Lessons From The Past
Robert Avery

Federal Housing Finance Agency

Artificial Intelligence in Consumer Finance: Defining and Insuring Fairness
Federal Reserve Banks of Cleveland and Philadelphia
November 9-10, 2021

Outline of my remarks

- The focus of this conference is on the use of analytic tools for decisioning in the consumer finance market and how they impact fairness. Question for my talk is: What does history tell us about this topic?
- There are 4 general components of the environment governing the use of analytic tools: (1) technology used to develop models (2) legal structure (3) data inputs (4) products where tools are applied. Each of these are germane to our topic.
- Although the consumer finance market of 1960 was very different than it is today, I would argue that of its evolution took place between 1960 and the mid 1990s—changes since the 1990s have only been incremental.
- First third of talk will focus on the development of this market structure.
- Second third will report on a 2007 FRB Congressional study of disparate impact that I co-authored with Glenn Canner and Ken Brevoort. I will modestly argue that this report is the most relevant existing study addressing the issues of disparate impact and fairness focused on in this conference.
- Last third, talk about the applicability of the 2007 study results to today's issues.

1960 to 1980

- Technology
  - Earliest models were for finance companies. In 1958 Fair and Isaac built their first commercial score. Geared toward default as price was fixed & one-time loans.
  - Way before its time
    - Logistic model – log odds.
    - Multiple “scorecards” (interactions)
    - Categorical variables created by using decision tree models to break up continuous variables.
    - Stepwise regression used to add variables. Tested with holdout samples.
- Legal structure established in 1970’s
  - Dominated previously by state usury laws. No interstate banking limiting competition
  - FCRA (1970)
    - Right to dispute/informed when adverse action taken
    - Defined permissible purposes – means users and suppliers of data likely to be the same. Dramatically increases leverage of credit bureaus and incentives for clean data and resolving disputes
  - ECOA (1975) as implemented by Reg B
    - Identifies protected classes
    - Rules for credit scoring –empirically derived and statistically sound (Reg B)
    - Defined rules governing choice of reasons for adverse action (less mechanical)
    - Related laws for mortgage (HMDA and FHA) and fraud (Reg E)

1980 to 1995

- Major changes in data and in products
  - Consolidation of 3 credit bureaus into national entities in early 1990s. Availability of online reports. Universal reporting of mortgages mandated by GSEs. Data in bureaus reflect models used in 1960s. Circularity to process.
  - Shift from closed-end to open-ended products
    - Consolidation of the credit card market into large players facilitated by the ability to take advantage of national credit bureau data and avoid state usury laws.
    - Development of interstate commercial banking and expanded role of mortgage secondary market did the same with other credit.
    - Failure of the thrift industry in the 1980s & consumer finance industry in the late 1990s led to the further consolidation of their products into commercial banks.
- Technology
  - First generic credit history scores – MDS Bankruptcy Score 1987 and FICO Prescore in 1987. Made possible because of national credit bureaus.
  - Models estimated and scored entirely from credit bureau data and only available through one of the 3 bureaus. Models are bureau specific.
  - MDS and FICO technology built on custom score experience & culture.
- Legal environment little changed since early 1980’s. Major changes were the development of statistical fair lending support within Justice/FFIEC and the decision under Reg B to prohibit collection of racial data except for mortgage.

2007 Federal Reserve Congressional Report

- Fact Act (2003) mandated a Fed/FTC study of the effects of credit scoring on the availability and affordability of credit and insurance and certain issues of adverse impact (which I will term differential effect).

- Study released in August 2007
  - Addressed the potential differential effect of scoring itself
  - Addressed the potential differential effect of factors embedded within the scoring model
  - Examined the extent to which scoring system could achieve comparable results using factors with less negative impact

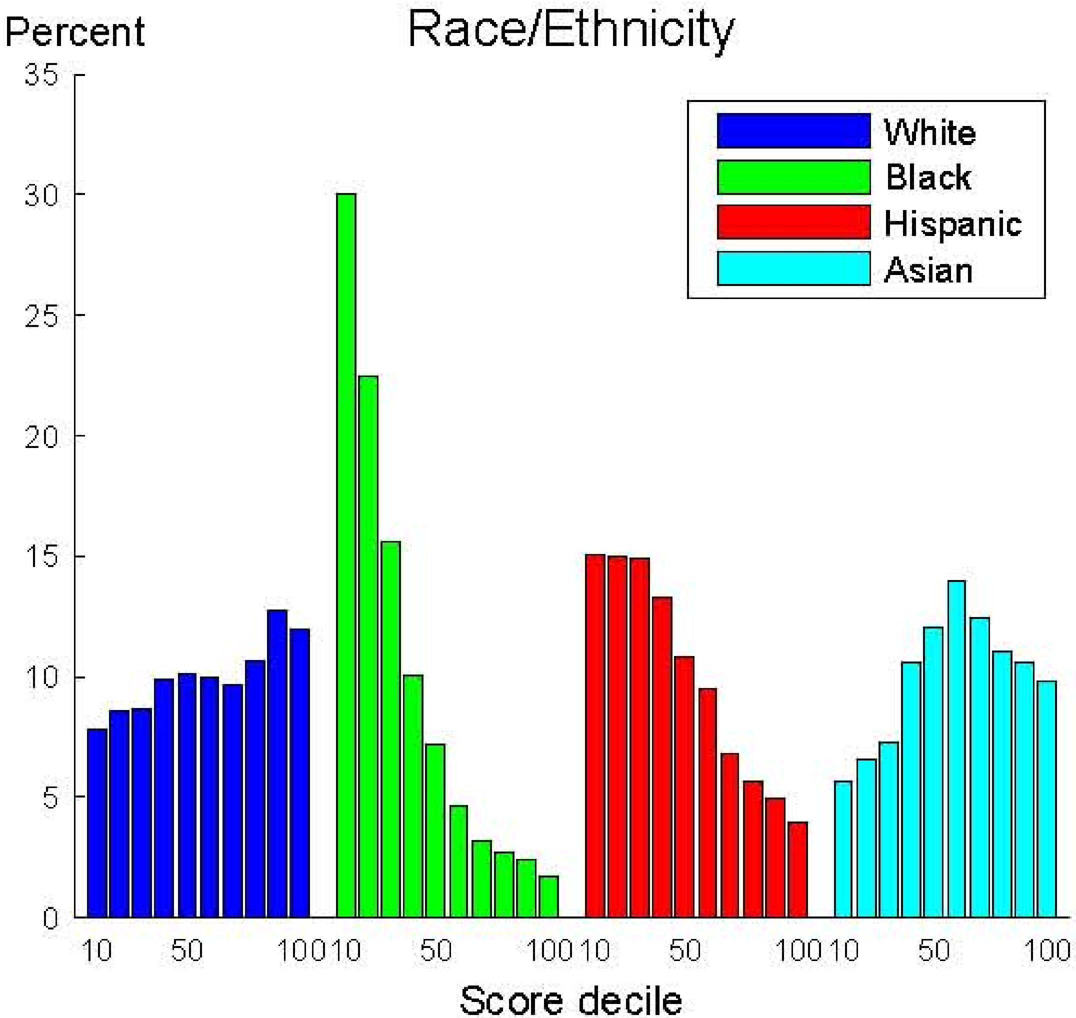
- Unique dataset combining Social Security data on race/gender/age with panel of 301,000 random TransUnion customers with full credit records at 2 points in time in early 2000s
  - TransRisk (generic TransUnion account maintenance score) and VantageScore plus we estimated our own FRB model built with 312 credit attributes
  - Normal dataset that would be used to estimate a generic credit history score

Distribution of Scores

- Scores vary wildly by Race/ethnicity and Age
  - Blacks and Hispanics have significantly lower scores
  - Individuals under 30 have significantly lower scores
  - Black score differential is constant over lifetime.
  - These differences are reduced once other factors are accounted for, but do not go away.
  - <u>One-half</u> of the gross difference in score between blacks and whites remains after controlling for age and census tract. Similar results controlling for age and income.
  - Implies that race of neighborhood is an imperfect substitute for race of individual.

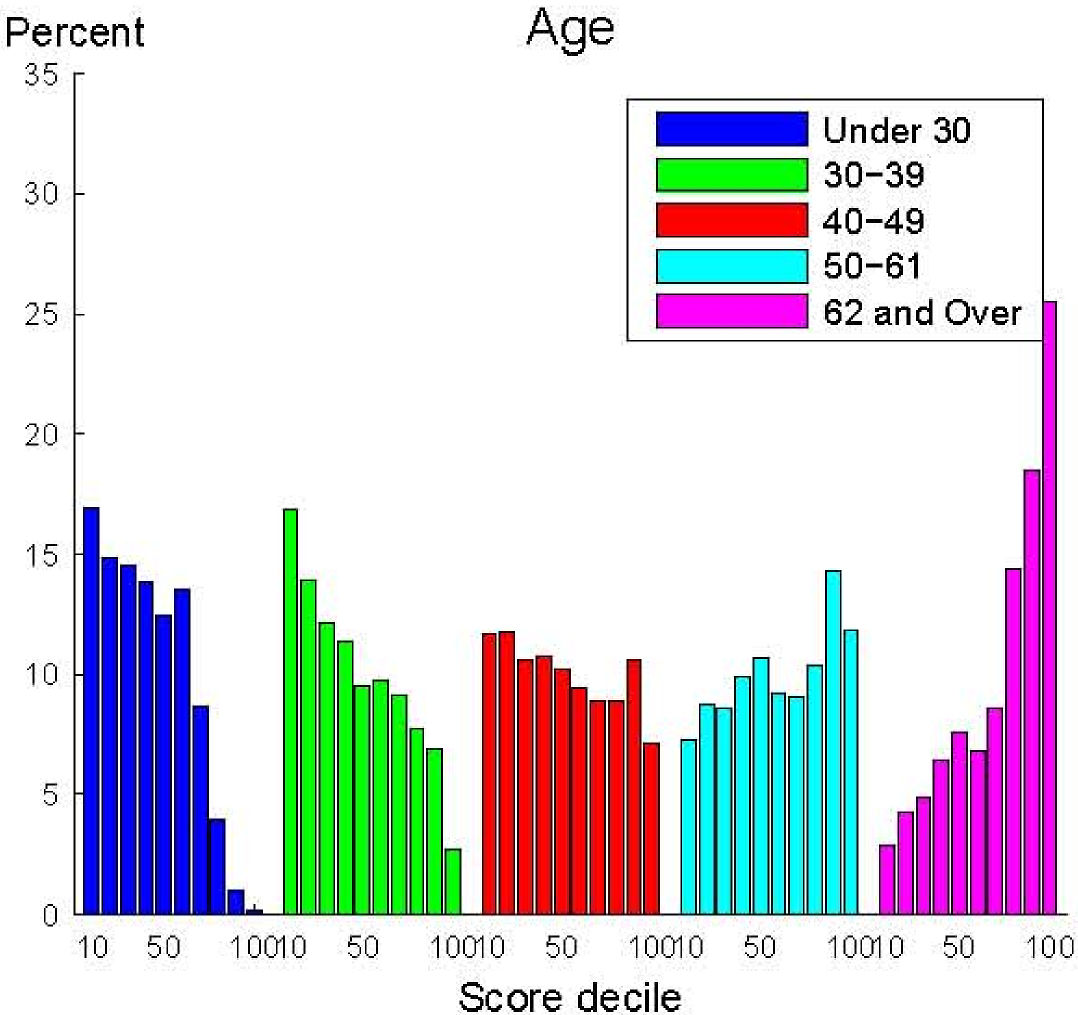
- Scores have modest variation by gender and marital status

Distribution of Score (Normalized TransRisk): Race



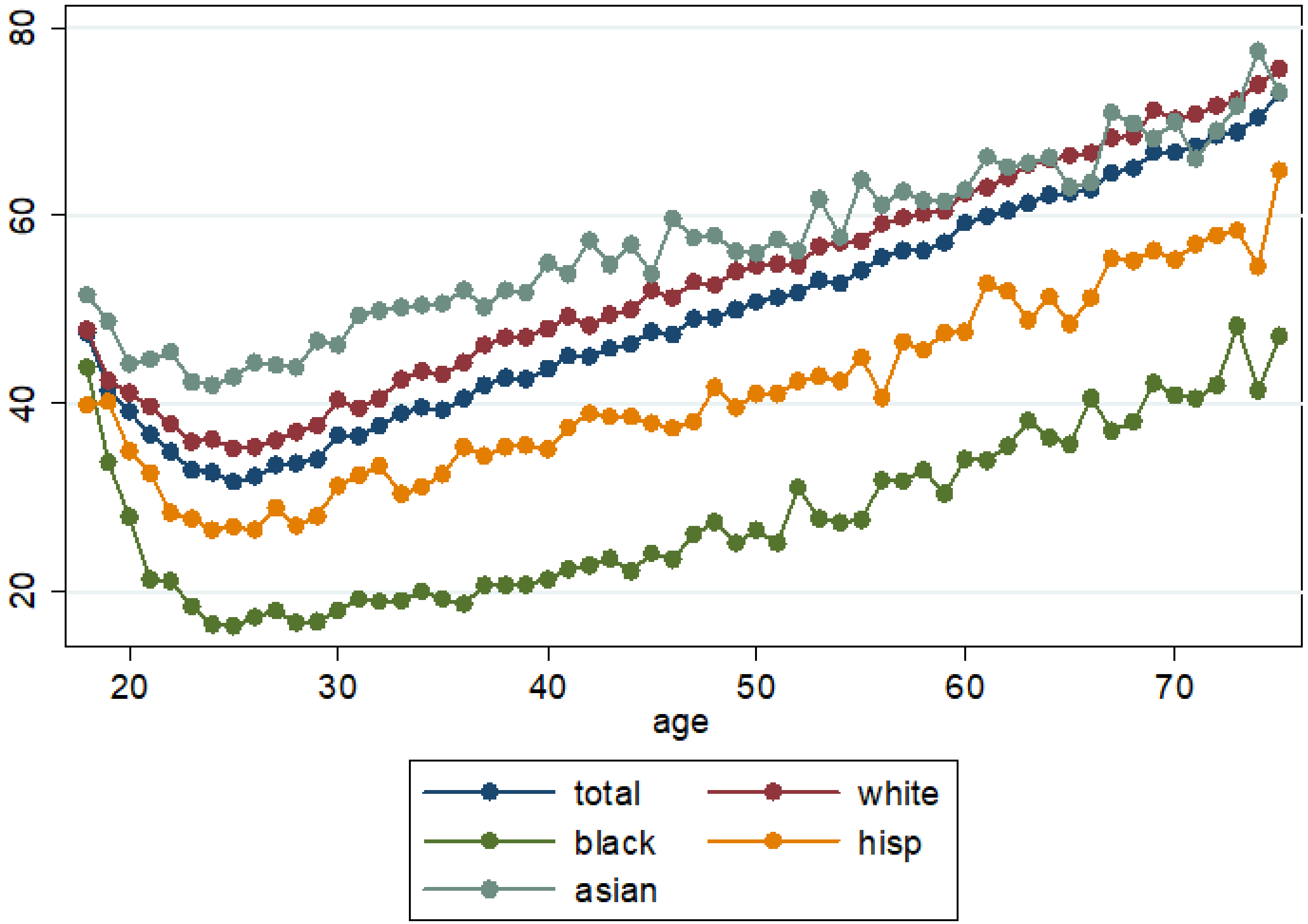
	Mean Score
White	54.0
Black	25.6
Hispanic	38.2
Asian	54.8

Distribution of Score (Normalized TransRisk): Age

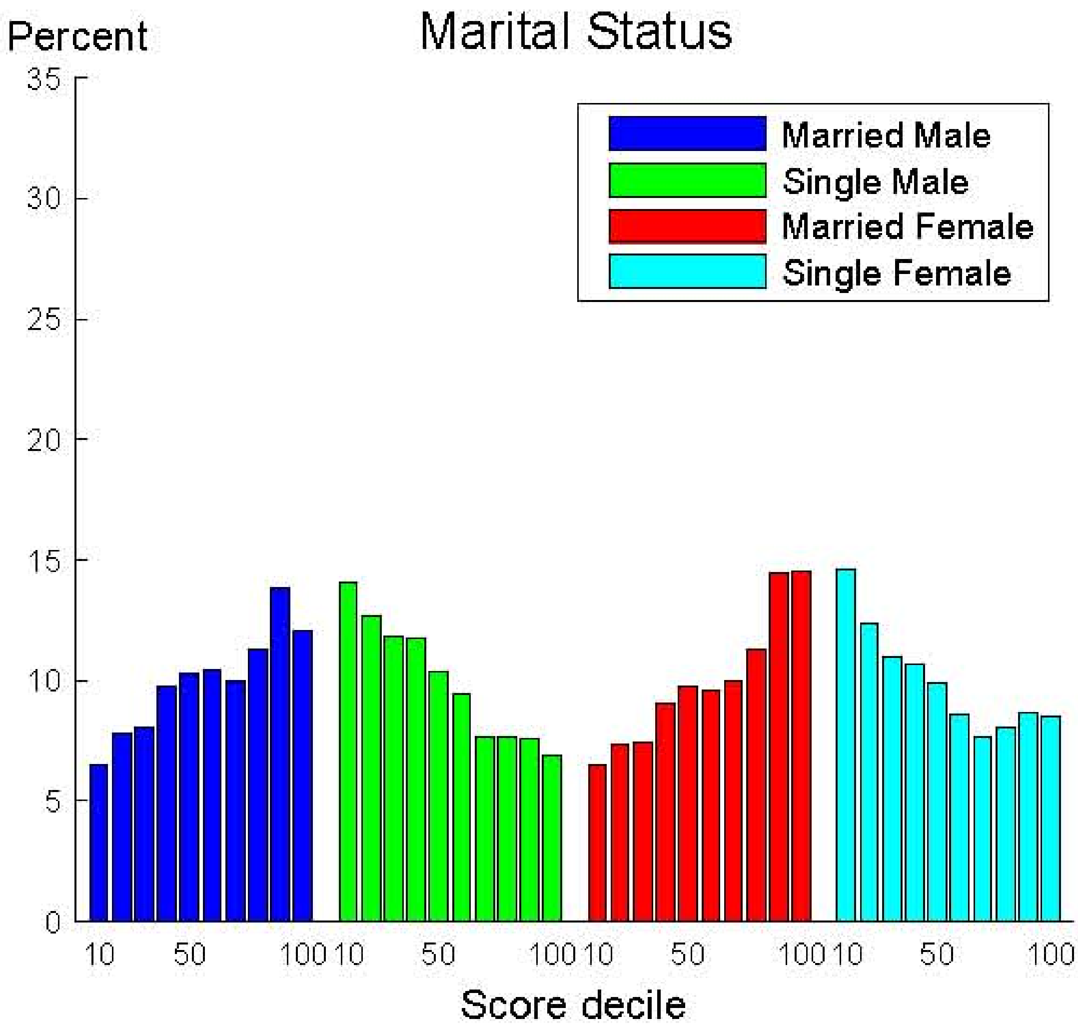


	Mean Score
Under 30	34.3
30 – 39	39.8
40 – 49	46.9
50 - 61	54.5
62 and over	68.1

Distribution of Score (Normalized TransRisk): Race and Age



Distribution of Score (Normalized TransRisk): Gender/Marital Status



	Mean Score
Married Male	55.7
Single Male	43.4
Married Female	57.5
Single Female	44.8

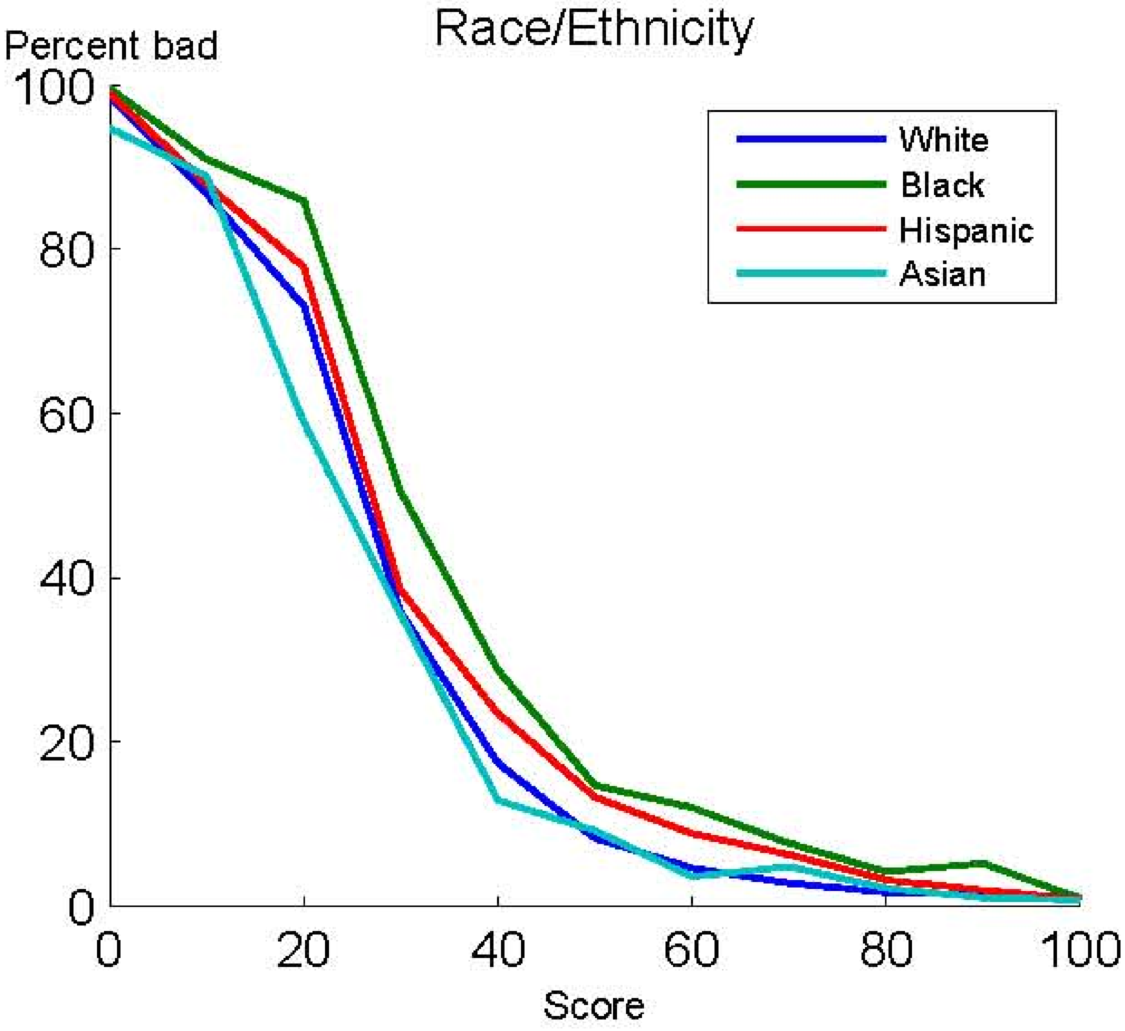
Summary of Results – Differential Effect of Score Itself

- If there were “differential effect” of score on protected group then the group should “over-perform.” Might also be the case that the score predicted poorly within the group.

- However, in study the credit scores rank ordered individuals by credit risk for all populations fairly evenly.

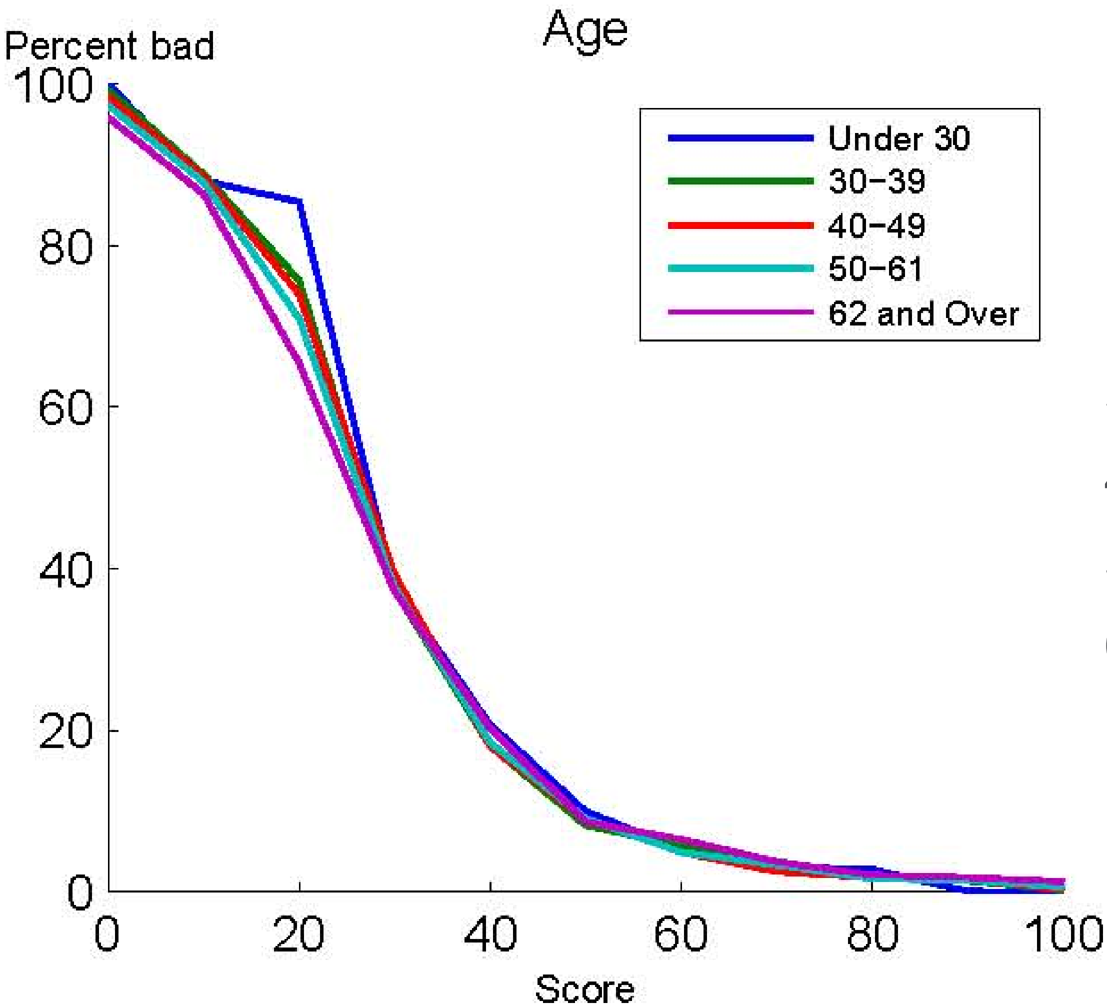
- Across populations, actual performance conditional on credit score was approximately the same.
  - Blacks and single individuals underperformed
  - Married Individuals and foreign-born individuals (particularly recent immigrants) overperformed.
  - These differences are narrowed somewhat with controls for product etc. but do not go away completely.

Differential Effect of Score: Prediction of "Any Account Bad" by Race



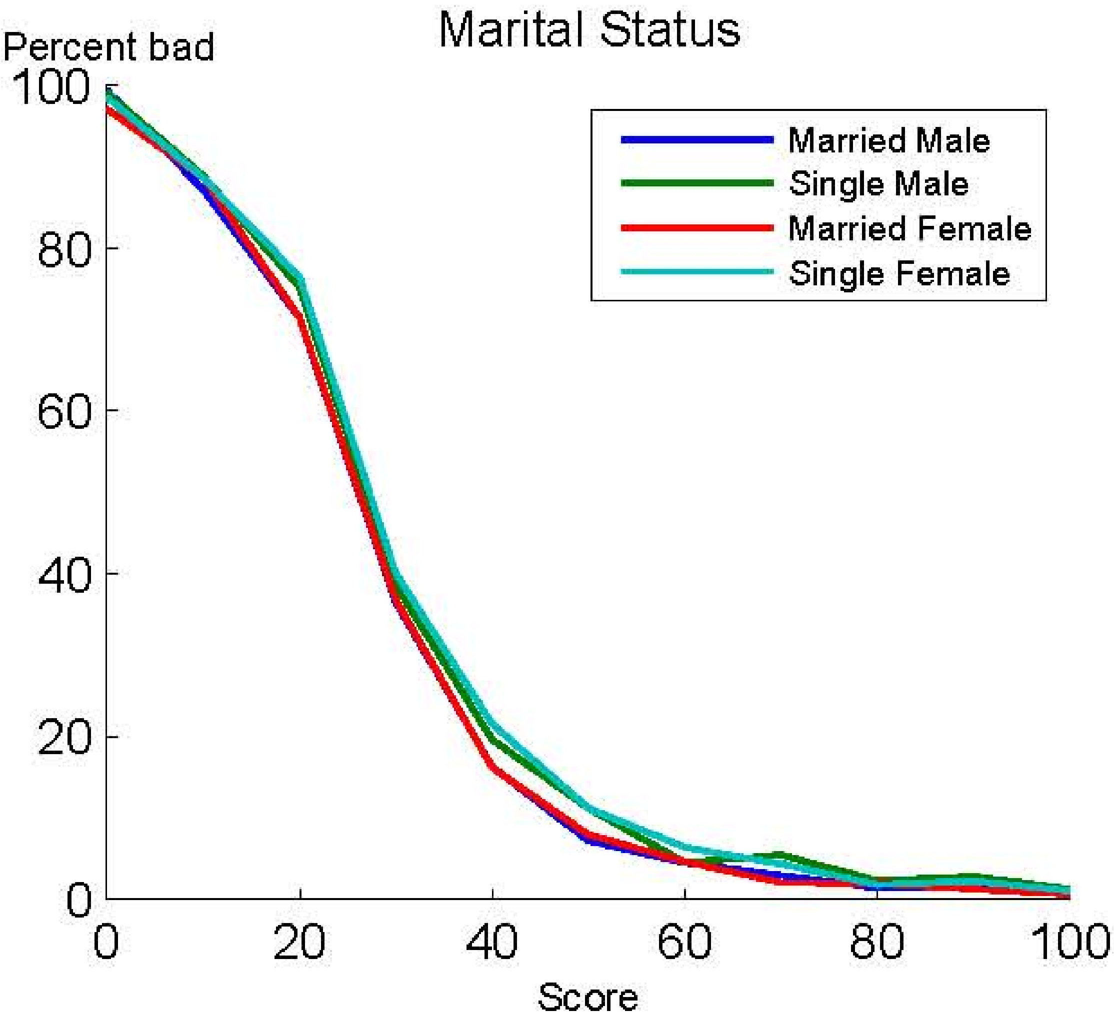
	Mean Residual
White	-1.0
Black	5.6
Hispanic	1.7
Asian	-2.1

Differential Effect of Score: Prediction of "Any Account Bad" by Age



	Mean Residual
Under 30	1.5
30 – 39	-0.2
40 – 49	-0.4
50 - 61	-0.7
62 and over	-0.3

Differential Effect of Score: Prediction of "Any Account Bad" by Gender/Marital Status



	Mean Residual
Married Male	-1.2
Single Male	0.4
Married Female	-1.1
Single Female	0.8

Differential Effect of Individual Factors

- A credit characteristic could be said to have a “differential effect” if the weight assigned to the characteristic in a model differs from the weight that would be assigned in a model estimated in a “demographically neutral” environment.

- Or a model embeds a “differential effect” if the mean credit scores for subpopulations change markedly when re-estimated in a demographically neutral environment.

- For FRB study many “demographically neutral” models were estimated for tests (e.g. a model developed using only white non-Hispanic individuals or with dummy variables for race.)

Example of Race-neutral Model testing of Individual Factors

	FRB Base		White		Race Indicator	
	Mean	KS	Mean	KS	Mean	KS
White (NH)	54.0	72.8	54.1	72.7	54.2	72.8
Black	25.8	69.4	25.9	69.3	25.9	69.4
Hispanic	38.3	66.2	38.5	66.1	38.5	66.3
Asian	54.8	66.9	54.9	66.4	54.9	66.4

Summary of Results – Differential Effect of Individual Factors

- No evidence any factors have differential effect by gender/marital status. Demographically neutral models have equal predictiveness of overall model

- No evidence any factors have differential effect by race/ethnicity. Demographically neutral models had slightly less predictiveness than overall model.

- Certain credit characteristics (related to age of oldest tradeline) serve as limited proxies for age.
  - Older persons scores are somewhat lower, younger persons and recent immigrants scores are a little higher than would be implied by age-neutral models because age tradeline more predictive than age.
  - However, scores of recent immigrants are lower than implied by their performance. This is because their credit profiles are similar to the young whose performance tends to be relatively poor.
  - No easy fix since omitting these variables reduces predictiveness.

Questions that remained after 2007 study

- Several "grey" areas which were not answered by the study.
  - Use of "authorized user" history. Under ECOA spouses are supposed to benefit from credit history of spouse. But should consumers be "punished" for bad credit history? Also, not legal to use spousal status. How does one identify appropriate ECOA compliance?
  - Treatment of Income – imperfectly studied in report. Quite complicated. Strong univariate relationship between income, race, gender and age. Conditioned on score, however, the relationship can reverse.
  - How to treat solicitations—are they "marketing" or offers of credit? Do consumers have right of dispute if not solicited for best products? Most redlining cases bought under FHA not ECOA.
  - Definition of "fairness." Study measures impact on protected group as a whole with the outcome metric of credit default. Is that the right definition? Very relevant for use of age in scoring model.

How Relevant is the Study for Today?

- Little reason to believe that the results would not still hold for models designed for one-time offers of credit/pricing of closed-end products--particularly those based entirely on information contained in the credit bureaus.
  - Information added to the credit bureaus in recent years (e.g. trendview, use of subprime products, rental/utilities) is likely marginal. Though models often contain information beyond credit bureau data "reasons for denial" still give consumers right of dispute and support transparency.
  - Huge advantage of generic credit risk scores in that they are built on very large populations with fewer problems of censoring and needs for adjustments such as reject inference. Allows for more scorecards and interactions.
  - With a full range of testing, neither FHA nor Freddie Mac were able to develop models based only on their own customer base, which out-performed those using embedded generic FICO scores.
  - In the FRB study, the FRB model performed about the same as the commercial scores. This suggests that differences among these products is marginal in terms of predictiveness.
  - Marginal benefits of predictiveness (at least for credit history component) may not be enough to overcome large fixed costs of development and lack of transparency and the potential for gaming that come with custom scoring.

Open-ended Credit and Account Management

- Potentially different story for open-ended credit where "account management" is much more important.
  - Such models are likely to rely much more on a continuous data collection process.
  - Risk is also more incremental & much less a function of the original decision to extend credit.
  - Lender objective function likely to involve more than managing credit default. Lenders might benefit from ability to "holdup" borrowers, for example (although Card Act limits the ability of lenders to take advantage of such situations).

Issues with Account Management Models

- If account management models are based entirely on information in the credit bureaus it is likely that FRB results apply.

- However, it is much less clear if such models rely on other data such as how and what the customer buys (such as used in fraud models).

- No one would be surprised if such behavior varied across demographic groups. Use of such variables is complex and it is the multivariate relationships that matter. Intuition may be misleading.

- Tools to combat disparate impact have real limitation. Other than mortgage credit, Reg B prohibits the collection of information on race/ethnicity.
  - FRB study showed that neighborhood is very imperfect substitute for race of individual.
  - How does a responsible lender test for bias then?
  - How does a regulator conduct enforcement?