# Supplemental estimation methodology for institutional coverage and market-level cost estimates in the small business lending rulemaking

Consumer Financial
Protection Bureau

# Introduction

On March 30, 2023, the Consumer Financial Protection Bureau (CFPB) released a final rule to amend Regulation B to implement changes to the Equal Credit Opportunity Act (ECOA) made by section 1071 of the Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank Act). Consistent with section 1071, the CFPB is requiring covered financial institutions to collect and report data on applications for credit for small businesses, including those that are owned by women or minorities. The CFPB's final rule also addresses its approach to privacy interests and the publication of section 1071 data; shielding certain demographic data from underwriters and other persons; recordkeeping requirements; enforcement provisions; and the final rule's effective and compliance dates. The CFPB's final rule and related materials can be accessed at https://www.consumerfinance.gov/1071-rule/.

The final rule uses the term "covered financial institution" to refer to those financial institutions that are required to comply with the rule's data collection and reporting requirements.[1] The rule provides that a covered financial institution is a financial institution that originated at least 100 covered credit transactions[2] for small businesses[3] in each of the two preceding calendar years.

In order to estimate how many institutions will be covered under the final rule, we need comprehensive data on originations of credit transactions made to small businesses for all financial institutions. However, market-wide data on small business lending are currently limited, and we are unaware of any such comprehensive data.

Existing data for banks either do not include origination-specific information or do not cover all institutions. For example, the Federal Financial Institutions Examination Council (FFIEC) Call

---

[1] The final rule defines a "financial institution" as any partnership, company, corporation, association (incorporated or unincorporated), trust, estate, cooperative organization, or other entity that engages in any financial activity. The rule thus applies to a variety of entities that engage in small business lending, including depository institutions (*i.e.*, banks, savings associations, and credit unions), online lenders, platform lenders, community development financial institutions, Farm Credit System lenders, lenders involved in equipment and vehicle financing (captive financing companies and independent financing companies), commercial finance companies, organizations exempt from taxation pursuant to 26 U.S.C. 501(c), and governments or governmental subdivisions or agencies.

[2] The final rule defines a "covered credit transaction" as an extension of business credit under existing Regulation B, with certain exceptions. Loans, lines of credit, credit cards, and merchant cash advances are all covered credit transactions within the scope of the rule. The rule excludes trade credit, transactions that are reportable under the Home Mortgage Disclosure Act (HMDA), insurance premium financing, public utilities credit, securities credit, and incidental credit. Factoring, leases, consumer-designated credit used for business or agricultural purposes, credit transaction purchases, purchases of an interest in a pool of credit transactions, and purchases of a partial interest in a credit transaction are also not covered credit transactions.

[3] The final rule defines a "small business," about whose applications for credit data must be collected and reported, by reference to the definitions of "business concern" and "small business concern" as set out in the Small Business Act (15 U.S.C. 631 *et seq.*) and Small Business Administration (SBA) regulations. However, in lieu of using the SBA's size standards for defining a small business concern, the definition looks to whether the business had $5 million or less in gross annual revenue for its preceding fiscal year.

Reports are the primary source of information about the financial condition of banks and savings associations (hereafter, banks) in the United States. All banks are required to regularly report, among other things, bank-level information on outstanding balances for various loan products. These data generally do not include information on originations. Meanwhile, the Community Reinvestment Act (CRA) data provide information on annual originations for small loans to businesses and farms, but only relatively large banks are required to report these data.[4] However, in order to estimate the number of banks that will be required to report under the final rule's 100 originations threshold (or at alternative threshold levels considered by the CFPB), we need an estimate of bank-level originations. We use the relationship in the data between originations and outstanding small business credit transactions among required CRA reporters to estimate the originations for banks that are not required to report CRA data. We then use these estimated originations to estimate coverage for banks under the final rule.

In this document, we describe our methodology for estimating how many banks will be required to report under the final rule and for producing market-level estimates of the costs associated with implementing the final rule.[5]

# Data

Loans to small businesses and farms are not directly identified in either the FFIEC Call Report or CRA data. Instead, small loans to businesses or farms of any size are used, in whole or in part, as a proxy for loans to small businesses or small farms. For the purposes of estimating the impacts of the final rule on banks, we follow this convention of using small loans to businesses as a proxy for loans to small businesses and small loans to farms as a proxy for loans to small farms in the bank data.[6]

The FFIEC Call Report captures each bank's total outstanding number and dollar volume of small loans to businesses (that is, loans originated under $1 million to businesses of any size;

---

[4] The CRA Agencies have proposed eliminating this reporting requirement once 1071 data become available. Bd. of Governors of the Fed. Reserve Sys.; Fed. Deposit Ins. Corp.; Off. of the Comptroller of the Currency, *Community Reinvestment Act, Joint notice of proposed rulemaking*, 87 FR 33884, 33930 (June 3, 2022) ("Importantly, the agencies' proposal to leverage the CFPB's definitions would reduce bank data collection and reporting burden under CRA regulations. The agencies would intend to eliminate the current CRA small business and small farm data collection and reporting and replace it with the CFPB's section 1071 data, once available, which covered banks would be required to collect and report under section 1071.").

[5] We do not need to estimate originations for credit unions because they report originations on the NCUA Call Reports. See part IX.D of the final rule for additional information. We discuss the methodology we use for estimating the number of originations by nondepository institutions in part II.D of the final rule.

[6] Fed. Deposit Ins. Corp., Staff Study, *Measurement of Small Business Lending Using Call Reports: Further Insights From the Small Business Lending Survey* (July 2020), https://www.fdic.gov/analysis/cfr/staff-studies/2020-04.pdf.

small loans to farms are those originated under $500,000).[7] All banks report the outstanding number and volume of small loans to businesses and farms as of the end of each calendar year.

The CRA requires banks with assets over a specified threshold[8] to collect and report data on small loans to businesses and farms according to the same definition that is used for the Call Report described above. The FFIEC publishes aggregate numbers and values of annual originations at a bank level and at various geographic levels.

For banks that report under the CRA, we have annual data on both the outstanding number and dollar volume of their small loans to businesses from the Call Report and the annual number and dollar volume of their originations from the CRA data. These banks have what we call complete data. However, for banks that do not report under the CRA, we only observe outstanding values from the Call Report. These banks do not have complete data. In the next section, we describe how we use information from the CRA reporters to estimate the number and dollar volume of originations for banks that are not required to report CRA data.[9]

# Methodology

Statistically, the lack of data on originations for a subset of banks is a missing data problem. We observe a large amount of information about every bank every quarter, but some key variables, such as originations of small loans to businesses, are missing for some banks. To address this, we employ established methods for handling missing data to impute, or fill in, the missing originations.

Missing data can be imputed a single time or multiple times. Single imputation methods generate one complete dataset by replacing the missing data with one value. Multiple imputation methods generate multiple complete datasets.[10] We use multiple imputation instead of single imputation to systematically account for the uncertainty about the missing data. We

---

[7] For what products are included, see Fed. Fin. Insts. Examination Council, *Instructions for Preparation of Consolidated Reports of Condition and Income*, at Schedule RC-C, Part II, https://www.ffiec.gov/pdf/FFIEC_forms/FFIEC031_FFIEC041_202106_i.pdf (last updated June 2021). Note that these are called "loans to small businesses" but include credit transactions that are not loans.

[8] The threshold is $1.384 billion as of 2022. For annual reporting criteria, see Fed. Fin. Insts. Examination Council, *Explanation of the Community Reinvestment Act Asset-Size Threshold Change*. https://www.ffiec.gov/cra/pdf/2022_Asset_Size_Threshold.pdf (last visited Oct. 14, 2022).

[9] The CRA data include data for banks that voluntarily report but are not required to. As noted below, though we have complete data for these voluntary reporters, we exclude data on the number and volume of originations for the purposes of estimating the number and dollar volume of originations for banks that are not required to report under CRA.

[10] Roderick J. A. Little & Donald B. Rubin, *Statistical Analysis with Missing Data* at ch. 4, 10 (3d ed. 2019).

use a standard multiple imputation model that is appropriate for the structure of the missing data. In particular, we use a Bayesian independent univariate conditional multiple ordinary least squares (OLS) regression model. We can use a Bayesian multiple OLS regression model because the data are missing at random (MAR).[11] We need to impute data for multiple variables, origination number and dollar volume. Because the missing variables are monotone, we can use an independent univariate conditional model to generate the multivariate imputations. Below, we detail the concepts of MAR and monotone missing and show how the data meet these conditions.

Missing data are considered missing at random if "the probability of being missing is the same only within groups defined by the observed data."[12] That is, if the probability that data are missing only depends on observable data, not on unobservable data, then the missing data are considered MAR. The origination data are missing if a bank falls below the CRA reporting threshold and decides not to voluntarily report. After we set observations of voluntary reporters equal to missing, the missingness of the data become fully predicted by the observed data, namely asset size. We do not know the underlying mechanism for why some banks voluntarily report so we exclude those observations from the imputation model so the data more plausibly satisfy the MAR assumption.[13] The key assumption is that the estimated relationships in the data for those who are required to report is the same as that which would be estimated if everyone were required to report. That is, we can extrapolate the model for required reporters to all banks. Furthermore, we are concerned that including voluntary reporters could introduce selection bias into our regression estimates.

A dataset is considered monotone missing if the variables with missing values $X_1, X_2, \ldots, X_J$ can be ordered such that if $X_l$ is missing in an observation then $X_k$ is also missing in that observation for all $k > l$. In the originations data, both the number of originations and dollar volume of originations are missing for some banks, and both are variables of interest for our eventual analysis. Additionally, both variables are always missing for the same observations. That is, no bank reports numbers of originations but not dollar volume, or vice versa. As such, the data are

---

[11] *See* Stef van Buuren, *Flexible Imputation of Missing Data* (2d ed. 2018), https://stefvanbuuren.name/fimd/.

More broadly, Bayesian multiple imputation is valid when the missing data mechanism is ignorable. The data mechanism is ignorable when: (1) the data are missing at random and (2) the parameters of the data model and the parameters of missing-data mechanism are distinct. However, it is difficult to test ignorability. As discussed in Section 2.2, in general, the missing at random condition for ignorability is considered more important than the distinctness condition. For the purposes of this analysis, we will only focus on the MAR condition for practical purposes.

[12] *See id.* at Section 1.2.

[13] As discussed below, we will use the observed data for voluntary reporters to calculate the number of covered institutions, just not in the imputation model.

a type of monotone missing and we can use a monotone data imputation, such as an independent univariate conditional model.

We calculate 200 imputations for this analysis because of the high degree of missingness in the data. In general, multiple imputation can generate unbiased results even with a low number of imputations. However, recent scholarship has recommended more imputations when the fraction of missingness is high.[14] For example, White *et al.* (2011) suggest the rule that the number of imputations should be equal to the percentage of missing data.[15] Once we set voluntary reporters equal to missing, the share of missing for these data is 91 percent. We conservatively choose an even higher number of imputations. We believe this is the best option given the available data, but we acknowledge that we are imputing values for a very large share of the data. We discuss several robustness checks in the final section of this document that give us confidence that this model predicts originations reasonably well.

# Estimation

We begin by running two OLS regressions using data on banks that were required to report under the CRA:

$$log(q_{0it}) = \alpha_0 + \beta_1 * log(outq_{0it}) + \beta_2 * log(outa_{0it}) + \sum_{\tau=2012}^{2019} \delta_\tau * I\{t = \tau\} + e_{it} \qquad (1)$$

$$log(a_{0it}) = \alpha_0 + \gamma_1 * log(outq_{0it}) + \gamma_2 * log(outa_{0it}) + \gamma_3 * log(q_{0it}) + \sum_{\tau=2012}^{2019} \delta_\tau * I\{t = \tau\} + u_{it} \quad (2)$$

Where $q_{oit}$ is the number of originations of small loans to businesses and farms (SLBF), $a_{oit}$ is the dollar value of originations of SLBF for institution $i$ in year $t$ observed in the CRA data, and $outq_{oit}$ and $outa_{oit}$ are the outstanding number of SLBF and outstanding dollar value of SLBF, respectively, for institution $i$ in year $t$ observed in the FFIEC Call Report data.[16] The appendix lists the FFIEC Call Report codes for the variables. The $o$ subscript indicates that these values

---

[14] For more discussion see Univ. of Cal. Los Angeles, Inst. for Digit. Rsch. & Educ., *Multiple Imputation in Stata*, https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/ (last visited Aug. 16. 2021).

[15] Ian R. White *et al.*, *Multiple imputation using chained equations: Issues and guidance for practice*, Stat. in Med. 30(4) 377-99 (Feb. 11, 2011), https://pubmed.ncbi.nlm.nih.gov/21225900/.

[16] We note that the independent variables in these regressions are endogenous. The imputation step is purely about prediction, not estimating the causal relationship between variables. Hence, it is valid to include these variables as predictors, but it would not be valid to interpret the coefficients from these regressions as causal estimates.

We also note that equation (1) does not include log of dollar volume of originations because we use an independent univariate conditional model instead of an iterative model that jointly estimates the two variables.

are observed in the data. We also include a full set of year fixed effects. The error term $e_{it}$ is assumed to be normally distributed with mean 0 and variance $\sigma_e^2$ and the error term $u_{it}$ is assumed to be normally distributed with mean 0 and variance $\sigma_u^2$. The terms $e_{it}$ and $u_{it}$ are assumed to be independent.

We run these regressions on bank-year observations for which the bank assets for that year exceeded the CRA reporting threshold and the bank reported under the CRA.[17] That is, as discussed above, we exclude observations for which the bank voluntarily reported. Table 1 reports summary statistics on the number of banks, the number of mandatory CRA reporters, and the share of data missing (calculated as the share of banks that are not mandatory CRA reporters) for each year. We define $C = \{i_t\}$ as the set of institutions $i_t$ at time $t$ that are not mandatory reporters.

**TABLE 1:**   SUMMARY STATISTICS

| Year | Banks | Voluntary CRA Reporters | Mandatory CRA Reporters | Share Missing (percent) |
|------|-------|-------------------------|-------------------------|-------------------------|
| 2012 | 7,083 | 298 | 516 | 92.7 |
| 2013 | 6,812 | 272 | 508 | 92.5 |
| 2014 | 6,509 | 226 | 524 | 91.9 |
| 2015 | 6,182 | 201 | 533 | 91.4 |
| 2016 | 5,913 | 160 | 549 | 90.7 |
| 2017 | 5,671 | 141 | 561 | 90.1 |
| 2018 | 5,407 | 133 | 555 | 89.7 |
| 2019 | 5,177 | 111 | 563 | 89.1 |
| Total | 48,754 | 1,542 | 4,309 | 91.2 |

---

[17] For this analysis, we proxy for mandatory reporting. The actual requirement for CRA reporting is that the bank exceeds the asset threshold for the previous two years. *See* Fed. Fin. Insts. Examination Council, *Who Is Required to Report CRA Data*, https://www.ffiec.gov/cra/reporter.htm (last updated Dec. 16, 2020). For example, a bank would be required to report in 2019 if the merger adjusted assets in 2018 and 2017 exceeded $1.284 billion. Instead, we say a bank was required to report if its assets exceeded the CRA threshold in the base year (2019 in this example) and the bank actually reported under CRA. In 2019, we include 17 banks out of 563 that were not actually required to report.

We use a log linear model because it fits the data the best. [18] Figures 1 and 2 plot $log(q_{it})$ versus $log(outq_{it})$ and $log(a_{it})$ versus $log(outa_{it})$, respectively, for 2019. [19]

From these regressions, we obtain the vector of estimated coefficients $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ and the estimated variances $\widehat{\sigma_e^2}$ and $\widehat{\sigma_u^2}$. Table 2 presents the estimated coefficients from these two regressions. Note that we estimate the regression model across multiple years to get a more precise estimate of the coefficients on outstanding loans and amounts and on originations, assuming these coefficients are constant across years. In order to get the most current view of the market not affected by the COVID-19 pandemic, we use 2017–2019 for our institutional coverage and cost estimates.

**FIGURE 1:**    LOG OF OUTSTANDING NUMBER VERSUS LOG OF NUMBER OF ORIGINATIONS FOR MANDATORY CRA REPORTERS IN 2019
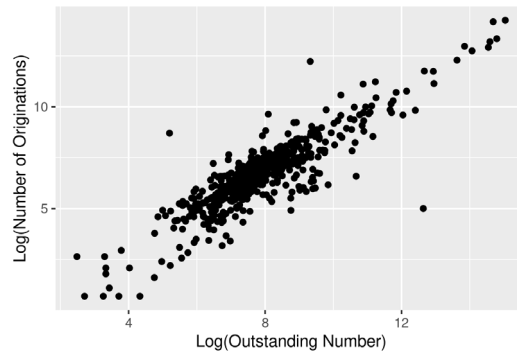


**FIGURE 2:**    LOG OF OUTSTANDING DOLLAR VOLUME VERSUS LOG OF DOLLAR VOLUME OF ORIGINATIONS FOR MANDATORY CRA REPORTERS IN 2019



---
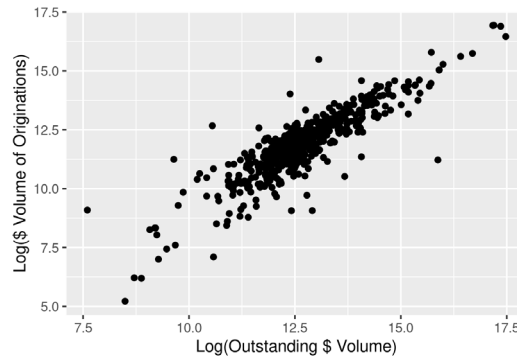
[18] We assume that banks that had no outstanding loans in a year also had no originations in that year. We drop banks that reported zero originations in a particular year from the regression model.

[19] We also tried using a Poisson model, but the log linear model performed better in out of sample prediction.

**TABLE 2:**    REGRESSION  RESULTS

| Independent variables | Dependent variable: $\log(q_{it})$ | Dependent variable: $\log(a_{it})$ |
|---|---|---|
| $log(out\,q_{it})$ | 0.790*** (0.019) | -0.481*** (0.01) |
| $log(out\,a_{it})$ | 0.261*** (0.026) | 0.722*** (0.012) |
| $log(q_{it})$ | | 0.744*** (0.007) |
| Year = 2013 | 0.065 (0.057) | 0.034 (0.025) |
| Year = 2014 | 0.057 (0.056) | 0.018 (0.024) |
| Year = 2015 | 0.061 (0.056) | 0.034 (0.024) |
| Year = 2016 | 0.053 (0.055) | 0.043* (0.024) |
| Year = 2017 | 0.001 (0.055) | 0.041* (0.024) |
| Year = 2018 | 0.004 (0.055) | 0.049** (0.024) |
| Year = 2019 | 0.038 (0.055) | 0.056** (0.024) |
| Constant | -2.881*** (0.211) | 1.470*** (0.093) |
| Observations | 4,046 | 4,046 |
| $R^2$ | 0.786 | 0.938 |
| Adjusted $R^2$ | 0.786 | 0.937 |
| Residual Std. Error | 0.873 (df = 4036) | 0.378 (df = 4035) |
| F Statistic | 1,650.768*** (df = 9; 4036) | 6,061.365*** (df = 10; 4035) |

Note: Standard errors are given in parentheses.

\* p<0.1; \*\*p<0.05; \*\*\*p<0.01

Next, we conduct the imputation step using one standard multiple imputation method. For each bank-year observation in the set $C$, we impute $M = 200$ sets of values, $\{q_{mit}\}_{m=1}^{M}$ and $\{a_{mit}\}_{m=1}^{M}$.

For each imputation $m$, we begin by simulating new parameters $\boldsymbol{\beta}^{\star}$ and $\sigma_e^{\star 2}$ from the joint posterior distribution under the conventional noninformative improper prior $Pr(\beta, \sigma_e^2) \propto 1/\sigma_e^2.$[20]

---

[20] *See* StataCorp, *Stata 17 Multiple-Imputation  Reference Manual* at 256, https://www.stata.com/manuals/mi.pdf (last visited Feb. 24, 2023) (Stata Multiple Imputation Manual).

Next, for every bank-year observations in the set $C$, we simulate $log(q_{mit})$ according to the equation

$$log(q_{mit}) = \alpha^\star + \beta_1^\star * log(outq_{0it}) + \beta_2^\star * log(outa_{0it}) + \sum_{\tau=2012}^{2019} \delta_\tau^\star * I\{t = \tau\} + e_{it}^\star \qquad (1')$$

where $e_{it}^\star \sim N(0, \sigma_e^{\star 2})$.

Next, we simulate values for $log(a_{mit})$ conditional on the imputed values of $log(q_{mit})$. As noted above, this method is valid because the data are monotone missing. First, we simulate another set of new parameters $\gamma^\star$ and $\sigma_u^{\star 2}$ and, for every bank-year observations in the set $C$, we simulate $log(a_{mit})$ according to the equation

$$log(a_{mit}) = a_0^\star + \gamma_1^\star * log(outq_{0it}) + \gamma_2^\star * log(outa_{0it}) + \gamma_3^\star * log(q_{mit}) + \sum_{\tau=2012}^{2019} \delta_\tau^\star * I\{t = \tau\} + u_{it}^\star (2')$$

where $u_{it}^\star \sim N(0, \sigma_u^{\star 2})$ and $log(q_{mit})$ is the imputed value from equation (1'). [21]

After we have simulated the full set of values, we transform the simulated values from $log(q_{mit})$ to $q_{mit}$ and $log(a_{mit})$ to $a_{mit}$ by exponentiating. If an institution $i$ had complete data in year $t$, then we set $q_{mit} = q_{0it}$ and $a_{mit} = a_{0it}$ for all $m$. That is, we use the observed data for all voluntary and mandatory CRA reporters, even though we only estimated the model on mandatory reporters. The result is M complete sets of data for all institutions and years. Note that, for each set of imputed values, the errors are independently and identically distributed and not serially correlated. That is, imputed values for banks are only correlated across years through observable information, such as outstanding number and dollar value of loans.

Under the final rule, a financial institution is required to collect and report small business lending data in a given year if it originates 100 or more covered transactions in each of the each of the two preceding calendar years. If a financial institution merged during the previous two years, the surviving or newly formed institution reports if the preceding institutions collectively originated 100 or more covered transactions. In this estimation exercise, we need to: (1) account for merger and acquisition activity; and (2) keep track of the number of originations made by the institution in 2017, 2018, and 2019. The next goal is to create sets of values across 2017, 2018, and 2019 for all institutions based on the structure of the institution in 2019.

---

[21] For more information on multiple imputation with monotone missing data, see Stata Multiple Imputation Manual at 185.

Let $q_{mi} = \{q_{mi17}, q_{mi18}, q_{mi19}\}$ and $a_{mi} = \{a_{mi17}, a_{mi18}, a_{mi19}\}$ be the set of simulated values for institution $i$, imputation $m$, for the years 2017, 2018, and 2019.[22]

We merger adjust the values $q_{mi}$ and $a_{mi}$ to obtain $q'_{mi}$ and $a'_{mi}$. For each institution in 2017 and 2018, we determine the ultimate institution identifier as of the end of 2019.[23] Let $i$ be the ultimate institution in 2019 and let $j = 1,...J$, including $i$, be the set of institutions that precede $i$ between December 2017 and December 2019. Define the merger adjusted value $x'_{mit} = \sum_{j=1}^{J} x_{mjt}$.

For example, suppose that banks Y and Z were individual institutions as of December 31, 2017. Then, suppose that bank Y acquired bank Z in June 2018. Banks Y and Z would both file FFIEC Call Reports in December 2017 but only bank Y would file an FFIEC Call Report for the combined institution in December 2018 and December 2019. For all $m$, we would have simulated values for bank Y for three years $\{q_{mY17}, q_{mY18}, q_{mY19}\}$, but only one year of simulated values for bank Z ($q_{mZ17}$). We define the merger adjusted number of originations as $q'_{mY17} = q_{mY17} + q_{mZ17}, q'_{mY18} = q_{mY18}$, and $q'_{mY19} = q_{mY19}$.

Note, that we may have complete data on bank Y but not on bank Z. In this case, the value of originations for the ultimate institution Y will be partially imputed for 2017.

If a bank did not merge, then $q'_{mi} = q_{mi}$ and $a'_{mi} = a_{mi}$ for all $m$. If a bank existed in December 2019 but did not exist in year $t < 2019$, then we set $q'_{mit}$ and $a'_{mit}$ equal to 0.

We calculate which institutions would be required to report in 2019 based on the values of $q'_{mi}$ under each of the M sets of imputations. Let $r_{mi}$ be an indicator variable where

$$r_{mi} = \begin{cases} 1 \ if \ q'_{mi17} \geq 100 \ and \ q'_{mi18} \geq 100 \\ 0 \qquad\qquad\qquad else \end{cases}$$

That is, $r_{mi}$ indicates if, for imputation m, an institution $i$ would have been required to report under the final rule if it had been in effect in 2019. For imputation m, the total number of banks that would be required to report under the final rule is $r_m = \sum_i r_{mi}$. We construct a 95 percent confidence interval for the number of banks that would be required to report. We order the 200

---

[22] Note that, as described above, the values of $\{q_{mi17}, q_{mi18}, q_{mi19}\}$ are independently and identically distributed (i.i.d.) and not serially correlated but are imputed using the same values of the parameters $\boldsymbol{\beta}^{\star}$ and $\sigma_e^{\star 2}$. Similarly, $\{a_{mi17}, a_{mi18}, a_{mi19}\}$ are i.i.d. and not serially correlated and are imputed using the same values of the parameters $\boldsymbol{\gamma}^{\star}$ and $\sigma_u^{\star 2}$. The simulated values that correspond to different imputations are not drawn using the same parameter values but rather a different set of parameter values drawn from the same prior.

[23] Eric C. Breitenstein & Derek K. Thieme, *Merger Adjusting Bank Data: A Primer*, FDIC Quarterly 13(1) at 31-49 (2019), https://www.fdic.gov/analysis/quarterly-banking-profile/fdic-quarterly/2019-vol13-1/fdic-v13n1-4q2018-article.pdf.

values of $r_m$ from smallest to largest and find $r_L$, the fifth smallest value, and $r_H$, fifth largest value. We report these values in part IX.D of the final rule.

We also calculate what percent of originations by banks would have been covered in 2019. Let $totq_m = \sum_i q'_{mi19}$ be the estimated total number of originations and let $tota_m = \sum_i a'_{mi19}$ be the estimate total dollar volume of originations made by banks in 2019 in imputation $m$. For each imputation $m$, let $covq_m = \sum_i (r_{mi} q'_{mi19})/totq_m$ be the estimated share of the number of bank originations covered and let $cova_m = \sum_i (r_{mi} a'_{mi19})/tota_m$ be the estimated share of the dollar volume of bank originations covered. As with the number of institutions covered, we construct a 95 percent confidence interval for the share of the number of originations covered and the share of the dollar volume of originations covered. We report these ranges in the section-by-section analysis of § 1002.105(b) in part V of the final rule.

# Market-level cost estimates

Next, we describe how we use the imputations of originations, together with estimates of costs, to generate estimates of total market-level costs. As discussed in part IX of the final rule, we also estimate one-time and ongoing costs per application based on institution type. We define a bank's type according to its number of applications, discussed further in part IX.E of the final rule. We assume that low- and medium-complexity banks originate one loan for every two applications received and that high-complexity banks originate one loan for every three applications. For each of the M sets of imputations and for each institution, we calculate the type of the institution in 2019 based on $q'_{mi19}$ and $a_X$, the ratio of applications to originations for type $X \in \{A, B, C\}$. Let $\tau_{mi}$ indicate the type of an institution $i$ in imputation $m$ where

$$
\tau_{mi} = \begin{cases}
A \ if \ 0 \leq a_A q'_{mi19} < 150 \\
B \ if \ 150 \leq a_B q'_{mi19} < 1{,}000 \\
C \ if \ 1{,}000 \leq a_C q'_{mi19}
\end{cases}
$$

We calculate costs for each institution in 2019 for each imputation based on $\tau_{mi}$ and $q'_{mi19}$. For type $X \in \{A, B, C\}$, let $v_X$ be the estimated per application cost, $f_X$ be the estimated fixed ongoing cost, $c_X$ be the one-time cost under the final rule, and $a_X$ be the ratio of applications to originations.[24] For each institution $i$ and each imputation $m$, the total estimated one-time costs for the institution are

---

[24] See part IX.E.1 of the final rule for more detail on how we estimate one-time costs and part IX.F.3.i for the estimated values of one-time costs. See part IX.E.2 of the final rule for more detail on how we estimate ongoing costs and part IX.F.3.ii for the estimated values of ongoing costs.

$$onetime_{mi} = r_{mi} \left( \sum_X I(\tau_{mi} = X)c_X \right)$$

and the total estimated ongoing costs for the institution are

$$ongoing_{mi} = r_{mi} \left( \sum_X I(\tau_{mi} = X)(f_X + v_X a_X q'_{mi19}) \right)$$

Recall that $r_{mi}$ is an indicator variable for whether institution $i$ would be required to report in the $m^{th}$ imputation. If an institution $i$ would not be required to report in imputation $m$, then that institution has zero onetime and ongoing costs. For imputation $m$, the total estimated one-time costs across all institutions is $onetime_m = \sum_i onetime_{mi}$ and the total estimate ongoing costs across all institutions is $ongoing_m = \sum_i ongoing_{mi}$. We construct a 95 percent confidence interval for the estimated total one-time costs of banks that would be required to report. We order the 200 values of $onetime_m$ from smallest to largest and find $onetime_L$, the fifth smallest value, and $onetime_H$, fifth largest value. We similarly compute a 95 percent confidence interval for estimated total ongoing costs by finding $ongoing_L$ and $ongoing_H$. We report these values in part IX.F.3 of the final rule.

# Robustness checks

We conduct several robustness checks. First, we investigate how well the approach described in this document performs in out-of-sample prediction. Within the set of mandatory reporters, we identify the bottom 10 percent each year by assets. We then use the data from the remaining 90 percent of mandatory reporters to impute the number and dollar volume of originations for the bottom 10 percent of reporters, the held-out sample, 200 times. [25] For each bank-year in the held-out sample, we generate a 95 percent confidence interval for the institution's number of originations that year by finding the fifth smallest imputed value and the fifth largest imputed value. We similarly generate a 95 percent confidence interval for the dollar volume of originations for each bank-year in the held-out sample. We find that the true value of the number of originations falls in the 95 percent confidence interval for 97 percent of bank-year observations and that the true value of the dollar volume of originations falls in the 95 percent confidence interval for 99 percent of bank-year observations.

Second, for the same held out sample of the smallest 10 percent of mandatory reporters by asset size, we also test if the imputations accurately capture the true number of institutions that have

---

[25] We hold out smaller banks because the missingness mechanism in the true data is determined by asset size.

at least 25 originations each year.[26] For each year and each of the 200 imputations, we calculate how many of the institutions in the held-out sample have at least 25 originations. We generate a 95 percent confidence interval for the number of institutions that exceed the threshold each. We find that the true number of institutions in the held-out sample that had at least 25 originations lies in the confidence interval every year.

Finally, we test how well the model predicts originations for voluntary CRA reporters. As discussed above, we do not use data from voluntary CRA reporters to impute data for banks that do not report CRA data. We exclude voluntary reporters because we do not know the underlying mechanism for why these banks voluntarily report. Hence, we expect that the imputation model may not accurately predict originations for voluntary reporters. We find that the imputation model predicts the number of originations by voluntary reporters reasonably well but not the dollar volume of originations. We find that the true value of the number of originations falls in the imputed 95 percent confidence interval for 95 percent of bank-year observations for voluntary reporters. However, we find that the true value of dollar volume of originations falls in the imputed 95 percent confidence interval for only 76 percent of bank-year observations for voluntary reporters. Voluntary reporters may be systematically different from other banks in some ways, but these results show that, even for voluntary reporters, the model accurately predicts number of originations, the main variable of interest for the purposes of estimating institutional coverage and market-level costs.

---

[26] For the robustness checks, we focused on the 25 origination threshold proposed in the NPRM.

# Appendix: FFIEC Call Report Definitions

For banks that answer "yes" to RCON6999 ("whether all or substantially all of the dollar volume of your bank's 'Loan secured by nonfarm nonresidential properties' report in Schedule RC-C Part I ... and all or substantially all of the dollar volume of your bank's 'Commercial industrial loans' reported in schedule RC-C Part I, have original amounts of $100,000 or less"):

$outq$ = RCON5562 + RCON5563 + RCON5576 + RCON5577

$outa$ = RCON1766/RCON1763[27] + RCONF160 + RCONF161 + RCON1590 + RCON1420

For banks that answer "no" to RCON6999

$outq$ = RCON5564 + RCON5566 + RCON5568 + RCON5570 + RCON5572 + RCON5574 + RCON5578 + RCON5580 + RCON5582 + RCON5584 + RCON5586 + RCON5588

$outa$ = RCON5565 + RCON5567 + RCON5569 + RCON5571 + RCON5573 + RCON5575 + RCON5579 + RCON5581 + RCON5583 + RCON5585 + RCON5587 + RCON5589

---

[27] We use RCON1763 for banks with more than $300 million in total assets and RCON1766 for banks with less than $300 million in total assets.