Chapter 9

Fair Lending Transaction Testing

Jason Dietrich
March 2026

The views and opinions expressed in this paper belong solely to me and do not represent the views or opinions of any employer, institution, or organization with which I have been affiliated.

This paper is for informational and educational purposes only and is not intended to serve as professional or legal advice. I specifically disclaim all responsibility for any liability, loss, or risk, personal or otherwise, which is incurred as a direct or indirect consequence of the use or application of the contents of this paper. Every effort has been made to ensure that the information in this paper is correct. However, I assume no responsibility for errors, inaccuracies, or omissions. The use of this paper implies the reader's acceptance of this disclaimer.

I. Introduction

Historically, regulators conducting fair lending exams relied primarily on evidence from manual file reviews to draw conclusions about whether a lender violated the Equal Credit Opportunity Act (ECOA). These file reviews followed the transaction testing guidelines in the Interagency Fair Lending Examination Procedures (IFLEP).[1] Examiners would generate a sample of similarly-situated pairs where each pair consisted of two applicants from different demographic groups who had similar borrower and credit characteristics, but who received different credit decisions. For a given pair, if a review of the entire applications identified no additional information to explain the different credit decisions, this was taken as evidence of potential discrimination.

Beginning with the Financial Institutions Reform, Recovery, and Enforcement Act (FIRREA) amendments to Home Mortgage Disclosure Act (HMDA) data in 1989, electronic data on mortgages at the application level started to become increasingly available over time. To leverage the benefits of these data for exams, the Office of the Comptroller of the Currency developed its statistical fair lending program in the mid-1990s.[2] A statistical analysis is a more automated type of transaction testing focused on identifying patterns in data and estimating relationships between inputs and outputs. Specifically for fair lending exams, statistical analyses identify disparities in outcomes and provide evidence of whether prohibited bases factors are a significant driver of these disparities. These developments provided regulators with a second type of transaction testing to use during fair lending exams.

[1] 09-06_attachment.pdf
[2] See Stengel and Glennon (1995).

The availability of two different types of transaction testing approaches has created significant confusion during fair lending reviews over the years. This confusion is rooted in disagreement about whether statistical analyses alone can be used to draw conclusions about discrimination, or whether these statistical analyses only provide signals of risk and that evidence from a file review is necessary to draw conclusions about discrimination. One particularly difficult challenge that arises during fair lending reviews is that the conclusions from the two transaction testing approaches often differ with statistical analyses typically more likely to show evidence of discrimination and file reviews typically more likely to show no evidence of discrimination. Operationally, this raises questions about the best types of file review samples to generate, how best to utilize the file review, and how to weight the statistical evidence and file review evidence when drawing conclusions about whether discrimination occurred.

In this report, we focus on understanding why statistical analyses and file reviews often lead to contradictory conclusions about the existence of discrimination. We are particularly interested in when and why each transaction testing approach might be in error. Since statistical analyses are more likely to show evidence of discrimination, we focus on when and why this occurs in scenarios when in fact there is no discrimination. Similarly, since file reviews are more likely to show no evidence of discrimination, we focus on when and why this occurs in scenarios when in fact there is discrimination. We then discuss a detailed set of issues and questions economists need to consider when conducting statistical analyses and generating file review samples during fair lending reviews. As part of these discussions, we suggest possible best practices that minimize the likelihood of error for each transaction testing type and maximize the likelihood of generating accurate conclusions regarding discrimination during fair lending reviews.[3]

Throughout the report we focus primarily on the perspective of conducting a fair lending analysis as part of an examination as opposed to the perspective of overall compliance risk management, although many of the concepts presented are applicable to both. In addition, we also focus primarily on underwriting decisions and mortgages, although many of the concepts we discuss apply to other decisions and products as well. Finally, we focus solely on disparate treatment, using the terms "disparate treatment" and "discrimination" interchangeably.

The next section discusses how file review transaction testing might miss potential discrimination. We provide separate discussions of the two primary types of file review samples, similarly-situated pairs and individual applications. Section III then discusses how statistical analysis transaction testing might incorrectly suggest discrimination has occurred. Section IV discusses a detailed set of issues and questions that analysts need to consider when conducting transaction testing, along with potential best practices when possible. Section V concludes the discussion.

II. File Review Transaction Testing Errors

A common result during fair lending reviews is that the statistical analysis shows statistically significant and economically meaningful disparities suggesting evidence of discrimination, and the file review shows no evidence of discrimination. Given that the two

[3] Throughout all discussions in this report, we focus only on fair lending exam analyses, as opposed to compliance reviews, such as Reg C reviews, for which each Agency has formal procedures outlining sampling approaches, sample sizes, and test criteria.

pieces of evidence directly conflict with each other when this occurs, one must be correct and one must be in error. Specifically, when discrimination has truly occurred, the statistical evidence is accurate and when discrimination has truly not occurred, the file review evidence is accurate. In this section, we focus on when and why file reviews mistakenly indicate no discrimination in scenarios when discrimination has truly occurred. We present separate discussions for the two primary types of file review samples, similarly-situated pairs and individual applications. In the next section, we focus on when and why statistical analyses mistakenly indicate discrimination in scenarios when discrimination has truly not occurred.

Before beginning the discussion, there is one important issue to cover. If we know whether discrimination has truly occurred, then we would know which piece of evidence was accurate. However, for real-world fair lending analyses we rarely know whether discrimination has truly occurred, so a decision is needed on which piece of evidence to rely on when drawing conclusions. To inform this decision, a thorough assessment of the strengths and weaknesses of both the statistical analysis and file review should always be conducted. There is a significant amount of guidance available in this report, in the other chapters of this book, and from other sources, to assess the strengths and weaknesses of both statistical analyses and file reviews. If based on this assessment, one set of evidence clearly dominates, then that evidence should drive conclusions about the existence of discrimination. Alternatively, if both sets of evidence provide valuable information, which is typically the case, both sets should be considered as part of a holistic assessment of whether discrimination has occurred.

Similarly-situated Pairs

We begin the discussion of file review transaction testing errors by focusing on similarly-situated pairs. If discrimination has truly occurred, a fair lending review focused solely on similarly-situated pairs will erroneously not identify this discrimination if no pairs of applicants have similar borrower and credit characteristics. As a simple example, for a minority applicant with no other similarly-situated applicants, a bigoted loan officer could apply a 25 basis point premium to rate simply because the applicant is a minority. Relying on only similarly-situated pairs of applicants during the fair lending review would not identify the discrimination against this applicant.

This error is important because it is often difficult to find similarly-situated pairs of applicants during fair lending reviews, especially when a lender considers several factors, and in subjective ways, when making credit decisions. During many fair lending reviews, it is common for economists to have to relax the matching criteria used to identify similarly-situated pairs, just to be able to have some pairs for auditors to review. Not surprisingly, it is then typically easy to show that these weaker pairs are not truly similarly situated during the file review. If unexplained, similarly-situated pairs are a necessary condition for concluding disparate treatment has occurred, it therefore may often be difficult to identify instances when disparate treatment has truly occurred simply because there are no similarly-situated applicants.

To illustrate the difficulty in finding similarly-situated pairs, we use the 2024 HMDA data to identify pairs consisting of one minority (American Indian, Asian, Black, Pacific Islander, or Hispanic) applicant and one non-Hispanic white applicant.[4] For this exercise we keep only action types 1 (approved), 2 (approved but NA), and 3 (denied). Very importantly, however, we do not match on action taken, since we are only interested here in the difficulty in

[4] We classify an application into a minority racial or ethnic category if that minority race or ethnicity is reported in any of the HMDA race and ethnicity variables. With this approach, an application can be classified into more than one minority racial or ethnic category. We classify an application as non-Hispanic white if non-Hispanic is reported in at least one of the HMDA ethnicity variables, white is reported in at least one of the HMDA race variables, and all of the minority racial and ethnic flags are 0.

identifying pairs of applicants with similar borrower and credit characteristics. We discuss this issue in more detail in Section IV below. During an actual fair lending review, the matching factors would be specific to the lender and the focal point of the review. Here, we take a general approach and match on variables available in HMDA data that reflect five standard types of general matching factors: 1) lender (LEI); 2) transaction characteristics (loan purpose, loan type, occupancy, lien status, HOEPA status, reverse mortgage, balloon mortgage, commercial purpose mortgage, non-amortizing features, and construction method); 3) applicant characteristics (DTI, CLTV, and credit score); 4) geography (state); and 5) timing (action date). For each discrete variable, such as loan purpose or type, we require exact matches. For example, both applicants in a pair must have the same loan purpose and the same loan type. For each continuous variable, we use ranges. Specifically, for DTI, CLTV, and credit score, the values for the two applicants must be within +/- 10 points for the pair to be a valid similarly-situated pair. For action date, the dates for the two applicants must be within +/- 60 days.

Table 1 presents the results from this exercise separately for each racial/ethnic minority group. As an example of how to interpret the table, we walk through the results for Black applicants. The 2024 HMDA data contain a total of 844,440 Black applicants with an action taken of 1, 2, or 3. These Black applicants applied for mortgages at 3,884 different financial institutions (FIs). A total of 103,071 of these Black applicants (12.21 percent) had at least one non-Hispanic white applicant that met the matching criteria detailed above. These 103,071 Black applicants applied at 1,397 different FIs. Therefore, at 2,487 (= 3884 – 1397; 64.03 percent) of the FIs, none of the Black applicants had at least one similarly-situated non-Hispanic white applicant. Separately for each of the 1,397 FIs with at least one pair, we calculated the percentage of all Black applicants that had a similarly-situated non-Hispanic white applicant.

The median percentage across the 1,397 FIs was 4 percent. The median percentage across just the largest 859 FIs, defined as 50 or more Black applicants, was 5 percent.

Table 1: Similarly-Situated Pair Analysis Using 2024 HMDA Data

Minority	Total # of minority applicants	# of FIs with at least 1 minority applicant	# of minority applicants with a non-Hispanic white match	% of minority applicants with a non-Hispanic white match	# of FIs with a minority applicant with a match	Median % of minority applicants with a match across FIs	# of large FIs (>= 50 minority applicants) with a minority applicant with a match	Median % of minority applicants with a match across large FIs
American Indian	125,087	2,848	16,735	13.38%	916	2%	283	9%
Asian	642,529	3,760	140,125	21.81%	1,438	6%	813	12%
Black	844,440	3,884	103,071	12.21%	1,397	4%	859	5%
Pacific Islander	53,789	2,077	7,941	14.76%	625	2%	118	10%
Hispanic	1,290,612	4,256	181,837	14.09%	1,636	7%	1,141	7%

The results in Table 1 provide an initial glimpse into the difficulty finding similarly-situated pairs. Since fair lending reviews are lender specific, we focus on the FI results. Between 916 and 1,636 FIs had at least one minority applicant with a similarly-situated non-Hispanic white applicant. This is 30 to 38 percent of FIs with at least one minority applicant. Stated differently, 62 to 70 percent of FIs with at least one minority applicant had no similarly-situated pairs. For these FIs, relying on a file review of similarly-situated pairs would never lead to a conclusion of disparate treatment (regardless of whether it truly occurred) solely because there are no pairs of applicants with similar borrower and credit characteristics to review. Among the FIs with at least one pair, the median percentage of minority applicants with a similarly-situated non-Hispanic white applicant ranges from 2 to 7 percent. Focusing just on the largest FIs, this median ranges from 5 to 12 percent. Therefore, even for FIs that have at least one similarly-situated pair, the overall percentage of minority applicants with a similarly-situated non-Hispanic white applicant available for a file review is generally small for most FIs. Again, the ability to identify potential disparate treatment using a file review of similarly-situated pairs is limited for these FIs solely because of a lack of pairs of applicants with similar borrower and credit characteristics to review.

In general, the results in Table 1 convey how it can be difficult to identify similarly-situated pairs for fair lending reviews. In addition, since these results are based only on HMDA data, they likely significantly overestimate the number of similarly-situated pairs actually available for review. There are several additional pieces of information beyond what is available in HMDA data that lenders typically consider when making credit decisions and that economists would therefore consider when identifying similarly-situated pairs during fair lending reviews. First, product and program are two key matching variables since policies and procedures typically differ by product and program. For example, a DTI of 45% might be within policy guidelines for some programs but not others. HMDA data does not include either of these variables, so we were unable to match on product or program here. Second, credit score and CLTV are only proxies for all of the specific factors in an application that lead to denial reasons of credit history and collateral. HMDA data do not include these more granular variables, so we were unable to match on these variables here. Finally, per the IFLEP, fair lending risk tends to be higher for marginal applications, exceptions, and overrides. Again, HMDA data do not contain flags for these variables, so we were unable to incorporate these variables here. Based on past experience during fair lending reviews, adding these additional factors to the set of matching criteria when identifying similarly-situated pairs typically causes a significant reduction in the number of pairs available for review.

Pairs of applications with similar borrower and credit characteristics, but with different credit decisions, would clearly be compelling evidence of potential disparate treatment.

However, given the challenges in identifying pairs of applications with similar borrower and credit characteristics, this approach has some limitations. This is one gap that statistical analyses can potentially fill since these analyses compare all applications to all other applications and are not limited to comparing pairs of applications.

Individual Applicants

A common alternative to reviewing similarly-situated pairs during fair lending reviews is an audit review of individual applications where examiners essentially re-underwrite applications to determine whether the lender fairly and consistently applied its policies. Similar to reviews of similarly-situated pairs, audit reviews typically focus on a sample since file reviews are time and resource intensive. These samples typically include a few hundred applications.

There are two significant challenges with this type of file review transaction testing that create the potential to erroneously conclude no discrimination in scenarios when discrimination has in fact occurred. First, it is often easy to justify credit decisions when reviewing one application against policies in isolation, especially marginal applications, even when discrimination has actually occurred. Applications, especially for mortgages, are typically quite complicated, and contain a large number of documents, so there are typically many pieces of information to point to that justify any credit decision when reviewing one application against policies in isolation. This is made even more difficult since lenders directly control two key pieces of information that examiners heavily rely on during file reviews: adverse action notices and application notes. For fair lending and compliance reasons, lenders have significant incentive to ensure that these notices and notes are consistent with, and justify, the credit decision on the application.

A second challenge with an audit approach is that it is impossible to check for fair and consistent decisions across applications. Specifically, it is impossible to assess whether an underwriter seeing the same set of information in another application would have come to the same underwriting decision. During file reviews, examiners typically work closely as a team, comparing notes from reviews of each application in the sample. This collaboration mitigates the risk of potentially missing discrimination to some degree. However, given that examiners only review a relatively small sample of applications, it is not possible during an audit file review for examiners to do similar comparisons for applications that were not included in the file review sample.

Given these two challenges, when discrimination does in fact occur, focusing solely on an audit review to draw conclusions will erroneously be less likely to identify discrimination overall. Statistical analyses, on the other hand, which include all applications, are not impacted by either of these challenges, and could therefore potentially fill these gaps.

III. Statistical Analysis Transaction Testing Errors

As noted above, a common result during fair lending reviews is that the statistical analysis shows statistically significant and economically meaningful disparities suggesting evidence of discrimination, and the file review shows no evidence of discrimination. In this section, we focus on when and why statistical analyses erroneously indicate discrimination in scenarios when in fact no discrimination has occurred.

As a starting point for statistical fair lending analyses, economists generate raw disparities in credit decision outcomes across groups. These disparities, called unconditional disparities, provide an initial signal of fair lending risk. Unconditional disparities can be generated by computing differences in average outcomes for two groups or by estimating a regression model including only a treatment group flag. Since creditworthiness typically differs across demographic groups, unconditional disparities might occur simply because of differences in creditworthiness. To address this possibility, economists use statistical techniques such as regression analysis to determine how much of the unconditional disparity is explained by the policy factors a lender considered when making credit decisions. Any remaining disparity after accounting for the impacts of the lender's formal policy factors would be evidence of potential discrimination. This remaining disparity is called a conditional disparity.

When a conditional disparity estimate is statistically and economically meaningful, an assessment is needed of whether the disparity is accurately capturing actual discrimination or whether it is erroneously suggesting discrimination when no discrimination actually occurred. As we show in Chapter 5, "Statistical Analyses Testing for Pricing Discrimination," regression analyses can identify discrimination with a very high degree of accuracy, but only under certain conditions. When these conditions are not met, potential errors such as suggesting evidence of potential discrimination when in fact no discrimination has occurred, can arise. The primary driver of these errors is omitted variable bias.[5]

Omitted variable bias is statistical bias caused by omitting relevant factors from a regression model. Specific to fair lending analyses, if the economist does not include in the regression model every policy factor that the lender considered when making credit decisions, the estimated impact of belonging to a demographic group will partially reflect the impacts of the omitted policy factors. In other words, when there are omitted factors, the estimate of whether discrimination exists will be biased. The generally accepted view is that omitted variable bias

[5] Model mis-specification and poor data integrity are two other common drivers of these errors.

affecting disparities in fair lending statistical analyses is positive. As a result, when there are policy factors that are omitted from the regression model, the typical concern is that the estimated conditional disparities are overestimating the true amount of discrimination. Therefore, omitted variable bias is often used to support the claim that statistical analyses mistakenly indicate that discrimination has occurred when in fact there is no discrimination.

Some caution is needed when applying these assumptions and interpretations of omitted variable bias to draw conclusions. Although omitted variable bias in disparities in fair lending statistical analyses is often positive, it is not always positive as Dietrich (2005) shows. In instances when this bias is negative, the estimated conditional disparities would actually be underestimating the true amount of discrimination. Overall, the direction and magnitude of omitted variable bias is very complex, depending on all of the correlations between the credit decision outcome, demographic variables, omitted factor(s), and included factors, as well as the distributions of these factors. Unfortunately, it is typically not possible to quantify the omitted variable bias during actual fair lending analyses. From an empirical perspective, the direction and magnitude of omitted variable bias is impossible to estimate for the reason it exists in the first place, the electronic data for the omitted factor(s) were unavailable to incorporate into the analysis. It is also difficult to assess this bias from a theoretical perspective as well due to the complexities of all of the underlying correlations, especially for models with larger numbers of factors.

Given these challenges and uncertainties, when electronic data are not available for all relevant policy factors, focusing solely on statistical disparities to draw conclusions can erroneously lead to conclusions of discrimination when in fact discrimination has not occurred. Because of this, it is important to assess which policy factors are omitted, how important they are to the lender's overall decision-making process, and, to the extent possible, how they might be impacting disparities. As a point of comparison, file review transaction testing, which includes all information for all applications, is not impacted by these challenges.

IV. Considerations/Decisions When Conducting Transaction Testing

Having discussed some of the primary types of errors that can occur with transaction testing, we now pivot to discuss details of how to actually apply each transaction testing approach. As with any analysis, both transaction testing approaches require several very specific decisions to properly implement. In this section, we discuss several of these specific decisions and provide suggestions for best practices for fair lending reviews when possible and appropriate.

The objective here is not to develop an overall optimal approach to transaction testing. Transaction testing is too complex, and the optimal approach will be fair-lending-review-specific. Instead, the objectives here are to identify many of the primary issues that arise when conducting transaction testing, identify possible options for addressing these issues as well as the strengths and weaknesses of these options, and when appropriate and feasible recommend potential best practices. The hope is to facilitate more informed decisions about how to conduct transaction testing, and subsequently more accurate and defensible conclusions during fair lending reviews. In addition, this information will help inform decisions on what transaction testing evidence to rely on when the evidence is contradictory or potentially less reliable as discussed in Sections II and III.

IV.1 Reliability of Statistical Analysis and Results

The first consideration we discuss is the reliability of the statistical analysis and results. The reliability of the statistical analysis and results is a key consideration in determining whether they can be relied upon alone to draw conclusions about disparate treatment, or only provide signals of risk. This in turn is a key consideration in determining whether a file review is needed, the type of file review conducted, and the objective of the file review.

As a starting point, statistical analyses can always generate signals of risk for every fair lending review, although the strength of the signals will vary. Whether a statistical analysis can go further and be relied upon to draw conclusions about discrimination depends on the quality of available data and strength of the analysis. Most important is whether accurate electronic data are available for all, or most, of the policy factors the lender considered when making the credit decision that is the focus of the specific focal point. To facilitate this assessment, it is important to read all policies and procedures related to the decision-making processes the lender used for applications in the given focal point, make an exhaustive list of all factors the lender considered when making the given credit decisions on these applications, and then note for which factors accurate electronic data are available. This exercise can be very difficult, especially if the policies and procedures for a given focal point are broad and subjective, so the list of policy factors may not be short or highly specific in every instance. However, once completed, this information will be very useful in assessing the quality of the regression models and whether the statistical results are reliable enough to draw conclusions about discrimination. Additional types of information that are useful for this assessment include,

- Is there a sufficient volume of applications for the focal point?
- Are there sufficient volumes of applications for each prohibited basis group of interest?

- Do the coefficient estimates on each of the policy factors in the regression model have the expected sign and magnitude?
- Do the statistical goodness-of-fit measures suggest a high-quality regression model?
- Are the disparity estimates statistically and economically meaningful?
- Are the disparity estimates robust to different model specifications?
- Are the disparity estimates robust to outlier applications?

With all of this information at hand, economists and fair lending attorneys should discuss the strengths and weaknesses of the statistical analysis and results to determine whether they are reliable enough on their own to draw conclusions about whether the legal criteria for disparate treatment are met. The results of these discussions then determine the type and objective of the file review.

IV.2 File Review Objective

A second consideration for transaction testing is the objective of the file review. As noted above, there has been long-standing disagreement about whether statistical analyses alone can be used to draw conclusions about discrimination or whether these statistical analyses only provide signals of risk and that evidence from a file review is necessary to draw conclusions about discrimination. Because of this, all stakeholders should discuss, and come to agreement on, the objective of the file review early in the fair lending review.

Following item IV.1 above, if the statistical analysis and results are likely to be reliable enough on their own to draw conclusions about whether the legal criteria for disparate treatment are met, the file review should focus solely on gathering additional information to check and improve the statistical analysis. Specifically, the file review should be used to identify data integrity issues; verify that the statistical analysis and models include all of the policy factors the lender considered when making credit decisions; identify any policy factors the lender considered that were not incorporated into the analysis; and conduct general investigative checks to assess whether any policies, procedures, or data have been missed or mis-understood. Economists should then feed this information back into the statistical analysis to improve the accuracy of the disparity estimates. For this file review objective, the best type of files to review are outlier files that have high prediction errors or a significant impact on the disparity estimates, or a random sample of individual files. Both of these file review types will be discussed in more detail below.

If the statistical analysis is weak, primarily due to limited or bad data, and clearly not sufficiently reliable on their own to draw conclusions about discrimination, the objective of the file review should be to draw conclusions about discrimination. For this scenario, the file review should generally follow the guidance on transaction testing in the IFLEP, with a few modifications discussed below that address some of the concerns raised in Sections II and III. In this scenario, results from the file review would drive conclusions about discrimination, with the statistical results providing supporting evidence.

The two scenarios above are extreme cases where the statistical analysis is either clearly reliable or very unreliable. There is also an intermediate scenario where the statistical analysis and results are not reliable enough on their own to draw conclusions about whether the legal criteria for disparate treatment are met, but still compelling enough to raise significant concerns. This scenario is somewhat common during fair lending reviews and what often occurs is that evidence from the file review is relied upon to draw conclusions, and the statistical disparities are ignored. We believe that this is not the appropriate approach, especially given concerns raised above about the likelihood that file reviews will mistakenly miss some discrimination. Instead, we argue that in this intermediate scenario, additional statistical analysis and file review should be conducted focused on supporting or refuting the compelling statistical disparities, and that all evidence from both transaction testing approaches should be used holistically to draw overall conclusions.

IV.3 Type of Applications for File Review Transaction Testing

A third consideration, specific to file review transaction testing, is the type of applications to review. There are four general types of applications commonly used for file review transaction testing: similarly-situated pairs, randomly-drawn individual applications, individual marginal applications, and individual outlier applications. Randomly-drawn individual applications are discussed in detail in item IV.4 below, so we focus on the remaining three application types here. For each of these three types of applications, several specific analytical choices are needed when generating file review samples. We discuss these specific choices below for each application type in turn. For all discussions in this section we assume that the statistical analysis is not sufficiently reliable, so that the objective of the file review is to draw conclusions about discrimination.

Similarly-Situated Pairs

The first application type we discuss is similarly-situated pairs. We start with, and get into the most detail about, this application type since it is the primary focus of the IFLEP guidance on transaction testing. Generating file review samples consisting of similarly-situated pairs requires several specific analytical decisions. We discuss five specific considerations here.

1. General Approach to Using Similarly-Situated Pairs for Transaction Testing

The first consideration we discuss is the general approach to using similarly-situated pairs for transaction testing. The IFLEP includes a general approach to identifying and reviewing pairs, which regulators have used for fair lending exams for many years. As noted in Section II, however, there are some potential shortcomings to this approach. Here, we recommend a slightly modified version of the IFLEP approach that addresses some of these shortcomings, and potentially provides more information and leads to more accurate conclusions.

Following guidance in the IFLEP, a similarly-situated pair is a set of two applications with similar borrower and credit characteristics, but from different demographic groups and with different credit decision outcomes. Although presented as one concept, similarly-situated pairs actually consist of two distinct components. The first component is the requirement that applications have similar borrower and credit characteristics and belong to different demographic groups. This component focuses only on whether there are any similarly-situated applications from two groups and is completely unrelated to whether the lender discriminated. The second component is the requirement that credit decision outcomes differ for the two applicants in a pair. Different credit decision outcomes for similarly-situated applicants suggests evidence of disparate treatment.

This two-component perspective is important because viewing similarly-situated pairs as one concept ignores potentially useful information. Specifically, by matching on different credit decision outcomes initially, pairs of applications with similar borrower and credit characteristics and from different demographic groups, but with the same credit decision outcomes will never be identified or reviewed. These applications, to the extent they exist, provide evidence of a lender treating similarly-situated applications from different groups fairly and consistently. From an analytical perspective, to develop a full understanding of the overall patterns in a lender's treatment of applications, it is important to review sets of applications that might show potential disparate treatment and also sets of applications that might show fair and consistent treatment. As an example of why this matters, suppose we have a set of two treatment group and two control group applicants all with similar borrower and credit characteristics where one application from each group was denied. Reviewing only the pair of applications consisting of the one treatment group denial and the one control group approval would not accurately reflect the overall patterns for this set of similarly-situated applications.

Given the above concerns, we recommend modifying the IFLEP guidance into a three-step approach when generating and reviewing similarly-situated applications during a fair lending review. The first step consists of using available electronic data to identify all sets of applications that are similarly situated based on the borrower and credit characteristics the lender considered when making credit decisions and where there is at least one applicant from two demographic groups. It is important here to include only sets with applications that are as similarly situated as possible given the available electronic data and not to relax the matching criteria to increase the number of sets available for review. A given set might contain just one treatment group and one control group applicant or multiple applicants from one or both groups.

The second step is to conduct a full review of all applications in these sets to identify which sets contain truly similarly-situated applications. Based on results from Section II, we expect that the volume of sets from step 1 to review will generally be manageable even for larger lenders and focal points with larger volumes of applications. However, if the volume of sets for review is too large for available resources, then random sampling can be used to identify a smaller sample of sets to review. If after the full review, no sets remain, then a review of similarly-situated applications cannot be used to draw conclusions about disparate treatment, since there are no sets of applications with truly similar borrower and credit characteristics to analyze for differences in credit decisions. In this scenario, either the conclusion for the fair lending review is inconclusive or one of the other application types discussed below needs to be reviewed to gather other evidence. If some sets still remain after the full review, then step 3 of the analysis should be conducted.

Step 3 consists of analyzing the credit decision outcomes for each group's applications in each set. For a given set, if the credit decision outcomes for each group are exactly the same, or proportionately the same,[6] then the lender's treatment of applications in that set is consistent with no disparate treatment. Alternatively, if the credit decision outcomes differ for the two groups in a given set, then the lender's treatment of applications in that set suggests potential disparate treatment. If the number of sets remaining after step 2 is small, step 3 should focus on whether there is evidence of a practice of discrimination. Any conclusions about a pattern of discrimination should be inconclusive or other application types discussed below need to be reviewed to gather additional evidence. If the number of remaining sets is larger, then step 3 should include an analysis of patterns of treatment across all of the remaining sets of similarly-situated applications. This analysis should include summary statistics showing the number and percentage of approvals and denials for each group for each set of similarly-situated sets. Non-parametric tests, such as the Wilcoxon sign and ranked-sign tests, can then be used to formally

[6] As an example of what we mean by "proportionately the same," suppose a set of similarly-situated applications consists of two treatment group applicants and two control groups applicants. If one treatment group applicant (50 percent) and one control group applicant (50 percent) were denied, this would be a simple example of 'proportionately the same" treatment.

test for patterns in differences in credit decision outcomes between groups across all sets.[7] If the number of remaining sets is large enough, evidence from these analyses may be able to support conclusions about patterns of disparate treatment.

There are three important caveats here. First, whether the recommended three-step analysis is appropriate depends on the legal criteria for evidence of disparate treatment. The three-step approach takes a broad view of what could constitute evidence of discrimination by considering overall patterns of both discriminatory, and fair and consistent, treatment of similarly-situated applications. This approach would not be appropriate if from a legal perspective only similarly-situated pairs showing discrimination against the treatment group is necessary to meet the legal criteria for disparate treatment or other legal theory. We leave this question to fair lending attorneys to address. Second, care must be taken when using results from the proposed three-step approach to draw conclusions about overall patterns of discrimination, since it utilizes a unique, non-probability sampling strategy where applications are included in the sample for the sole reason that other similarly-situated applications exist. The next item below provides more details on this issue. Finally, as a reminder, and as discussed in Section II, a file review of similarly-situated applications, however conducted, will never identify discrimination against applications with no other similarly-situated applications.

[7] To provide the intuition behind these non-parametric tests, suppose there are 10 sets of similarly-situated applications. If there is not a pattern of discrimination, we would expect about the same number of sets to show better credit decision outcomes for the treatment group as for the control group. As an example, if 3 sets show better outcomes for the treatment group, 3 sets showed better outcomes for the control group, and 4 sets showed equal treatment, this would be consistent with no pattern of discrimination. The non-parametric tests formalize this intuition.

2. Approach to Identifying the Treatment Group Sample and Control Group Sample

A second, related consideration is how to identify the specific treatment and control group applications that comprise the sample of similarly-situated pairs for the file review. The two most common approaches for identifying these applications are the IFLEP approach, which is a hybrid approach combining both non-probability and probability sampling, and a purely data-driven approach where available electronic data are used to identify similarly-situated applications. For both approaches, the underlying premise is that the results from the review of the sample of applications can be used to draw conclusions about the entire population of applications.

The IFLEP approach to identifying the specific applications for a sample of similarly-situated pairs consists of three general components. First, since the risk of discrimination is higher when there are exceptions and overrides, the IFLEP recommends including all applications with an exception or override contingent on the lender being able to identify them. This is a non-probability, or targeted, sampling approach. Second, the IFLEP recommends focusing on marginal denials and approvals, since those applications tend to face a higher risk of discrimination as well. The Appendix of the IFLEP provides guidance on how to identify marginal applications. When initially pulling the file review sample, examiners rely on HMDA data, as well as any other available electronic data, to increase the likelihood of including marginal applications in the sample. Once the initial file review sample has been pulled and all of the granular application information is available, the sample is then narrowed to just the truly marginal denials and approvals based in large part on the criteria in the Appendix of the IFLEP. This is a non-probability, or targeted, sampling approach as well. Third, if there are insufficient numbers of overrides, exceptions, and marginal applications combined to meet the sample size guidelines for denials and approvals in the Appendix of the IFLEP, then random sampling is used to fill out the sample. This is a probability sampling approach.

At a general level, the non-probability, or targeted, sampling portion of the IFLEP approach is a standard and effective approach for fair lending reviews. If there is no evidence of discrimination in a sample of applications with the highest risk of discrimination then it is reasonable to conclude there is no discrimination for any applications. However, the IFLEP approach has two important limitations. First, since the samples of treatment group denials and control group approvals are identified independently of one another, and then similarly-situated pairs are identified from these two samples, this reduces the likelihood of finding pairs. As an example of this limitation, suppose there are a large number of marginal treatment group denials and control group approvals, such that sampling of these marginal applications is needed to stay within the sample size guidelines in the Appendix of the IFLEP. For the marginal treatment group denials, the likelihood of finding a similarly-situated marginal control group approval will be reduced since the pool of potential matches is just the marginal control group approvals in the sample and not all control group approvals. Overall, if the IFLEP approach is used, some steps should be taken to ensure that the entire set of approved control group applications relevant to the given type of denied treatment group applications is searched for similarly-situated applications, instead of just the approved control group applications in the sample.

A second limitation of the IFLEP approach is that it could result in a potentially inefficient use of resources. Unless there are very few overrides, exceptions, and marginal applications combined, the probability sampling portion of the sample, if it exists, will likely contain too few applications to take advantage of any of the statistical benefits of using a random sample to infer information about the entire set of applications. Therefore, reviewing these additional applications will have less value, and may not be the best use of limited resources.

The second common approach to identify the specific treatment and control group applications that comprise the sample of similarly-situated pairs for the file review is to search through all available electronic data to identify all possible pairs of similarly-situated applications. This approach does not use probability sampling, so formal statistical conclusions cannot be drawn. Instead, the approach uses non-probability sampling, where applications are included in the sample for the sole reason that other similarly-situated applications exist. If similarly-situated applications are necessary to meet the legal criteria for disparate treatment, then this approach would be an ideal targeted sample, since it focuses on identifying similarly-situated pairs. However, if other types of evidence also can meet the legal criteria for disparate treatment, then this non-probability sampling approach may miss potential discrimination, since it would likely not include many overrides, exceptions, or marginal applications, which are typically considered to be at higher risk of discrimination. As a general rule, if non-probability sampling is used to generate a sample of similarly-situated pairs, care must be taken to understand how the sampling approach relates to the likelihood that applications in the sample are at high risk of being subjected to discrimination, as well as how to properly interpret the results and draw conclusions.

Given the limitations of the two approaches, we recommend exploring a combination of the two approaches that minimizes these limitations. Specifically, the first step would follow the IFLEP approach and identify all denied treatment group applications with overrides and exceptions, as well as all marginal applications. The second step would then follow the data-driven approach to use all available electronic data to search all control group approvals for similarly-situated applications. This purely non-probability sample focuses on both the highest risk treatment group denials and the highest likelihood of finding similarly-situated control group approvals.

3. Definition of "Similarly-situated"

A third important consideration when generating a sample of similarly-situated pairs is how to define "similarly-situated." There are two primary components to this definition: what factors to match on and what specific matching criteria to use for each factor. Consistent with the discussion from above, we do not include credit decision outcomes as a matching factor for the discussion of definitions of similarly-situated here.

The first component, what factors to match on, is relatively straightforward. The objective when generating similarly-situated pairs is typically to match on every factor the lender considered when making the credit decision that is the focus of the fair lending review.[8] The recommendation above about reviewing all relevant policies and procedures and generating a list of all factors the lender considered when making credit decisions will be very useful for this component. Depending on the lender and the focus of the review, there could be just one or two matching factors, or there could be a large number of matching factors. To help facilitate the remaining discussion in this section, we use a simple example where a lender considers only two factors when making credit decisions, loan purpose and CLTV.

[8] One caveat here is that the predicted probabilities of denial from a regression model could be used as the "similarly-situated" metric. For examiners, these types of pairs are often challenging to review, because the underlying characteristics of two applications in a given pair may be very different, even though the predicted probabilities of denial are similar. In addition, since pairs are often used as tangible evidence to convey legal arguments, the effectiveness of pairs constructed in this way for this purpose is unclear.

The second component, choosing the specific matching criteria for each factor, is more complicated since there are several reasonable approaches. The types of factors the lender considers when making credit decisions will impact each of these approaches, so before discussing each approach in detail, we first define the two most common types of factors, categorical and continuous. A categorical factor takes on a discrete number of values. Loan purpose, which takes on three values, -- home purchase, home improvement, and refinance -- is an example of a categorical factor. A continuous factor takes on any value, including integers and fractions, within a given range. CLTV, which can range from 0 to over 100 is an example of a continuous factor.

We now present five commonly used matching criterion. Although we discuss each separately, in practice various components of different criterion are often combined to create a hybrid approach. The first possible matching criterion requires exact matches for the values of every matching factor. With this definition, a similarly-situated pair for our example would consist of two applications from different demographic groups and with the exact same values for loan purpose and CLTV. For example, both applications might be for a home purchase loan and have CLTVs of 75.35. If the credit decision outcomes differed for two applications from different demographic groups and with exactly the same values for every policy factor the lender considered when making the credit decision, this would be very compelling evidence of disparate treatment. Unfortunately, it is extremely unlikely that such pairs of applications exist, especially if the lender considers several continuous factors, and the focal point has smaller numbers of applications.

A second possible matching criterion ties the matching criteria directly to the lender's policies. This is our recommended approach since it anchors on which applications the lender viewed as similarly-situated per policy when making credit decisions. Continuing our example from above, suppose a lender charges one interest rate on home purchase loans to applicants with a CLTV between 0 and 80, and a higher interest rate on home purchase loans to applicants with a CLTV above 80. A similarly-situated pair would consist of two applications from different demographic groups who have the exact same values for loan purpose and CLTVs within the same policy bucket. This criterion relaxes the exact match requirement of the first criterion, which increases the likelihood of finding pairs. Also, pairs tied to policies provide compelling evidence of disparate treatment. However, it still might be difficult to find any similarly-situated pairs, especially if the overall volume of applications is small and the lender considers a large number of factors when making credit decisions. In addition, when a lender's policies and procedures contain broad criteria, such as, "we consider each applicant's overall credit history," it will be difficult to determine when two applications are similar for these policy factors. In these instances, some aspects of the other matching criterion discussed in this section can be helpful.

A third possible criterion requires the values of the continuous factors to be within a specified range, such as +/- 5 points for example. Continuing our example from above, a similarly-situated pair would consist of two applications from different demographic groups, with the same values for loan purpose, and with CLTV values within +/- a certain number of points. A significant shortcoming with this definition is that the values for a given factor might be in different policy buckets for the two applications in a pair. When this occurs, it would be expected that the credit decision outcomes would differ for the two applications. For example, suppose we use the same CLTV policy buckets of 0-80 and 80+ as above. A similarly-situated pair could consist of two applications from different demographic groups, with the exact same values for loan purpose, and with CLTVs of 78 and 82. Given the CLTVs are in different policy buckets, we would expect the rates for the two applications to differ. To avoid scenarios such as this, an additional constraint is typically applied to ensure that each of the values for each factor for each application in a pair are within the same policy bucket.

A fourth possible criterion requires the application in a pair with the worse outcome to have equal to or better values for each matching factor, as compared to the application with the better outcome. Often, a cap is also applied limiting how much better the values are for the application with the worse outcome. Continuing our example from above, a similarly-situated pair would consist of two applications from different demographic groups, with the same values for loan purpose, and with an equal or lower CLTV value for the application with the worse outcome. With a cap, the CLTV for the application with the worse outcome would be lower but not more than say 10 points lower. The lower CLTV may or may not be in the same policy bucket as the higher CLTV. This approach is consistent with the IFLEP guidance on file review transaction testing, but would not be suitable for the three-step approach to generating similarly-situated pairs that we recommended above.

A fifth possible matching criterion relies on a distance measure, which quantifies the aggregate distance between all of the matching factors for a set of two applications. There are several available distance measures, such as the Euclidean Distance, Manhattan Distance, Jaccard Distance, Minkowski Distance, cosine index, and more. The Euclidean Distance is the most common of these measures. For most analyses, similarly-situated pairs will consist of applicants that match exactly on each categorical variable, so any distance measure would typically only be constructed using the continuous factors. As an example of how to calculate the Euclidean Distance, we expand the policy criteria above to include a second continuous factor, credit score. Suppose application 1 is for a home purchase loan with a CLTV of 90 and a credit score of 700, and that application 2 is for a home purchase loan with a CLTV of 80 and a credit score of 720. The Euclidean Distance for this pair would be,

$$\sqrt{(90-80)^2+(700-720)^2}=\sqrt{100+400}=22.36. \quad (1)$$

The objective with this approach is to identify sets of two applications from different demographic groups that have the smallest distance measure. Although theoretically strong, this approach can cause confusion, since for the applications in a similarly-situated pair with a small distance measure, the values for each individual matching factor may not actually be very close, in the same policy bucket, or in the expected order. Given this potential confusion, some aspects of the first four matching criteria are typically also imposed when applying this matching criterion.

4. Sampling With or Without Replacement

A fourth consideration when generating similarly-situated pairs is whether to identify pairs with or without replacement. Generating pairs with replacement means a given control group application can be used as a match for multiple treatment group applications, and matching without replacement means a given control group application can only be used as a match for one treatment group application. Using sampling with replacement, one unusual control group application with poor borrower and credit characteristics but a favorable outcome due to idiosyncratic reasons or possibly data errors, might end up matching to several treatment group applications with worse outcomes. This would likely not be a very useful file review sample, since reconciling the control group application would explain away many of the pairs. One potential solution to this possibility is to search for multiple control group application matches for each treatment group application, instead of identifying pairs with just one treatment group and one control group application. Using sampling without replacement comes with a risk of lower-quality pairs as well. Since the pool of available control group applications will get smaller as the process of identifying pairs progresses, the quality of pairs may need to be relaxed just to identify a sufficient number of pairs for the review. This risk is difficult to address.

The three-step approach that we recommended above for file reviews of similarly-situated pairs utilizes sampling with replacement, so a decision on sampling with or without replacement is not needed. For the IFLEP approach, however, a decision on this issue is needed. If the IFLEP approach is used, we recommend matching without replacement, and searching for multiple control group application matches for each treatment group application to minimize concerns about a small number of idiosyncratic control group applications leading to easily explained pairs.

5. Quality vs Quantity

The final consideration we discuss related to generating file review samples of similarly-situated pairs is the tradeoff between the quality and quantity of pairs. We have already discussed various aspects of this consideration in a variety of places above. However, this is such a commonly occurring and important issue that it merits its own specific discussion here as well.

The typical objective when generating a file review sample of similarly-situated pairs is to match on every factor the lender considered when making the credit decision that is the focus of the fair lending review. When the lender considers only a small number of factors, it is generally easy to identify sets of applications that match on every factor. However, as the number of policy factors increases, two significant challenges arise. First, electronic data are often not available for every factor, making it impossible to match on those factors. Second, unless the volume of applications is extremely large, it quickly becomes impossible to find any sets of applications that match on every factor. The typical solution to these challenges is to relax the matching criteria by either matching on fewer factors or relaxing the criteria for specific factors, such as widening the range for an acceptable match for a continuous factor, from say +/- 5 points to +/- 10 points. By relaxing the matching criteria, however, the pairs will be lower quality and information from the file review will often readily show that applications in a given pair are not actually similarly situated. As a result, reviewing these pairs turns out not to be a good use of time or resources. Therefore, in these instances, instead of relaxing matching criteria just to have pairs to review, we recommend not focusing the fair lending review on similarly-situated pairs, but instead either reviewing other application types or relying more on the statistical analysis and results.

Marginal Applications

The second application type we discuss is marginal applications. A marginal application has borrower and credit characteristics that place it on the margin of the lender's decision-making process between a positive and an adverse credit decision. Since the concept of a marginal application is not as relevant for pricing, marginal applications are typically only used for fair lending reviews of underwriting decisions.[9] It is widely accepted that discrimination in underwriting decisions is more likely to occur for marginal applications, because of the potential impact of unconscious bias on applications where the underwriting decision is not obvious, and

[9] One possible exception here is applications on the border between risk-based pricing tiers when the lender allows some degree of pricing discretion.

because of the difficulty in reviewing for and ensuring fair and consistent decisions on these applications. Because these applications are at higher risk of discrimination, file reviews focused on marginal applications almost always use non-probability sampling. If the review of a sample of marginal applications identifies no evidence of discrimination, this implies there will be no discrimination in the overall set of applications for the entire focal point. In addition to being well-suited to draw conclusions about discrimination, focusing on marginal applications is also useful for identifying data integrity issues, as well as shortcomings of the statistical analysis, such as factors that should have been included in, or excluded from, the regression model.

Marginal applications can be incorporated into similarly-situated pairs for review or reviewed individually as part of an audit review. Therefore, many of the general, and application-type-specific, considerations and challenges mentioned elsewhere in this report are also applicable to reviews of marginal applications. An additional consideration specific to marginal applications is how to identify these applications. There are two primary approaches to identifying marginal applications, a manual approach outlined in the IFLEP and a regression-driven approach. Both of these approaches can effectively identify marginal applications for fair lending file reviews, with the IFLEP approach slightly better suited when the fair lending review focuses on file review transaction testing and the regression-driven approach slightly better suited when the fair lending review focuses on statistical analysis transaction testing.

The manual approach to identifying marginal applications follows general guidance in the IFLEP, as well as specific criteria in the Appendix to the IFLEP that characterizes both marginal denials and approvals.[10] Electronic data are typically available for some of these criteria and can be incorporated when generating the initial file review sample, while information for

[10] See pages 20 and 21 in Interagency Fair Lending Examination Procedures.

other criteria are only available in the complete application and are applied during the file review to weed out applications in the initial sample that are not truly marginal applications. The IFLEP approach to identifying marginal applications is a long-standing and accepted approach for fair lending reviews, especially those focused on file review transaction testing. One potential concern with this approach is that some of the criteria are subjective, which makes the identification process judgmental to some extent. A second potential concern about accuracy and small volumes of applications is covered below in the discussion of a secondary consideration.

The regression-driven approach to identifying marginal applications relies on regression modeling. When the fair lending review focuses on the statistical analysis transaction testing, using the regression model to identify marginal applications is particularly useful, since the file review will provide information that directly improves the accuracy of the statistical analysis and results. There are several steps to this approach. The first step is to generate the overall denial rate for all applications for the given focal point. To facilitate the discussion of the remaining steps, we use a specific denial rate of 10 percent. The second step is to identify and estimate the model specification that best reflects the lender's decision-making process conditional on the available electronic data. Demographic variables should not be included in this model. For an underwriting analysis, this regression model focuses on predicting the likelihood of an application being denied. Very importantly, the regression-driven approach should only be used if the model is high quality and accurate. The third step is to generate predicted probabilities of denial for each application using the coefficient estimates from the regression model. The fourth step is to rank order all applications by the predicted probability of denial from lowest to highest. The fifth step is to use the overall denial rate, which is 10 percent in our example, to identify the predicted probability of denial value where 10 percent of the applications are above this threshold in the rank-ordered list. Finally, the marginal files are the denied applications just below, but near this threshold and the approved applications just above, but near this threshold. The intuition here is that if the regression model perfectly predicted each credit decision outcome, in a list of applications rank-ordered by predicted probability from 0% to 100%, all of the approvals would be listed first (i.e. with the lower predicted probabilities of denial) and all of the denials would be listed later (i.e. with the higher predicted probabilities of denial). The threshold that separates the last approval from the first denial in this rank-ordered list is the predicted probability of denial that identifies the applications that were just on the lender's margin between being approved and denied.

A secondary consideration specific to marginal applications is potentially small volumes of applications. For both approaches to identifying marginal applications discussed above, generating the initial file review sample of marginal applications relies primarily on available electronic data. Although the amount of electronic data is always improving, it will almost always have some limitations, which will impact the accuracy of identifying the marginal applications. Once the file review sample has been pulled, and all of the granular application information is available, the sample is typically narrowed since some applications originally identified as marginal turn out not to be so. While this narrowing step more accurately identifies marginal applications, it reduces the size of the sample available for review. If the remaining sample of marginal applications becomes too small, an alternative approach should be taken for the fair lending review.

Outliers

The third application type we discuss is outlier applications. An outlier application is an application that is extreme in some way. One common example of an outlier is an application that clearly does not meet the lender's underwriting criteria yet was approved. Reviewing outlier applications is an especially effective approach for fair lending reviews that focus on statistical analysis transaction testing to draw conclusions. Applications with a credit decision outcome that is clearly unexpected given the borrower and credit characteristics, for whatever reason, can have a significant impact on the disparity estimates from a statistical analysis. Therefore, identifying the underlying reasons for outlier applications can help determine whether the data for these applications need to be modified in some way or whether these applications need to be excluded from the regression analysis. These modifications can significantly improve the accuracy of the statistical analysis and results, and improve overall conclusions for the fair lending review. Although less so, reviewing outlier applications can also be useful for fair lending reviews that focus on file review transaction testing to draw conclusions. Specifically, reviewing outliers can effectively identify instances of overt discrimination.

Outlier applications are almost always reviewed individually as part of an audit review, since it is typically difficult to find truly similarly-situated applications for these unusual applications.[11] Therefore, only the general and audit-review-specific considerations and challenges mentioned elsewhere in this report are also applicable to reviews of outlier applications. An additional consideration specific to outlier applications is how to identify or define outliers. There are many reasonable approaches to identifying outliers, such as analyzing summary statistics, using graphical techniques such as box and whisker plots, analyzing regression residuals, exploring influence statistics, and more. All of these approaches have strengths and weaknesses and all can be effective under certain circumstances. If time and

[11] Data errors and idiosyncratic application characteristics are two common reasons for outliers. Since these issues are typically easy to identify during the file review, it is typically easy to show that applications in a pair were not in fact similarly situated. Therefore, reviewing similarly-situated pairs of outliers is generally not useful.

resources to identify outliers are limited, we recommend focusing on influence statistics. Influence statistics identify applications that disproportionately affect the disparity estimates from a regression model. Since the primary objective of a statistical analysis is to generate accurate and robust disparity estimates, these are exactly the applications that should be reviewed to meet this objective.

There are many possible influence statistics available, such as DFBeta, Cookd, Dfit, Covratio, and more. As an example of how to use influence statistics during fair lending reviews, we provide an example of just one, the DFBeta. The DFBeta statistic quantifies how much a regression coefficient estimates change when a single application is removed from the data. The first step in generating DFBeta statistics is to identify and estimate the model specification that best reflects the lender's decision-making process conditional on the available electronic data. This model should include demographic variables. The next step is to drop the first application from the data and then re-estimate the same model to generate new coefficient estimates for each factor in the model. The DFBeta values capture how much each of the coefficient estimates changed by dropping the first application. For the first application, there will be one DFBeta value for each factor in the regression model. Continuing this process for each of the remaining applications generates a DFBeta value for each factor for each application. The final step is to examine the distribution of DFBeta values for the demographic variable of interest. In most cases, this distribution should generally be centered around 0 with most applications having a small impact on the estimated disparity. Applications that have a positive impact on the estimated disparity will be in the right part of the distribution and applications that have a negative impact on the estimated disparity will be in the left part of the distribution. If this distribution has one long tail, especially the right tail which consists of applications that have a positive impact on the estimated disparity, then the applications in that tail should be reviewed as part of the file review. These are the applications that are driving the estimated disparity up, so developing a better understanding of these applications will improve the accuracy and robustness of the statistical analysis. If this distribution is symmetric, then there are no applications of concern that need to be reviewed.

IV.4 Sampling Strategies

File reviews almost always rely on sampling, because manually reviewing applications is very time and resource intensive.[12] The strategy for generating the sample will determine how to interpret the results from the sample review and what conclusions can be drawn. There are several acceptable sampling strategies available, so a decision is therefore needed about which specific strategy to use. For any given fair lending review, an argument can typically made for multiple sampling strategies, so our only recommendations in this section are to make sure all stakeholders working on the fair lending review are clear about the chosen sampling strategy, and that the conclusions drawn are appropriate for the sampling strategy used.

At a general level, there are two primary types of sampling strategies: probability and non-probability. We discuss each in turn.

Probability Sampling (Random Sampling)

Probability sampling is a strategy that randomly selects a set of applications from the entire set of applications for the given focal point. Importantly, each application has a known and

[12] Unless electronic data is very limited or there are significant data integrity issues, statistical analysis transaction testing typically uses all applications, so no sampling is needed. However, file reviews focused on gathering information to support, understand, or improve the statistical analysis almost always relies on sampling as well.

non-zero probability of being included in the sample. Results from the sample of applications are then used to infer results about the entire set of applications for the focal point. Specifically, with probability sampling, we can estimate a rate of potential discrimination in the sample and then use that as an estimate of the rate of discrimination in the overall set of applications. Since this estimate is based on only a sample of applications, there will be some uncertainty in the accuracy of the estimate. However, since probability sampling is grounded in statistical theory, we can quantify this uncertainty, which makes it possible to estimate the level of confidence in the accuracy of the estimate.[13] These formal statistical results are valuable when drawing conclusions during fair lending reviews. In addition to these statistical benefits, probability samples are also useful for investigative purposes to identify areas of risk, for assessing data integrity, and for identifying shortcomings of the statistical analysis, such as factors that should have been included in, or excluded from, the regression model. The major shortcoming of probability sampling is that it is not as effective at identifying individual instances of discrimination.

One common type of probability sampling is simple random sampling, where each application in the entire focal point has the same likelihood of being included in the sample. This is akin to putting all applications into a hat and then drawing a set number of applications from the hat for review. When simple random sampling is used, the results from reviewing the sample can be directly used to infer results about the entire focal point. As an example, suppose 100 treatment group applications are sampled using simple random sampling and then reviewed. If the file review determines that 5 percent of these applications were subject to discrimination, then we can infer that 5 percent of the entire set of treatment group applications in the focal point

[13] The details of this uncertainty and these tests are beyond the scope of this paper.

were subject to discrimination. We can then quantify the uncertainty resulting from using a sample to estimate the level of confidence in the accuracy of the estimate. As an example, a conclusion based on this sample result would look something like, we are 95 percent confident that the rate of discrimination for the entire set of applications is between 4.1 and 5.9 percent.

A second common type of probability sampling is stratified random sampling, where the applications in the focal point are stratified into subsets of applications and then random samples are drawn from each stratum. For fair lending reviews, strata are typically defined by demographic group and outcome to ensure the sample includes sufficient minority applications with adverse outcomes when the overall volume of minority applications is small. For example, four common strata for underwriting analyses are minority denials, non-minority denials, minority approvals, and non-minority approvals. With stratified random sampling, the probability of being included in the sample typically varies across applications, so results from reviewing the file review sample cannot be directly used to infer results about the entire focal point. Instead, any statistical results from the file review need to be weighted so that they are representative of the overall focal point.[14] Once weighted, the same types of conclusions discussed above for simple random sampling can be drawn.

A probability sample of individual applications is a fourth application type (in addition to similarly-situated pairs, marginal applications, and outliers) to consider when conducting file review transaction testing. There are two primary scenarios where reviewing this application type can been particularly effective. First, if electronic data to identify overrides, exceptions, marginal applications, or similarly-situated pairs are very limited or poor, probability sampling

[14] There is a very large volume of available resources on sampling and weighting. See for example, Cochran (1977) and Lohr (2022).

can provide a random sample of individual applications for an audit-style review that can be used to generate estimates of the rate of potential discrimination. The types of conclusions that can be drawn with this type of sample are similar to the example discussed above for simple random sampling. Second, probability sampling can also be an effective strategy for assessing how relevant policy factors not included in the regression model (i.e., omitted variables) might be impacting the disparity estimates from the model. In this use case, the file review would focus on how often an omitted variable was salient to the lender's credit decision, and whether the omitted variable was more relevant for some groups. The conclusions here would be less statistically based, and more informative to the conclusions based on the statistical disparity estimates.

In addition to generating a separate application type, probability sampling can also be used to generate the other three application types discussed above. As noted above, the IFLEP approach for identifying a sample of similarly-situated pairs incorporates probability sampling. Also as noted above, however, the statistical benefits of probability sampling are generally not realized in this use case, since the random samples are typically too small. In addition, probability sampling can be used to identify a random sample of treatment group denials, which can be the first step in generating a sample of similarly-situated pairs. Searching through all control group applications to identify matches for these treatment group denials yields the sample of pairs for review. If similarly-situated control group approvals are readily available, with this type of sample we can estimate the overall rate of discrimination and take advantage of the statistical benefits of probability sampling to draw formal statistical conclusions about potential discrimination. Finally, probability sampling can also be used to reduce the number of individual applications when the volumes of overrides, exceptions, or marginal applications exceed desired sample sizes. In this use case, the sample statistical results should be used to draw statistical conclusions specifically about the entire set of overrides, exceptions, or marginal applications. Then, because these application types are at higher risk of discrimination, these initial statistical conclusions can be used to draw judgmental conclusions about the entire set of applications for the focal point, similar to a non-probability sample as discussed next.

Non-Probability Sampling (Targeted Sampling)

Non-probability sampling is a sampling strategy that selects applications with particular characteristics that can provide insight into specific questions. Similar to probability sampling, we can use results from the sample of applications to infer results for the entire set of applications for the focal point. Unlike probability sampling, we cannot use the estimated rate of potential discrimination from the sample as an estimate of the rate of discrimination in the overall set of applications, because the sample is not representative of the entire focal point. Instead, we can only generate evidence of whether discrimination is likely to have occurred. Specifically, in a fair lending context, non-probability sampling is typically used to identify applications most likely to be subjected to discrimination. If a review of these high-risk applications shows no discrimination, then it is reasonable to conclude that applications at lower risk of discrimination were not subjected to discrimination either. The statistical benefits of probability samples do not apply to non-probability samples, so these conclusions are judgmental.

Unlike probability sampling, non-probability sampling is good for identifying individual instances of discrimination, since it focuses on applications with the highest likelihood of being subjected to discrimination. It can also be very effective at identifying shortcomings of the statistical analysis, especially if the sample focuses on outlier applications which have the biggest impact on the statistical disparities. The major drawbacks of non-probability sampling are that it does not allow for estimates of the rate of discrimination in the population as a whole or for formal statistical tests of whether discrimination has occurred, so all conclusions will be judgmental. In addition, as compared with probability sampling, it is not as useful for identifying general areas of risk or assessing data integrity, since targeted samples are typically not representative of the entire set of applications that comprise the focal point.

Currently, non-probability sampling is used extensively during fair lending reviews. Reviews of overrides, exceptions, and marginal applications are classic examples of using non-probability sampling to identify applications most at risk of discrimination. Similarly, reviews of outliers are examples of non-probability sampling to identify applications that have the biggest impact on the statistical results. Identifying similarly-situated pairs is a third common use of non-probability sampling, but there are some nuances here. For this use case, applications are included in the sample because other similarly-situated applications exist. If similarly-situated pairs are necessary to meet the legal criteria for disparate treatment, then these applications are the highest risk applications by definition. However, if other evidence can also meet the legal criteria for disparate treatment, then some applications with higher risk of discrimination likely will have no chance of being sampled. As a result, some care must be taken when concluding there is no discrimination across all applications if the review of a non-probability sample of similarly-situated pairs identifies no discrimination.

IV.5 Is the Sample Appropriate?

A fifth consideration for transaction testing is the appropriateness of the sample. We have already discussed various aspects of this consideration in previous items above. However, given the importance of sampling for most fair lending reviews, it merits its own specific discussion here as well.

For probability sampling, a sample is appropriate if it is random and representative. To assess appropriateness when using probability sampling, the following steps should always be taken,

- Clearly specify the exact set of applications upon which you want to draw conclusions. This will typically be the definition of the focal point.
- Ensure that some form of random sampling is used to draw the sample from just the set of applications upon which you want to draw conclusions.
- Ensure that every application has a non-zero probability of being included in the sample.
- Determine the probability of each application being included in the sample. If these probabilities vary across applications, then weighting will be needed when constructing sample statistics.
- Check whether the sample is generally representative of the entire set of applications in the focal point by comparing the distributions of key factors using the sample data to their corresponding distributions using data for the entire focal point. If the sample is clearly not representative, do not use the sample of applications to infer any information about the entire set of applications for the focal point.

For non-probability sampling, a sample is appropriate if it focuses on the characteristics of interest to the analysis. Specifically for fair lending reviews, these samples typically consist of applications with the highest risk of potential discrimination. For these samples, conduct checks to verify that the applications in the sample do in fact contain the desired characteristics. As one example, after generating an initial sample of marginal applications using available electronic data, review additional information from the complete applications to determine whether these applications are truly marginal applications. If a meaningful percentage of applications are not in fact marginal, then the sample should not be used to draw conclusions about the entire set of applications in the focal point.

IV.6 Two or Four Quadrants

A sixth consideration for transaction testing is whether to generate file review samples using applications from two or four quadrants. On a general level, a similarly-situated pair is two applications with similar borrower and credit characteristics, from different demographic groups, and with different credit decision outcomes. This general definition encompasses two important and distinct types of pairs: treatment group applications with a worse outcome matched to control group applications with a better outcome (quadrants 1 and 2), and the reverse, control group applications with a worse outcome matched to treatment group applications with a better outcome (quadrants 3 and 4). Together, these two types of pairs are referred to as a four-quadrant file review sample.[15]

File review transaction testing often focuses on only the first two quadrants, since the goal of these reviews is typically to understand why treatment group applications received worse outcomes. In addition, given that file reviews are very time and resource intensive, focusing on just two quadrants is often the only feasible option given limited resources. However, focusing on only two quadrants does not provide a complete picture and understanding of overall patterns in a lender's decision-making. As a simple example, suppose a review of applications cannot explain 10 pairs consisting of a treatment group application that received a worse outcome than a similarly-situated control group application. Alone, this would likely be compelling evidence of disparate treatment. Now, in addition to these 10 unexplained pairs, suppose also that the review of applications cannot explain 10 pairs consisting of a control group application that received a worse outcome than a similarly-situated treatment group application. Adding this evidence may

[15] Although the concept of four quadrants applies to reviews of individual applicants as well, it is typically discussed in terms of similarly-situated pairs.

lead to different discussions and conclusions about disparate treatment. As a point of comparison, statistical analysis transaction testing typically includes all applications in the entire focal point, so these analyses use applications from all four quadrants.

On this issue, we recommend conducting a four-quadrant review if possible, since it provides the most comprehensive assessment of patterns in a lender's decision-making. However, if time and resources are limited, then focusing on just quadrants 1 and 2 is acceptable.

IV.7 Sample Size

The final consideration when conducting transaction testing is the sample size. There is extensive guidance available on determining appropriate sample sizes. Directly related to fair lending reviews is Appendix A of the IFLEP, which provides sample size tables for prohibited basis group denials and control group approvals for file review samples focused on underwriting decisions, as well as for prohibited basis group approvals and control group approvals for file review samples focused on terms and conditions decisions.[16] Guidance applicable more broadly to other types of analyses is widely available as well.[17] Depending on the specific purpose and objective of a given analysis, these sources provide many acceptable options for determining appropriate sample sizes.

To augment the extensive technical guidance available on how to choose appropriate sample sizes we discuss three additional high-level considerations. First, most sample size tables are grounded in statistical theory[18] and therefore are designed specifically for probability

[16] See page 19 of Interagency Fair Lending Examination Procedures - Appendix.
[17] Three examples include Cohen (2013), Krejcie and Morgan (1970), and the Research Advisors (Sample Size Table).
[18] See Ahmed (2024) for examples of common statistical approaches to determining sample sizes.

sampling where samples are randomly drawn and the probability that each application is included in the sample is non-zero and known. These sample size tables are not specifically intended for non-probability sampling, since this is a judgmental sampling approach with no statistical basis. However, since non-probability sampling is purely judgmental there is no correct sample size, so there is no harm in using statistically-based sample size tables to generate non-probability samples.

The second high-level point is the general rule of thumb that larger samples are typically better than smaller samples, because they provide more information. This rule of thumb applies to both probability and non-probability sampling. There is one important caveat to this rule of thumb, however. With probability sampling, samples that are too large can impact statistical tests and how they are interpreted. Whether an estimated disparity is statistically significant depends on the precision of the disparity estimate, which in turn depends on the sample size. Specifically, larger sample sizes increase precision, which increases the likelihood that a given disparity will be statistically significant, all-else-equal. If the sample size is very large, just about every disparity will be statistically significant, even disparities that are not economically significant. As a result, when sample sizes are really large, care must be taken when interpreting both the statistical and economic significance of disparity estimates. We note, though, that it would be extremely rare for a file review sample during any fair lending review to be so large to create this issue. With non-probability sampling, as the sample size gets larger, the marginal benefit of reviewing additional applications decreases, and will eventually become less than the marginal time and resource costs needed to review the application. Using marginal applications as an example, by increasing the sample size, eventually the additional applications added will become less marginal, and therefore less likely to be subjected to discrimination. Therefore, the benefit of reviewing these applications falls. As noted above, during the discussion of the random sample portion of the IFLEP approach to generating similarly-situated pairs, this issue is not uncommon during fair lending reviews.

The final high-level point is the general rule of thumb that, when using probability sampling, a minimum sample size of 30 files is sufficient to generate reliably accurate statistical tests for discrimination. For example, 30 applications from the treatment group and 30 applications from the control group would be the minimum sample if underwriting decisions are of interest, and 30 originations from the treatment group and 30 originations from the control group if pricing decisions are of interest. A sample size of 30 is often viewed as the point at which the central limit theorem begins to apply. The central limit theorem states that the distribution of sample means will be approximately normal, regardless of the distribution of the population from which the samples are drawn, as long as the sample size is large enough. This is important because many statistical tests rely on the assumption that the sample means are normally distributed. If the sample size is too small, the distribution of sample means may not be normal, and the results of these statistical tests may be unreliable. This is not an issue for non-probability sampling since there are no formal statistical tests with this sampling strategy. Non-probability sampling is much more focused on gathering information, and useful information can be gathered even with very small non-probability samples.

V. Conclusion

In this report, we attempted to clarify when and why statistical analyses and file reviews often lead to contradictory conclusions about the existence of discrimination during fair lending reviews. Specifically, we explored why statistical analyses are generally more likely to generate evidence of potential discrimination when no discrimination actually occurred and file reviews are generally more likely to generate evidence showing no discrimination when discrimination actually did occur. We then provided a detailed discussion of key analytical issues and considerations that arise when conducting transaction testing. The overall objective of this report was to highlight the analytical complexities and nuances of transaction testing to help create more informed decisions about how to conduct transaction testing, as well as more accurate and defensible conclusions during fair lending reviews.

References

Ahmed, Sirwan Khalid. 2024. “How to Choose a Sampling Technique and Determine Sample Size for Research: A Simplified Guide for Researchers,” Oral Oncology Reports, Volume 12, pps. 1-7.

Cochran, William G. 1977. Sampling Techniques, 3rd Edition, John Wiley & Sons.

Cohen, Jacob. 2013. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition, Cambridge Academic Press.

Dietrich, Jason. 2005. “Under-specified Models and Detection of Discrimination: A Case Study of Mortgage Lending,” The Journal of Real Estate Finance and Economics, Vol. 31, pp. 83-105.

Krejcie, R.V. and D. W. Morgan. 1970. “Determining Sample Size for Research Activities,” Educational and Psychological Measurement, 30(3), pps. 607-610.

Lohr, Sharon L. 2022. Sampling: Design and Analysis, 3rd Edition, CRC Press.

Stengel, Mitchell, and Dennis Glennon. 1995. “Evaluating Statistical Models of Mortgage Lending Discrimination: A Bank-Specific Analysis,” Office of the Comptroller of the Currency Economic & Policy Analysis Working Paper 95-3, May 1995.