View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

NBER WORKING PAPER SERIES

CATEGORICAL COGNITION: A PSYCHOLOGICAL MODEL OF CATEGORIES AND IDENTIFICATION IN DECISION MAKING

Roland G. Fryer, Jr.
Matthew O. Jackson

Working Paper 9579
http://www.nber.org/papers/w9579

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2003

Given the nature of this work, it has been helpful to us to have comments and reactions from a diverse group of researchers to whom we are very grateful: Michael Alvarez, Josh Angrist, John Bargh, Gary Becker, John Cacioppo, Colin Camerer, Glenn Ellison, Susan Fiske, Dan Friedman, Drew Fudenberg, Bengt Holmstrom, Philippe Jehiel, Vijay Krishna, Steven Levitt, Glenn Loury, Robert Marshall, Barry Mazur, Scott Page, Tom Palfrey, Michael Piore, Andre Shleifer, Tomas Sjostrom, and Steve Tadelis. Fryer is at the Department of Economics, University of Chicago, and American Bar Foundation, 1126 East 59-th Street, Box 9, Chicago, Illinois 60637, USA, (roland@midway.uchicago.edu); and Jackson is at the Division of the Humanities and Social Sciences, 228-77, California Institute of Technology, Pasadena, California 91125, USA, (jacksonm@hss.caltech.edu). Fryer gratefully acknowledges financial support from the National Science Foundation under grant SES-0109196. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

©2003 by Roland G. Fryer, Jr. and Matthew O. Jackson. All rights reserved. Short sections of text not to exceed two paragraphs, may be quoted without explicit permission provided that full credit including ©notice, is given to the source.

Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making

Roland G. Fryer and Matthew O. Jackson

NBER Working Paper No. 9579
March 2003
JEL No. D81, J15, J71

ABSTRACT

There is a wealth of research in psychology demonstrating that agents process information with the aid of categories. In this paper we study this phenomenon in two parts. First, we build a model of how experiences are sorted into categories and how categorization affects decision making. Second, we analyze the personal biases that result from categorization, in economic contexts. We show that discrimination can result from such cognitive processes even when there is no malevolent taste to do so and workers' qualifications are fully observable. The model also provides a framework that is equipped to investigate the social psychological concept of identity, where identity is viewed as self-categorization.

Roland G. Fryer, Jr.
Department of Economics
1126 East 59th Street
University of Chicago
Chicago, IL 60637
and NBER
roland@uchicago.edu

Matthew O. Jackson
Division of the Humanities
and Social Sciences
228-77
California Institute of Technology
Pasadena, CA 91125
jacksonm@hss.caltech.edu

"People will be prejudiced so long as they continue to think." Billig (1985, p.81)

1 Introduction

People categorize others in order to effectively navigate their way through the world of murky social interactions and exchange. Gordon Allport memorably noted, "the human mind must think with the aid of categories. We cannot possibly avoid this process. Orderly living depends upon it." To this, most (if not all) psychologists would agree. Perhaps even more importantly, there is a long tradition in social psychology that treats stereotyping and prejudice as inevitable consequences of categorization (for example, see Allport (1954), Hamilton (1981), Tajfel (1969), or Fiske (1998) for a recent review). While ideas of categorical thinking and stereotyping have been at the forefront of social psychology for five decades (Macrae and Bodenhausen, 2002, Markman and Gentner, 2001), their potential has yet to be realized in economics.

This paper introduces a notion of categorization into economics and derives its implication for economic decision making. In particular, based on a wealth of research from psychology, we build a model of social cognition centered on the basic principle that humans process information with the aid of categories. We derive some basic results regarding what our model of categorization implies about biases in decision making, and then apply this model to understand some simple forms of discrimination in economic contexts, including labor and lending markets, and identity choice; though we envision even broader implications.

A short synopsis of our approach is as follows. We construct a model where a decision maker stores past experiences in a finite set of bins or "categories." The number of categories is limited, and so there is some grouping of experiences based on their likeness. The decision maker, then, forms prototypes for prediction based on some aggregate memory or statistic from each category. When encountering a new situation, the decision maker matches the current situation to the most analogous category, and then makes predictions based on the prototype from that category. We show that an "optimal" categorization (i.e., one that minimizes the total of within category variation) necessarily lumps less frequent types of experiences into categories that end up being more heterogeneous. This has implications regarding discrimination. An important insinuation being that interactions with minority groups (which for most decision makers are necessarily less frequent due to the minority nature of the group) will generally be sorted more coarsely into categories than interactions with larger groups. In a labor theoretic example, this means that minorities will not

be as finely sorted based on their investments in human capital, for instance. This in turn provides minorities with less of an incentive to invest in human capital, which then further reinforces the coarse sorting, and those minorities who have invested are still not viewed on equal footing with others who have made similar investments.

We think of our contribution in two parts. The first is developing a model of how experiences are sorted into categories and how categorization affects decision making. The results derived are based on understanding what an "optimal" categorization would imply about which experiences are grouped together. The second is developing the deeper implications of this in terms of personal biases that result from categorical cognition, which only become evident once we place the model in specific economic contexts.

In particular, our contribution may be described as follows: (1) we develop a formal model of social cognition based on categorization of experiences, (2) we show that such categorization results in particular biases based on the frequency with which an observer encounters similar situations, (3) we provide a theoretical basis for the experimental evidence in social psychology attesting that people tend to characterize others by race (Brewer, 1988; Bruner, 1957; Fiske and Neuberg, 1990); (4) we show that discrimination can exist in environments where there is no taste for discrimination and personal attributes (such as employment qualifications) are fully observable; (5) we show that this discrimination is unique to minority groups and not the consequence of some "coordination failure"; and (6) we discuss implications of the model regarding social identity, which we view as self-categorization.

As we mentioned before, we are certainly not the first to provide a model of categorization. In particular, there is a rich literature in psychology on categorization, and a number of models that have been developed for use in analyzing data to understand how categorization works (for example, Ashby and Maddox, 1993; Ashby and Waldron, 1999, McKinley and Nosofsky, 1995; Reed, 1972; Rosseel, 2002). However, our analysis is the first to provide a model for which one can prove results regarding the properties of optimal categorization; and in particular showing that it necessarily implies differential treatment of groups based on their size. This model is thus particularly well-suited to use in analyzing how categorization results in specific and predictable biases in decision making.

The closest study to ours in the economics literature is Mullainathan (2001)[1]. Mullainathan

[1]Some other related studies are Loury (2002), Alavarez and Brehm (2002), and Barberis and Shleifer (forthcoming). Related to some of the points that we make about the implications of our model of categorization with regards

studies the implications of a model of categorization for the estimation of probabilities. He demonstrates biases in estimating high and low probabilities, and runs of events, based on the lumpiness common to categorization models. However, Mullainathan's model is more in the tradition of the probabilistic psychological models referred to above, where a likelihood is estimated from linking to an established (or in some models, an estimated) set of categories. Here instead, we examine how the categories themselves are built in terms of optimally storing information, and how that leads to particular groupings of past experiences based on their frequency. As such, the results and applications have little in common with Mullainathan's (or the previous work in psychology).

Discrimination

Our initial curiosity in the workings of categorization was motivated in thinking about how people's preferences manifest themselves in discriminatory behavior. Rather than simply assume preferences for one's own type, the model we develop here provides a foundation in which such behavior might emerge and persist over time.

Relative to previous theories of discrimination, our contribution can be viewed in the following manner. There are two main theories of discrimination in the economics literature: one attributed to a "taste" for discrimination (e.g., Becker (1957)); and one based on an informational asymmetry between a principal (employer, creditor, etc.) and an agent (worker, borrower, etc.) (e.g., Arrow (1973)). In the taste framework, agents from a particular group (say whites) are averse to interacting with another group (say blacks) because they experience a psychic utility cost. While this may indeed be the case, there is no discussion of why such taste exists, much less persists.[2] Using a different approach, Arrow (1973) contends that labor market discrimination is a result of an informational asymmetry between employers and workers regarding capital investments made by the worker. Because of this asymmetry and employers' initial pessimistic beliefs about blacks, they set hiring standards higher for blacks. This, however, induces blacks to invest at a lower rate, which confirms the employers initial pessimism. Discrimination arises because of multiple

to discrimination is Loury (2002). He discusses the existence and persistence of racial inequality, viewing race as a socially constructed trait. The implications of our categorization model with regards to discrimination is quite complementary to Loury's work, as we show how race as a visible attribute can become salient through categorization. Alvarez and Brehm (2002) (and the references cited there) discuss how uncertainty or inexperience with other groups can influence opinions and voting behavior. Also see Barberis and Shleifer (forthcoming) for implications of categorization ("style investing") in financial markets.

[2] There is some discussion of this in the sociology literature (see Goffman (1963)) and the psychology literature (see Allport (1954) and the references in Fiske (1998)).

equilibria. Therefore, from a purely theoretical point of view, blacks are just as likely as whites to be disadvantaged (face a high employment standard and end up with a low investment rate) in equilibrium.[3] The statistical discrimination literature, therefore, relies on an unspecified equilibrium selection mechanism that somehow always ensures that the equilibrium is one that is bad for blacks. There are many papers in the literatures on discrimination that have followed the seminal contributions of Becker (1957) and Arrow (1973).[4] The model that we outline here provides a basis for a theory of discrimination that brings the past five decades of progress in social psychology into a formal decision and game theoretic environment without assuming some a priori taste for discrimination or any form of informational asymmetry[5].

Before moving to a full description of the categorization model and our results, we present a stark example that previews some of the ideas, intuitions, and subtleties in the general modeling.

2 An Example

Consider a population of employers and a population of workers. The population of workers consists of a fraction λ of "white" workers and $1-\lambda$ of "black" workers, where λ is greater than $\frac{1}{2}$. Thus, the black workers are the "minority" group. Workers come in two human capital (say education) levels: high and low. So, overall, workers come in four flavors: black-high, black-low, white-high, and white-low. For now, let us assume that black and white workers are both just as likely to be of high human capital levels as low. We can represent a worker's type by a vector in $\{0,1\}^2$, where $(0,0)$ represents black-low, $(0,1)$ represents black-high, $(1,0)$ represents white-low, and $(1,1)$ represents white-high.

Let us suppose that an employer has fewer categories available in her memory than there are types of people in the world, and start by examining the case where the employer has three categories available. Suppose also that the employer has interacted with workers in the past roughly in proportion to their presence in the population.

How might the employer sort the past types that s/he has interacted with into the categories?

[3] Moro and Norman (forthcoming) provide a notable exception. They show that group inequalities can arise even when the corresponding model with a single group has a unique equilibrium. Again, however, in their model whites are just as likely as blacks to be in the disadvantaged group, as the formal results do not depend on their population proportions.

[4] See Fryer (2002a) for a recent review of theoretical models of discrimination in the economics literature.

[5] Our model can also be used to justify the assumptions on signaling technologies in papers on statistical discrimination (Phelps, 1972) or screening discrimination (Cornell and Welch, 1996).

Let us suppose that this is done in a way so that the objects (experiences with types of past workers in this case) in the categories are as similar as possible. To be more explicit, let us assume that the objects are sorted to minimize the sum across categories of the total variation about the mean from each category. For instance, consider a case where $\lambda = .9$ and the employer has previously interacted with 100 workers in proportion to their presence in the population. So the employer has interacted with 5 workers of type (0,0); 5 of type (0,1); 45 of type (1,0) and 45 of type (1,1). Let us assign these to three categories. The most obvious way, and the unique way to minimize the sum across categories of the total variation about the mean from each category, is to put all of the type (1,1)'s in one category, all of the type (1,0)'s in another category, and all of (0,$\cdot$)'s in the third category. This means that the white workers end up perfectly sorted, but the black workers end up only sorted by race and not by their human capital level! To get an idea of why this is the optimal sorting, let us examine the total variation (within-categories) that it generates, and compare it to some other possible assignments to categories.

The variation in category 1 (all (1,1)'s) is 0, the variation in category 2 (all (1,0))'s is 0, and the variation in category 3 (containing $5 \times$ (0,0) and $5 \times$ (0,1)) is $10 \times \frac{1}{2}$, for a total variation of 5; where the distance between either type (0,0) or (0,1) and the category 3 average of $(0,\frac{1}{2})$ is $\frac{1}{2}$ and there are ten such objects stored in category 3. To see why this leads to the least variation, consider another assignment of objects to categories where the low human capital types were all assigned to one category and the high human capital types were sorted into two categories (by race). Here the variation in category 1 (all (1,1)'s) is 0, the variation in category 2 (all (0,1))'s is 0, and the variation in category 3 (containing $45 \times$ (1,0) and $5 \times$ (0,0)) is $45 \times .1$ and $5 \times .9$ for a total variation of 9 (noting that the average in that category is (.9,0)). In total, objects are further from their category means in the second assignment. This gives us an idea of how categorization can lead to a sorting where some group members are more coarsely sorted than others. Note, it is in particular *minority* group members that are more coarsely sorted, due to their lower frequency in the population.

This example is consistent with the experimental evidence in social psychology and neuroscience that agents tend to categorize others by race (Brewer, 1988; Bruner, 1957; Fiske and Neuberg, 1990). Once we couple this with the observation that prototypes are important in forming expectations, we can see how discrimination can result. Under the optimal categorization, the prototype for the third category is the average of that category of $(0, \frac{1}{2})$. This prototype works against the high human capital blacks, as the expectation from the prototype of their category is lower than

their type. It is due to the fact that the mind of the employer has stored them in a category that we can label "black" rather than "black-high". This can result in high human capital blacks not being hired for positions that require high human capital levels, and also in offers of wages that are below their productivity levels. This form of discrimination is not malevolent, nor is it derived from some primitive preference or taste for race. It is the result of a minority population being sorted more coarsely due to the categorical way in which experiences are stored. This also contrasts with statistical discrimination since it is not a multiple equilibrium phenomenon where it could equally as well be the majority that is discriminated against, but rather it results from an inherent bias against minority interactions in the process of categorization of human memory, even when qualifications are fully observable. The latter deserves considerable emphasis, as it provides a second vital distinction between a model of categorical discrimination and that of statistical discrimination.

Our example might seem to be ambiguous in terms of the outcomes for blacks, as the black-lows are benefiting from being stored as "black" rather than "black-low". However, let us now go one step further and endogenize the decision to acquire human capital. Given that blacks expect to be categorized as "black", they have less incentive to invest in high levels of human capital since such investments are under-appreciated by employers. Hence, this can lead to lower investment rates in human capital by minority group members. So, in the end we end up with more "black-low" types in the black population.[6]

Let us make a couple of remarks about the example. First, why are employers interested in keeping track of race at all? In this example, there is no need for them to do so as race is not tied to productivity, even statistically. Nevertheless, humans keep track of race in categorizing memories (there is substantial evidence in this regard[7]). The explanation is that categorization and memories are used for many tasks. Race is one of the most easily and directly identifiable traits, it is immutable, and in many situations will correlate with other attributes. An optimal categorization is used for many functions, thus, race can be a useful attribute to keep track of for a variety of interactions. Furthermore, to the extent that one sees endogenous choices to acquire human capital influenced by race, keeping track of race becomes a useful predictor of productivity,

[6] More generally, the effects of coarse sorting depend on the context. For instance, one might have a "Kennedy Coattail Effect" (thanks to Colin Camerer for this example) where being categorized as a "Kennedy" leads to certain perceptions about one's political capital.

[7] For instance, see Hart et. al., (2000) and Phelps et. al., (2000), though there is evidence that this may be context dependent (see Wheeler and Fiske (2002)).

echoing the ideas of statistical discrimination.

Second, why won't some employers benefit from categorizing differently? There are profits to be had if one employer is able to overcome their categorical bias while others do not.[8] Indeed, this is a possibility. The question is whether there are sufficiently many such employers to give incentives to minority group members to make efficient investments in education and human capital. This point is clearly developed by Becker (1957) in a model with tastes for race. Moreover, if there are frictions in the market, for instance any search costs in finding employment, even having such employers around might not be sufficient to induce efficient investment in human capital by minorities. At the least, to the extent that there are fewer such employers than those who do not overcome the bias, there is still a tilting in favor of majority group members in finding full recognition in the hiring process. This bias affects the choices regarding investment, which then reinforce the bias.

As this example has pointed out, categorization can lead to biases against minorities and to discrimination. It also has interesting predictions for how minorities might group themselves. To the extent that they can group themselves in ways that make interaction more likely with decision makers who have had relatively large numbers of interactions with minorities (for example more interactions with minorities than majorities), they can be made better off as this may change the categorization. We also see in this example that there are some subtleties to these claims. For instance, if the population of black-high was larger than the population of white-low, then it would make sense to sort differently, so that it was low types who are coarsely sorted rather than blacks. Also, many complications arise and things become a bit more clouded when we move to many dimensions of attributes and a large number of categories. Here with only two dimensions and two types on each dimension things are fairly unambiguous. With large numbers of categories and attributes, it can be that certain subgroups are treated differently, so that conclusions are not so clear cut. Nevertheless, we still can show some basic biases. We will see more of this when we return to a more detailed discussion of categorization of minority groups, after the presentation of the model of categorization.

[8] But note that non-discriminating employers may need to borrow from a discriminating banker who might view a diverse workforce as having lower human capital. Thus, profits from overcoming a categorical bias are not so obvious.

3 A Model of Categorization

Before presenting the details of the model, let us first discuss its overall edifice and some of the evidence which led us to structure it in this way.

In understanding how a human groups memories and stores them in the brain, psychologists have developed ideas of how categories are important. There is substantial experimental evidence that when faced with an object or person, a given individual's brain "automatically" activates a category that, according to some metric, best matches the given object (and at times context) in question.[9] There is also new evidence in social psychology and cognitive neuroscience that the brain pays particular attention to racial identity[10]. While the reasons behind the use of categories are not yet completely understood, there are theories based on the efficiency of storage and retrieval of information (much like the organizing of a file system on a computer) as well as speed in being able to react.[11] Effectively, this is a bounded rationality story in which there are both costs to storing details of every past interaction separately, and costs to and delays in activating stored information based on how finely it is stored. So, the first piece of the puzzle from our perspective is that a given decision maker will store information from their past experiences in a finite set of bins to be called "categories." We sidestep the interesting question of how the number of such categories might be selected, and for now simply take it to be some given number n.

The second question regards both how new information is stored in categories as well as how it is called up.[12] The activation of a category when faced with a new object is accomplished through a matching of the "attributes" of the object with the attributes associated with the category. In particular, an "attribute" is one of the observable characteristics of the object. Associated with each category is an idea of which attributes something in that category should have. For instance if we think of a category of "bird", then it would have "beak", "feathers", "wings", etc., as attributes associated with it. We call the list of attributes associated with a category a "prototype." The given

[9] For example, see Allport (1954), Bargh (1994, 1997, 1999) for views on the automaticity of categorical thinking, and Dovidio et. al. (1986) for some of the experimental evidence. Note that under automaticity subjects are often not even aware of the process, much less the biases that are inherent in it.

[10] For instance, see Hart et. al. (2000) and Phelps et. al (2000).

[11] Rosch (1978) is perhaps the most precise. She argues that humans are searching for "cognitive efficiency" by minimizing the variation in attributes within each category for a fixed set of categories.

[12] There is evidence that the storage of information and the categorization structure is quite different in young children during their "developmental stages" than when they are adults (see Hayne, 1996, and Quinn and Eimas, 1996). While understanding the development of categories is an important question, we will focus on the behavior of adult decision makers, whose categorical structure is largely in place.

object's attributes are then compared to the prototypes of different categories until a best match is found. The precise process by which such matching is made is not completely understood at present based on what we have seen in the psychology literature.[13] We assume that decision makers act in rough congruence with a sort of "cognitive efficiency," which we take as the minimization of a distance metric between the given object and the matching prototype. This matching process is what is often called "identification."[14]

The third piece of the puzzle is what is then to be done with the categorical information once it is activated. For instance, once a decision maker has activated a category for a new object, say "bird", what happens next? This is where the theory of "stereotypes" comes into play. The idea of a stereotype is that it is an association of a given category with a series of different possible behaviors or other characteristics[15]. The priming of the category leads to an activation of the stereotype, which is the basis for the prediction of future behavior. The formation of stereotypes is another place where the understanding in the psychology literature is still a bit nebulous.[16] Here we model the stereotype as built on past interactions with objects in a given category. This black-boxes the issue of whether this is entirely built on a person's own interactions or also through what they might have heard or vicariously experienced, as we can treat such vicarious information as being stored in the same way. We shall, however, refer to a representative type for a category as a "prototype" rather than a "stereotype," for reasons that we come back to discuss later.[17]

On top of this automatic process of identification with a category and calling of a prototype, there is evidence that people then go through a thinking process where the conscious mind reasons through the information it has at hand.[18] In situations where an individual has time to think (which are often more relevant in economic applications) the reaction may move beyond reacting to

[13] For example, see Sternberg and Ben-Zeev (2001), Chapter 3.

[14] See Rosch (1978).

[15] In the psychology literature these are also often referred to as attributes (Hamilton and Sherman, 1994; Hamilton, Sherman, and Ruvolo, 1990; Stangor and Ford, 1992; and Stangor and Lange, 1994). Here we separate readily identifiable attributes used in first activating a category, like "beak", "wings", etc., with those things such as characteristics or behaviors that we might try to predict, like, "is difficult to catch", "is frightened of cats", etc. This distinction is somewhat artificial, but will be very useful from our perspective.

[16] See Hilton and Von Hippell (1996). For instance, part will be based on personal experience, but part can also be based on public information. There is evidence that people can accurately describe the "stereotype" associated with a given group or category that they believe others to have, even if it is not exactly what comes up in their own minds.

[17] For a discussion of some of the standard uses of these terms, see Hilton and von Hippel (1996).

[18] See Bargh (1984), Devine (1989), and Leopore and Brown (1997).

the prototype. Here individuals more carefully review the past situations that they can recall based on the category that has been activated and bring in other considerations. This part of the puzzle is probably the most complicated and the least well understood from the psychological perspective. This is, therefore, the part of the model where our treatment will be the most ad hoc.

Based on these different pieces of the puzzle, the crux of this paper is to put together a formal model of a decision maker which is consistent with what psychologists know about how the human mind stores and retrieves information and uses it to form predictions about behavior that are relevant to economists.

The Basic Building Blocks

Categories

$C = \{C_1, \ldots, C_n\}$ is a finite set of categories. We take these as given and discuss endogenizing them in Section 6. These categories can be thought of as "file folders" in our decision maker's brain that will be useful for the storing of information.

Objects

O is the potentially infinite set of objects that are to be sorted or encountered. These will generally be the agents with whom our decision maker might interact, such as the workers they may hire or have hired if they are an employer. We should emphasize that an object is not simply a physical object, but is in effect a particular experience or view of an interaction. Thus, a number of different interactions with the same person under different circumstances would be viewed as different objects. Further, an object might also be a vicarious interaction, such as viewing a movie or a news report, rather than a direct personal interaction.

Attributes

There is a finite set of attributes. Let m be the number of attributes. Attributes are the easily identified traits that may be possessed by an object. These might be race, sex, hair color, nationality, education level, which schools they attended, their grades, age, where someone lives, the pitch of their voice, etc. Different attributes might be observable in different situations. If I meet someone in a cocktail party I might see some easily observed attributes, and not observe some such as their grades, work experience, etc. In contrast, if I am interviewing them for a job, I may observe their transcripts and resume, but may not know whether they are married or like to bike

ride. For simplicity in our modeling, we will assume that each object has the same set of observable attributes, but the model is very easily altered to allow for the more general case.[19]

Let $\theta : O \rightarrow [0,1]^m$ denote the function, written as $(\theta_1(o), \ldots, \theta_m(o))$, which describes the attributes that each object has. For instance $\theta_k(o) = 1$ means that object o has attribute k. More generally, $\theta_k(o) = .7$ would indicate that object o has some intensity (.7) of attribute k. If, for instance, the attribute is "blond", then this might be a measure of "how blond" the person's hair is. For some attributes it might be that $\theta_k(o) \in \{0,1\}$ (for instance gender), but for others the possession of an attribute might lie between 0 and 1. There are some attributes that come in many flavors, such as race or ethnicity. These can simply be coded by having a dimension for each race. Then if a person is coded as having a 1 in the attribute "Black", they would get a 0 in the attribute "Asian". This also allows for the coding of mixed races, etc.

Categorization

The basic building blocks above are simply descriptions of objects. Once an object is encountered, then it is stored in memory by assigning it to a category. For simplicity, we will assume that each object is assigned to just one category, although we realize that in some settings this is with some loss of generality. Let $f : O \rightarrow C$ denote the function that keeps track of the assignment of each object to a category, where $f(o) = C_i$ means that object o has been assigned to category C_i. This is how objects are stored in the decision maker's memory.

Prototypes

Given some set of objects that have been categorized, O, and a categorization f, the decision-maker will find it useful to capture the essence of a category through a *prototype*. This is essentially a representative object. Prototype theory in the social psychology literature (Posner and Keele, 1968, and Reed, 1972) was designed to show that people create a representation of a category's central tendency in the form of a prototype. A prototype, according to this view, is judged to be prototypical of a category "in proportion to the extent to which it has family resemblance to, or shows overlapping attributes with, other objects in the category" (e.g. robin shares the highest

[19] Simply extend the range of the θ function, defined below, to have a $\emptyset$ possibility on various dimensions that mean that the dimension is not observed. In terms of sorting, there are many different ways to treat unobserved dimensions - simply ignoring them works, as well as imputing some average value, or trying to estimate them based on past correlations with other dimensions.

number of features with other birds). More generally, a prototype for a category might also be developed in other ways, for instance through some statistic other than mode, such as min if the decision maker cares about worst case scenarios. A very natural prototype of some category is simply the average across attribute vectors of objects in the category.

For some group of objects O let the mean attribute vector be given by

$$\overline{\theta}(O) = \frac{\sum_{o \in O} \theta(o)}{\#\{O\}}. \tag{1}$$

Let us emphasize that $\overline{\theta}(O)$ is a vector: the average of the attribute vectors of all the objects in O. The mean of a category C_i under a categorization f is then simply

$$\overline{\theta}^f(C_i) = \overline{\theta}(\{o : f(o) = C_i\}). \tag{2}$$

For now, let us think of $\overline{\theta}^f(C_i)$ as being the prototype for category C_i, although this is not essential in what follows.

Measuring Variation

Let us begin with an initial set of objects that our decision maker has interacted with in the past, O. The decision maker has categorized these according to some f. In some situations it will be useful for us to think about an "optimal" method of categorization. There are many possible ways to do this, and we pick an obvious one. We define an optimal categorization as categorizing past objects in a way to minimize the total sum (across objects) of within-category variance. In order to do this, we need to be explicit about how variation is measured.

First, let d be some measure of the distance between two vectors of attributes. It can make a difference how one keeps track of the distance between two attribute vectors. In some situations, it will be easy, natural, and salient to use the "city-block" metric. That is, when comparing two vectors, one simply looks at how far apart they are on each dimension and then adds up across dimensions. Another natural measure of distance would be the Euclidean metric which measures the magnitude of the vector difference. In the mathematical psychology literature, however, it has been argued for some time that when the attributes of objects are obvious or separable, spatial or geometric models should be constructed using the city-block metric rather than a Euclidean metric (Arabie, 1991; Attneave, 1950; Householder and Landahl, 1945; Shephard, 1987; Torgerson, 1958). As will be clear, this choice will not have much impact on our results. Nevertheless, unless indicated otherwise, we will stick with the city-block metric.

Let the variation of a group of objects simply be the total sum of distances from the mean:

$$Var(O) = \sum_{o \in O} d\left(\theta(o), \overline{\theta}(O)\right), \tag{3}$$

The total sum of within category variance under a categorization f is then simply summing the variation across the categories of objects:

$$Var(f, O) = \sum_{C_i \in C} Var(\{o : f(o) \in C_i\}). \tag{4}$$

Prediction

Now let us suppose that the decision maker faces an object and must choose an action from a set of actions A. One can think of the object as a worker, and the decision is whether or not to hire the worker. Let us also suppose that the decision maker has experienced some of the different actions before with some past objects and has a categorization of past experiences in place. Define $U(a, \theta)$ as the utility that the decision maker obtains from using action a against an object with attributes θ.

If o is assigned to category $f(o)$, then the expected utility of taking action a when faced with object o is

$$EU(a, o) = U(a, \overline{\theta}(f(o))). \tag{5}$$

The decision maker calls upon past experiences as a guide for predicting future payoffs in a boundedly rational manner. The decision maker views an object only through the prototype of the category that the object is identified with. In situations where the object is completely new to the decision maker, and not yet categorized, the relevant f will be the category to which the object is newly assigned. We discuss the assignment of new objects to categories (the process of "identification") below.[20]

An Optimal Categorization

An *optimal categorization function* relative to O is a categorization f^* that minimizes $Var(f, O)$[21].

[20] One can find many alternative methods for making predictions for a given categorization. An alternative to what we propose is an an adaptation of case-based decision theory, developed by Gilboa and Schmeidler (2001), to the categorical model. To see this, let a function $s : O \to O$ keep track of how similar two objects are. An example of a similarity function might be 1 minus the distance between the attributes of the objects: $s(o, o') = 1 - d(\theta(o), \theta(o'))$. A prediction, then, for what utility one might expect from action a against object o can be made based on: $EU(a, o) = \sum_{o' \in f^{new}(o)} s(o, o')U(a, \theta(o'))$. See also Jehiel (2002) for another approach, based on analogies in the context of game-theoretic decisions.

[21] There may be multiple solutions to this problem, but there is always at least one for any finite set of objects.

Let us comment on why minimizing the variance is a sensible objective. The ultimate goal of the decision maker is to use their categorization to form the most accurate predictions for expected future interactions. A best guess at the distribution of future interactions is based on the frequency of past interactions. So what does this imply? When a decision maker sees a new object it is identified with the category to which it most closely matches the prototype (as described below) in order to form an expectation. The error in prediction is then related to the distance between the new object and the prototype. Thus, we would like to minimize the expected distance between new objects and the prototypes that they will be matched with. Since this expectation is formed using the past frequency of observation, this is equivalent to minimizing the average distance between past objects and the prototypes that they are closest to. This is exactly minimizing the average distance between objects and the prototypes of their categories – which is minimizing the sum of variances across categories[22].

One shortcoming of this approach is that it does not take into account that some interactions may be more important than others, and might have higher utility impacts than others; which might lead one to weight attributes differently. We come back to discuss this in Section 6. And again, to the extent that attributes like race are (possibly endogenously) correlated with attributes like human capital, they end up being useful and cheap attributes to pay attention to.

Identification: Assignment of New Objects to Categories

When encountering a new object o, given an existing categorization f and past set of objects O, it will also be useful for the decision maker to assign the new object to a category. Generally, a decision-maker will not re-categorize all objects every time he or she encounters a new object. While at a developmental phase, it is clear that children's categorizations are much more temporary and flexible, evidence suggests that adult categorizations are more stationary (e.g., Hayne, 1996, and Quinn and Eimas, 1996).

Let $f^{new}(o)$ denote the category to which this new object is assigned. There are many ways in which this might be done. For instance, it might be that we fix the categorization of previous objects and simply assign the new object to the category with the best fit; that is, the category for which the distance between the object and prototype of the category is minimized. Instead, one might use a slightly more sophisticated manner of assigning the new object to a category. It takes

[22] This is reminiscent of principal component analysis, factor analysis, and other sorts of statistical procedures for decomposing data. However, working with a finite set of categories whose number is exogenously given, results in a different mathematical structure.

into account the fact that the new object will affect the category to which it is assigned, as it may change the prototype of the category. One could minimize the total sum of variation, subject to holding the previous objects in their previous categories, and accounting for the possible change in prototypes due to the addition of the new object. If a sufficiently large number have already been categorized, then these two methods of defining f^{new} will be approximately the same. In either case, we are assuming that f is not necessarily re-optimized, but simply that o is added to a sensible category now. Hence, f^{new} just represents the concatenation of the old f with the assignment of the new object. Note also that we are not requiring the f to have been optimal to begin with. It might be that f is only periodically "re-optimized."

3.1 A Categorization Theorem

For simplicity, for the remainder of the paper let us suppose that attributes are such that an object either has an attribute or does not.[23] That is, for each $o \in O$ and each $k \in \{1, \ldots, m\}$, $\theta_k(o) \in \{0, 1\}$. We assume throughout that $2^m > n$, so that there are fewer categories than types. This rules out the degenerate case where each object gets its own category. When faced with a limited number of categories, a decision maker will be forced to assign some different types of objects into the same category. That is the primary source of impact of categorical cognition. The question of which types end up grouped together has important implications, as we have already seen in the labor market example in Section 2. In that example, we saw that it was the smallest groups of types that were categorized together. The intuition being that this has the least impact on the within category variation. We can now develop this idea more generally, and show how the categorization model operates. First, a few useful definitions.

Homogeneous and Heterogeneous Groups

Let us say that a group of objects is homogeneous if all objects have the same vector of attributes, and heterogeneous otherwise. That is, O is *homogeneous* if $o \in O$ and $o' \in O$ implies that $\theta(o) = \theta(o')$. O is *heterogeneous* if there exist $o \in O$ and $o' \in O$ such that $\theta(o) \neq \theta(o')$.

A Balanced Splitting of a Heterogeneous Group of Objects

A group of heterogeneous objects might be split up in various ways, in cases where it consists of a number of different types. Let us consider splitting a group up into two parts. A natural way

[23] This has some impact, since it may prohibit scaling of attributes. The reason that scaling might matter is that it would allow for differential importance assigned to different attributes when, for instance, measuring how far apart two types are.

to try to do this (and more on why this is "natural" below), is to try to divide it in such a way that comes as close to being slit into equal parts as possible. So, for instance, let us suppose that we are looking at a set of objects that we call "birds". We might think of splitting that group up into birds that are "red" and birds that are "not red". This might end up with a relatively small group of birds that are red and a much larger group that are not. We might also think of dividing it according to whether the birds can fly or not, or by size, or where they live, etc. Specifically, as we look across different attributes, we might find the attribute that splits the group most evenly. Let us define an attribute that splits a group of objects most evenly as being a *balanced splitting attribute*. [24] More formally, an attribute k is a *balanced splitting attribute* for a group of objects O if it is the attribute k' that minimizes

$$|\#\{o \in O|\theta_{k'}(o) = 1\} - \#\{o \in O|\theta_{k'}(o) = 0\}|.$$

A splitting of a group of heterogeneous objects O into groups O_A and O_B is said to be a *balanced splitting* if there is a balanced splitting attribute k such that

$$O_A = \{o \in O|\theta_k(o) = 1\} \quad \text{and} \quad O_B = \{o \in O|\theta_k(o) = 0\}.$$

Let us emphasize that this definition of balanced splitting is a crude one, as it does not account for how close the various types involved are in terms of the number of attributes on which they differ. This definition allows for a fairly easy statement of our results. With a more complex definition of balanced, one can improve the bounds in the theorems below.

Frequency of Interaction and Categorization

Optimal categorizations are sensitive to the number of attributes, the relative numbers of different types of objects, the number of categories, and other features of the environment. This makes the general results on a characterization of optimal categorization quite complex, and so we have relegated them to the appendix. Nevertheless, in the text, we provide some basic intuition behind how objects are categorized, by looking at problems that are the building blocks of an optimal categorization. We examine which groups of objects should be lumped together and which ones should be separated.

Consider two groups of objects O_1 and O_2 with corresponding cardinalities n_1 and n_2, and let h be the number of attributes on which they differ at all (which may even be 0). Consider a group of objects O_3, with an balanced splitting into O_A, O_B and corresponding cardinalities n_A and n_B.

[24] It is possible to have more than one such attribute.

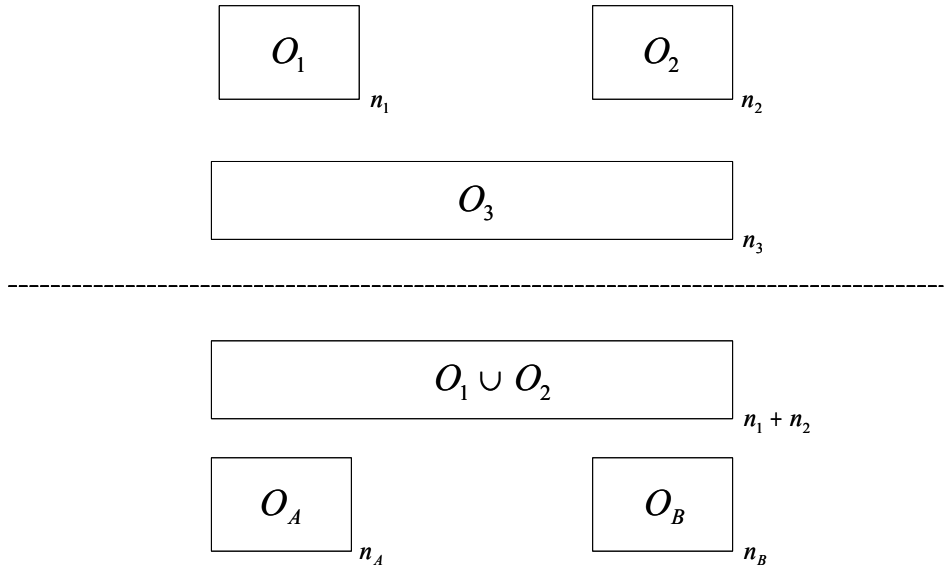

Figure 1:

Suppose that we have categorized things so that O_1, O_2 and O_3 each have their own categories. When would we do better by re-categorizing so that we split up O_3 and instead put O_1 and O_2 together? The answer is given in the following theorem.

Theorem 1 *Let f be a categorization that assigns objects O_1 to one category, O_2 to another category, and O_3 to a third category. If*

$$\frac{1}{n_1} + \frac{1}{n_2} > \frac{h}{n_A} + \frac{h}{n_B}, \tag{6}$$

then there exists another categorization which leads to lower total variation than f, where O_1 and O_2 are assigned to the same category and O_3 is a balanced split into O_A and O_B, which are assigned to two different categories.

The proof of Theorem 1 appears in the appendix as a corollary of the full characterization (Theorem 3).

Theorem 1 shows that rather than have O_1,O_2 and O_3 assigned to their own categories, it is better to split up O_3 and instead put O_1 and O_2 together provided:

- the sizes of O_1 and O_2 are relatively small (so, this gives a large $\frac{1}{n_1} + \frac{1}{n_2}$ in inequality (6)),
- O_1 and O_2 are fairly similar in their attributes (so h is small on the right hand side of inequality (6)), and

- O_A and O_B are relatively large and so it is optimal to assign them to their own categories (so $\frac{1}{n_A} + \frac{1}{n_B}$ is small in inequality (6)).

In terms of returning to the intuition of the labor market example, groups of objects that are minority objects are smaller in size, and so this theorem tells us that optimal categorizations will generally group them together before grouping types of majority objects together.

Let us examine our discrimination in labor markets example in more detail. Suppose we started with a categorization f where we assigned black-high and black-low to their own categories and assigned all whites to the same category. In the notation of the theorem, O_1 would be the black-high types with $n_1 = 5$; O_2 would be black-low with $n_2 = 5$; and O_3 would be all white types with $n_3 = 90$, and the balanced splitting of low types into O_A and O_B being according to the other attribute, human capital. So, O_A is white-high with $n_A = 45$ and O_B is white-low with $n_B = 45$. Here, clearly $\frac{1}{5} + \frac{1}{5} > \frac{h}{45} + \frac{h}{45}$, and so it is optimal to categorize the blacks into one category and separate the whites into two.

The condition in Theorem 1 is a sufficient condition, but not a necessary one. We do provide a full characterization, which appears in the appendix as Theorem 3. However, the necessary and sufficient conditions are complex, and simplify in the case of Theorem 1. In the case where the heterogeneous group consists of exactly two types, then the optimal categorization further simplifies in the following manner.

Theorem 2 *Consider four different homogenous groups of objects O_1, O_2, O_3 and O_4, with corresponding cardinalities n_1, n_2, n_3, and n_4. Let h_{12} be the number of attributes on which O_1 and O_2 differ, and h_{34} be the corresponding number for O_3 and O_4.*

$$Var(O_3 \cup O_4) > Var(O_1 \cup O_2)$$

if and only if

$$h_{34}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) > h_{12}\left(\frac{1}{n_3} + \frac{1}{n_4}\right).$$

The proof of Theorem 2 also appears in the appendix.

To paraphrase Theorem 2, suppose that we have three groups of objects and we want to know which two groups, when put together, will result in the smallest total variation. The two groups i and j which minimize $n_i n_j h_{ij}$, are the best ones to put together. Thus, as intuition suggests, groups are more likely to be (optimally) categorized together if they are smaller and more similar. In our discrimination example, $n_i n_j h_{ij}$ is clearly minimized when black-low and black-high are put

together, as then $n_i = n_j = 5$ and $h_{ij} = 1$. Instead, placing low types together would lead to $n_i = 45$ and $n_j = 5$ and a higher value of $n_i n_j h_{ij}$.

4 Categorization and Minority Groups

The previous section provided a detailed model of categorization in decision making. We now illustrate it by analyzing categorization in the presence of a "minority" group. This is a more general development of the example in Section 2, and illustrates in more detail some of the issues that arise and assumptions that are needed to conclude something about the categorization of minority groups more generally.

Consider a decision-maker facing a finite set of objects O. For simplicity, we shall also assume that every type of object has at least one representative in O. That is, every possible vector of 0's and 1's exists in the population. This is easily relaxed but leads to complications in the proofs.

Minority Groups

Let us now define what a "minority" group is. Consider a set of objects O and some attribute k with respect to which we are defining a group. That is, suppose that we are interested in the group of objects that have attribute $\theta_k(o) = 0$.[25] This might be race, or say left-handed individuals.

A group of objects having attribute $k = 0$ is a *minority group* of objects in O if for every $\theta_{-k} \in \{0,1\}^{m-1}$:

$$\#\{o \in O \mid \theta_k(o) = 0 \text{ and } \theta_{-k}(o) = \theta_{-k}\} < \#\{o \in O \mid \theta_k(o) = 1 \text{ and } \theta_{-k}(o) = \theta_{-k}\}.$$

The definition of minority group requires that whatever type of object having that attribute are in a smaller number in the population than objects with the same type except for not having that attribute. For instance, let us suppose that there are three possible attributes, so that the attributes of an object are represented by vectors (0,0,0), (1,0,0), (0,1,1), etc. Moreover, suppose that it is the first attribute we are interested in, so we want to check whether the population of objects of the form (0,·,·) is a minority population. The definition requires that there are fewer (0,0,0)'s than (1,0,0)'s; fewer (0,1,0)'s than (1,1,0)'s; fewer (0,1,1)'s than (1,1,1)'s; etc.

A *strict minority group* of objects in O is such that

$$\max_{\theta_{-k}} \#\{o \in O \mid \theta_k(o) = 0 \text{ and } \theta_{-k}(o) = \theta_{-k}\} < \min_{\theta_{-k}} \#\{o \in O \mid \theta_k(o) = 1 \text{ and } \theta_{-k}(o) = \theta_{-k}\}.$$

[25] The definitions have obvious analogs for a group of objects having attribute $\theta_k = 1$.

The definition of strict minority group is even stronger. It means that every type of object that falls in the minority group has a lower frequency in the population than any type of object that falls in the majority group. Going back to our example from above, it requires that there are fewer (0,0,0)'s than (1,0,0)'s, (1,1,0)'s, (1,0,1)'s, and (1,1,1)'s; and the same for (0,1,0) and so forth. This really requires the minority group to have fewer members of every type in a strong sense.

While the definition of strict minority group is demanding, keep in mind that this will be in reference to the set of objects that a given observer will have encountered. In many cases, this set may have strong selection biases, that result in seeing more objects with certain attributes than with others, as the observer will generally not be seeing a random selection of objects.

A Corollary on Categorizations of Strict Minority Groups

In order to establish a result analogous to that of the example in Section 2 we need some bounds on the relative frequencies of different types both within the minority group and across the minority and majority group. For a strict minority group, let the *external ratio* of the group be denoted

$$r_E = \frac{\text{size of smallest group of majority objects}}{\text{size of largest group of minority objects}};$$

in symbols,

$$r_E = \frac{\min_{\theta_{-k}} \#\{o \in O \mid \theta_k(o) = 1 \text{ and } \theta{-}k(o) = \theta{-}k\}}{\max_{\theta_{-k}} \#\{o \in O \mid \theta_k(o) = 0 \text{ and } \theta_{-k}(o) = \theta_{-k}\}}.$$

The external ratio keeps track of how large the smallest group of majority objects (in terms of type) is relative to the largest group of minority objects. This will always be a number greater than 1 for a strict minority, and gives a rough idea of the extent to which majority members outnumber minority members.

Let the *internal ratio* of the group be

$$r_I = \frac{\text{size of largest group of minority objects}}{\text{size of smallest group of minority objects}};$$

that is,

$$r_I = \frac{\max_{\theta_{-k}} \#\{o \in O \mid \theta_k(o) = 0 \text{ and } \theta_{-k}(o) = \theta_{-k}\}}{\min_{\theta_{-k}} \#\{o \in O \mid \theta_k(o) = 0 \text{ and } \theta_{-k}(o) = \theta_{-k}\}}.$$

This is a similar ratio except that it keeps track of how large the biggest group of minority objects is compared to the smallest group of minority objects. This might be thought of as a very rough measure of heterogeneity of the minority population. If it is close to 1, then the minority group is divided into equally sized subgroups of every possible type. If this ratio becomes larger then there are some types that are much more frequent and some that are less frequent in the minority

population. In our discrimination example, the external ratio $r_E = \frac{45}{5} = 9$, and the internal ratio $r_I = \frac{5}{5} = 1$.

Corollary 1 *Consider a set of objects that are optimally categorized. If a strict minority group defined relative to an attribute k has external and internal ratios that satisfy $r_E(2 - r_I) > 1$, and the number of categories n satisfies:*

$$\frac{\text{number of categories}}{\text{number of types}} > \frac{7}{8} - \frac{1}{\text{number of types}},$$

then minority objects are strictly more coarsely sorted than majority objects; and in particular

- *objects are segregated according to attribute k (objects from the minority group are never placed in a category with any majority objects); and*
- *majority types are perfectly sorted (any two objects from the majority group that are in the same category must have the same type).*

The proof of Corollary 1 appears in the appendix.

Corollary 1 provides sufficient conditions for a more coarse sorting to occur for a strict minority group. These produce an unambiguously coarser categorization for the minority group. The very strong conclusions of this theorem in terms of every majority type getting its own category necessarily requires strong conditions. Clearly we need more categories than the number of majority types, requiring that the overall ratio of categories to types exceed $\frac{1}{2}$. To get such a clean sorting we need even a much higher ratio, approaching $\frac{7}{8}$. While the proof is fairly complicated, some of the intuition is already conveyed in the example in Section 2 and the proof works through the challenges posed by the added dimensions of attributes, which are substantial. Rather than detail the proof, let us simply outline the role of the different conditions in the corollary and why they are useful.

First, the strictness of the minority group overcomes the problem that some less frequent types might be grouped together regardless of the minority/majority characteristic. Most importantly, depending on the relative frequency in the two populations, the grouping could take some forms that combine the types from the groups in different ways so that an unambiguous characterization is no longer possible. Next, the bounds on n play a role as follows. If n is at least 2^m, then each type has its own category so the categorization is degenerate. If n is too small, then it can be that various majority group types are grouped together as well as minority group types. For instance, it might

be that (1,1,1)'s are grouped with (1,1,0)'s, while under the minorities it is the (0,1,0)'s are grouped with (0,0,0)'s. The comparison of how they are grouped is no longer unambiguous. Interestingly, as n tends toward $\frac{7}{8}2^m$ (the lower bound as m becomes large), minorities are grouped in fewer and fewer categories, while the majority continues to be perfectly sorted. There is an interesting implication of this analysis. To the extent that the number of categories n correlates with some measure of "intelligence" (there is no direct evidence on this) we would expect agents with lower "intelligence" to be more likely to think of minorities as homogeneous. Finally, the internal and external ratios are also important in ensuring that the majority types each are assigned to their own category, for the same reason as mentioned above.

The condition that there be $\frac{7}{8}$ as many categories as types is one that might often be violated. The Corollary is merely meant to be suggestive and to give an idea of how it is that categorization can lead to a different treatment of minority members. This also shows the tractability of the model of categorization. More generally, one can expect the categorization to be more ambiguous and complex, as there are likely to be differences in the types one observes across different populations. For instance, there are more blacks than whites attending inner-city public schools in most U.S. cities. Nevertheless, the basic insight that smaller groups will tend to be more coarsely sorted in some rough sense carries through, as we can see through applications of Theorem 1.

5 Choice of Attributes, Self-Categorization, and Identity

Many attributes that an individual possesses are actually actively chosen, especially those that are easily observed such as clothing, hair style, tattoos, etc. Given that such attributes will be noticed by others, and will often play a role in categorization, these choices are important and can provide information and signals to others. In short, we can view the choice of attributes as a choice of identity, and it is clear that choosing one's identity is an important economic decision.

Identity has been the subject of a wealth of research in sociology (Goffman, 1963) and psychology (see Ellemers, et. al., 2002, for a review), and a couple of papers in economics. In particular, a recent paper by Akerlof and Kranton (2000) shows how individual preferences relating to identity can have important implications for a wide variety of decisions. Here we try to develop a more primitive understanding of identity than that which has appeared before. The model we have put forth in the previous sections allow us to do just that, viewing *identity as self-categorization*. In essence, categorization involves sorting attributes (objects) into categories in a cognitively optimal

manner, whereas, identity is about the endogenous choices of those attributes; realizing that other decision makers are categorizing. In what follows, we briefly describe how identity, as viewed from our model, can be coupled with two existing models in economics, though these applications are far from exhaustive.

Signaling and Impression Management

One obvious way in which the choice of identity might matter, is in signaling. This brings up a distinction between the questions of "Who am I?" and "Who do I want others to think I am?". This is a distinction between self-categorization and categorization by others; which are both related to identity.[26]

In terms of impression management or signaling to others, the choice of attributes might be viewed as a variation on Spence's (1974) famous signaling model. Agents choose some observable attributes, such as education, attitude and clothing, and have some attributes which are difficult to observe, such as productivity. To the extent that the cost of some observable attributes correlate with productivity, they can serve as a separating or sorting device in some situations. When separation occurs in Spence's setting, it is the high productivity types who invest in sufficient amounts of the observable attributes so that the low productivity types do not find it beneficial to mimic them, even given that the high productivity types will be recognized as such by their observable attributes and receive higher wages. Effectively, this choice of observable attributes is a choice of identity. In our model this would work not through any direct inference of employers, but simply through the fact that they will come to categorize those who have high cost investments in a category which has high average productivity, and those without the investments in a category that ends up having lower average productivity. Thus, Spence's signaling ideas provide a direct impetus for the choice of identity. Identity, then, is an equilibrium phenomenon that captures the fact that high productivity types are investing in seemingly irrelevant activities and these identities become endogenous characteristics of high productivity types[27].

Culture and Identity

Consider a stylized illustration of a community in Catalonia, Spain. Each agent decides whether to invest their time in learning Catalan (a romance language spoken mainly by the local population) or computer programming. Investments in computer programming are valued in the global

[26] See Goffman (1959). We are grateful to Glenn Loury for pointing us to Goffman's work.

[27] A similar argument is made by Fang (2001) in a statistical discrimination model with endogenous group choice. Our model of categorization can also be used to extend Fang's work.

labor market, whereas, Catalan is only valued in the small local community. Agents observe each other's investment portfolio, and can calculate the conditional probability of any agent being in the community in the future. Investments in Catalan yield a relatively high probability of being in the community in the future, since it is not valued elsewhere, and investments in computer programming yield a relatively low probability. Agents prefer to interact with those with whom they know they will have a lasting interaction. One can then envision that agents are more likely to want to interact or cooperate with others if they observe sufficient investment in Catalan skills (i.e. the likelihood of being in the community in the future is relatively high). Agents face a tension between being successful in the global labor market and cooperating with their local peers.[28]

Consider, a three attribute example where the unobserved attribute is one's willingness to cooperate in repeated play. Assume that the observed attributes allow the community to calculate the likelihood of any agent being around in the future. For example, good computer programming skills and low literacy in Catalan may imply that you are likely to leave Catalonia. Hence, the community will not cooperate with agents who invest too much in computer programming or too little in Catalan. The general point is that when one is deciding whether or not to invest in a particular identity (albeit "ghetto", "black bourgeoisie", "white yuppie", etc.) they realize that this investment (in language, clothing, etc.) will be used by their community to infer the potential payoff from repeated social interactions with them. This does not depend on any complicated calculations by the community, but simply categorization of experiences. Thus, coupling the models of categorization and cultural capital provides an explanation of why particular attributes are associated with particular communities.

Correlation in Attributes

When individuals choose an identity, "being from the streets," "being tough," or whatever, it is curious why they do not invest in just one attribute that signals their type. Instead, they seemingly invest in extreme behaviors. For instance, when choosing to be identified with "being tough," an individual may invest in tattoos, body piercings, clothing, language, attitude, hair style, and the like, instead of just picking one attribute.[29] Why? An answer comes directly out of our model, when there is sufficient heterogeneity in the population of observers.

As an example, suppose there are three decision makers, A, B, and C, who believe that being tough is associated with directly observable attributes (1,1,0); (0,1,1); and (1,0,1) respectively,

[28] See Fryer (2002b) for a more elaborate discussion.

[29] This is casual empiricism. We have no direct evidence of this.

based on their past individual experiences. By associated we mean that they have some category with a prototype of such a vector, that also is a category where they have seen past "tough" behavior. Given this variation, if an individual were to choose an identity of (1,1,0), then s/he would be recognized as "tough" by decision maker A. However, this vector differs in *two* attributes from the prototypes of each of B and C. So, it is quite possible that this may not lead to a "tough" categorization by decision makers B and C. If instead, the person chooses an identity of (1,1,1), while not matching the "tough" prototype of any single decision maker, the person is within *one* attribute of the prototype of each. Thus, we might see attributes becoming linked. Note also, that this reinforces itself. As more hopeful "toughs" choose (1,1,1), more of these types will appear and the categorizations will be further skewed.

These examples provide a flavor for the types of applications that are likely to be influenced by our model of social categorization, but one may think beyond these to include such things as conformity, gang behavior, and voting.

6 Some Further Remarks

Endogeneity of Categories

We have treated the set of categories as a given. We know from the developmental psychology literature that this is not true of children (see Hayne, 1996, and Quinn and Eimas, 1996). More generally, there may still be some flexibility in categorization even as adults. Effectively there is a trade-off between the benefits that a new category brings in terms of a finer sorting of experiences, and the cost that a new category entails in terms of identifying new objects with categories and searching across a larger number of categories when making predictions.

We simply observe that there will be some interesting non-monotonicities that pose significant challenges for such work, and then leave an analysis of the endogeneity of categories for further research. To see such a non-monotonicity, let us revisit our leading example once again. Under the sorting with three categories things are imperfectly sorted in terms of human capital which leads to inefficient hiring and discrimination. If instead we actually *decrease* the number of categories to two, the unique optimal categorization is then by human capital level. That is, with only two categories the optimal sorting is to have all high human capital types in one category and all low human capital types in the other. This leads to no discrimination and efficient hiring.

Salience and Importance of Attributes

In our model all attributes are on an equal footing. It is clear that some attributes are more easily identified, that some attributes are more relevant for decision making, that the importance of an attribute can be context-dependent, and even that the perception of attributes might be biased by an existing categorization (see Rabin and Shrag (1999)). One way to handle relative differences in importance without altering our model is simply to code important attributes a number of times. So, for instance, in our leading example, if we code a vector of attributes as (race, human capital, human capital, human capital), then the type of a black-high becomes (0,1,1,1). Here race becomes relatively less important in the optimal categorization.

What this might miss is the context-dependence of attributes. For instance, Fiske (1993) has shown people tend to more finely categorize individuals who are above them in a hierarchy and more coarsely categorize individuals who are below them in a hierarchy. To the extent that this actually proxies for relative numbers of interactions, it is already captured in the model. However, to the extent that it reflects some relative importance of interactions, it is not directly accounted for in our model. A way to adapt the model to deal with this is similar to handling the relative importance of attributes, as we discussed above. Relatively more important objects can be treated as multiple objects, and more important objects receive larger weights.

Stereotypes

In our model, we have been careful to use the term "prototype" for the representative of a category, rather than the term "stereotype".[30] There is evidence that individuals can quite accurately identify a "stereotype" for a given vector of attributes that will be common to others or possibly even to a cultural history, even without having that as their own belief.[31] While this is a bit beyond our model - a stereotype might be thought of as knowing something about other people's categorizations and prototypes. While this makes it possible to view stereotypes as prototypes

[30] As with any term that has been used as much as stereotype or prototype, there are many working definitions. We realize that the word "prototype" also has working definitions that differ from what we have defined here. For instance, the term is sometimes used to identify certain objects as "prototypes" if they seem to fit into a category more naturally than other objects.

[31] See Hilton and von Hippel (1996) for an overview of some of the literature on stereotyping. Generally, prototype models are thought of as a particular type of stereotyping, while we are arguing that stereotypes might best be viewed as a different object than a prototype.

coming through some indirect or vicarious experiences, it seems to put them on a different (meta-) level from prototypes and this explains our distinction in the use of the term.

Testing the Model

While we have paid close attention to the evidence in the psychology literature in constructing our model, it still puts enough pieces together that direct tests of the model would be of interest. In particular, whether less frequent types of objects are more coarsely sorted, is something that could be directly tested to see whether such types of objects have more biased predictions associated with them. In particular, an important implication of our model is not simply that less frequent types of objects will have less accurate predictions associated with them, but that they will have biased predictions associated with them. This suggests a direct test of the discrimination implications of the model. By a comparison of human capital investment decisions by blacks in different locations where their percentage of the population varies from minority to majority, holding all else equal. Of course, there are self-selection issues and endogeneity of population to location that present significant challenges to such an approach.

More indirect testing is also possible. To the extent that there is simply a taste for discrimination, one might see similar levels of discrimination in, for instance, whether or not one goes to a restaurant that employs black workers and whether or not one chooses a black doctor from a medical plan. Our model, would predict that to the extent that one has fewer experiences with black doctors relative to black fast-food workers, the behavior might be very different. Also, our model can be distinguished from statistical discrimination by examining to what extent observable skill levels matter. In our model, discriminating behavior can exist even when skills are observable, while a statistical discrimination model would not allow for such discrimination.

Categorization and Social Policy

Social cognition and categorization are inextricably linked. Because of this, prejudice and discrimination may be inevitable consequences of our cognitive processes. The resulting implications for policy makers and academicians interested in, for instance, racial inequality can be complicated.

Given our model, it seems that a critical goal ought to be integrating students in a potpourri of races and ethnic groups early in life while their categorization structure is still flexible. Given the lack of housing integration among the races (Massey and Denton, 1993), kids are significantly more likely to only interact with others of their same race. In fact, Fryer and Levitt (2002) report that 35% of white students attend a kindergarten where there are no black students. Having

sufficiently many minorities in schools with other non-minorities might go a long way in changing their categorization structures.

The categorization model also provides another pointed prediction. Minority group members will benefit from congregating together. This is consistent with interesting patterns of segregation by race and income, as documented for instance by Jargowsky and Bane (1991) and Massey and Denton (1993). To understand this, note that if minorities live in a location with a relatively large minority population or apply to schools, firms, etc., which are more frequented by other minorities, then they are more likely to interact with people with sufficiently many experiences with minorities so that minorities are more finely sorted in memory.[32]

Another area in which the effects of categorical cognition could be felt is in the design and implementation of equal opportunity laws. As it stands, Title VII's disparate treatment model of discrimination is premised on the notion that intergroup bias is malevolent in origin. Our model, however, shows how discrimination can arise even when agents have no a priori motivation to do so. Regulating cognitive processes, on the other hand, is an impossible assignment. Krieger (1995) proposes several solutions and extensions to the current Title VII legislation to account for this. Most fundamentally, she argues that "courts should reformulate disparate treatment doctrine to reflect the reality that disparate treatment discrimination can result from things other than discriminatory intent." To establish liability for disparate treatment discrimination, a Title VII plaintiff would simply be required to prove that his group membership played a role in causing the employer's action or decision. While these ideas are promising, they have yet to be investigated in a formal model.

A New View of Role Models

Allen (1995) reports three different types of influences of role models: (1) moral - effects on preferences, perhaps through conformity effects; (2) information - provision of information on the present value of current decisions; and (3) mentors - resources through which human capital can be augmented. Most research, in economics, is aligned with the informational repercussions of role models.[33] In those analyses the role model provides information that similar types have the ability to succeed at a given task. In particular, it is future emulators who are learning from the role model. While that may be an important aspect of a role model, our analysis also provides

[32] As the president of a major state university indicated, "the best way to teach students that not all blacks think alike, is to admit more black students so the other students can see that not all blacks think alike."

[33] See Chung (2000) or Jackson and Kalai (1997).

another new view of a role model: teaching the decision makers (e.g., employers) and not just the potential emulators. In essence, a black Supreme Court Justice not only shows black children that blacks can obtain the highest ranking judicial appointment, but just as importantly it shows this to majority group members as well. Furthermore, because optimal categorization depends on the frequency of interaction (which comes with visibility and repeated instances), our model makes it easy to understand why Tiger Woods or the Williams' sisters (as role models) have larger impacts on minority participation in particular sports than Ken Chenault (CEO of American Express) or Stanley O'neal (COO of Merrill Lynch) has on minority business majors in college.

Beyond Cognition

In closing, let us point out that there are also other potential applications beyond cognition. In particular, the categorization of objects into different groups arises in a variety of areas ranging from computer science to marketing. Understanding optimal categorizations and how minority objects are grouped, potentially has implications in such applications as well. To give an illustration, let us discuss one particular example[34]. Consider a firm that wishes to advertise its product. There are many people it may wish to reach and it may find it useful to categorize them. For instance, if it is planning to advertise on television, through print, or on the radio, then an important attribute to keep track of is what newspapers they read, what shows they watch, etc. People may end up categorized based on this attribute, and possibly also on other observables. Based on prototypes from the different categories, an advertiser might then adjust its message to best communicate or sell its product, to the extent that it can target its message to specific categories. While this description is very superficial, it is still clear that the potential for the model extends beyond cognition to a variety of situations where categories are useful either from a computational or search perspective, or do to the fact that they define some natural communication classes.

References

[1] Akerlof, G. and Kranton, R. 2000. "Economics and Identity." *Quarterly Journal of Economics*, 715-753.

[2] Allen, A. 1995. "The Role Model Argument and Faculty Diversity," in S.M. Cahn, ed., *The Affirmative Action Debate*. New York: Routledge, 121-134.

[34] Thank you to Josh Angrist and Tom Palfrey for, independently, pointing this out.

[3] Allport, G.W. 1954. *The Nature of Prejudice*. Reading MA: Addison Wesley.

[4] Alvarez, M.R. and Brehm, J. 2002. *Hard Choices, Easy Answers*. Princeton University Press: Princeton.

[5] Arabie, P. 1991. "Was Euclid an Unnecessarily Sophisticated Psychologist?" *Psychometrica*, 56, 567-587.

[6] Arrow, K.J. 1973. "The Theory of Discrimination" in Discrimination in Labor Markets, Orley Ashenfelter and Albert Rees, eds. Princeton University Press, 3-33.

[7] Ashby, F.G. & Maddox, W.T. 1993. "Relations Between Prototype, Exemplar, and Decision Bound Models of Categorization. *Journal of Mathematical Psychology*, 37, 372-400."

[8] Ashby, F.G. & Waldron, E. M. 1999. "On the Nature of Implicit Categorization." *Psychonomic Bulletin and Review*, 6, 363-378.

[9] Attneave, F. 1950. "Dimensions of Similarity." *American Journal of Psychology*, 3, 515-556.

[10] Barberis, N., and Shleifer, A. forthcoming. "Style Investing." *Journal of Financial Economics*.

[11] Bargh, J.A. 1999. "The Cognitive Monster: The Case Against the Controllability of Automatic Stereotype Effects" in *Dual Process Theories in Social Psychology*. New York: Guilford

[12] Bargh, J.A. 1997. "The Automaticity of Everyday Life." in R.S. Wyer, Jr. (Ed.), *The Automaticity of Everyday Life: Advances in Social Cognition*. Vol 10, 1-61. Mahwah, NJ: Erlbaum.

[13] Bargh, J.A. 1994. "The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition." in T.K. Srull and R.S. Wyer, Jr. (Eds.), *Handbook of Social Cognition*. Vol 1, 1-40. Hillsdale, NJ: Erlbaum.

[14] Bargh, J.A. 1984. Automatic and Conscious Processing of Social Information, in T.K. Srull and R.S. Wyer, Jr. (Eds.), *Handbook of Social Cognition*. Vol 3, 1-43. Hillsdale, NJ: Erlbaum.

[15] Becker, Gary. (1957) *The Economics of Discrimination*. Chicago: University of Chicago Press.

[16] Billig, M. 1985. "Prejudice, Categorization, and Particularization: From a Perceptual to a Rhetorical Approach." *European Journal of Social Psychology*, 15, 79-103.

[17] Brewer, M.B. 1988. "A Dual Process Model of Impression Formation." in T.K. Srull and R.S. Wyer, Jr. (Eds.), *Advances in Social Cognition*. Vol 1, 1-36. Hillsdale, NJ: Erlbaum.

[18] Bruner, J.S. 1957. "On Perceptual Readiness." *Psychological Review*, 64, 123-152.

[19] Chung, K. 2000. "Role Models and Arguments for Affirmative Action." *American Economic Review*, 90 (3), 640-648.

[20] Cornell, B., and Welch, I. 1996. "Culture, Information, and Screening Discrimination." *Journal of Political Economy,* Vol. 104 (3), 542-571.

[21] Devin, P.G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Responses." *Journal of Personality and Social Psychology*. 56:5-18.

[22] Dovidio, J.F., Evans, N., and Tyler, R.B. 1986. "Racial Stereotypes: The Contents of Their Cognitive Representations." *Journal of Experimental Social Psychology*, 22, 22-37.

[23] Ellemers, N., Spears, R. and Doosje, B. 2002. "Self and Social Identity," *Annual Review of Psychology*, 53, 161-186.

[24] Fang, H. 2001. "Social Culture and Economic Performance." *American Economic Review*, 91(4), 924-937.

[25] Fiske, S.T. 1993. "Controlling Other People: the Impact of Power on Stereotyping," *American Psychologist*, 48, 621-638.

[26] Fiske, S.T. 1998. "Stereotyping, Prejudice, and Discrimination" Chapter 25 in *Handbook of Social Psychology*, 2, eds: Gilbert, D.T., Fiske, S.T. and Lindzey, G., Oxford University Press: Oxford.

[27] Fiske, S.T., and Neuberg, S.L. 1990. "A Continuum of Impression Formation, from category based to individuating processes: Influences of Information and Motivation no Attention and Interpretation." *Advances in Experimental Social Psychology*, 23, 1-74.

[28] Fryer, R. 2002a. *Economists' Models of Discrimination*. monograph. The University of Chicago.

[29] Fryer, R. 2002b. "An Economic Approach to Cultural Capital." unpublished manuscript. The University of Chicago and American Bar Foundation.

[30] Fryer, R., and Levitt, S. 2002. "Understanding the Black-White Test Score Gap in the First Two Years of School." *NBER Working Paper #8975.*

[31] Gilboa, I., and Schmeidler, D. 2001. *A Theory of Case-Based Decisions.* Cambridge University Press.

[32] Goffman, E. 1963. *Stigma: Notes on the Management of Spoiled Identity.* NY: Simon and Schuster.

[33] Goffman, E. 1959. *The Presentation of Self in Everyday Life.* Garden City: Doubleday.

[34] Hamilton, D.L. (Ed.). 1981. *Cognitive Processes in Stereotyping and Inter-group Behavior.* Hillsdale, NJ: Erlbaum.

[35] Hamilton, D.L., and Sherman, J.W. 1994. "Stereotypes", in T.K. Srull and R.S. Wyer, Jr. (Eds.), *Handbook of Social Cognition.* Vol 2, 1-68. Hillsdale, NJ: Erlbaum.

[36] Hamilton, D.L., Sherman, S.J., and Ruvolo, C.M. 1990. "Stereotype-Based Expectancies: Effects on Information Processing and Social Behavior." *Journal of Social Issues*, 46, 35-60.

[37] Hart, A., Whalen, P., Shin, L., McInerney S., Fischer, H., and Rausch, S. 2000. "Differential Response in the Human Amygdala to Racial Outgroup vs Ingroup Face Stimuli." *Neuroreport*, 11, 2351-2355.

[38] Hayne, H. 1996. "Categorization in Infancy," in Rovee-Collier, C., and Lipsitt, L (Eds.), Advances in Infancy Research, 10.

[39] Hilton, J.L., and von Hippel, W. 1996. "Stereotypes." *Annual Review of Psychology*, 47, 237-271.

[40] Householder, A.S., and Landahl, H.D. 1945. *Mathematical Biophysics of the Central Nervous System.* Bloomington, IN: Principia Press.

[41] Jackson, M.O., and E. Kalai, 1997. "Social Learning in Recurring Games." *Games and Economic Behavior*, 21, 102-134.

[42] Jargowsky, Paul, and Bane, Mary Jo., 1991 "Ghetto Poverty in the United States, 1970-1980." in *The Urban Underclass*, Christopher Jencks, and Paul Peterson eds. The Brookings Institution. Washington D.C.

[43] Jehiel, P. 2002. "Analogy-Based Expectation Equilibrium," mimeo: CERAS.

[44] Krieger, L.H. 1995. "The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity." *Stanford Law Review*, 47:116, 1161-1248.

[45] Lepore, L., and Brown, R. 1997. "Category and Stereotype Activation: Is Prejudice Inevitable?" *Journal of Personality and Social Psychology.* 72: 275-87

[46] Loury, G.C. 2002. *The Anatomy of Racial Inequality*, Harvard University Press: Cambridge MA.

[47] Macrae, N., and Bodenhausen, G. 2000. "Social Cognition: Thinking Categorically About Others." *Annual Review of Psychology.* 51: 93-120.

[48] Markman, A.B. and Gentner, D. 2001. "Thinking." *Annual Review of Psychology*, 52, 223-247.

[49] Massey, D., and Denton, N. 1993. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, MA: Harvard University Press.

[50] McKinley, S.C., and Nosofsky, R. M. 1995. "Investigations of Exemplar and Decision Bound Models in Large, Ill-defined Category Structures." *Journal of Experimental Psychology*, 21, 128-48.

[51] Moro, A., and Norman, P. forthcoming. "A General Equilibrium Model of Statistical Discrimination." *Journal of Economic Theory.*

[52] Mullainathan, S. 2001. "Thinking through Categories" working paper: MIT.

[53] Phelps, E. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review*, Vol. 62 (4), 659-661.

[54] Phelps, E., O'Connor, K., Cunningham, W., Funayama, E., Gatenby, J., Gore, John., and Banaji, M. 2000. "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation." J*ournal of Cognitive Neuroscience*, 12:5, 729-738.

[55] Posner, and Keele, . 1968. "On the Genesis of Abstract Ideas." *Journal of Experimental Psychology*, 77, 3,1, 353-363.

[56] Quinn, P.C., and Eimas, P.D. 1996. Perceptual Organization and Categorization in Young Infants, in Rovee-Collier, C., and Lipsitt, L (Eds.), Advances in Infancy Research, 10, 1-36.

[57] Rabin, M. and Shrag, 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics*, Vol. 114 (1), 37-82.

[58] Reed, S. K. 1972. "Pattern Recognition and Categorization." *Cognitive Psychology*, 3, 382-407.

[59] Rosch, E. 1978. "Principles of Categorization", in E. Rosch and B.B. Lloyd (Eds.), *Cognition and Categorization*, 27-48. Hillsdale, NJ: Erlbaum.

[60] Rosseel, Y. 2002. "Mixture Models of Categorization." *Journal of Mathematical Psychology*, 46, 178-210.

[61] Shepard. R.N. 1987. "Toward a Universal Law of Generalization of Psychological Science." *Science*, 237, 1317-1323.

[62] Spence, M. 1974. *Market Signaling*. Cambridge: Harvard University Press.

[63] Stangor, C., and Ford, T.E. 1992. "Accuracy and Expectancy-Confirming Orientations and the Development of Stereotypes and Prejudice." *European Review of Social Psychology*, 3, 5-89.

[64] Stangor, C., and Lange, J.E. 1994. "Mental Representations of Social Groups: Advances in Understanding Stereotypes and Stereotyping." *Advances in Experimental Social Psychology*, 26, 357-416.

[65] Sternberg, R., and Ben-Zeev, T. 2001. *Complex Cognition: The Psychology of Human Thought*. NY: Oxford University Press.

[66] Tajfel, H., 1969. "Cognitive Aspects of Prejudice". *Journal of Social Issues*. 25(4) 79-97.

[67] Torgerson, W.S. 1958. *Theory and Methods of Scaling*. New York: Wiley.

[68] Wheeler, M.E. and Fiske, S.T. 2002. "Controlling Racial Prejudice and Stereotyping: Social Cognitive Goals Affect Amygdala and Stereotype Activation," unpublished.

7 Appendix

The following lemmas are useful in the proof of Theorems 1, 2, and Corollary 1. We work with the city-block metric and again assume that attributes take on values in $\{0, 1\}$, throughout.

When dealing with objects o and o', we will often write $d(o, o')$ to represent $d(\theta(o), \theta(o'))$.

Lemma 1 *For any group of objects O with cardinality n,*

$$Var(O) = \sum_{o \in O} d(\theta(o), \overline{\theta}(O)) = \frac{1}{n} \sum_{o \in O} \sum_{o' \in O} d(o, o').$$

Proof of Lemma 1:

Write

$$\sum_{o \in O} d(\theta(o), \overline{\theta}(O)) = \sum_{k} \sum_{o \in O} d(\theta_k(o), \overline{\theta}_k(O))$$

Given the fact that $\theta_k(o) \in \{0, 1\}$ for each k and o, we can write

$$d(\theta_k(o), \overline{\theta}_k(O)) = \sum_{o' \in O} \frac{1}{n} d(\theta_k(o), \theta_k(o'))$$

Then

$$\sum_{o \in O} d(\theta(o), \overline{\theta}(O)) = \sum_{k} \sum_{o \in O} \sum_{o' \in O} \frac{1}{n} d(\theta_k(o), \theta_k(o'))$$

which rearranging, leads to required expression. ∎

Lemma 2 *Consider any group of objects O with cardinality n and for any attribute k let $n_k^+ = \#\{o \in O | \theta_k(o) = 1\}$ and let $n_k^- = \#\{o \in O | \theta_k(o) = 0\}$. Then*

$$Var(O) = \frac{2 \sum_{k=1}^{m} n_k^+ n_k^-}{n}.$$

Proof of Lemma 2: By Lemma 1,

$$Var(O) = \frac{1}{n} \sum_{o \in O} \sum_{o' \in O} d(o, o').$$

Thus, by the additive separability of the city block metric

$$Var(O) = \frac{\sum_k \sum_{o \in O} \sum_{o' \in O} d(\theta_k(o), \theta_k(o'))}{n}.$$

The lemma then follows immediately. ∎

Lemma 3 *If the number of types of objects is at least as large as the number of categories, then under an optimal categorization, objects of the same type are assigned to the same category. That is, if* $\theta(o) = \theta(o')$*, then* $f_d^*(o) = f_d^*(o')$*.*

Proof of Lemma 3: Consider a group of objects O of cardinality n that are all of the same type. Suppose that a portion $\delta \in [0,1]$ of them is in one category and $1-\delta$ in another category. The lemma can be established by showing that the sum of the two categories variation is minimized at either $\delta = 0$ or $\delta = 1$. (Applying this iteratively then handles the case where a group of objects of the same type is categorized into more than two categories.)

Denote the categories by C_1 and C_2. Let O_i be the set of objects in C_i that are not in O, n_i be cardinality of O_i, and d_o denote the distance between o and an object in O. Then for a given choice of δ, by Lemma 1 we can write the total variation of categories C_1 and C_2 as

$$\frac{2\delta n \sum_{o\in O_1} d_o + \sum_{o\in O_1}\sum_{o'\in O_1} d(o,o')}{\delta n + n_1} + \frac{2(1-\delta)n \sum_{o\in O_2} d_o + \sum_{o\in O_2}\sum_{o'\in O_2} d(o,o')}{(1-\delta)n + n_2}.$$

The derivative of this expression with respect to δ [35] is (after simplifying some terms)

$$n\left[\frac{2n_1 \sum_{o\in O_1} d_o - \sum_{o\in O_1}\sum_{o'\in O_1} d(o,o')}{(\delta n + n_1)^2} - \frac{2n_2 \sum_{o\in O_2} d_o - \sum_{o\in O_2}\sum_{o'\in O_2} d(o,o')}{((1-\delta)n + n_2)^2}\right]. \tag{7}$$

For any o and o' in O_i, note that by the triangle inequality

$$d(o,o') \leq d_o + d_{o'},$$

with strict inequality when $o = o'$. This implies (after some rearrangement of summations, and noting that we will have at least one strict inequality) that

$$2n_i \sum_{o\in O_i} d_o - \sum_{o\in O_i}\sum_{o'\in O_i} d(o,o') > 2n_i \sum_{o\in O_i} d_o - 2\sum_{o\in O_i} d_o \;=\; 0. \tag{8}$$

From the expression for the derivative in (7), it follows that the second derivative of the total variation is

$$-2n^2 \left[\begin{array}{c} (\delta n + n_1)\frac{2n_1 \sum_{o\in O_1} d_o - \sum_{o\in O_1}\sum_{o'\in O_1} d(o,o')}{(\delta n + n_1)^3} \\ +((1-\delta)n + n_2)\frac{2n_2 \sum_{o\in O_2} d_o - \sum_{o\in O_2}\sum_{o'\in O_2} d(o,o')}{((1-\delta)n + n_2)^3} \end{array}\right]. \tag{9}$$

By the inequality (8), the second derivative is negative. This implies that the total variation is strictly concave in δ, and so the minimum over $\delta \in [0,1]$ must then be achieved at an endpoint of the interval. ∎

[35] Even though δ will need to be chosen in multiples of $1/n$, we show that the max of this equation over any $\delta \in [0,1]$ is achieved when δ is at one of the endpoints.

Given Lemma 3, we can think of the categorization of objects in terms of which types (θ's) are assigned to which category. The following lemma is also useful. Say that two attribute vectors are *adjacent* if they differ in terms of one and only one attribute.

Lemma 4 *If $n > \frac{7}{8}2^m$, and some majority type does not get its own unique category, then there exist (at least) two minority types that are adjacent to each other and each get their own category.*

Proof of Lemma 4: We use the following fact. If a hypercube has 2^x vertices, then any subset of more than 2^{x-1} vertices contains at least two that are adjacent.[36]

If $n > \frac{7}{8}2^m - 1$ (collecting the terms $2^{m-1} + 2^{m-2} + 2^{m-3}$), and not every majority item gets its own category, then minority items occupy more than $\frac{3}{8}2^m$ categories which have no majority items in them. This means that more than half of the minority objects are in categories that have only one type of object in it. The lemma then follows from the fact mentioned above. ∎

Now, let us return to the proof of the theorems. Theorems 1 and 2 follow from the following characterization.

Given a group of objects O_j, let n_j denote its cardinality; and for an attribute k let n_k^{j+} and n_k^{j-} be the number of objects in O_j with $\theta_k = 1$ and $\theta_k = 0$, respectively, as defined in the proof of Lemma 1.

Theorem 3 *Consider four groups of objects O_A, O_B, O_C, and O_D that we are considering categorizing into three categories. Assigning O_A and O_B each to their own category and grouping O_C and O_D together in one category leads to lower total variation than assigning O_C and O_D each to their own category and grouping O_A and O_B together in one category if and only if*

$$\frac{\sum_k (n_k^{A+} n_k^{B-} - n_k^{A-} n_k^{B+})^2}{n_A n_B (n_A + n_B)} > \frac{\sum_k (n_k^{C+} n_k^{D-} - n_k^{C-} n_k^{D+})^2}{n_C n_D (n_C + n_D)}. \tag{10}$$

Proof of Theorem 3: We need to show that

$$Var(O_A \cup O_B) + Var(O_C) + Var(O_D) > Var(O_C \cup O_D) + Var(O_A) + Var(O_B)$$

[36] It is easily checked that this bound is tight - that is, one can always find a subset of exactly 2^{x-1} vertices such that no two are adjacent.

holds if and only if (10) holds. Lemma 2 implies that this boils down to showing that

$$\begin{array}{ccc} \frac{2\sum_k (n_k^{A+}+n_k^{B+})(n_k^{A-}+n_k^{B-})}{n_A+n_B} & + & \frac{2\sum_k (n_k^{C+})(n_k^{C-})}{n_C} + \frac{2\sum_k (n_k^{D+})(n_k^{D-})}{n_D} > \\ \frac{2\sum_k (n_k^{C+}+n_k^{D+})(n_k^{C-}+n_k^{D-})}{n_C+n_D} & + & \frac{2\sum_k (n_k^{A+})(n_k^{A-})}{n_A} + \frac{2\sum_k (n_k^{B+})(n_k^{B-})}{n_B}. \end{array} \tag{11}$$

Cross multiplication, some cancelling of terms, and factoring allows us to rewrite (11) as (10). ∎

Proof of Theorem 1: We apply Theorem 3 as follows. Given that O_A and O_B form a balanced splitting of O_3 we know that for the balanced splitting k with respect to which O_A and O_B are defined,

$$(n_k^{A+}n_k^{B-} - n_k^{A-}n_k^{B+})^2 = (n_A n_B)^2.$$

Thus,

$$\frac{\sum_k (n_k^{A+}n_k^{B-} - n_k^{A-}n_k^{B+})^2}{n_A n_B (n_A+n_B)} \geq \frac{n_A n_B}{n_A+n_B}. \tag{12}$$

Also note that for any k

$$(n_k^{1+}n_k^{2-} - n_k^{1-}n_k^{2+})^2 \leq (n_1 n_2)^2.$$

and this equals 0 for all but h of the k's. Thus,

$$\frac{\sum_k (n_k^{1+}n_k^{2-} - n_k^{1-}n_k^{1+})^2}{n_1 n_2 (n_1+n_2)} \leq \frac{h n_1 n_2}{n_1+n_2}. \tag{13}$$

By inequalities 12 and 13,if

$$\frac{n_A n_B}{n_A+n_B} \geq \frac{h n_1 n_2}{n_1+n_2}$$

(which is equivalent to inequality 6), the result follows from (10). ∎

Proof of Theorem 2: This follows directly from Theorem 3 noting that for h_{12} of the k's that

$$(n_k^{1+}n_k^{2-} - n_k^{1-}n_k^{1+})^2 = (n_1 n_2)^2,$$

and the for remaining k's this is 0. Similarly for groups O_3 and O_4. (10) then simplifies to

$$h_{34}\left(\frac{1}{n_1}+\frac{1}{n_2}\right) > h_{12}\left(\frac{1}{n_3}+\frac{1}{n_4}\right),$$

which concludes the proof. ∎

Proof of Corollary 1 : We establish the theorem by showing that each of the majority types gets its own unique category. That is, if $\theta_k(o) = 1$ and $f_d^*(o) = f_d^*(o')$, then $\theta(o) = \theta(o')$.

Consider some f such that $\theta_k(o) = 1$ and $f(o) = f(o')$, and yet $\theta(o) \neq \theta(o')$. We need only show that such an f is not a solution to $f_d^*(o) = f_d^*(o')$.

By Lemmas 3 and 4, if some majority type does not get its own category we know that there are at least two adjacent minority types that are assigned to their own categories. Let the types of the two adjacent minority types be denoted θ^1 and θ^2, and the majority type be θ^3, and denote the corresponding groups of objects by O^1, O^2, and O^3 with corresponding cardinalities n^1, n^2, and n^3. Let O^4 be set of the remaining objects that are in the same category as O^3. By the adjacency of O_1 and O_2, by Theorem 1, it is enough to show that

$$\frac{1}{n_1} + \frac{1}{n_2} > \frac{1}{n_A} + \frac{1}{n_B},$$

where n_A and n_B correspond to a balanced splitting of $O_3 \cup O_4$. Note that by the definition of balanced splitting it follows that

$$\frac{1}{n_A} + \frac{1}{n_B} \geq \frac{1}{n_3} + \frac{1}{n_4}.$$

Thus, we need to show that

$$\frac{1}{n_1} + \frac{1}{n_2} > \frac{1}{n_3} + \frac{1}{n_4}.$$

Without loss of generality, assume that $n_1 \geq n_2$. Then it is sufficient to check that

$$\frac{2}{n_1} > \frac{1}{n_3} + \frac{1}{n_4},$$

or

$$\frac{2n_3}{n_1} - \frac{n_3}{n_4} > 1.$$

Noting that $\frac{n_3}{n_1} > r_E$, and $\frac{n_3}{n_4} < r_E r_I$ then leads to inequality. ∎