Chapter 11

Disparate Impact Analyses: Building a Bridge from Legal Theory to Economic Analysis

Jason Dietrich
March 2026

I. Introduction

The foundational principle behind most credit decisions is that past behavior/performance is predictive of future behavior/performance. Based on this principle, lenders try to identify what characteristics are specific to consumers who paid their loans on time and what characteristics are specific to consumers who did not pay their loans on time. Lenders then use this information to make a variety of credit decisions, such as where to market their products, whether to approve applications for credit, and what terms and conditions to offer. Due in part to historical discrimination, some demographic groups consistently have worse performance on past credit. As a result, relying on past performance to predict future performance when making credit decisions generates significant disparate impact on some demographic groups. This raises questions about whether this disparate impact is illegal and what evidence is needed to prove disparate impact is illegal.[1]

From an analytical perspective, economists have a variety of well-established and easy to apply statistical tools and approaches to conduct disparate impact analyses during fair lending reviews. However, the legal criteria for what constitutes evidence of illegal disparate impact are opaque at best. As a result, economists are often unclear about what analytical approaches to apply, what tests to conduct, and what statistical evidence indicates illegal disparate impact. The objective of this report is to highlight the questions economists have when attempting to develop and conduct statistical analyses that are consistent with legal theories of disparate impact. The hope is to clarify and strengthen the relationship between legal language and theories of disparate

[1] Another very important question is, if this disparate impact is not illegal then what policy options are available to reduce or mitigate it.

impact, and the statistical analyses economists conduct during fair lending reviews to identify illegal disparate impact.

To identify the specific analytical questions economists have, we need a specific legal perspective of disparate impact to use as a framework. Unfortunately, there is significant variation in legal perspectives of disparate impact, which adds another layer of uncertainty for economists. As one example, the series of HUD rules on disparate impact from 2013, 2020, and 2023, as well as the proposed rule from 2026, alternate between more and less stringent legal criteria for disparate impact, as well as between whether disparate impact is a viable legal theory or not. For our report we use two specific documents as our legal framework, HUD's Discriminatory Effects Standard Rule from 2023 (HUD Document)[2] and the DOJ's guidance, "Section VII - Proving Discrimination — Disparate Impact."[3] Both of these documents view disparate impact as a viable legal theory and have relatively less stringent legal criteria. On a general level, it is much more interesting to discuss statistical analyses of disparate impact when disparate impact is a viable legal theory. In addition, each of these documents contains specific information that is useful for the objectives of this report. The HUD Document re-affirmed the three-prong, burden-shifting framework from HUD's 2013 rule, which is currently the standard approach for disparate impact analyses, and what we use as the structure of our report. The DOJ document provides guidance on proving a legal violation of disparate impact, which provides a direct roadmap for developing and conducting a statistical analysis of disparate impact. Although we rely on these two documents as our legal framework, we are not saying that this is the correct legal perspective of disparate impact. We leave that issue to fair lending attorneys to decide. In

[2] See Federal Register :: Reinstatement of HUD's Discriminatory Effects Standard
[3] The DOJ removed this document from their website in 2026. However, a copy of this document is available at Title VI Legal Manual- Disparate Impact.

addition, although some of the analytical issues and questions we raise may not fully apply to different legal perspectives on disparate impact, most of the issues and questions we raise are framework agnostic and will be relevant for most legal perspectives.

Within this legal framework, we now discuss a variety of analytical issues economists face when conducting disparate impact analyses, as well as questions that arise that require legal guidance. To make the discussion more concrete, we focus on one common use case, potential disparate impact that lenders' use of credit scoring models generates. To structure the discussion, we use the three prongs of a disparate impact analysis, presenting issues specific to each prong. There is substantial overlap in many of the issues discussed, and the possible options for many issues implicate the possible options for other issues. None-the-less, we present each issue separately in the hope that presenting similar issues from multiple perspectives will provide the most comprehensive understanding of the analytical challenges that economists face when conducting statistical analyses of disparate impact.

II. Estimating Disparate Impact

Under the burden-shifting framework for disparate impact, the first step is for plaintiffs to demonstrate that a policy or practice has a discriminatory effect on a protected group, meaning it has a disproportionately adverse impact on individuals based on race, color, religion, sex, national origin, or other protected class. Within this prong of the disparate impact framework, economists need legal guidance on several analytical issues.

Issue 2.1: What is the appropriate facially-neutral policy?

The DOJ Document states that disparate impact analyses begin with accurately and completely defining the facially-neutral policy in question. For most analyses, there are typically

several potentially interesting definitions of the facially-neutral policy. Since the specific analytical approach, as well as the possible conclusions, depend on the specific definition of the facially-neutral policy, economists need legal guidance early in the analysis on the appropriate definition to use.

In general, economists take a broad and flexible view of what constitutes an appropriate facially-neutral policy to analyze. Essentially, any policy a lender uses that impacts consumers in any way could be appropriate for a statistical analysis of disparate impact. Specific to the example we focus on in this report, there are many potential definitions of a facially-neutral policy for a disparate impact analysis focused on a credit scoring model, including,

- Representativeness of the development data
    - Are any demographic groups underrepresented in the development data compared to the overall population or the populations of previous or expected future applicants?
    - Does data quality vary by demographic group?
    - For each demographic group, is loan performance in the development data consistent with other available sources of information on loan performance?
- One specific characteristic[4] in the scoring model
- Groups of characteristics in the scoring model
- The scoring model as a whole
- How the lender uses the scoring model
    - Thresholds the lender applies to the scoring model
    - Other factors the lender considers in combination with the scoring model
- Overlays, exceptions, and overrides to the scoring model
- The overall decision-making process in which the scoring model is one factor
- Others

Real-world fair lending reviews typically focus on one specific facially-neutral policy. There are compelling arguments for and against each of the options listed above, so this typically

[4] A characteristic is a factor or variable (such as "number of collections") that predict the outcome and attributes are values that a characteristic can take on for particular individuals (such as 0 collections, 1-3 collections, 4+ collections).

becomes a policy choice. While this is a very important decision, it is more of a policy and legal decision than a statistical decision, so we leave this debate to attorneys and policymakers. For this report, the details of many of the issues discussed below vary depending on the specific facially-neutral policy. To avoid getting lost in all of the potential scenarios and details, and to keep the discussion manageable, we focus on just one facially-neutral policy, the scoring model as a whole. Different choices may raise different issues and questions.

Issue 2.2: What is the relevant population for the analysis?

Once we have defined the facially-neutral policy, the next fundamental question is, what is the population for the analysis? The population refers to the specific group of individuals, objects, or data points we want to focus on and draw conclusions about. A second, related concept is the sample. A sample is a subset of the population that we can use to infer information about the population. Economists often use samples to reduce costs, and because data for the entire population are often not available for analysis. Throughout most of this report, we assume that data for the entire population are available for analysis, and only discuss the additional complexities that sampling introduces into disparate impact analyses in a few instances.

From an economist's perspective, there is generally considerable flexibility in defining the population. For example, we could define the population very broadly, such as all potential applications for credit at a given lender, or very narrowly, such as applications from Asian borrowers for 30-yr, fixed rate, conventional mortgages at a given branch office during 2025. The most important concern for economists is making sure to define the population very precisely to just the set of applications of interest, since the definition of the population impacts most of the analytical decisions and conclusions that follow.

There are several analytical issues that economists consider when defining the population for a disparate impact analysis of credit scoring models. We discuss each of these in turn, keeping in mind, as noted above, that there is substantial overlap in issues in some instances.

Issue 2.2a: Should the population include actual applicants impacted by the policy or applicants that could be impacted by the policy?

The DOJ Document points out that, " ... the legally relevant 'population base' for a statistical measure of adverse disparate impact is all persons the policy or practice affects or who could possibly be affected by some change in (or the elimination of) the policy or practice."[5] This language suggests two very different populations for analysis. The phrase, "all persons the policy or practice affects," suggests defining the population as all individuals who were **actually** impacted by the lender's use of the facially-neutral policy. Alternatively, the phrase, "or who could possibly be affected by," suggests defining the population as all individuals who **could be** impacted by the facially-neutral policy if they applied for credit. The language in the DOJ Document is consistent with language in the HUD Document which states, "Among other things, the 2013 Rule codified a three-part burden-shifting framework consistent with frameworks on which HUD and courts had long relied: (1) The plaintiff or charging party is first required to prove as part of the prima facie showing that a challenged practice **caused or predictably will cause** a discriminatory effect."[6]

From an analytical perspective, focusing on individuals who were **actually impacted** by the lender's use of a policy is relatively straightforward. In our example using the entire credit scoring model as the facially-neutral policy, the population would include all, and only, applications for which the lender used the credit scoring model to make credit decisions. Even

[5] See page 21 of the DOJ Document.
[6] See page 19451 of the HUD Document.

though this is relatively straightforward, economists still need to make some analytical choices. As one example, if a lender uses a scoring model for underwriting decisions for one product and for pricing decisions for a different product, we need to decide whether to include applications for both products in the population. We discuss this specific example in more detail in item 2.2c below. This simple example highlights the overlap in the analytical issues discussed in this report, since instead of using the population definition to specify exactly which applications the disparate impact analysis will focus on, we could instead include a product or decision-type component to the definition of the facially-neutral policy.

Alternatively, focusing on individuals who **could be impacted** by the lender's use of a policy is much more challenging, because this population would include any individual who could potentially submit an application of credit to the lender, which is potentially all individuals. In our example using the entire credit scoring model as the facially-neutral policy, the analysis would attempt to assess how the credit scoring model would impact all consumers from different demographic groups if they applied for credit at the given lender and the lender used the scoring model to make credit decisions on those applications. With this population definition it would not be possible to analyze the impact of the scoring model on actual outcomes, such as credit decisions, since outcomes for potential applicants would not be available. Instead, the disparate impact analysis would focus on differences in the distributions of score values across demographic groups using either the development data or data on actual applicants, with the inference being that differences in score distributions would imply disparate impact in outcomes for potential applicants. We discuss this type of analysis in more detail in the impacts section below.

Issue 2.2b: What time period should the population cover?

Over time, lenders continually develop, implement, modify, and retire policies. These changes introduce a timing dimension to disparate impact analyses. Specific to our example of credit scoring models, lenders often utilize a given scoring model for several years, since the cost of developing a new model is high. Over time, a scoring model's predictiveness deteriorates, and as this occurs, the disparate impact the scoring model generates typically changes as well. Economists need to consider both of these intertemporal characteristics of scoring models when defining the population for disparate impact analyses.

There are three general approaches to incorporating time into the population definition.[7] First, we could focus on drawing conclusions about the disparate impact the scoring model generates over the entire life span of the model. Here the definition of the population is all applicants for which the lender used the credit score to make credit decisions over its entire life span. To include this entire population in the disparate impact analysis we would need to wait until after the lender retires the scoring model. This timing might be unpalatable, since the scoring model could generate disparate impact for many years before being identified. In addition, providing restitution to harmed consumers becomes increasingly difficult the further the analysis is removed from the actual disparate impact, since it becomes increasingly difficult to locate individuals as time passes. Alternatively, instead of waiting until the lender retires the scoring model, we could use a sampling approach to generate disparate impact estimates for a specific time period and then use these estimates to infer the disparate impact for the entire population. The key here is that the sample needs to be random and representative. Given that a

[7] In addition to direct implications for the population definition, these intertemporal characteristics also impact other components of disparate impact analyses. For example, for a scoring model that generates disparate impact, reduced predictiveness over time weakens the business justification for that scoring model.

scoring model's disparate impact typically varies over time, and in unknowable ways, sampling is likely not a viable option with this population definition. Given these challenges, this population definition is likely not viable for disparate impact analyses.

Second, we could focus on drawing conclusions about disparate impact the lender's use of a scoring model generates during a specific time period. Here the definition of the population includes a specific time period. This approach comes with several challenges as well. We present a simple example to illustrate these challenges. Suppose a lender intends to use a scoring model for five years, and that we conduct a disparate impact analysis of that scoring model at the end of year 2. Further, suppose that the scoring model created disparate impact benefiting group 1 in year 1 and disparate impact harming group 1 in year 2, and that these two impacts offset each other so there is no overall disparate impact over the combined two-year period. If we took the standard approach used for fair lending reviews and focused on the most recent two-year period, we would conclude that the scoring model did not generate any disparate impact. This approach would miss the disparate impact in year 2, which could be an early signal of potential increases in disparate impact in years 3 through 5 to come.[8] In addition, there is the question of how to interpret the beneficial disparate impact that occurred for group 1 in year 1. For this approach, it is not clear what the appropriate time period should be when defining the population for analysis, or how to interpret the variation in disparate impact estimates over time.

A third general approach to incorporating time into the population definition is a hybrid combination of the first two approaches. Under this approach, economists would conduct disparate impact analyses of the scoring model on regular intervals (for example, annually, quarterly, or monthly) over the entire life of the model. For each time period, the goal is to draw

[8] Model Risk Management plans should include continuous disparate impact monitoring to identify these types of intertemporal patterns and risks.

conclusions about the disparate impact the scoring model generates only during the specific time period. The overall objective of this approach is to identify trends in disparate impact over time. One or two periods showing evidence of disparate impact would provide signals of risk that lenders should monitor. Three or more consecutive periods showing evidence of disparate impact would support a conclusion that the scoring model generated disparate impact. This approach is most effective at monitoring and detecting disparate impact that the scoring model generates, but it could be costly to implement. Regulatory Agencies should expect lenders to explore these types of analyses as part of their regular model risk management conditional on manageable costs. For Regulatory Agencies, however, this is a difficult approach to apply, since it requires on-going analyses, and fair lending reviews are point-in-time analyses. At a minimum, fair lending reviews should include a review of lenders' on-going testing of the disparate impact that scoring models generate over time.

Issue 2.2c: What is the appropriate level of granularity for the population definition?

Closely related to the previous two issues is the question of the appropriate level of granularity for the population definition. Lenders often use a given credit score for several products and programs, and they often use the score in substantially different ways across products and programs. When a lender uses a credit score in significantly different ways across subsets of applications, the optimal analytical approach is to designate each subset as a separate population and then conduct separate analyses of, and draw separate conclusions for, each population. This segmentation approach yields more accurate disparate impact estimates than an aggregate analysis since aggregate disparate impact estimates mask variation in the disparate impact estimates across subsets. However, this approach does not work well within the three-prong framework of disparate impact analyses. For the third prong, one of the primary options

for a less discriminatory alternative to a scoring model that creates a disparate impact is to develop a completely new scoring model. If the original scoring model created a disparate impact for only some subsets of applications, the new scoring model would presumably reduce disparate impact for those subsets, but at the same time may generate new disparate impact for the other subsets. Instead of replacing the original scoring model for every use, the lender could replace the scoring model with alternative models for just the subsets of applications where disparate impact was identified. This approach would likely result in several additional scoring models, which would likely not be cost effective for the lender.

There is an important caveat to this specific issue regarding less discriminatory alternatives. If how the lender uses the scoring model (which thresholds it applies for example) causes the disparate impact instead of the scoring model itself (disparate impact occurs for any threshold for example), then the search for a less discriminatory alternative would focus on different strategies for using the scoring model (different thresholds for example). If this is the case, then we could realize the increased accuracy of disparate impact estimates with the segmentation approach within the three-prong framework for disparate impact analyses.

Overall, it is not clear what the appropriate level of granularity should be when defining the population for analysis, how to interpret multiple disparate impact estimates for a given scoring model, or what might be an appropriate less discriminatory alternative to a scoring model that creates disparate impact for only some uses. Economists need legal guidance on these issues.

Issue 2.3: What is the appropriate measure of "impact" for a given facially-neutral policy?

The next issue we discuss is the appropriate measure of a facially-neutral policy's impact. Facially-neutral policies can affect consumers in various ways, so we need to choose which

impacts to address or consider when conducting a disparate impact analysis. Economists consider several issues when making these decisions, and we discuss each of these in turn.

Issue 2.3a: Should the analysis focus on impacts on final outcomes or intermediate outcomes?

Many facially-neutral policies directly impact final credit outcomes for consumers. In addition, depending on how we define the facially-neutral policy, some policies also have intermediate impacts on consumers. The DOJ Document frames this as an issue of disparities in outcomes versus disparities in access/opportunities. As one example, for the entire scoring model, which we use as the facially-neutral policy example in this report, economists could focus on how the scoring model directly impacts final outcomes, such as denial rates for two groups. This approach is consistent with defining the population as all applicants who were impacted, as discussed in issue 2.2b above. Alternatively, economists could focus on differences in score values for two groups (either differences in average scores or score distributions), which is an intermediate outcome. This definition of impact is consistent with DOJ's statement about disparities in access or opportunity, as well with defining the population as all applicants who could be impacted, as discussed in issue 2.2b above.

The main takeaway here is that economists need legal guidance on whether to focus on intermediate outcomes or final outcomes when constructing disparities during statistical analyses of disparate impact.

Issue 2.3b: Which specific outcomes should economists focus on during disparate impact analyses?

Lenders often use credit scoring models for more than one purpose, and as a result, a model may impact multiple outcomes for consumers. For example, a credit score might impact whether a lender sends marketing materials or offers an invitation to apply for credit, the amount

of level of assistance and guidance they give during the application process (specifically, lenders might provide more assistance to applicants with higher scores), the underwriting decision on applications, whether they make a counter-offer, whether they grant an exception or override, the terms and conditions of the credit, and more.

When a lender uses a scoring model for multiple purposes, economists can often separately estimate the amount of disparate impact the scoring model creates for each use. For example, if a lender uses a scoring model to make underwriting and pricing decisions, economists can estimate the amount of disparate impact the scoring model generates separately for underwriting decisions and for pricing decisions. However, it is typically very difficult to generate one overall estimate of disparate impact across all uses. This limitation is particularly salient to disparate impact analyses, since one primary less discriminatory alternative is to develop a new scoring model.[9] If the original scoring model creates a disparate impact harming consumers for one use (underwriting for example) and benefiting consumers for another use (pricing for example), replacing the original scoring model with an alternative model for all uses would likely have some unintended consequences. Alternatively, the lender could replace the original scoring model with an alternative model for just the narrowly defined uses where the scoring model created a disparate impact, but this could potentially result in several scoring models, which would likely not be cost effective for the lender.

Overall, economists need legal guidance on what specific outcomes to focus on when conducting statistical analyses of disparate impact.

[9] See the caveat about less discriminatory alternatives in issue 2.2c above.

Issue 2.3c: How should economists consider partial or contributing impacts?

When conducting disparate impact analyses, economists attempt to isolate the impact of the facially-neutral policy on the outcome of interest, all-else-equal. There are situations, however, where multiple policy factors partially contribute to disparate impact. Therefore, when estimating the disparate impact that a given facially-neutral policy generates, economists need to decide whether to include all applications where the policy was *one* of the drivers of the adverse outcome or only applications where the policy was the *only* driver of the adverse outcome. A common example of this scenario occurs for underwriting decisions. Lenders often consider multiple pieces of information when underwriting applications for credit. In these instances, there often are multiple factors that drive a decision to deny an application for credit. For example, the lender might deny an application because the score value did not meet the policy threshold, DTI did not meet the policy threshold, the applicant had a bankruptcy five years ago, and there was a fraud alert or risk. When isolating the disparate impact that the scoring model created on denial rate disparities, economists would typically exclude applications denied for several reasons, under the assumption that the lender would have denied these applications regardless of whether the scoring value met the policy threshold.[10] In other words, economists would typically only include applications where the scoring model was the sole driver of the denial. This is a conservative approach to estimating the disparate impact that a policy creates because it excludes all the policy's partial contributions. Economists need legal guidance on whether this analytical approach is appropriate.

[10] This approach has important implications for fair lending risk management, since changes in policies over time (for example no longer considering a given factor when making credit decisions), might alter the set of applications used to generate estimates of disparate impact. As a result, estimates of disparate impact may differ pre- and post-policy change.

Issue 2.3d: Should economists consider the impact of other policies or components of policies?

As noted above, lenders typically consider multiple pieces of information when making credit decisions. When conducting disparate impact analyses, economists typically try to isolate the impact of just the facially-neutral policy. Once isolated, economists need to decide whether to end the analysis and focus on this estimate of disparate impact, or to continue the analysis and consider whether other components of the decision-making process might offset this disparate impact and result in no overall, bottom-line disparate impact.

The DOJ Document summarizes several legal cases related to this issue. We discuss one of the cases here to show how, for this issue, a court's ruling relates to how economists conduct statistical analyses of disparate impact. Greater New Orleans Fair Housing Action Center v. HUD, 639 F.3d 1078 (D.C. Cir. 2011) focused on whether one specific part of HUD's formula for awarding hurricane relief grants created a disparate impact. In that case, "The court held that while that one part of the formula, viewed in isolation from the rest, may have had an adverse impact on African Americans, other parts of the formula may have disproportionately benefitted African Americans."[11] One could argue that the court's ruling is both consistent with, and not consistent with, economists' all-else-equal analytical approach discussed in issue 2.3c above. If the facially-neutral policy is defined as one specific part of the HUD formula, economists would assess whether that specific part of the policy created a disparate impact, all-else equal. Given that the court argued that other parts of HUD's policy, as well as the total impact of the entire policy, should also be considered, the court essentially disagreed with the all-else-equal approach that economists use. This ruling suggests that economists should consider the impacts of other policies on overall outcomes. On the other hand, economists would likely view this as a

[11] See page 12 of the DOJ Document.

definitional issue. Specifically, based on the court's ruling, economists would argue that the facially-neutral policy should have been defined as the entire HUD formula. With this facially-neutral policy, economists would estimate the disparate impact this policy created all-else-equal, which would appear to be consistent with this court's ruling.

Economists need legal guidance on three questions related to this issue. First, for a given facially-neutral policy, should economists isolate the disparate impact that policy creates all-else-equal, or also consider how other policies might enhance or offset the specific disparate impact that the policy creates? Second, if a facially-neutral policy creates a disparate impact (based on isolating the impact of just the policy), is it possible to conclude this disparate impact is illegal if there is no overall disparity in final outcomes? Third, building off of the legal case above, for a given disparate impact analysis, are there alternative definitions of the facially-neutral policy that would be legally acceptable and align with economists' approach of isolating the impact that the policy creates, all-else-equal?

Issue 2.3e: What impact should economists focus on when a lender uses a sequential decision-making process?

The final issue in this section focuses on one specific and common use case that is closely related to issues 2.2a and 2.3a above. Specifically, how should economists measure the impact of a facially-neutral policy when a lender uses a sequential decision-making process? In a typical sequential decision-making process, lenders consider a credit score only if the application first meets other eligibility criteria and knockout rules. Lenders typically use this type of decision-making process to avoid the costs of pulling a credit bureau report for applications they are likely to deny.

When analyzing the impact of a scoring model in a sequential decision-making process, one approach is to include only the applications for which the lender considered the credit scoring model. This approach is consistent with the phrase, “all persons the policy or practice affects" from issue 2.2a above. One important consideration with this approach is the number of applications that the lender scored. The statistical analysis could show that the scoring model created a large disparate impact, but if the lender used the scoring model for only a small number of applications, the overall importance of this finding could be relatively small. One important risk with this approach is that it could miss disparate impact that some factors in the scoring model generated. In many instances, the factors lenders use as eligibility criteria and knockout rules are the same as, or highly correlated to, the factors in the scoring model. Focusing on just the set of applications for which the lender formally considered the scoring model would therefore miss any disparate impact that factors in the scoring model that the lender also used for eligibility and knockout rules generated. In these instances, if there is interest in assessing the disparate impact that the factors in the scoring model generated more broadly, it would be important to also assess the disparate impact that the eligibility criteria and knockout rules generated using all applications.

Alternatively, when analyzing the impact of a scoring model in a sequential decision-making process, economists could include all applications in the analysis. Since lenders pull credit bureau reports only for applications that meet eligibility criteria and knockout rules, credit scores are typically not available for all applications, assuming some factors in the scoring model come from the credit bureau report. Therefore, it is typically not possible to analyze the direct impact of the scoring model on outcomes for all applications. As a result, this approach focuses on estimating the disparate impact that the scoring model generated relative to the disparate

impact generated by all other factors that the lender considered conditional on the overall decision-making process the lender used. The first step of this type of analysis is to estimate the overall disparate impact for a given demographic group, which is simply the unconditional disparity in the outcome of interest using all applications. The next step is to decompose this disparity into the portion that the credit scoring model determines and the portions that each of the other factors the lender considered determines. Using an underwriting analysis as an example, suppose the unconditional denial rate disparity for two groups is 10 percentage points (pps). Identifying all applications the lender denied solely due to the credit score, re-coding these applications as approvals, re-calculating the denial rate disparity, and then taking the difference of this modified disparity and the original disparity (10 pps) provides an estimate of the portion of the disparate impact that the scoring model determines. Using the same approach provides similar estimates for each of the other factors the lender considered when making underwriting decisions. Since lenders deny some applications for multiple reasons, we also need to calculate the portion of the disparate impact that combinations of factors determines. The sum of all of these estimated portions will equal the denial rate disparity, which is what we meant by this approach decomposes the disparity. Very importantly, this approach reflects the overall decision-making process the lender used (sequential in this example) and how it weighed different factors. In general, if the lender formally considered the credit scoring model for only a small percentage of applications (i.e., the eligibility criteria and knockout rules drove most of the adverse outcomes), most likely the scoring model's contribution to overall disparate impact would tend to be smaller. Alternatively, if the lender formally considered the credit scoring model for a larger percentage of applications (i.e., most applications met the eligibility criteria and knockout

rules), then the potential is higher that the scoring model contributed more to the overall disparate impact.

For this issue economists need legal guidance on whether to focus on just the applications for which the lender applied the facially-neutral policy or all applications, as well as whether to separately analyze potential disparate impact that specific factors in a scoring model that the lender also uses as eligibility criteria or knock-out rules generated.

Issue 2.4: What protected classes should economists focus on for the disparate impact analysis?

The Equal Credit Opportunity Act (ECOA) includes several protected classes, including race, color, religion, national origin, sex, marital status, and age. For each of these broad protected classes, economists typically generate granular subsets to use for analyses. For example, economists typically separate "race" into African Americans, Asians, Whites, and others. In addition, economists also often combine categories for analysis, such as combining all racial minorities into one group, or combining gender and race. Given the large number of possible definitions of protected class, a given facially-neutral policy will almost always generate disparate impacts that disadvantage some protected classes and simultaneously advantage other protected classes. This issue is particularly relevant in a fair lending review context, since a violation of law would likely require the lender to change or eliminate its facially-neutral policy, which would likely benefit some protected classes and harm other protected classes.

Economists need legal guidance on which protected classes to focus on, and how to interpret statistical results showing that a facially-neutral policy generates disparate impact simultaneously benefiting some groups and harming other groups.

Issue 2.5: What disparity measure should economists use?

Legal materials on disparate impact typically include only general references to "disparities," but do not provide details about how to specifically measure disparities. For example, the HUD Document references "disparity" or "statistical disparity" 39 times, but does not provide any details about how to specifically measure a disparity. The DOJ Document provides more detail, including a separate section called "Establishing Disparity," acknowledging that multiple specific disparity measures exist, and discussing the 4/5th rule from employment cases, which states that a selection rate for a protected group should be at least 80 percent of the rate for the group with the highest selection rate. However, it provides no further details, but instead uses general language such as, "To establish a disparity, an investigating agency must use an 'appropriate measure.'" and "There is no one-size-fits-all measure for disparity."[12]

Economists have several specific measures of "disparity" that they can use for disparate impact analyses. The three most common measures are the ratio (which is what the 4/5th rule is based on), difference, and odds ratio. Much of economic analysis has foundations in calculus, so economists tend to focus on the difference (marginal effects) measure. However, each measure can be a reasonable approach for disparate impact analyses depending on the legal, policy, and economic characteristics and objectives of the analysis.[13]

The choice of disparity measure is important because different measures can lead to different interpretations and conclusions. As a simple example, suppose we are analyzing the disparate impact of a facially-neutral policy, and denial rates are the outcome measure of interest. Suppose further that the approval rates for two groups are 1 percent and 2 percent. The ratio is

[12] See pages 18 and 19 of the DOJ Document.
[13] See chapter 4, "Disparity Measures for Fair Lending Analyses," for more details on disparity measures.

0.5 (= 1/2), which violates the 4/5th rule. However, the difference is only 1 percentage point, which some might view as small. From a purely analytical perspective, it is not clear which of these disparity measures to use or what conclusions to draw, so we need to consider additional information, such as legal, policy, and economic implications. As a second example, suppose the approval rates for the two groups are 65 percent and 80 percent. In this example, the ratio is 0.8125 (= 65/80), which satisfies the 4/5th rule. However, the difference is 15 percentage points, which some might consider large. Again, it is not clear which of these disparity measures to use or what conclusions to draw, so we need to consider additional information.

An additional scenario that creates significant analytical challenges is generating disparity estimates when group membership is unknown and the economist needs to use proxies. There are multiple reasonable approaches that economists can use in this scenario, and different approaches can lead to different results, interpretations, and conclusions. The appropriate approach for a given analysis is an open question for economists.[14]

Overall, economists need legal guidance on whether there is one specific disparity measure they should use for disparate impact analyses, both when race data are available and not, as well as if there are any disparity measures they should not use. In addition to this guidance, for any given disparate impact analysis, it is critical that all stakeholders know which specific disparity measure the economist used and understand how to appropriately interpret the disparity results for that measure, so that they can draw accurate conclusions.

[14] This topic is too extensive to address here, so we refer to chapter 8, "Marginal Effects Estimates for Continuous Race Proxies in Limited Dependent Variable Models" for details.

Issue 2.6: Does it matter which specific individuals a facially-neutral policy harms or helps?

Any given facially-neutral policy will benefit some specific individuals and harm other specific individuals. The question here is whether the disparate impact analysis should include a search for specific individuals, or subsets of individuals, who a policy harmed even when the statistical analysis shows that the policy created no disparate impact at the group level for the entire population. The case of Betsey v. Turtle Creek Association, 736 F.2d 983, 987 (4th Cir. 1984) provides the impetus for raising this issue. In a summary of that case, the DOJ Document notes that, "Bottom-line' considerations of the number and percentage of minorities in the rest of the complex or community are 'of little comfort' to those minority families evicted from Building Three."[15] This statement suggests that, even if there is no evidence of overall disparate impact for a specified population, it may still be of interest whether there is harm to specific subsets of the population.

Economists need legal guidance on whether they should include these types of drill-down analyses into their overall statistical analyses of disparate impact.

Issue 2.7: What is statistical evidence of causality for disparate impact analyses?

The DOJ document is very clear on the importance of causality for disparate impact analyses stating, "To establish a violation of its disparate impact provision, an investigating agency must determine that the impact is causally linked to a recipient's policy or practice."[16] However, the document is not clear on exactly what statistical evidence would meet the legal definition of a causal link. To add to this uncertainty, the economic definition of causality may not completely align with the legal definition of causality. Economists define causality based on

[15] See page 13 of the DOJ Document.
[16] See page 28 of the DOJ Document.

statistical criteria and econometric theory, and the bar for causality is generally quite high and difficult to meet.[17] The legal definition relies more on case law and precedent, which yields more subjective criteria that is open to interpretation. The discussion of causality in the HUD Document provides an example of these differences. That document includes phrases such as, "robust causality," "practice caused or *predictably will cause* a discriminatory effect," and "link a specific practice to a current or *predictable disparity*." Many economists would view these statements as vague and therefore need additional legal guidance to determine whether a given set of statistical results meets these more subjective legal criteria. Although these definitional differences add an additional level of uncertainty to statistical analyses, they also provide economists with some flexibility to consider whether statistical results provide legal evidence of causality even though the results may not meet the economic bar for evidence of causality.

For automated decision-making processes based on a small number of factors, it is fairly easy for economists to establish both an economic and legal causal link between a facially-neutral policy and an outcome of interest. For discretionary decision-making processes where a lender considers a larger number of factors in complex ways, identifying economic causal links is much more challenging. For these disparate impact analyses, economists typically use some form of multivariate analysis (often regression analysis) to isolate the impact of the facially-neutral policy on the outcome of interest by first controlling for the impacts of all other factors the lender considered when making credit decisions. For highly complex decision-making processes, the data to control for all other factors are typically unavailable, so it is difficult to identify economic causal relationships using regression techniques. As noted above, however,

[17] One primary example of these criteria is a concept called, identification, which refers to the process of ensuring that a causal effect can be measured from observed data. Economists utilize a variety of specific econometric techniques such as randomized controlled trials, instrumental variables, regression discontinuity designs, and difference-in-difference estimators to achieve identification so that the results reflect causal relationships.

depending on what data are available and the overall strength of the analysis, the statistical results may still meet the legal criteria for causality.

Overall, economists need legal guidance on what specific statistical evidence meets the legal definition of causality, especially for more complex regression analyses where the statistical evidence may not be strong enough to meet the economic criteria for causality but still might meet the legal criteria.

Issue 2.8: When assessing causality for composite variables, should economists focus on the composite variable or on the underlying factors that generate the composite variable?

Causality involves identifying the specific factors responsible for an observed disparity in outcomes. For a composite variable such as a credit score, which is comprised of many individual factors, economists could focus on whether the scoring model as a whole is responsible for any part of an observed disparity or on whether the underlying factors in the scoring model are responsible for any part of an observed disparity. From a purely analytical perspective, a credit score is just like any other policy factor a lender considers when making credit decisions, so it is reasonable to focus on the disparate impact that the scoring model as a whole creates. This is the approach we have taken throughout this report. On the other hand, focusing the disparate impact analysis on the individual factors in the scoring model is consistent with aspects of Reg B. Specifically, Appendix C to Part 1002 of Regulation B includes a sample Adverse Action Notice form (Form C-3) specifically for adverse outcomes based on a credit score, which requires lenders to identify and share the specific factors in the scoring model that led to the adverse outcome. Economists need legal guidance on whether to focus on a scoring model as a whole, the underlying factors in the scoring model, or both when conducting a statistical analysis of disparate impact.

This issue creates two additional questions as well that also require legal guidance. First, scoring models are not the only composite variable that lenders consider when making credit decisions. Debt-to-income (DTI) for example, is comprised of a wide variety of individual debt obligations and sources of income. Unlike for scoring models, however, Form C-1 in Appendix C to Part 1002 of Regulation B explicitly includes the aggregate DTI measure (specifically, "Excessive obligations in relation to income") as an acceptable reason for adverse outcomes. This raises the question of whether or not economists should treat scoring models the same as other composite variables, such as DTI, when conducting disparate impact analyses. Second, focusing on the underlying factors of a scoring model introduces analytical challenges, especially for the relatively new scoring models developed using machine learning (ML) classifiers. ML scoring models tend to have very large numbers of factors and include interactions of factors. As a result, it can be difficult to isolate and estimate the impact of each individual factor in an ML scoring model. This issue is part of the on-going debate about explainability for these models. In addition, given the large volume of factors in these models, each individual factor typically has only a small impact on the overall scoring model, which could make it difficult to identify meaningful amounts of disparate impact. One alternative here is to analyze groups of factors for disparate impact, but the dimensionality of this analysis would be challenging.

Given all of these challenges, economists need legal guidance on whether it is acceptable to treat composite variables differently, as well as how to conduct disparate impact analyses of ML scoring models overall given their unique characteristics.

III. Business Justification

In the second step of the three-prong, burden-shifting framework for disparate impact, defendants must show that the challenged policy or practice is necessary to achieve a substantial, legitimate, non-discriminatory interest. We discuss two issues related to this step where economists need legal guidance.

Issue 3.1: What is acceptable evidence of business justification?

There are several arguments lenders could make as business justification to support using a given policy. These include economic arguments focused on profitability, revenue, loss, market share, and others; statistical arguments focused on predictiveness and accuracy; legal arguments, such as compliance with various regulations and laws; and process/policy arguments, such as ease of use or understanding, and processing speed. In most instances, lenders will likely incorporate some components of each of these arguments to make the business justification for using a given policy.

For this issue economists need legal guidance on questions such as, are there any specific arguments that are not acceptable, are certain types of arguments given more weight than others, and is one strong argument acceptable or must the lender provide multiple arguments?

Issue 3.2: What if the business justification changes over time?

Lenders typically create business justifications for a new policy in the initial stages of policy development. For a credit scoring model, which modelers often start developing several years prior to implementation, this creates two challenges, business interests may change over time and predictions at time of model development about how the scoring model will meet business interests may turn out to be inaccurate. During a disparate impact analysis post-

implementation, the lender should address each of these issues, as well as demonstrate how the scoring model meets current business interests.

IV. Less Discriminatory Alternative (LDA)

Under the three-prong, burden-shifting framework for disparate impact, even if the defendant successfully justifies the facially-neutral policy in step 2, the plaintiff can still prevail by demonstrating that the defendant could have served its legitimate business interests with an alternative policy or practice that would have had a less discriminatory effect. Operationalizing this step of the disparate impact framework raises several analytical issues. These issues are particularly interesting and challenging for a credit scoring model, which is the facially-neutral policy we focus on in this report, because lenders need significant amounts of time and resources to develop, test, implement, and re-develop scoring models. There are two LDA searches relevant to scoring models, one at time of model development and one during post-implementation disparate impact testing. We discuss the analytical issues with LDA searches separately for each of these time periods.[18]

Issue 4.1: At time of model development, what are the legal requirements/expectations for LDA searches?

We begin with a discussion of LDA searches at time of model development. There are several analytical issues here that we discuss in turn.

[18] www.paceanalyticsllc.com contains several reports exploring the statistical details of LDA searches.

Issue 4.1a: Are model developers legally required to conduct LDA searches during model development?

Economists often review the LDA searches that model developers conduct when developing scoring models. To help inform these reviews, it is important to understand whether model developers are legally required to conduct these searches and what these searches are legally required to include. Unfortunately, we were unable to identify any statute requiring these searches or detailing the specific requirements for these searches. We were able to locate guidance the CFPB has released on the importance of including LDA searches into model development as part of effective overall fair lending compliance testing and model risk management.[19] However, the CFPB recently removed the Supervisory Highlights reports that officially conveyed these positions, so current expectations are unclear. Overall, economists need guidance on the legal requirements for model developers conducting LDA searches.

Issue 4.1b: What is a sufficient LDA search?

When a model developer builds a credit scoring model, an LDA search involves exploring whether alternative scoring models exist that may have less of a discriminatory effect yet still meet the legitimate business interests of the lender. The potential dimensions of this search create significant challenges. When building a scoring model, model developers can explore a variety of different data sources, measures of bad performance, variable constructions, combinations of variables, and estimation techniques. All of these options create a large volume of possible alternative scoring models. In addition, as noted throughout this report, lenders use scoring models for a variety of purposes, such as underwriting and pricing for example, and

[19] See CFPB Puts Lenders & FinTechs On Notice: Their Models Must Search For Less Discriminatory Alternatives Or Face Fair Lending Non-Compliance Risk » NCRC and CFPB Highlights Fair Lending Risks in Advanced Credit Scoring Models | Consumer Financial Services Law Monitor.

disparate impact can affect a variety of demographic groups. This creates a large volume of estimates of potential disparate impact that each scoring model might generate. Overall, it would be difficult to explore every possible alternative scoring model and its potential discriminatory effect given the volume of alternatives, as well as time and resource costs.

If a search of all possible alternative models is not feasible, then economists and model developers need legal guidance on what constitutes a sufficient search. Since the criteria for a sufficient search is likely subjective, economists and lawyers need to work closely during these reviews to determine whether the model developer's LDA search is sufficient.

Issue 4.1c: What is an appropriate strategy to estimate potential disparate impact?

Since model development typically occurs several years before economists can test whether the scoring model generated a disparate impact on actual applicants, the best model developers can do at time of development is provide estimates of the likelihood that the scoring model will create disparate impacts when the lender applies it to actual applicants. Generating these estimate can be very challenging, since the model developer likely will not know exactly which products, programs, or applicants the lender will apply the scoring model to; what decision-making processes the lender will incorporate it into; exactly how the lender will implement it (policy score thresholds for example); or the characteristics of applicants that apply. All of these challenges affect the accuracy of any estimates of potential disparate impact.

When conducting LDA searches during model development, there are three common approaches model developers use to estimate the likelihood that a credit scoring model will create a disparate impact when the lender implements it. First, modelers can analyze differences in average scores and score distributions across demographic groups using the development data and hold-out sample. Similar averages and score distributions across groups suggest the

likelihood is lower that the scoring model will create a disparate impact when implemented. Given that modelers build credit scoring models on the premise that past performance is a good predictor of future performance, and that minorities often have worse performance on past credit, score distributions for minorities are often shifted in a lower creditworthiness direction. In this scenario, to obtain similar score distributions across groups, modelers would need to build scoring models that systematically over-predict risk for non-minorities and under-predict risk for minorities. Economists need guidance on whether this would be acceptable from a legal perspective.

Second, model developers can focus on prediction errors across groups using the development data and hold-out sample. With this approach, if the scoring model does not systematically over- or under-predict bad performance for any demographic group, this suggests a higher likelihood that when the lender implements the scoring model the distributions of credit decision outcomes will accurately reflect the actual performance risks for each demographic group as well. This approach is more focused on the business justification prong of the disparate impact framework, since one desirable statistical characteristic of scoring models is that they do not make systematic prediction errors. There are two significant concerns with this approach. First, since performance on past credit typically differs across demographic groups, score distributions will likely differ across groups, which means access to credit and favorable terms will likely differ across groups as well. Therefore, although this approach might be justified from a statistical perspective, it may be undesirable from a policy perspective. Second, this approach takes the development data model developers use to build scoring models as exogenous. If the development data show differences in performance on past credit for some demographic groups due to past discrimination, this approach implicitly accepts this past discrimination and builds it

into the scoring model. Again, it is not clear whether this would be acceptable from a policy perspective.

Finally, model developers can simulate how the scoring model might impact credit decision outcomes. At time of model development, there is obviously no data available on how the scoring model impacted outcomes for actual applicants. This is why the first two approaches focus on distributions of scores and prediction errors using only data available at time of model development. However, using the development data and hold-out sample, model developers can attempt to simulate how a lender's use of the scoring model might impact credit decision outcomes. As one example, developers can essentially re-underwrite the loans in the development data and hold-out sample by applying thresholds to the score values to classify applications as approved or denied and then analyze disparities in these outcomes by demographic group. This approach is closer to the type of disparate impact analysis that economists conduct post-implementation, which might yield estimates of potential disparate impact that are more accurate than compared to the first two approaches. However, there are two significant limitations that affect the accuracy of these estimates. First, the approximate policies the model developer applies during this analysis will likely differ significantly from the policies the lender actually applies post-implementation. This includes the programs/products which the lender uses the scoring model for, the thresholds the lender applies, and the other variables the lender considers with the scoring model when making credit decisions. Second, this analysis typically only focuses on how a scoring model might impact approve/deny decisions, but lenders often use scoring models for many purposes, including marketing, pricing, and servicing. Focusing this analysis only on potential disparate impact in underwriting decisions would miss these additional risks.

Overall, economists need legal guidance on which of these three strategies, or other strategies, is most appropriate for model developers to use to generate estimates of the likelihood that the scoring model will create disparate impacts when the lender applies it to actual applicants.

Issue 4.1d: What are the implications of allowing model developers to directly consider protected class information during model development?

Historically, regulatory agencies have not allowed model developers to include protected class information directly into the model development process. As a result, model developers have used an iterative approach to LDA searches during model development. As a first step of this iterative approach, developers build the most predictive scoring model possible. Next, a separate compliance or legal group conducts disparate impact testing and makes suggestions to the model developers on how to reduce potential disparate impact. These suggestions often entail dropping or modifying specific variables. Model developers then incorporate these suggested changes, build a modified model, and provide the new model to the compliance or legal group for additional review. This iterative process continues until the compliance or legal group is comfortable that the latest model meets all business interests and that no alternative model would likely create less disparate impact.

Recently, the CFPB relaxed these restrictions and advocated that model developers incorporate protected class information directly into the model estimation process as part of LDA searches.[20] One common estimation approach simultaneously maximizes model predictiveness and minimizes potential disparate impact. This automated statistical approach is much more

[20] See for example, CFPB Highlights Fair Lending Risks in Advanced Credit Scoring Models | Consumer Financial Services Law Monitor. The CFPB recently removed guidance on this topic, so their current position on this issue is unclear.

efficient than the manual iterative approach that lenders have historically used. However, it raises several important issues. We discuss one question, several analytical challenges, and one risk here.

Starting with the question, are model developers legally allowed to directly consider protected class information when developing scoring models? Although historical guidance from prudential regulators would suggest no, the recent guidance from the CFPB would suggest yes. Model developers and economists need clear and formal guidance on this question, as well as on the parameters of acceptable usage of this information if model developers can use it. In addition, legal guidance is also needed on whether lenders can consider protected class information to develop policies more broadly. DTI is a good example since, similar to scoring models, it is comprised of many underlying components. One implication of allowing model developers to directly consider protected class information when developing scoring models is that it might alter the factors included in the scoring model. Similarly, allowing lenders to consider protected class information when exploring DTI measures that are more predictive of bad performance might alter the specific incomes and debts lenders use to generate the DTI measure. Using different underlying components to construct DTI already occurs with the front-end and back-end measures of DTI, so using protected class information for this purpose would just be an extension of current practices.

Next, we discuss several analytical challenges related to incorporating protected class information into model development. We focus this discussion on the simultaneous estimation approach discussed above. The first analytical challenge relates to protected class groups. As noted in issue 2.4 above, there are several reasonable definitions of protected classes. However, the simultaneous estimation approach can typically only minimize potential disparate impact for

one or two of these groups. Therefore, even if this estimation approach generated accurate estimates of the likelihood that the scoring model will create disparate impact for these groups, the likelihood of the scoring model creating disparate impact on every other group is unclear. This is a significant risk. On this specific issue, incorporating components of the iterative approach that lenders have historically used might be effective, since that approach is more flexible with regards to assessing potential disparate impact for multiple groups. Overall, model developers need legal guidance on whether the simultaneous estimation approach limited to one or two protected class groups is a sufficient LDA search.

The second analytical challenge relates to specific credit decisions. Model developers using the simultaneous estimation approach to build scoring models typically focus on predictiveness and disparate impact related to approve/deny decisions. Given that lenders use scoring models for a variety of credit decisions, this approach creates only a limited picture of the overall disparate impact that the lender's use of the scoring model might create. Even if this approach generated accurate estimates of the likelihood that the scoring model will create disparate impact for underwriting decisions, the likelihood of it creating disparate impact for other credit decisions is unclear. Again, this is a significant risk. Model developers need legal guidance on whether this would be a sufficient LDA.

Third, what predictiveness / disparate impact tradeoff is acceptable? During LDA searches, model developers often generate scatterplots showing the tradeoffs between model predictiveness and estimates of potential disparate impact across alternative scoring models. Lenders and developers then decide which specific tradeoffs are acceptable, and which specific model to finalize. The underlying question here is, how much of a reduction in predictiveness is acceptable for a given reduction in disparate impact? This question creates significant

uncertainty. Should regulatory agencies decide what is an acceptable tradeoff? Is one tradeoff acceptable in every situation, or will the acceptable tradeoff be case-specific? If case-specific, what criteria determine acceptable tradeoffs? How will these tradeoffs impact the safety and soundness of the scoring model? Would it be acceptable if a lender argued that no tradeoff is acceptable due to safety and soundness or business interest reasons? Overall, model developers, lenders, and economists need legal guidance on these, and other, questions related to acceptable tradeoffs.

Fourth, statistical approaches such as the simultaneous estimation approach often suffer from dimensionality challenges. The simultaneous estimation approach focuses on the tradeoffs between model predictiveness and potential disparate impact. There are many specific statistical measures of predictiveness that the model developer could use, such as the Gini coefficient, AUC/ROC curve, KS-statistic, and many others. Importantly, different measures can lead to different conclusions, so it is important to know which measure the developer used, the strengths and weaknesses of the measure, and how to properly interpret the measure. Evidence of whether the results are robust to the use of alternative measures is valuable information as well. Very similarly, as discussed in issue 4.1c above, there are several statistical approaches that model developers can use to estimate potential disparate impact. Again, different approaches can lead to different conclusions, so it is important to understand the details of the specific measure model developers used. Overall, economists need legal guidance on whether there are specific measures of model predictiveness and potential disparate impact that are acceptable, as well as how to interpret LDA searches when the estimated tradeoffs differ across measures.

Finally, it is unclear how the simultaneous estimation approach will affect estimates of potential disparate impact overall. As noted in the discussion of issue 4.1c above, the LDA

search at time of model development only provides estimates of the likelihood that the scoring model will create disparate impact once the lender implements the scoring model, which could be several years after model development. Regardless of any benefits or efficiency gains of utilizing a simultaneous estimation approach, it will still only provide an estimate of the likelihood of future disparate impact and does not guarantee that the scoring model will create no disparate impact. Given all of the difficult analytical choices model developers must make when applying this estimation approach, it is unclear how this approach will affect the accuracy of the estimates of potential disparate impact.

The last issue we discuss in this section is the risk that bad actors will misuse protected class information when building scoring models. One of the main reasons why prudential regulators have historically created a firewall between model developers and demographic data is a fear that model developers would use these data inappropriately. Much of this concern stems from an incentives problem. Model developers' primary incentive is to build predictive scoring models, since these models directly impact the lender's bottom line. Minimizing disparate impact is typically not a concern, or is only a secondary concern. Given that performance on past credit almost always varies by demographic group, group membership will be predictive of bad loan performance. As a result, from a purely analytical perspective, model developers have incentive to include demographic variables, or variables that are highly correlated with demographic variables, into scoring models. This creates incentives for model developers to use protected class data inappropriately. The variety of analytical options developers have when developing scoring models, especially regarding the specific measures of disparate impact, protected classes, and credit decisions, creates opportunities to use protected class information to maximize predictiveness, and at the same time influence to some extent the estimates of potential disparate

impact. We expect that most model developers would likely use these data appropriately, especially if there are reasonable guardrails, monitoring, and controls in place. However, the expectation should be that some model developers will use these data inappropriately, especially since their primary incentive is to build predictive models and not to minimize disparate impact risk.

Given that incorporating protected class information directly into the model development process is a departure from past guidance regulators have provided to industry, and creates several questions, analytical challenges, and risks, model developers, lenders, and economists need clear legal guidance from regulators on all of these issues.

Issue 4.1e: Can a sufficient LDA search create a safe harbor for lenders?

The last issue we discuss in this section is safe harbors. Conducting an LDA search at time of model development can be an effective strategy for mitigating disparate impact risk as part of overall fair lending compliance and risk management programs, but can it also provide lenders with a safe harbor during disparate impact analyses of scoring models post-implementation? As discussed in issue 4.1c, LDA searches at time of model development only provide an estimate of the likelihood that a scoring model will create a disparate impact when the lender implements it. This means that a scoring model can create disparate impact post-implementation regardless of the quality of the LDA search at time of model development. As discussed in issue 4.1b, the large volume of alternative scoring models makes it impossible for model developers to explore every alternative model during LDA searches. Combining these two results, even if the quality of the model developer's LDA search at time of model development is high, economists conducting disparate impact analyses of scoring models post-implementation might still identify disparate impact in prong 1, as well as an LDA scoring model during prong 3

from among the models the developer did not explore. This suggests that the model developer can conduct a sufficient LDA search (however defined) conditional on available data, time, and resources, but the post-implementation disparate impact analysis can still potentially generate evidence meeting the legal criteria for disparate impact. When a model developer puts forth a good-faith effort and conducts a sufficient LDA search, can this provide lenders a safe harbor? If the answer is no, then what additional steps, if any, can model developers take to ensure that the scoring model does not meet the legal threshold for disparate impact post-implementation? Legal guidance is needed on these issues.

Issue 4.2: During disparate impact analyses of a facially-neutral policy post-implementation, what are the legal requirements/expectations for LDA searches?

We now turn to a discussion of LDA searches during prong 3 of disparate impact analyses post-implementation. There are several analytical issues here that we discuss in turn.

Issues 4.2a: Do economists need to show that an alternative policy would literally have created less disparate impact on actual applicants?

In issue 2.2a above, we discussed whether the relevant population for a disparate impact analysis should consist of actual applicants that a facially-neutral policy impacted. If so, then in the first step of the three-prong disparate impact analysis economists would assess whether a lender's use of a facially-neutral policy generated a disparate impact on actual applicants. If evidence of disparate impact exists, one possible definition of an LDA is an alternative policy that would have with certainty generated a smaller disparate impact on the same set of actual applicants had the lender used that policy instead. This strict interpretation of an LDA is consistent with the discussion of prong 3 of the disparate impact framework in the HUD Document, which states, "if the defendant or respondent meets its burden at step two, the

plaintiff or charging party may still prevail by proving that the substantial, legitimate, nondiscriminatory interests supporting the challenged practice could be served by another practice that has a less discriminatory effect."[21] Importantly, this statement includes the language, “that has” and does not include any language similar to, “or predictably will cause.”

For some definitions of the facially-neutral policy, such as policy score thresholds, this strict definition of an LDA is relatively easy to apply. We would simply vary the policy score thresholds, determine what the outcome for each application would have been under these new thresholds all-else-equal, generate new estimates of disparate impact, and compare these estimates to the original estimates. However, for other definitions of the facially-neutral policy, such as the scoring model as a whole, this LDA definition can be much more difficult, if not impossible, to apply. For this policy, we would need to show that if the lender had used an alternative scoring model to make credit decisions on actual applicants, the disparate impact would have been lower than what it was based on the actual scoring model the lender used. This is a challenging bar to meet, because we would need to develop an alternative scoring model using the same development dataset as for the original scoring model (which is costly and takes time), gather all of the data for the actual applicants necessary to score those applicants using the new scoring model, and re-decision each actual applicant using this new scoring model together with all other factors the lender considered, including possible discretion. Except for completely automated decision-making processes that consider only a small number of factors, this strict definition of an LDA would be very challenging to meet. If the lender allowed exceptions or some amount of discretion during the decision-making process, this LDA definition would be impossible to meet.

[21] See page 19451 of the HUD Document.

On this issue, economists need legal guidance on whether adhering to the strict LDA definition is necessary to determine if an alternative policy qualifies as an acceptable LDA.

Issue 4.2b: What are the legal criteria for evidence that a facially-neutral policy is a less discriminatory alternative when conducting a disparate impact analysis post-implementation?

Given the analytical challenges of applying the strict definition of an LDA discussed in issue 4.2a, are there alternative definitions of an LDA that are more analytically feasible and still legally acceptable? As noted above, the discussion of the first prong of the disparate impact framework in the HUD Document includes the phrase, "**or predictably will cause** a discriminatory effect." Although the HUD Document includes this language only in the discussion of prong 1 and not prong 3, it still suggests that one possible alternative definition of an LDA might be a policy that predictably will reduce the disparate impact that the original policy created, conditional on continuing to satisfy the lender's business interests. This less stringent definition of an LDA creates opportunities for a wide variety of statistical approaches and analyses that could provide evidence of an LDA.[22] As one very simple example, if an alternative policy is less correlated with group membership than the facially-neutral policy of interest, this is one signal that the alternative policy might predictably reduce disparate impact had the lender used it. Could this statistical evidence be sufficient to conclude the alternative policy is an LDA, again assuming the policy also satisfied the lender's business interests? Since these statistical approaches come with more uncertainty and are more subjective, economists and lawyers would need to work closely together to assess whether a given statistical analysis and set

[22] Unlike for the discussion of issue 4.2a, this LDA definition would be applicable to populations defined as actual applicants and populations defined as individuals who could be impacted by the facially-neutral policy.

of statistical results are sufficient to conclude that an alternative policy would predictably reduce the disparate impact that the original policy created.

Issue 4.2c: Do economists need to use the same development data and estimation approach that the model developer used to build the scoring model?

This issue, which is closely related to issue 4.1e above, focuses on what information, data, and estimation approaches economists can use to search for LDAs. Since post-implementation disparate impact analyses typically occur several years after model development, economists have significantly more data and information available for LDA searches compared to model developers at time of model development. Can economists use this information when conducting LDA searches? As one example, economists will have data on the actual applicants to which the lender applied the scoring model. This is valuable information when searching for a scoring model that would have created less disparate impact. Can economists use these data as part of the LDA search, or can they only use the development data that the model developer had available at time of model development? As a second example, researchers are always developing new estimation approaches over time. If the model developer was not aware of, or did not have expertise in, a newly developed machine learning classifier, can economists explore this estimation strategy as part of the LDA search? Economists need legal guidance on these questions.

Issue 4.2d: Do economists need to show that an LDA scoring model meets the lender's same business interests?

Discussions of LDA searches often focus on the "less discriminatory" portion of the definition with less emphasis given to the requirement that the alternative policy meets the business interests of the lender. When business interests for scoring models are discussed, the

focus is often on just model predictiveness. As noted in the section on business justifications above, however, a lender's business interests are often much more expansive than just model predictiveness, and can include a variety of credit risk, model risk, and operational risk management components. Per prong 2 of the three-prong, burden-shifting disparate impact framework, it is the lender's responsibility to specifically define these business interests. In many instances, it is difficult for lenders to accurately quantify all of these interests. As a result of all of these issues, when exploring an alternative policy as part of an LDA search, it can be challenging for economists to evaluate the policy against the lender's business interests, since the lender developed these interests internally, they might be expansive, and some may be difficult to quantify.

Given these difficulties, economists need legal guidance on what constitutes sufficient evidence that an alternative policy meets the lender's business interests. In addition, since these analyses will be somewhat subjective, economists and lawyers need to work closely during these analyses to determine if an alternative policy meets this requirement.

Issue 4.2e: What is an acceptable tradeoff between predictiveness and disparate impact?

The final issue we discuss is the tradeoff between model predictiveness and disparate impact. Similar to how model developers analyze this tradeoff during LDA searches at time of model development, economists conduct a similar analysis during disparate impact analyses post-implementation. In issue 4.1d above, we discussed the issues related to analyzing this tradeoff, so we raise just one additional issue here. At time of model development, model developers analyze tradeoffs between model predictiveness and estimates of the likelihood that scoring model will create a disparate impact when the lender implements. Post-implementation, economists can analyze tradeoffs between model predictiveness and estimates of actual disparate

treatment on the actual applicants for which the lender applied the scoring model. This analysis provides a more accurate assessment of these tradeoffs, but it is unclear how this analysis fits within the disparate impact framework. Economists therefore need legal guidance on whether this type of analysis has value.

V. Conclusion

From a purely analytical perspective, disparate impact analyses are fairly straightforward as economists have a suite of standard statistical tools available to conduct these analyses. The challenge comes from the lack of clarity on the legal criteria for evidence of disparate impact. In this report we have raised many of the legal issues and questions that economists have when conducting a statistical analysis of disparate impact within the three-prong, burden shifting framework. Unfortunately, we have raised many questions but offered few solutions. Hopefully, however, this report will help create the necessary discussions between relevant stakeholders to answer some of these questions and fill some of these gaps.