 CCP archive to calculate credit performance for the 2-year period immediately following the rest of the data in this study. When estimating or validating generic credit history models, 2 years is a commonly used performance period (Board of Governors of the Federal Reserve System, 2007). Performance is measured using an "any account" performance measure, which includes performance on accounts that were open and in good standing at the start of the performance period or that were opened during the first 3 months of the performance period. For consumers with observable performance, we determine whether the person was 90 or more days past due on any credit obligation during the performance period. We use this performance definition to construct our measure of delinquency.

[15] The population with unobservable performance will also include those consumers who did not want credit during the performance period. Because consumers who do not demand credit are unlikely to start when the model is applied, for simplicity, we ignore the portion of consumers who lacked observable performance because of a lack of demand for credit. This population also, however, should raise some concern about the accuracy of models estimated on observable performance.

[16] An alternative method is to employ "reject inference" methods when developing the scoring model. It is unclear, however, how successful these efforts can be (Crook and Banaskik, 2004; Hand and Henley, 1993), and to the best of our knowledge these methods are not widely used in estimating the generic scoring models explored in this study.

Exhibit 8, which illustrates the performance measures calculated for the population of consumers with credit records, shows that nearly 90 percent of consumers with scored credit records had observable performance during the ensuing 2 years. The delinquency rate for consumers with observable performance was about 12 percent. Exhibit 8 also shows the performance for the scored population broken down by the number of tradelines contained in the credit record. “Thick” files are those with at least three tradelines and “thin” files are those with two or fewer tradelines. For both groups, performance is available for most records and delinquency rates are similar to those of the overall scored population (although the delinquency rate for the thin-file population is a bit lower).

Consumers with unscored credit records were much less likely to have observable performance. Only 21.8 percent of insufficient-unscored and 12.3 percent of stale-unscored consumers had observable performance. Delinquency rates for these consumers were also notably higher than they were for consumers with scored credit records. The relatively high delinquency rates are not necessarily a problem, providing that the alternative data can adequately predict the likelihood of delinquency, although they do suggest that any model estimated for this population will likely produce scores for these populations that are below those of consumers with currently scored records.[17] These numbers, particularly the relatively lower shares of consumers with observable performance, also help explain why the people who built the credit-scoring model that produced the scores used in this study considered these records to be unscorable.

Absent from the numbers in exhibit 8 are the credit invisibles. Because we had no data on these consumers from December 2010, it is not possible to determine how many of these consumers had performance during the next 2 years. One approach would have been to identify all credit records that appear in the December 2012 archive that did not exist in December 2010 and assume that these were the credit records of previously credit-invisible consumers. The problem with this approach is that we expect most of the newly created records in December 2012 to be fragments. As such, this approach would overcount the number of credit invisibles who obtain credit records in the ensuing 2 years. Although we could have attempted to filter out the fragments using methods similar to those we used for the December 2010 data, at the time of this study we did not have access to a comparable 4 years of data.

Instead, we take the population of consumers with a credit record in December 2010 and identify which of those records did not exist 2 years earlier. Of the 4.3 million records from 2010, 156,269

Exhibit 8

Incidence of Performance for Scored and Unscored Consumers

Population	Number (millions)	Share With Performance (%)	Delinquency Rate (%)
Scored	188.7	89.5	12.2
Thick file	180.7	90.2	12.4
Thin file	8.1	72.9	6.3
Stale-unscored	9.6	12.3	26.0
Insufficient-unscored	9.9	21.8	22.4

[17] Indeed, the average delinquency rate for these consumers is consistent with a sub-600 credit score, suggesting that any scores generated for these consumers will be disproportionately subprime. For these consumers, credit access may still be very limited, even if their credit records were to become “scorable” with alternative data.

records did not exist in 2008, and, of those records, only 11,738 (7.5 percent) had observable performance during the 2-year performance period from January 2009 to December 2010, with a delinquency rate of 17.8 percent. Using these sample percentages to construct population estimates, we estimate that only about 0.5 million consumers who were credit invisible in December 2008 had observable performance. Unless the number of credit invisibles in 2008 was substantially lower than the 26 million who were credit invisible in 2010, observable performance is potentially available for only a very small portion of credit invisibles.

The lack of performance data for credit invisibles and consumers with unscored credit records suggests that efforts to expand the universe of scored credit records will likely be hampered by a lack of observable performance data with which to estimate credit-scoring models.

Discussion

The data assembled for this analysis suggest that about 45 million adult consumers in the United States are credit invisible or have a credit record that is considered to be unscorable by a widely used credit-scoring model. As a result, these consumers likely face impaired credit access. In addition to the direct consequences that impaired credit access has for these consumers, it will also make establishing a credit history more difficult, potentially perpetuating the problem. For those consumers who are able to obtain credit despite their limited credit histories, credit costs will likely be higher as a result of the limited history, which could increase their likelihood of default and increase the likelihood of establishing a negative credit history.

Efforts to help consumers with limited credit histories have focused on forms of alternative data that might be used to supplement NCRA credit records. In general, these studies have sought to establish that specific forms of alternative data are predictive of future credit performance, which would indicate that alternative data provide valuable additional information. Although these studies are useful in establishing the types of data that might help alleviate the problems associated with limited credit histories, they have largely ignored two issues that might limit the effectiveness of these sources of alternative data.

The first issue is that alternative data will be useful only in alleviating the problems of limited credit history to the extent that people with limited credit histories have utility accounts or rental agreements in their own names (in the case of rental or utility payment histories) or have otherwise engaged in the activities (such as remittance histories, checking accounts, or even social media) that alternative data may reflect. Our results suggest that these forms of alternative data (rental or utility payments) may be able to supplement NCRA information for many consumers; nevertheless, our results also suggest that a significant number of consumers may be in housing situations that would not generate rental or utility histories for these consumers.

The second issue is specific to the use of credit records in credit-scoring models. Scoring models estimate conditional correlations between credit-record information (plus any alternative data) and subsequent credit performance. Having observable performance with which to build and validate a model using alternative data is difficult, given that this performance is observed in credit records. Although alternative data can expand the scope of information that might be related to performance, it cannot expand the number of consumers for whom performance is observed (at least without altering the type of credit performance the score is meant to predict). When credit performance is not observed for a sufficiently representative sample of consumers, scoring models can produce biased estimates of creditworthiness. If lenders respond to this bias by employing credit overlays, as many do today, then producing scores for consumers with limited credit histories will do little to enhance their access to credit.

To date, these two issues have gone largely unmentioned by studies that examine the potential of alternative data to alleviate the problems of limited credit histories. Future research on these topics should confront these issues directly. If the goal is to expand credit access to a significant portion of consumers with limited credit histories, our results suggest that it is not enough for alternative data to produce high goodness-of-fit measures when added to a credit-scoring model. Studies also need to evaluate how widely available those data are for the population of consumers with limited credit histories and establish that any statements about the predictiveness of alternative data are based on a sufficiently representative sample of consumers with limited credit histories. Our results suggest that these limitations may be significant hurdles for most types of alternative data.

Appendix A

This appendix presents the results of estimating equation 1 for each of the three types of limited credit history: credit invisible, insufficient-unscored, and stale-unscored. Like the results for all types of limited credit history, we present the results graphically. Exhibit A-1 shows results for the incidence of being credit invisible, exhibit A-2 shows results for insufficient-unscored, and exhibit A-3 shows the results for stale-unscored.

Looked at separately, these results can be more difficult to interpret. Any factor that is positively correlated with being insufficient-unscored must be negatively correlated with the (sum of) other types of limited credit history. As a result, factors that appear to be positively related to one of the types of limited credit history will tend to have the opposite effect on at least some of the other types, which is the reason we focused on the results for all types of limited credit history earlier.

Nevertheless, these results may be helpful in identifying specific characteristics that lead to particular types of limited credit history.

Exhibit A-1

Coefficient Estimates, Credit Invisible

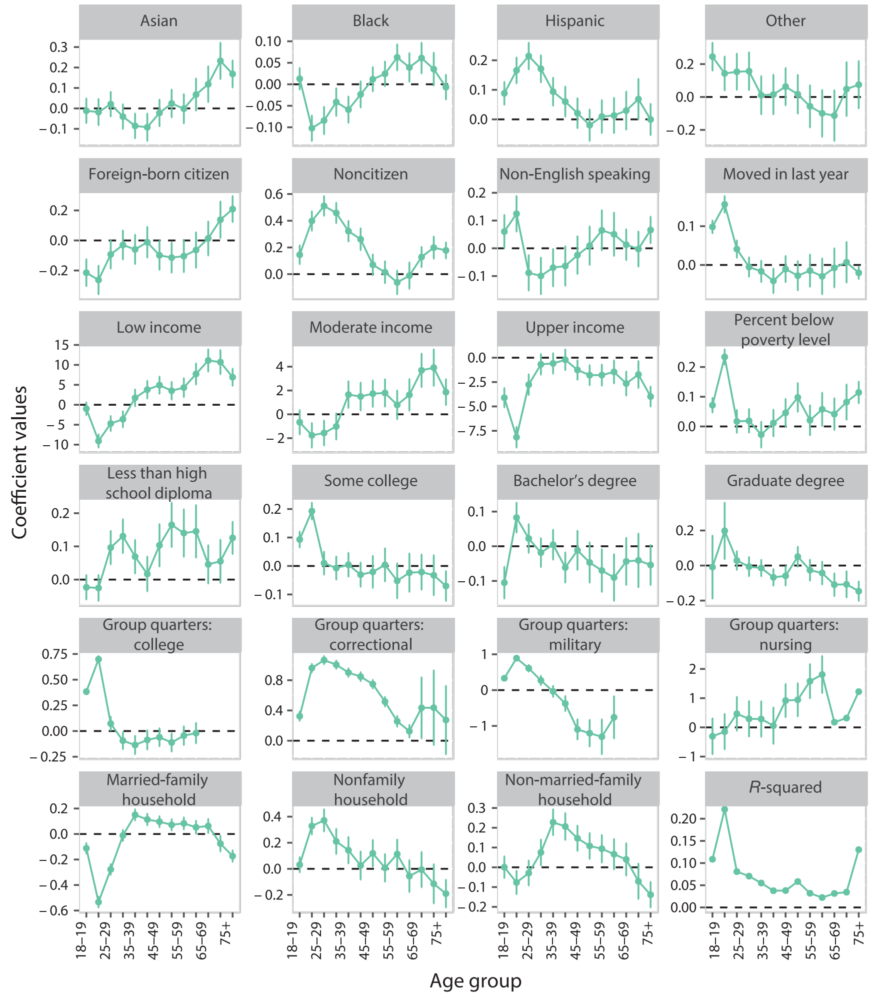


Exhibit A-2

Coefficient Estimates, Insufficient-Unscored

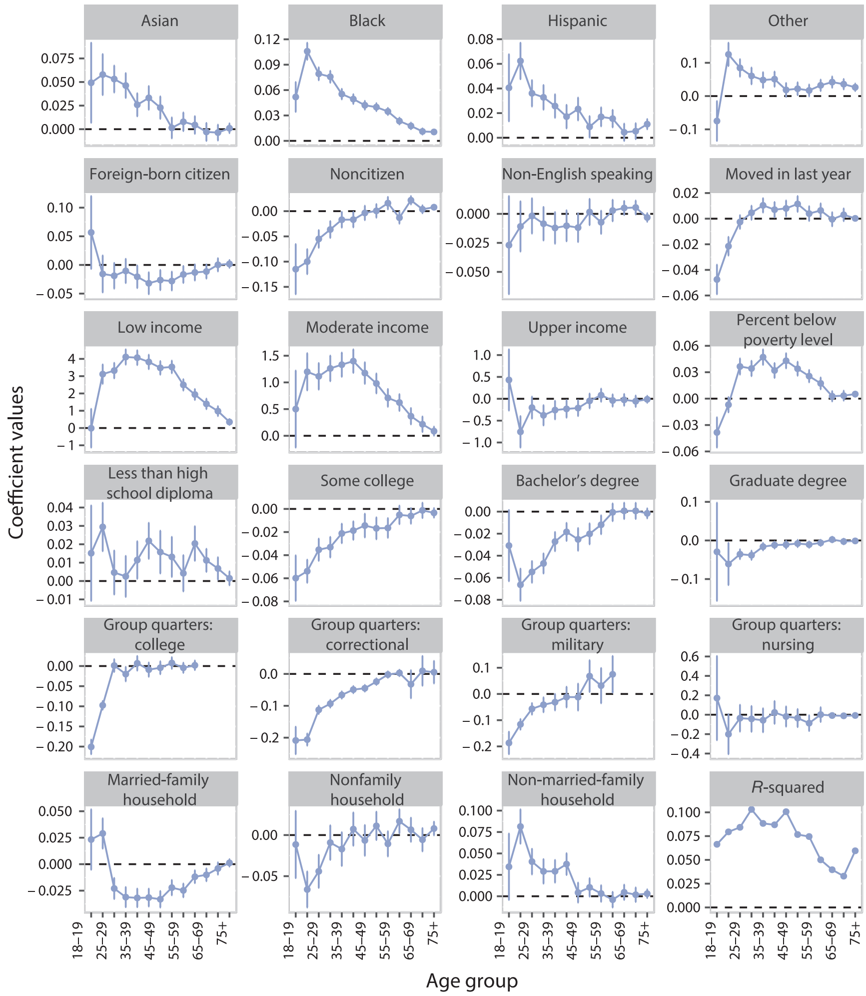
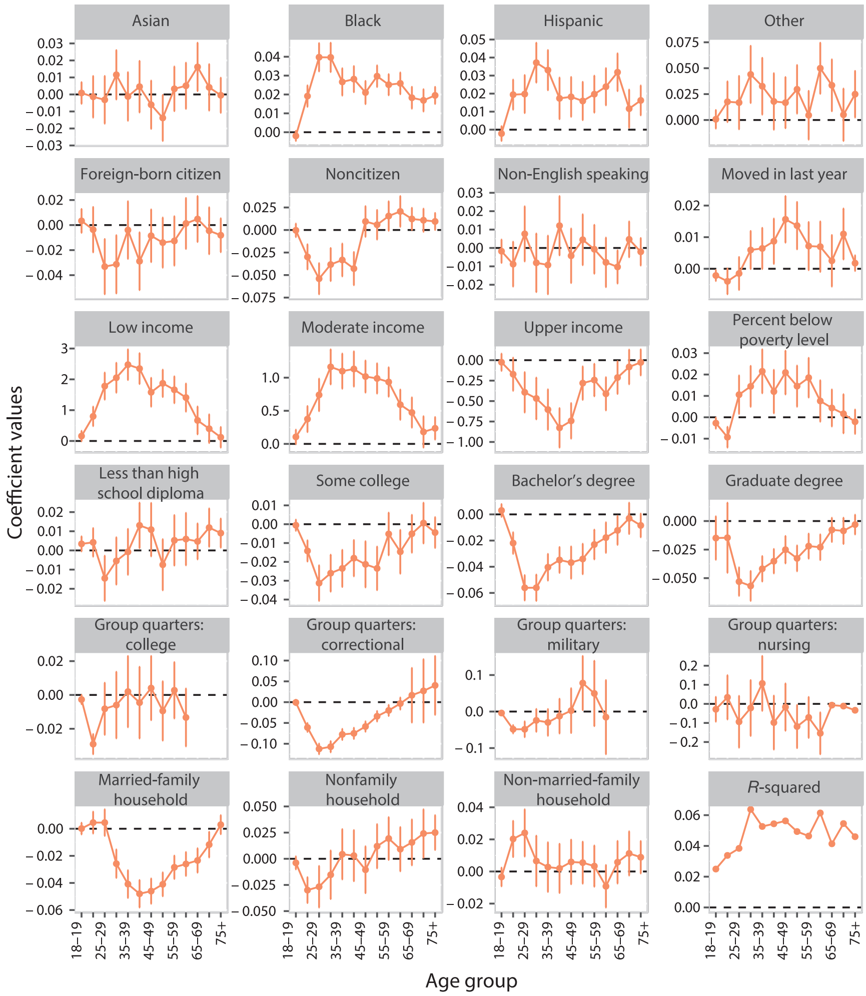


Exhibit A-3

Coefficient Estimates, Stale-Unscored



Acknowledgments

The authors thank Matias Barenstein and colleagues at the Consumer Financial Protection Bureau for helpful comments and advice. Any errors are those of the authors.

Authors

Kenneth P. Brevoort is a section chief in the Office of Research at the Consumer Financial Protection Bureau.

Philipp Grimm is a research assistant in the Office of Research at the Consumer Financial Protection Bureau.

Michelle Kambara is a financial analyst in the Office of Research at the Consumer Financial Protection Bureau.

References

Avery, Robert B., Kenneth P. Brevoort, and Glenn Canner. 2012. "Does Credit Scoring Produce a Disparate Impact?" *Real Estate Economics* 40 (s1): S65–S114.

Avery, Robert B., Paul S. Calem, Glenn B. Canner, and Raphael W. Bostic. 2003. "An Overview of Consumer Data and Credit Reporting," *Federal Reserve Bulletin* 89 (2): 47–73.

Board of Governors of the Federal Reserve System. 2007. *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*. Washington, DC: Federal Reserve Board.

Brevoort, Kenneth P., Robert B. Avery, and Glenn B. Canner. 2013. "Credit Where None Is Due? Authorized-User Account Status and Piggybacking Credit," *Journal of Consumer Affairs* 47 (3): 518–547.

Brevoort, Kenneth P., Philipp Grimm, and Michelle Kambara. 2015. *Credit Invisibles*. Consumer Financial Protection Bureau Data Point. Washington, DC: Consumer Financial Protection Bureau.

Bucks, Brian K., Arthur B. Kennickell, Traci L. Mach, and Kevin B. Moore. 2009. "Changes in U.S. Family Finances From 2004 to 2007: Evidence From the Survey of Consumer Finances," *Federal Reserve Bulletin* 95 (5): A1–A55.

Consumer Financial Protection Bureau (CFPB). 2014. *Report on the Use of Remittance Histories in Credit Scoring*. Washington, DC: Consumer Financial Protection Bureau.

Crook, Jonathan, and John Banaskik. 2004. "Does Reject Inference Really Improve the Performance of Application Scoring Models?" *Journal of Banking and Finance* 28 (4): 857–874.

Experian. 2014. *Let There Be Light: The Impact of Positive Energy-Utility Reporting on Consumers*. Experian white paper. Costa Mesa, CA: Experian Information Services.

Experian RentBureau. 2014. *Credit for Renting: The Impact of Positive Rent Reporting on Subsidized Housing Residents*. Experian RentBureau white paper. Costa Mesa, CA: Experian Information Services.

Hand, David J., and William E. Henley. 1993. “Can Reject Inference Ever Work?” *IMA Journal of Management Mathematics* 5 (1): 45–55.

Howat, John. 2009. *Full Utility Credit Reporting: Risks to Low Income Consumers*. National Consumer Law Center working paper. Boston: National Consumer Law Center.

Schneider, Rachel, and Arjan Schutte. 2007. *The Predictive Value of Alternative Credit Scores*. Chicago: Center for Financial Services Innovation.

Turner, Michael A., Alyssa Stewart Lee, Ann Schnare, Robin Varghese, and Patrick D. Walker. 2006. *Give Credit Where Credit Is Due: Increasing Access to Affordable Mainstream Credit Using Alternative Data*. Political and Economic Research Council and The Brookings Institution Urban Markets Initiative Report. Washington, DC: The Brookings Institution.

VantageScore. 2015. *Maximizing the Credit Universe*. VantageScore white paper. Stamford, CT: VantageScore Solutions, LLC.