ELSEVIER

Journal of Public Economics 92 (2008) 944–961

JOURNAL OF PUBLIC ECONOMICS

www.elsevier.com/locate/econbase

Understanding preferences for income redistribution ☆

Louise C. Keely [a], Chih Ming Tan [b,*]

[a] *Department of Economics, University of Wisconsin, United States*
[b] *Department of Economics, Tufts University, 8 Upper Campus Road, Medford, MA 02155, United States*

Received 11 July 2005; received in revised form 21 June 2007; accepted 8 November 2007
Available online 15 December 2007

Abstract

Recent research suggests that income redistribution preferences vary across identity groups. We employ statistical learning methods that emphasize pattern recognition; classification and regression trees (CART™) and random forests (RandomForests™), to uncover what these groups are. Using data from the General Social Survey, we find that, out of a large set of identity markers, only race, gender, age, and socioeconomic class are important classifiers for income redistribution preferences. Further, the uncovered identity groupings are characterized by complex patterns of interaction amongst these salient classifiers. We explore the extent to which existing theories of income redistribution can explain our results, but conclude that current approaches do not fully explain the findings.

JEL classification: C45; C49; H50; H53
Keywords: Data mining; Classification and regression trees; Random forests; Redistribution preferences; Identity

1. Introduction

The idea that an individual's identity potentially plays a key role in determining preferences that affect economic decisions and outcomes has gained increasing acceptance in the economics literature since the seminal work by Akerlof and Kranton (2000). This emerging literature should rightly be seen as an extension of the large body of work in sociology examining how people make sense of their world and how identity plays into their views of themselves and others (see, for instance, Lamont (2000)). In this paper, we focus our attention on one such preference — an individual's preferred level of income redistribution.

☆ We thank Jim Andreoni, Buz Brock, Steven Durlauf, Carol Graham, Yannis Ioannides, Wei-Yin Loh, Larry Samuelson and seminar participants at the London School of Economics, University of North Carolina-Chapel Hill, and the University of Wisconsin Institute for Research on Poverty for comments. We also thank two anonymous referees for their valuable suggestions and insights. We are grateful for funding under the Robock Award in Empirical Economics from the University of Wisconsin. Keely thanks WARF and the Brookings Institution for generous research support, and Brookings for their hospitality. Tan thanks the generous research support provided by the Program of Fellowships for Junior Scholars, MacArthur Research Network on Social Interactions and Economic Inequality. We greatly appreciate the excellence and diligence of our research assistant, Zhiguo Xiao.
* Corresponding author.
E-mail address: chihming.tan@tufts.edu (C.M. Tan).

0047-2727/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2007.11.006

Existing theoretical treatments of how an individual's identity determines her income redistribution preferences highlight various mechanisms. In *preference-based* theories, identity matters because people care, in an exogenous fashion, about the actions or outcomes of others in the same or across identity groups. The relevance of identity to economic decision-making is modeled via modifications to the preference structure. For instance, Alesina et al. (1999) and Alesina et al. (2001) both propose simple models which capture an individual's utility being dependent on the utilities of members of other ethnic groups. They conclude that this awareness of ethnic heterogeneity, or "racism" (Alesina et al. (2001)), could be responsible for the divergence in views on redistribution across groups.

In *information-based* theories, identity provides information about an individual's future economic circumstances in an environment with uncertainty. Identity groupings may correspond to a set of initial conditions that have persistent implications for income mobility (Benabou (1996) and Benabou and Ok (2001)) or for shaping mobility beliefs (Loury (1998) and Piketty (1995)) that in turn determine redistribution preferences. Loury (1998), for instance, argues that people are 'socially located' — they are part of social and cultural networks that exert strong influence on behavior. Behavior may be ex-post rational in that it is self-fulfilling and persistent. As a result, initial differences across groups can have long-run effects on outcomes such as income or preferences for income redistribution.

The aim of this paper is twofold. The first and primary objective of this paper is to propose a new way to explore the empirical implications of information-based theories using data from the General Social Survey (GSS). To do so, we exploit the fact that these theories imply a mapping between identity variables and various outcome variables that are in turn related to income redistribution preferences. These mappings imply restrictions on the data that we propose to investigate. We make clear the exact restrictions we are investigating and how we do so in Section 2.

The second objective of this paper is to collect a (broad) set of stylized facts regarding redistribution preferences. For instance, as noted above, preference-based theories tend to focus on ethnicity as the important identity marker for determining redistribution preferences. It is of interest, however, to ascertain whether there are other prominent dimensions of identity that matter. We also extend our analysis to investigate the relationship between American's views on helping the poor more generally and their views on welfare policy in the United States over the past two decades.

To address the above two objectives, and to uncover the role of identity in driving differences in redistribution preferences, we employ statistical learning methodologies that emphasize pattern recognition; i.e., classification and regression trees (CART™) and random forests (RandomForests™). These statistical learning methods, which are widely used in other disciplines, provide insight into these views that linear regression could easily miss.

Existing empirical work in this area typically employ the following strategy: (i) a linear relationship between redistribution preferences and other covariates is assumed, (ii) investigations of heterogeneity in redistribution preferences are carried out using pre-specified identity groups; for instance, the existing empirical literature tends to emphasize the (a priori) importance of race and gender, and (iii) typically, only a small number of alternative specifications (such as interactions between covariates, and polynomial terms) are explored before settling on a particular specification that is then reported.

For instance, in a work that is closest to ours in spirit, Alesina and La Ferrara (2005) employ a linear ordered probit model to study the determinants of redistribution preferences. They consider a large number of such determinants, including respondents' age, race, gender, socioeconomic class, etc., but do not report any results for possible interaction effects for these variables. On the other hand, Fong (2001), which explores support for redistribution, does consider the interaction effect between race and gender. However, her choice of interaction effect is made a priori; there is no systematic attempt to explore other possibilities.

An important consequence of the above empirical strategy is that, because nonlinearity and heterogeneity are not systematically investigated, this work in effect makes strong prior claims about the correct (econometric) model for preferences. Researchers essentially focus their attention on a small number of models out of the very large set of possible models that could have been generated if one considered nonlinearity as well as the full range of possible interactions between the covariates. However, as pointed out by Brock and Durlauf (2001) in the economic growth context, there are many instances in economics where theories about particular outcomes are "open-ended". By open-ended, Brock and Durlauf are referring to the possibility that in these instances, the fact that one theory (e.g., gender) may be important to redistribution preferences, does not automatically exclude some other theory (such as race or parental income or any of the many other possible alternative explanations for variations in such preference) from also being important. It also does not exclude the possibility that their interaction may be important.

The important point is that theory open-endedness implies model uncertainty, and therefore, any assessment of the likely effect of an explanatory variable on redistribution preferences should be made with the full universe of possible

alternative models in view. We should not a priori restrict ourselves to only exploring a small subset of models in the model space. Furthermore, as Manski (1993) argued, using identity groups that are pre-specified by the econometrician potentially leads to serious estimation consequences and incorrect inference when those groups are defined differently by the individuals whose responses compose the data.

The statistical learning methods we employ in this paper allow us to better address these points. CART™ is a flexible estimation method that automatically detects nonlinearities and interactions supported by the data. It does so by recursively partitioning the set of responses into increasingly homogenous subsets. At each stage, an identity variable is chosen to facilitate the splitting of the sample. In this paper, we consider identity markers such as age, race, gender, initial birthplace, religious upbringing, and socioeconomic background using data from the GSS. The final set of groupings is selected according to a generalized information criterion. CART™ effectively allows the researcher to consider a very large model space consisting of the main and interaction effects of a wide range of identity markers.

The result of the CART™ procedure is to classify responses to various GSS questions into groups according to similar shared identity characteristics of respondents. These results are reported in the form of a regression tree. It should be noted that this is accomplished without the need for the researcher to impose any a priori structure on the number or nature of these groups; nonlinearity and interaction detection is automatic in this sense. The results are also structurally interpretable in the sense that they reveal the relative importance of particular identity characteristics to responses, such as the preference for redistribution.

CART™ has been shown to be consistent in the sense that as the number of observations gets large, the algorithm reproduces the "true" set of sample splits (see Breiman et al. (1984)). Their weakness, however, lies in the lack of available asymptotic results that would be useful for conducting inference on split variable choices and split value estimates.[1] Our method, therefore, does not allow for a straightforward hypothesis test of, for instance, the Benabou and Ok or Piketty predictions with the classification patterns uncovered in the data. We therefore attempt to assess the validity of our CART™ tree results in terms of prediction performance. Specifically, we compare them with those obtained using Breiman's (2001) RandomForests™ (RF) algorithm.

RF is an adaptive classification method which combines bootstrap aggregation ("bagging") with pooling information from a multiplicity or ensemble of randomly built trees to obtain classifications of the outcome responses with lower mean prediction error compared to CART™. In fact, Breiman (2001) has shown that the prediction performance of RF is currently unmatched beating other leading adaptive learning methods like boosting. However, because RF pools information from a multiplicity of (randomly generated) trees, the results lack the sort of structural interpretability that CART™ is able to offer in the form of a tree diagram. Because the uncovering of such structure is a main goal of this paper, we limit RF's role to two aspects. RF offers guidance on which identity markers are salient in the classification of outcome responses into groups; we wish to compare the identity markers found to be important by RF with those in our CART™ tree results. Also, we want to see how much better RF does in terms of reducing mean prediction error when compared to CART™ in order to assess the validity of the latter's results. We refer the reader to the Technical Appendix for more details of these classification methods.

In terms of our findings, we present new evidence from the GSS that views on whether there should be governmental administration of income redistribution are found to differ along racial, gender, and class lines in the United States. That is, identity groups are found to be salient in describing individual views regarding government's role in the reduction of income inequality. However, the uncovered pattern of interaction between these identity markers suggests a more complex relationship between identity and redistribution preferences than previously found by past studies. Our exploration of these empirical restrictions also leads us to conclude that existing information-based theories are potentially inadequate for explaining redistribution preferences.

In Section 2, we provide a discussion and formal framework for investigating the set of restrictions that link existing information-based theories of income redistribution preference determination and their empirical implications. We present our findings in Section 3. Finally, Section 4 concludes.

[1] It should be noted that there have been recent advances on this front in the context of test-based sequential sample splitting and threshold regression (as opposed to classification) models (see Hansen (1999, 2000)). However, results such as confidence intervals derived in these settings are typically restricted to the single split variable–single split case. There is, however, some comfort from the fact that studies comparing classifications obtained by CART™ with those gathered using sample splitting methods tend to be identical (see, in particular, Duffy and Engle-Warnick (2006) as well as Hansen's (2000) replication of the results in Durlauf and Johnson (1995).

2. Evaluating information-based theories

2.1. Discussion

We consider two classes of information-based theories for redistribution preferences. In one set of theories, identity corresponds to a set of initial conditions for the individual, and these have persistent effects. In this way, outcomes across individuals can be classified according to these initial conditions.

The set of identity variables we consider in this paper are exogenous identifiable characteristics from the GSS. It is comprehensive and includes the respondent's age in years (AGE), her gender (SEX), her self-reported race[2] (RACE); the region of the US in which she was living at 16 (REGION16), whether the respondent was born in the US (BORN), whether the respondent's parents were born in the US (PARBORN), the respondent's mother's highest educational degree as a proxy of socioeconomic background (MADEG), what religion in which the respondent was raised (RELIG16), and the respondent's description of her religious upbringing as fundamentalist, moderate or liberal[3] (FUND16). A trend variable (YEAR) is also included. A summary of these variables is provided in Table 1.

Benabou (1996) surveys the literature on inequality and its immediate implications for, among other things, redistribution preferences. The basic idea is to link income heterogeneity with variation in private tolerance for inequality, and in turn with differences in the preference for income redistribution.

We consider two measures of redistribution preferences,[4] EQWLTH and NATFARE. Our main redistribution preference measure will be EQWLTH which asks whether the respondent thinks that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Respondents are asked to choose a response on a scale of 1 to 7 where a score of 1 means that the government ought to reduce the income differences between the rich and poor, and a score of 7 means that the government should not concern itself with reducing income differences. EQWLTH is asked in each wave of the GSS between 1978 and 2000. We also compare our results for EQWLTH with those for NATFARE which asks the respondent whether she thinks America is spending too much (scored as 1), too little (scored as 2), or about the right amount on welfare (scored as 3)? For consistent comparison, the sample considered is also each survey wave between 1978 and 2000.

An interesting extension of Benabou's mechanism is proposed by Benabou and Ok (2001). They formalize a "prospect of upward mobility" (POUM) hypothesis in order to understand why individuals with less than the population mean income may vote against income redistribution. They show that with a single, commonly-known, concave function that links current to future individual income, a group of voters with incomes less than the mean but above some threshold will vote against redistribution. They do so because the concavity of the mobility process leads them to expect a higher than average income in the next period.

This model predicts that patterns of heterogeneity in income, prospects for upward mobility, and preference for income redistribution should be related, with a one-to-one correspondence between the latter two. From this framework we expect to find that any classification of responses to redistribution preferences (EQWLTH or NATFARE) according to identity matches those for perceived prospects of upward mobility. We measure perceived prospects for upward mobility with the variable GOODLIFE[5] that asks whether the respondent agreed with the statement that given the way things are in America, people like her and her family had a good chance of improving their standard of living.

In a second set of theories, identity is viewed as a source of information about one's outcomes in an environment with uncertainty. Initial differences across groups can have long-run effects on outcomes such as income or preferences

[2] This question asks the respondent to identify herself as White, Black, or other. While we would have preferred a question with more ethnic detail, this was the best question that the GSS offered over many waves.

[3] Unlike the other identity variables that we use, FUND16 is not objective. We include this variable because of an a priori hypothesis that religious background may impact one's view of income redistribution. The variable RELIG16, that classifies the denomination of religious upbringing, does not distinguish between, say, different ideologies within Protestantism. We use FUND16 as an attempt to allow for such distinction. We ran the trees for EQWLTH, the main question of interest on income redistribution preferences, as well as NATFARE with and without FUND16 as an explanatory variable. In fact, we find that neither RELIG16 nor FUND16 appears in a robust manner as a classification variable except for some trees classifying socioeconomic status.

[4] These questions are used in related empirical studies. Alesina and La Ferrara (2005) use both EQWLTH and NATFARE. Luttmer (2001) employs NATFARE in his work.

[5] See also Alesina and La Ferrara's (2005) discussion of this subjective measure of upward mobility expectations.

Table 1
Summary statistics

	Years	Mean	Std. Dev.
Identity markers			
SEX	1978–2000	1.56	0.50
RACE	1978–2000	1.16	0.44
REG16	1978–2000	4.37	2.46
BORN	1978–2000	1.06	0.24
PARBORN	1978–2000	1.24	0.60
MADEG	1978–2000	0.81	0.94
RELIG16	1978–2000	1.47	0.73
FUND16	1978–2000	1.90	0.74
AGE	1978–2000	44.78	16.98
Dependent variables			
EQWLTH	1978–2000	3.76	1.95
NATFARE	1978–2000	2.32	0.77
INCGAP	1987, 1996,2000	2.34	1.13
WHYPOOR4	1990	1.62	0.63
GOODLIFE	1987,1994,1996,2000	2.42	1.03
GETAHEAD	1980–2000	1.45	0.70
OPHRDWRK	1987	1.75	0.69
LFEHRDWK	1993	1.49	0.64
PADEG_ABS_DIF	1978–2000	0.94	1.03
PARSOL	1994–2000	2.21	1.11
KIDSSOL	1994–2000	2.79	1.55
REALINC	1978–1996	31,075.67	26,563.77
REALRINC	1978–1996	20,299.39	18,686.24
DEGREE	1978–2000	1.43	1.17
CONFED	1978–2000	2.16	0.67
CONLEGIS	1978–2000	2.17	0.62

for income redistribution if expectations lead to self-reinforcing behavior. As an example of such information-based models, Piketty (1995) presents a model in which there is a single mobility process that is unknown to agents. Agents learn from past mobility experience to form beliefs about the true mobility process. In this framework, mobility beliefs (but not restricted to just beliefs about upward mobility) directly inform preferences for income redistribution. Further, mobility beliefs are parameterized to correspond to views on the relative importance of luck and hard work in determining one's future income. We consider various measures of mobility beliefs such as questions that ask about the role of hard work in getting ahead (OPHRDWK, LFEHRDWK, and GETAHEAD) as well as those that ask about (perceived) intergenerational mobility between parents and children (PARSOL and KIDSSOL).

According to this model, long-run differences in preference for redistribution (EQWLTH or NATFARE) and mobility beliefs are a result of two forces. Initial differences in the priors over the true mobility process are one factor. A second is that individual learning about the mobility process uses incomplete information that varies across individuals. Specifically, individuals use information only from their own past experience and the population's average experience, and individuals do not experiment in order to learn. A role for identity, akin to that suggested by Loury, is introduced into this framework by allowing individuals to extend their learning to a reference group that is defined by identity. In this setting, heterogeneity in mobility beliefs and income redistribution preferences (EQWLTH or NATFARE) across individuals will both correspond to these reference groups.[6]

It would also be interesting to see how classification patterns with regard to socioeconomic status compare with those for redistribution preferences. Note that in the Benabou and Ok model, relatively poorer people may nevertheless vote for higher income distribution. We would therefore expect to see more complex classification patterns for redistribution preferences than for income. In Piketty's framework, mobility beliefs within reference groups can

[6] For long-run heterogeneity in mobility beliefs, we require that these reference groups vary in their priors regarding a true mobility process and that there is heterogeneity in the income distribution history across reference groups.

converge over time, although they may differ across groups. Income heterogeneity will not disappear because it is determined in part by a stochastic process that is exogenous to beliefs. In this case, the model suggests that we should observe identity groupings for socioeconomic status that are at least as complex as that for redistribution preferences. We measure socioeconomic status using a measure of the respondent's education level (DEGREE) as well as her real family income (REALINC).

We use the above predictions of the Benabou and Ok and Piketty models to structure our empirical study and we evaluate the consistency of those predictions against the data. We provide a more formal presentation of our empirical strategy in the next subsection.

2.2. Framework

Formally, let $y \in Y$ denote an outcome variable of interest that takes on K categorical values $\{y_1,\ldots, y_K\}$ and let $x \in X$ be a vector of $\boldsymbol{M}$ identity markers (which might be discrete or continuous variables or a mixture of both). We model the population of individuals as being classified by their identity markers into an unknown number b of subpopulations indexed by j. Within each subpopulation j, individuals are expected to return a response of y_j^* for the outcome variable of interest. The classification of individuals into identity subgroups corresponds to the partitioning of the support of identity markers, X, into b partitions, $\Lambda = \{A_j\}_{j=1}^{b}$. The partitions A_j are mutually exclusive and their union is X. That is, $A_j \cap A_l = \varnothing$ and $\cup_{j=1}^{b} A_j = X$. We use CART™ to provide us with estimates for the number and nature of the identity partitions, as well as the predicted responses within each group; i.e., $\left(b, \Lambda, \left\{y_j^*\right\}_{j=1}^{b}\right)$.

For example, suppose y measures redistribution preferences, and x=(Race, Sex) where Race takes on values $\{B, W\}$ and Sex takes on values $\{M, F\}$. Then, a possible set of identity partitions reported by CART™ may be three optimal groupings $\Lambda = \{A_1, A_2, A_3\}$, $\{(BF), (BM), (WM), (WF)\}$, with corresponding expected responses $\{y_B^*, y_{WM}^*, y_{WF}^*\}$. That is, in this example, if this was the set of identity groupings that we uncovered in the data, we would conclude that redistribution preferences differ systematically across subgroups in the population depending on whether respondents are Black, White-male, or White-female.

Suppose there are two outcomes of interest, y_1 and y_2, where y_2 measures the redistribution preference (EQWLTH or NATFARE). A theory of redistribution preference may imply a mapping f of a partition Λ_1 that corresponds to y_1 into a partition Λ_2 that corresponds to y_2.

As discussed above, under Piketty's theory, f implies that $\Lambda_1 = \Lambda_2$ where y_1 represents mobility beliefs (OPHRDWK, LFEHRDWK, GETAHEA, PARSOL, and KIDSSOL). Under the framework of Benabou and Ok, f implies $\Lambda_1 = \Lambda_2$ where y_1 represents perceived prospects for upward mobility (GOODLIFE). Further, a partition Λ_1 that corresponds to current income as y_1 should potentially have at least as many elements as Λ_2 in the Piketty case but should exhibit potentially less heterogeneity than Λ_2 in the Benabou and Ok case. We proceed to investigate these implications.

3. Results

The classification trees and random forests were constructed using pooled data for all years between 1978 and 2000 in which the relevant dependent variable was asked. A summary of all dependent variables used is provided in Table 1. Key results[7] discussed in this section are summarized in Tables 2–12.

In general, the CART™ and RF results are consistent. In particular, the variables that RF identifies as the most important classifiers generally reflect the splitting variables chosen by CART™. The difference in misclassification error rates between RF and CART™ are marginal at around 5% (with the former being the lower of the two as expected). However, the RF error rates are relatively high at above 60%. This is not entirely surprising since misclassification rates tend to increase with greater number of categories for the outcome response variable. Further, this error rate should be compared to an error rate between predicted response and actual response in a multinomial regression context; the typical R^2 number for studies employing GSS data is around 0.6 or (sometimes much) lower. Nonetheless, given that the aim of the classification exercise is the identification of homogenous groupings, the

[7] Some results described are not summarized in tables in order to keep the number of tables manageable. All results are available from the corresponding author upon request.

Table 2
EQWLTH classification tree and random forest results

Classification variable	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
RACE	Black and others	White	White	White	White	White	White	White
MADEG	All	Less than high school	Less than high school	Less than high school	High school to graduate	High school to graduate	High school to graduate	High school to graduate
AGE	All	<44	>43	>43	<26	>25	<37	>36
SEX	All	All	Male	Female	Male	Male	Female	Female
Predicted classification	1	1	7	4	6	7	3	6,7*

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, EQWLTH. EQWLTH measures redistribution preferences as defined by the following GSS question, "Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Others think that the government should not concern itself with reducing this income difference between the rich and the poor. Here is a card with a scale from 1 to 7. Think of a score of 1 as meaning that the government ought to reduce the income differences between rich and poor, and a score of 7 meaning that the government should not concern itself with reducing income differences. What score between 1 and 7 comes closest to the way you feel?" *For identity grouping A_8, there are two further terminal nodes split by years: 84,88,89,90,91,96,00 and 78,80,83.86,87,93,94,98. Number of observations = 13,024.

Variable	Score
RACE	100.00
AGE	82.93
SEX	69.23
MADEG	48.78
FUND16	22.19
REGION16	12.03
PARBORN	8.93
RELIG16	6.29
BORN	2.59
YEAR	1.25

This table shows variable importance scores for the dependent variable, EQWLTH, derived from a large set of trees using RandomForests™.

residual heterogeneity within such groupings strongly suggests that we need to be careful in avoiding strict, monolithic interpretations of our results.

3.1. Regarding redistribution preferences

We turn first to our results for redistribution preferences. The CART™ tree and RF results for EQWLTH have the following robust features (see Table 2). The RACE variable is the most important splitting variable, and it splits into Whites and non-Whites.[8] AGE, SEX, and mother's education (a proxy for socioeconomic background; MADEG) are also important splitting variables within Whites only. AGE splits the sample into young-to-middle aged adults and older adults. This split corresponds to lifecycle effects on income and wealth. Older adults, having accumulated wealth and higher incomes, may be expected to be less in favor of income redistribution than younger adults. The split by MADEG separates respondents with mothers who did not complete high school (MADEG = 0) from the rest of the population. Men and women are also classified distinctly.

Overall, non-Whites and young Whites with low maternal education (MADEG = 0) are classified as having strong preferences for redistribution (EQWLTH = 1). All other White men and older White women not from low socioeconomic backgrounds are classified as having preferences against redistribution (EQWLTH = 6 or 7). Older White women from low socioeconomic backgrounds and younger White women from higher socioeconomic backgrounds are classified as having neutral preferences (EQWLTH = 3 or 4). Non-Whites have a strong preference for

[8] Because the non-Black, non-White group consists of a small number of observations and are such a heterogeneous group, we focus on White–Black differences here and elsewhere in the paper.

Table 3
NATFARE classification tree and random forest results

Classification variable	A_1	A_2
RACE	White and others	Black
Predicted classification	2, 3*	1

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, NATFARE. NATFARE measures redistribution preferences as defined by the following GSS question, "We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount… Welfare… are we spending too much, too little, or about the right amount on welfare? (1 = too little, 2 = about right, 3 = too much)". *For identity grouping A_1, there are two further terminal nodes split by years: 83,84,86,87,88,89,90,91,98,00 and 78,80,93,94,96. Number of observations = 13,024.

Variable	Score
RACE	100.00
AGE	23.15
REG16	11.17
MADEG	5.32
RELIG16	4.96
SEX	4.58
FUND16	3.80
PARBORN	2.64
BORN	0.90
YEAR	0.00

This table shows variable importance scores for the dependent variable, NATFARE, derived from a large set of trees using RandomForests™.

governmental redistribution, while White men who are not young or who do not have a low-status socioeconomic background have a strong preference against governmental redistribution. White women are classified across a range of views depending on age and socioeconomic background.

Table 4
CONFED classification tree and random forest results

Classification variable	A_1	A_2	A_3	A_4	A_5
YEAR	1978, 1983	1980, 1993, 94, 96, 98 2000	1980, 1993, 94, 96, 98 2000	1984, 86, 87, 88, 89, 90, 91	1984, 86, 87, 88, 89, 90, 91
PARBORN	All	Neither parent born in US	At least one parent born in US	All	All
RACE	All	All	All	White and others	Black
Predicted classification	2	1	3	1	3

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, CONFED. CONFED measures confidence in government as defined by the following GSS question, "I am going to name some institutions in this country. As far as the people running these institutions are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them? Executive branch of the federal government (1 = A great deal, 2 = Only some, 3 = Hardly any)". Number of observations = 17,982.

Variable	Score
REGION16	100.00
AGE	84.05
SEX	71.84
FUND16	39.88
PARBORN	35.40
BORN	28.82
MADEG	27.43
RELIG16	25.56
RACE	21.93
YEAR	14.37

This table shows variable importance scores for the dependent variable, CONFED, derived from a large set of trees using RandomForests™.

Table 5
INCGAP classification tree and random forest results

Classification variable	A_1	A_2
YEAR	1987, 2000	1996
Predicted classification	3	5

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, INCGAP. INCGAP measures tolerance for inequality as defined by the following GSS question, "Do you agree or disagree. Differences in income in America are too large. (1=Strongly agree, 2=Agree, 3=Neither agree nor disagree, 4=Somewhat disagree, 5=Strongly disagree)". Number of observations=3502.

Variable	Score
YEAR	100.00
SEX	68.23
AGE	31.35
REGION16	26.91
RACE	13.34
MADEG	10.21
RELIG16	9.09
FUND16	7.62
PARBORN	3.31
BORN	2.61

This table shows variable importance scores for the dependent variable, INCGAP, derived from a large set of trees using RandomForests™.

By allowing for the possibility of interactions between identity variables, therefore, we obtain results that provide a more nuance picture of redistribution preferences than the findings by, for instance, Alesina and La Ferrara (2005) who argue that women, younger persons and Blacks are generally more supportive of redistribution policies, while more educated individuals are instead less favorable. While our findings certainly agree with their conclusions generally, they are also more specific in terms of identifying particular subgroups of individuals characterized by combinations of these variables that correspondent to very different degrees of tolerance for redistribution.

Table 6
WHYPOOR classification tree and random forest results

Classification variable	A_1	A_2
REGION16	2–7	0,1,8,9
Predicted classification	1	3

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, WHYPOOR. WHYPOOR measures tolerance for inequality as defined by the following GSS question, "Now I will a list of reasons some people give to explain why there are poor people in this country. Please tell me whether you feel each of these is very important, somewhat important, or not important in explaining why there are poor people in this country. Lack of effort by the poor themselves (1=Very important, 2=Somewhat important, 3=Not important)". Note that the independent variable, REGION16, corresponds to the GSS question, "In what state or foreign country were you living when you were 16 years old? (Coded by region) (1=New England, 2=Middle Atlantic, 3=East North Central, 4=West North Central, 5=South Atlantic, 6=East South Central, 7=West South Central, 8=Mountain, 9=Pacific, 0=Foreign)". Number of observations=1180.

Variable	Score
REGION16	100.00
RACE	46.31
RELIG16	43.15
MADEG	42.73
AGE	40.37
PARBORN	27.39
FUND16	19.19
SEX	12.89
BORN	1.81
YEAR	0.00

This table shows variable importance scores for the dependent variable, WHYPOOR, derived from a large set of trees using RandomForests™.

Table 7
GOODLIFE classification tree and random forest results

Classification variable	A_1	A_2
YEAR	87	94,96,00
Predicted classification	3	4

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, GOODLIFE. GOODLIFE measures (predicted) prospects for upward mobility as defined by the following GSS question, "The way things are in America, people like me and my family have a good chance of improving our standard of living — do you agree or disagree? (1=Strongly agree, 2=Agree, 3=Neither, 4=Disagree, 5=Strongly disagree)". Number of observations=3756.

Variable	Score
YEAR	100.00
REALINC	57.74
REG16	45.66
SEX	43.87
AGE	34.17
MADEG	24.78
FUND16	12.91
RACE	8.33
RELIG16	8.11
PARBORN	6.89
BORN	1.61

This table shows variable importance scores for the dependent variable, GOODLIFE, derived from a large set of trees using RandomForests™.

Table 8
PARSOL classification tree and random forest results

Classification variable	A_1	A_2	A_3	A_4	A_5	A_6
AGE	<48	>47 and <62	<27	>26 and <62	>26 and <62	>61
MADEG	Less than high school	Less than high school	High school to graduate	High school to graduate	High school to graduate	All
REGION16	All	All	All	2–6	0, 1, 7–9	All
Predicted classification	5	1	2	4	4, 5	1

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, PARSOL. PARSOL measures perceived income mobility as defined by the following GSS question, "Compared to your parents when they were the age you are now, do you think your own standard of living now is much better, somewhat better, about the same, somewhat worse, or much worse than theirs was? (1=Much better, 2=Somewhat better, 3=About the same, 4=Somewhat worse, 5=Much worse)". Note that the independent variable, REGION16, corresponds to the GSS question, "In what state or foreign country were you living when you were 16 years old? (Coded by region) (1=New England, 2=Middle Atlantic, 3=East North Central, 4=West North Central, 5=South Atlantic, 6=East South Central, 7=West South Central, 8=Mountain, 9=Pacific, 0=Foreign)". Number of observations=5939.

Variable	Score
AGE	100.00
MADEG	79.41
REGION16	16.66
YEAR	8.98
FUND16	8.60
SEX	8.16
RELIG16	6.10
PARBORN	3.55
BORN	2.66
RACE	2.61

This table shows variable importance scores for the dependent variable, PARSOL, derived from a large set of trees using RandomForests™.

Table 9
REALINC classification tree and random forest results

Classification variable	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
MADEG	Less than high school	Less than high school	Less than high school	Less than high school	High school to graduate	High school to graduate	High school to graduate	High school to graduate	High school to graduate
AGE	<33	>32, <65	>32, <65	>64	<31	>30, <66	>30, <37	>36, <66	>65
RACE	All	Black and others	White	All	All	Black	White and others	White and others	All
Predicted REALINC	$21K	$23K	$33K	7K	$27K	$28K	$38K	$45K	$28K

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, REALINC. REALINC measures family income on 1972–1996 surveys in constant dollars (base=1986). Number of observations=18,019.

Variable	Score
YEAR	100.00
AGE	62.42
REG16	51.33
MADEG	49.29
SEX	41.51
FUND16	40.51
RELIG16	32.26
RACE	29.70
PARBORN	10.08
BORN	8.24

This table shows variable importance scores for the dependent variable, REALINC, derived from a large set of trees using RandomForests™.

The robust groupings for NATFARE (see Table 3) correspond primarily to race, with a split between Blacks and others. Although this variable is the same as that for EQWLTH, there are more subtle groupings for EQWLTH that are not present for NATFARE. Therefore, though responses to the variables may be related, we conclude that responses to the EQWLTH question do not simply reflect views on welfare. We more thoroughly explore the joint views on redistribution preferences and welfare in Subsection 3.4 below.

Since both of these redistribution preference measures, EQWLTH and NATFARE, include explicit reference to a governmental role in redistribution, there are two possible interpretations for the variation in responses across identity groups. First, this variation could be attributed to differences in individuals' general confidence in government. Second, this variation could be due to individual differences in tolerance for inequality.

To investigate the possibility that the variation in responses to our redistribution preference measures, EQWLTH and NATFARE, across identity groups, could be due to variation in the general confidence in government, we consider the classification of responses to two questions that ask about the respondent's confidence in federal governmental institutions (CONFED and CONLEGIS) and compare them with those obtained for EQWLTH and NATFARE. In the interest of space, we show only the results for CONFED (see Table 4).

We find that the nature of the identity groups responsible for variations in responses to CONFED and CONLEGIS are not the same as those for EQWLTH or NATFARE. For instance, views on confidence in government as measured by CONFED are classified according to YEARS where in some years, differences in views appear to be attributed to differences in parental origin (PARBORN) whereas in other years, they appear to be further differentiated according to RACE. The picture for CONLEGIS is more complicated with views on confidence in government being differentiated according to YEARS and then to AGE, whether the respondent was born in the US (BORN), and RACE. Crucially, the classifying variables for CONFED and CONLEGIS are not the same across the CART™ and RF analyses. With this lack of robustness, the classifying variables do not appear to provide an informative prediction of opinions. Overall, there appears to be a relationship between confidence in government and views on welfare spending via classification by RACE, but the evidence is suggestive at best. We have to look elsewhere to understand the identity groupings that delineate redistribution preferences.

Table 10
EQWLTH classification tree and random forest results (with REALINC as classification variable)

Classification variable	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
RACE	2,3	1	1	1	1	1	1	1
REALINC	All	<34,003	<34,003	<34,003	<14,171	>14,170 and <34,003	>34,002 and <62,250	>62,249
AGE	All	<39	<39	<39	>38	>38	All	All
MADEG	All	Less than high school	High school to graduate	High school to graduate	All	All	All	All
SEX	All	All	Male	Female	All	All	All	All
Predicted Classification	1	1	6	3	1	7	6, 7*	6, 7**

Each column of this table corresponds to an identity grouping uncovered by CART™ for the dependent variable, EQWLTH. EQWLTH measures redistribution preferences as defined by the following GSS question, "Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Others think that the government should not concern itself with reducing this income difference between the rich and the poor. Here is a card with a scale from 1 to 7. Think of a score of 1 as meaning that the government ought to reduce the income differences between rich and poor, and a score of 7 meaning that the government should not concern itself with reducing income differences. What score between 1 and 7 comes closest to the way you feel?" This table differs from Table 2 in that REALINC was included as a covariate here but not in the exercise for Table 2. *For identity group A_7, there are two terminal nodes split by years: 1980, 84, 89, 90, 91, 96. **For identity group A_8, there are two terminal nodes split by age: less than 49 and over 48. Number of observations = 13,024.

Variable	Score
REALINC	100.00
AGE	81.99
RACE	59.79
MADEG	58.69
SEX	51.01
YEAR	49.97
FUND16	42.25
REG16	41.96
RELIG16	20.33
PARBORN	16.52

This table shows variable importance scores for the dependent variable, EQWLTH, derived from a large set of trees using RandomForests™.

We next ask whether variation in redistribution preference can be attributed to differences in tolerance for inequality using two questions that ask the respondent's view on the fairness of income differences (INCGAP and WHYPOOR4). The robust finding is that the splits for tolerance for inequality variables are different from EQWLTH and NATFARE.

Table 11
Public redistribution and welfare variable (EQWLTH–NATFARE) Summary

ALL years	NATFARE			Total
EQWLTH	Too little	About right	Too much	
Gov't reduce diff 1	[1] 6%	[8] 5%	[15] 7%	18%
2	[2] 3%	[9] 3%	[16] 5%	11%
3	[3] 3%	[10] 6%	[17] 7%	17%
4	[4] 3%	[11] 7%	[18] 10%	20%
5	[5] 2%	[12] 4%	[19] 7%	12%
6	[6] 1%	[13] 2%	[20] 5%	8%
No Gov't action 7	[7] 1%	[14] 3%	[21] 10%	14%
Total	19%	31%	50%	100%

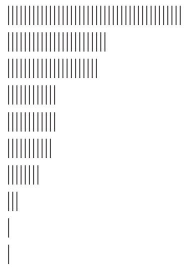
This table describes the composite dependent variable in Section 3.4 of the text. This "public redistribution and welfare" variable (EQWLTH–NATFARE) using joint responses to the GSS questions, EQWLTH and NATFARE. There are twenty-one possible pairs of responses to EQWLTH and NATFARE as seen above; each response is coded within the square brackets "[.]". The years of the data include GSS waves between 1978 and 2000. We find very little variation in the composition of responses across waves. The total sample size is 11,249.

Table 12
EQWLTH–NATFARE classification tree and random forest results

Classification variable	A_1	A_2	A_3	A_4	A_5
RACE	Black and others	White	White	White	White
YEAR	All	78,80,93,94,96	83,84,86,87,88,89,90,91,98,00	83,84,86,87,88,89,90,91,98,00	83,84,86,87,88,89,90,91,98,00
AGE	All	All	>45	<46	<46
REGION16	All	All	All	2,7	0,1,3,6,8,9
Predicated Classification	[1]	[19]	[14]	[6]	[3]

Each column of this table corresponds to an identity grouping uncovered by CART™ for the "public redistribution and welfare" dependent variable (EQWLTH–NATFARE) which was constructed using joint responses to the GSS questions, EQWLTH and NATFARE. Please see Table 11 for a detailed description of this variable. Note that REGION16 corresponds to the GSS question, "In what state or foreign country were you living when you were 16 years old? (Coded by region) (1 = New England, 2 = Middle Atlantic, 3 = East North Central, 4 = West North Central, 5 = South Atlantic, 6 = East South Central, 7 = West South Central, 8 = Mountain, 9 = Pacific, 0 = Foreign)". Number of observations = 13,024.

Variable	Score
RACE	100.00
FUND16	58.99
SEX	54.60
AGE	29.54
REGION16	29.28
MADEG	27.08
RELIG16	21.13
PARBORN	8.47
YEAR	3.78
BORN	3.56



This table shows variable importance scores for the "public redistribution and welfare" dependent variable, EQWLTH–NATFARE, derived from a large set of trees using RandomForests™.

For one dependent variable, INCGAP (see Table 5), there is a split by years. The split of 1996 from other two years may be reflective of welfare reform that was legislated that year. The other variable, WHYPOOR (see Table 6), is split by region in a way that is not readily interpretable. Crucially, the splits are not the same as each other, nor the same as EQWLTH or NATFARE.

In sum, there is some systematic heterogeneity in Americans' concerns about inequality and beliefs regarding the ability to escape poverty. But the key features of this heterogeneity do not correspond to the particular groupings uncovered for redistribution preferences.

3.2. Comparison with mobility beliefs

As described in Section 2 above, the Benabou and Ok model predicts matching identity groups for redistribution preferences and prospects of upward mobility. Therefore, we would expect the trees for EQWLTH and NATFARE to be similar to that of GOODLIFE. However, we do not find evidence to suggest that this is the case (see Table 7). Views on whether people and their family had a good chance of improving their standard of living essentially differ by years. Respondents appear to be more optimistic about their prospects for upward mobility in the mid-1980's, but become more pessimistic from the mid-1990's onwards. The RF results confirm the CART™ findings and also suggest that gender differences may play an additional though marginal role in differences in such views. It should be noted that the number of observations for GOODLIFE is small and therefore comparisons between the identity groupings for GOODLIFE and those for redistribution preferences (EQWLTH and NATFARE) should be interpreted with care. Nevertheless, there appears to be at least no clear-cut support for the Benabou and Ok hypothesis from the classification tree findings in the sense that the factors that segregate views on the prospect of upward mobility appear very different from those that explain redistribution preferences.

Similarly, an implication of Piketty's model and the hypothesis of endogenous interactions is that heterogeneity in mobility beliefs drives heterogeneity in redistribution preferences. That is, Piketty's theory can imply corresponding

identity groupings for the redistribution preference dependent variables and those that describe mobility beliefs, particularly views on hard work versus luck. Therefore, we examine whether identity classifications for redistribution preferences match those for mobility belief variables. The results suggest that this is not the case.

We first consider the classification of responses to two questions that ask only about the role of hard work in getting ahead (OPHRDWK and LFEHRDWK) and compare them to those for EQWLTH and NATFARE. The variables OPHRDWRK and LFEHRDWK produce no splits in the classification trees. The RF results suggest some importance of age, sex, and the region in which one was raised. Next, we consider the classification of responses to a question that asks about the relative importance of hard work for 'getting ahead' (GETAHEAD). The CART™ results for GETAHEAD suggest that views vary according to YEARS, AGE, and RACE, but these results have the same type of problem as CONFED and CONLEGIS described above; i.e., the CART™ and RF results for GETAHEAD do not agree with each other and are therefore not robust.

At this point, one might question the generality of the questions on hard work as proxies for mobility beliefs. Perhaps a respondent's mobility beliefs are influenced by evaluation of her past or future mobility. In that case, identity classifications for past mobility should inform identity groupings for redistribution preferences. We therefore consider the classification of responses to alternative proxies for mobility beliefs, and consider the classification of responses to two questions that provide an evaluation of the respondent's actual mobility, and compare these to those obtained for EQWLTH and NATFARE. This approach is based on a presumption that actual mobility informs mobility beliefs.

One variable we construct is the absolute value of the difference between the respondent's education level and that of his or her father (PADEG_ABS_DIFF). The second is a variable that measures the respondent's perceived standard of living now relative to his parents at the same age[9] (PARSOL). A third question provides an evaluation of expected dynastic mobility, and asks the respondent to compare his standard of living to that expected for his children at a similar age (KIDSSOL).

When asked to compare one's standard of living to one's parents' at the same age (PARSOL) the robust classifications are by AGE and MADEG (see Table 8). There is a split at middle age, similar to EQWLTH, but also at retirement age. The MADEG split is qualitatively the same as for EQWLTH. However, the classifications by MADEG do not run in the direction one would expect to explain the classification by MADEG for EQWLTH. That is, those from low-education backgrounds are more likely to consider themselves better off than their parents, but are also classified as more strongly in favor of income redistribution. There is no split by race.

The variable that measures comparison with one's children's standard of living (KIDSSOL) is classified differently from PARSOL and EQWLTH. There is a split by RACE into Whites and non-Whites, but only for some regions, which is hard to interpret. More importantly, the classifications by race do not run in the direction one would expect, from Piketty's framework, to explain the classification by RACE for EQWLTH.

Using the dependent variable measuring the difference between respondent's education and his father's (PADEG_ABS_DIFF), we find that AGE and MADEG are important splitting variables. Again, RACE is conspicuous in its absence.

In contrast to what one would expect from theory, we do not find a concurring set of identity groupings for mobility beliefs and redistribution preferences. Rather, whatever forces drive heterogeneity in mobility beliefs do not appear to be the same as those at work for redistribution preferences.

3.3. Comparison with socioeconomic status

As noted above, we would expect from information-based theories on the determination of redistribution preferences that heterogeneity of identity groupings uncovered for redistribution preferences be less complex than those for variables measuring socioeconomic status. To investigate this implication of the theory, we first consider classification of responses to a measure of the respondent's education level (DEGREE) and compare it with those obtained for redistribution preferences. The results indicate that the classification tree for DEGREE is highly complex,

[9] We do not include results for PADEG_DIFF which is the pure difference between the respondent's degree level and his father's. The results are similar to those for PADEG_ABS_DIFF. However, Fields and Ok (1999) provides an axiomatic justification for PADEG_ABS_DIFF that does not hold for PADEG_DIFF. Also, PADEG_DIFF will inevitably result in an un-interpretable distribution of responses since the education variables are by construction censored above and below. We also do not employ a question that asks about the respondent's job status relative to his or her father's. This question seems difficult to interpret in that perceptions of job status potentially vary over time and across individuals.

with 39 terminal nodes. The tree does not produce interpretable structure at that level of complexity. The important classification variables in this tree are MADEG and, secondarily, AGE. These variables are also the most important ones for explaining variation in responses to DEGREE according to the RF results. There is therefore more complexity present but it does not include RACE as an important classifying variable. That is, the salient classifiers of EQWLTH and NATFARE are not nested in the classifications for DEGREE.

We next compare the classification of responses for EQWLTH and NATFARE to a measure of the respondent's real family income (REALINC). We find that the regression tree for REALINC is not more complex than the analogous tree for EQWLTH (see Table 9). Similar splits are present, though here RACE is not the most important variable. Rather, MADEG and AGE are. All else equal, being younger, coming from a low-status socioeconomic background, or being Black is associated with a lower predicted household income. There is also a split by AGE around retirement that is not present in the EQWLTH tree described above.[10]

Given this similarity in classifying variables, a valid question is whether responses to EQWLTH, or preferences for income redistribution, are determined entirely by the respondent's income. If the classification tree for EQWLTH were to be run using the same set of identity markers plus REALINC, how does the classification tree change? We report the classification tree and random forest results for this exercise in Table 10. In a classification tree for EQWLTH that includes REALINC as a classifying variable, RACE remains an important classifier of redistribution preferences, independent of REALINC. In fact, the RF results show that RACE is as important as REALINC. Comparing this tree to that without REALINC, it appears that REALINC partly takes the place of MADEG and AGE as classifying variables. This displacement is not surprising in light of the REALINC tree results.

These results suggest that differences in respondent's income cannot fully explain differences in redistribution preferences. Crucially, variations in responses attributed to differences in race remain even when respondent's income is controlled for.

3.4. A further look at views on welfare and assisting the poor

To complete our analysis of redistribution preferences using GSS data, we present one more stylized fact. A larger percentage of Americans think the government should redistribute income from the rich to the poor than the percentage who think the government should not redistribute. However, it is also the case that a majority (or near-majority) of Americans think too much is spent on welfare (see, Table 11).

One explanation for this apparent "discrepancy" could be variations in the composition of joint views on redistribution preferences and welfare across social groups that make up the population. To investigate which demographic variables are useful predictors of such joint views, we construct a composite dependent variable using responses to EQWLTH and NATFARE. We refer to this joint EQWLTH–NATFARE variable as our "public redistribution and welfare" variable. There are twenty-one possible pairs of responses to EQWLTH and NATFARE. We characterize these respective 21 responses in Table 11. Our goal is to identify the most important predictive demographic characteristics of each response grouping for this variable and to contrast them with one another.

The classification trees and random forest results are presented in Table 12. We find that a key classification variable is race. Non-Whites are overwhelmingly in favor of a government role in redistribution, and are also more strongly in favor of increasing spending on welfare compared to Whites.

The joint views of Whites are more complex. Overall, views on welfare change temporally. Across all White, the least support for increased welfare spending is evident at the end of the Carter administration and in the years immediately preceding the 1996 reform. Outside of those periods, younger Whites are more pro-welfare than older Whites.

In unreported results, when we included family income (REALINC) as a potential classifier, we obtained the same findings as we did before for backs. However, for Whites, we found that those with the lowest socioeconomic status (MADEG=0) were more in favor of a government role for redistribution than other Whites (but not as much as Blacks), but were neutral towards increased welfare spending. Lowest socioeconomic status Whites were also more in favor of increasing assistance to the poor, toward the levels of Blacks. Whites of higher socioeconomic status

[10] Results obtained using real respondent's income, REALRINC, was also analyzed. The results do not tell us more than REALINC except that sex is a major component in REALRINC. This is expected since the income variable corresponds to a respondent's income rather than a household's. Also, Jewish men are classified as making significantly more than other men, which is interesting but peripheral.

(MADEG = 1 through 4) were neutral-to-opposed to a governmental role in redistribution and were opposed to increased welfare spending.

It seems, therefore, that race, and to an extent class, have important roles to play in explaining the divergence in views between the need to redistribute income to the poor and the need for increased welfare spending.

4. Conclusion

We provide a new set of stylized facts regarding salient heterogeneity patterns for preferences regarding government provision of income redistribution and related variables. We find that general views on redistribution are heterogeneous according to race as well as income determinants including socioeconomic background, age, and gender. Specific views on welfare are heterogeneous primarily according to race.

We cannot explain these patterns by variation in overall confidence in government, nor by differences across identity groups in their abstract tolerance for inequality. These results raise theoretical challenges. How can it be that we have no systematic correspondence between inequality tolerance or confidence in government and variation in preference for government-administered income redistribution? Why is race an important classifying variable for views on income redistribution independent of income?

Existing information-based theoretical models do not appear to completely explain our empirical results. The empirical patterns of systematic heterogeneity for mobility beliefs and abstract inequality tolerance are not consistent with patterns predicted by theory. We conclude that while these models provide important insight into the process of redistribution preference determination, they do not tell the whole story. This is a potentially important area for future research.

Our results also imply that the salient groupings relevant for preference-based theories of redistribution preferences go beyond ethnicity, except perhaps when talking about welfare policy specifically. In general, we find these groupings are more complex, and also reflect differences based on lifecycle considerations and class background. Perhaps surprisingly, religious background, both in terms of denomination and ideology, does not play a role in describing systematic heterogeneity in redistribution preference or household income. Religious background and its influence on individual income differences, as well as cross-country growth differences, has been the subject of many studies.[11]

In our view, the results of this paper constitute a puzzle to be resolved in future research. We see at least two avenues of theoretical ideas that are potentially useful toward such resolution. One is related to the ideas of Loury (1998). Redistribution preference classifications may reflect expected income classifications, as in Benabou and Ok (2001). Expected income groupings may differ from those of current income for the following reason. Expected income may be determined using information about others in one's identity group. Such information may be costly to gather. Thus, these identity groups may be determined using a few historically important variables such as race, class background, age, and gender. In addition, the determination of expected income may vary little with individual mobility beliefs. Individuals may reason that the combination of individual effort and institutional constraints that hold for others like one's self will, in expectation, hold for one's self. Expected income may largely be determined by information regarding institutional constraints that vary across identity groups, rather than views on the marginal effect of effort in determining outcomes that do not vary in the same way.

Redistribution preferences may also be determined based not only on current income but also on the ability to smooth consumption. There exists empirical evidence that Blacks face more volatile income, have less wealth, and are more credit-constrained than Whites. These differences may also provide an explanation for race's independent salience that is grounded in rational expectations.

A second idea is related to Roemer's (1999) analysis of the implications of people voting on a range of issues, only some of which are directly identity-relevant. Some issues are directly relevant to race, gender, or class. Examples are affirmative action and civil rights policies. Other issues are less directly relevant, such as those regarding income redistribution or public education funding. Because people vote on a range of issues at once, such as when voting for a candidate, views on policies that are not directly related to identity may be highly correlated with identity.[12] In this way,

[11] For examples of work on religion and its effect on income, see Sander (1992) and Tomes (1984). For an example of work on religion and its effect on growth, see Barro and McCleary (2003) and Durlauf et al. (2005).

[12] This hypothesis is also discussed in Lee and Roemer (2006) and empirical tests are offered.

redistribution preferences may vary significantly across identity groups, even if theoretically related variables do not vary similarly.

Technical Appendix

The main tools we use in the empirical analysis are classification and regression trees (CART™) and RandomForests™ (RF).

CART™ delivers a set of identity partitions by carrying out essentially two algorithms: (1) *recursive binary splitting* of the set of all observations, and (2) *cost complexity pruning* to address over-fitting. The recursive binary splitting algorithm starts with the set of all observations and then partitions it into two sub-samples – the *Left* and *Right* children nodes – by choosing an identity marker, j, and a corresponding split value, s, in the support of j so as to produce the largest decrease in diversity in the outcome responses within each sub-sample. This is achieved by minimizing the joint impurity across the two sub-samples; i.e., $\min_{j,s}(Q_L(j,s)+Q_R(j,s))$. The choice of identity marker, j, and split value, s, is obtained using exhaustive search.[13] This process is then repeated iteratively on each of the subsequent sub-samples, and so on, until the number of observations in each sub-sample is too small for further splitting to occur (or a preset minimum number of observations for nodes is reached). Different impurity measures $Q(.)$, such as the Gini and Twoing indices, are standard in the literature. It has been found that the Twoing index tends to give considerably better prediction performance than the Gini index when the dependent variable is a higher-level categorical variable (i.e., with 10 or more categories). We therefore emphasize results that employ the Twoing index as the impurity measure in Section 3, but note that we find no substantive differences using the Gini index (unreported results).

The result of the recursive binary splitting algorithm is a full set of partitions of the original sample, or "tree". In order to avoid over-fitting, this tree is then "pruned". The objective of the pruning algorithm is to locate the (nested) subset of partitions within the full set of partitions that minimizes a generalized information criterion where the complexity penalty parameter is chosen by V-fold cross-validation.[14] A key theorem in Breiman et al. is that a process known as "weakest link pruning" achieves the best tree given by the generalized information criterion. The final set of partitions (the "pruned" tree) is then reported by CART™. Therefore, the end result of CART™ is to deliver a set of homogeneous groupings of outcome responses and a pattern of identity partitions that characterizes these groupings, subject to not over-fitting the data.

We now briefly describe the RF algorithm and state key results. We refer the reader to Breiman (2001) for further details on random forests methods and implementation. RF generates a multiplicity of trees, and then pools information from these trees to obtain the best classification of responses in the following way. First, RF obtains L bootstrap samples (with replacement) from the data. Then, for each bootstrap sample, one-third is left aside ("out-of-bag") while two thirds are used to generate a tree (fully grown without pruning) using CART™. To generate each tree, RF randomly selects a subset of identity markers of fixed size $m<M$ from the set of all identity markers to be used as split variables. Therefore, as a result, an outcome response assignment is obtained for each observation in about one-third of the trees.

Each tree now "votes" for the final outcome assignment for each observation. That is, at the end of the L iterations, take j to be the outcome response that was most frequently assigned to observation n when it was "out-of-bag". This is then the RF predicted classification for that observation. In this way, each observation in the original sample is classified as corresponding to a particular outcome response depending on the modal classification accorded to it by the L trees. The "out-of-bag" misclassification estimate is then the proportion of times that j is not equal to the actual outcome response of observation n given by the data averaged over all observations. Breiman (2001) shows that this misclassification estimate is unbiased.

Finally, RF obtains a measure of variable importance for each identity marker by randomly permuting the values of each particular identity marker for the "out-of-bag" observations and then classifying these scrambled observations using the "in-bag" trees. RF defines the importance score for each identity marker as the average difference between the

[13] Loh and Shih (1997) point out that there may be variable selection bias towards identity markers which take on more values in CART™'s exhaustive search algorithm. To get around this problem, we impose a penalty on high categorical variables in CART™. We calibrate the penalty to ensure that categorical variables have no inherent advantage in being selected for splitting over a continuous variable with unique values for each observation.

[14] In our exercises, we set $V=10$.

number of votes for the correct (i.e., observed) outcome response in the permuted "out-of-bag" data from the number of votes for the correct outcome response in the untouched "out-of-bag" data across the L trees. The idea is simple and compelling. If it is possible to substitute incorrect values for an identity marker and still obtain accurate predictions for outcome response classifications, then that identity marker cannot have been very important for classifying outcome responses in the first place.

References

Akerlof, G.A., Kranton, R., 2000. Economics and identity. Quarterly Journal of Economics 115 (3), 715–753.

Alesina, A., La Ferrara, E., 2005. Preferences for redistribution in the land of opportunities. Journal of Public Economics 89 (5–6), 897–931.

Alesina, A., Baqir, R., Easterly, W., 1999. Public goods and ethnic divisions. Quarterly Journal of Economics 114 (4), 1243–1284.

Alesina, A., Glaeser, E., Sacerdote, B., 2001. Why doesn't the US have a European-style welfare system? Brookings Paper on Economics Activity 187–278.

Barro, Robert J., McCleary, Rachel M., 2003. Religion and economic growth. American Sociological Review 68, 760–781.

Benabou, R., 1996. Inequality and growth. NBER Macroeconomics Annual 11–73.

Benabou, R., Ok, E.A., 2001. Social mobility and the demand for income redistribution. Quarterly Journal of Economics 116 (2), 447–487.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J.H., Olsen, R.A., Stone, C.J., 1984. Classification and Regression Trees, Wadsworth, Belmont.

Brock, W.A., Durlauf, S.N., 2001. Growth empirics and reality. World Bank Economic Review 15 (2), 229–272.

Duffy, J., Engle-Warnick, J., 2006. Multiple Regimes in U.S. Monetary Policy? A Nonparametric Approach. Journal of Money, Credit, and Banking 38 (5), 1363–1377.

Durlauf, S.N., Johnson, P.A., 1995. Multiple regimes and cross country behavior. Journal of Applied Econometrics 10 (4), 365–384.

Durlauf, S.N., Kourtellos, A., Tan, C.M., 2005. How Robust Are the Linkages Between Religiosity and Economic Growth? Tufts University, Dept. of Economics Working Paper No. 2005–10.

Fields, G.S., Ok, E.A., 1999. The measurement of income mobility: an introduction to the literature. In: Selber, J. (Ed.), Handbook of Income Inequality Measurement. Kluwer Academic Publishers, pp. 557–598.

Fong, C., 2001. Social preferences, self-interest, and the demand for redistribution. Journal of Public Economics 82 (2).

Hansen, B.E., 1999. Threshold effects in non-dynamic panels: estimation, testing and inference. Journal of Econometrics 93, 345–368.

Hansen, B.E., 2000. Sample splitting and threshold estimation. Econometrica 68, 575–603.

Lamont, M., 2000. The Dignity of Working Men: Morality and the Boundaries of Race, Class, and Immigration. Harvard University Press.

Lee, W., Roemer, J.E., 2006. Racism and redistribution: a solution to the problem of american exceptionalism. Journal of Public Economics 90 (6–7), 1027–1052.

Loh, W.Y., Shih, Y.S., 1997. Split selection methods for classification trees. Statistica Sinica 7, 815–840.

Loury, G.C., 1998. Discrimination in the post-civil rights era: beyond market interactions. Journal of Economic Perspectives 12 (2), 117–126.

Luttmer, E.F., 2001. Group loyalty and the taste for redistribution. Journal of Political Economy 109 (3), 500–528.

Manski, C.F., 1993. Dynamic choice in social settings: learning from the experience of others. Journal of Econometrics 58 (1–2), 121–136.

Piketty, T., 1995. Social mobility and redistributive politics. Quarterly Journal of Economics 110 (3), 551–584.

Roemer, J.E., 1999. The democratic political economy of progressive income taxation. Econometrica 67 (1), 1–20.

Sander, W., 1992. Catholicism and the economics of fertility. Population Studies 46 (3), 477–489.

Tomes, N., 1984. The effects of religion and denomination on earnings and the returns to human capital. The Journal of Human Resource 19 (4), 472–488.