# A Categorical Model of Cognition and Biased Decision-Making*

Roland G. Fryer, Jr. and Matthew O. Jackson†

First Draft: October 17, 2002
Revision: November 12, 2004

**Abstract**

There is a wealth of research 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 show that specific biases emerge from categorization. For instance, types of experiences and objects that are less frequent in the population are more coarsely categorized and more often lumped together. As a result, decision makers make less accurate predictions when confronted with such objects. This can result in discrimination against minority groups even when there is no malevolent taste for discrimination.

*This supercedes a previous version with title "Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making.

†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, Douglas Bernheim, John Cacioppo, Colin Camerer, Gerald Clore, Glenn Ellison, Daniel Gilbert, Edward Glaeser, Susan Fiske, Dan Friedman, Drew Fudenberg, Claire Hill, Bengt Holmstrom, Philippe Jehiel, Vijay Krishna, Steven Levitt, Glenn Loury, George Lowenstein, Robert Marshall, Barry Mazur, Scott Page, Tom Palfrey, Michael Piore, Antonio Rangel, Andre Shleifer, Tomas Sjöström, and Steve Tadelis. Fryer is at the Harvard Society of Fellows and NBER, 1875 Cambridge Street, Cambridge MA, 02138 (rfryer@fas.harvard.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, and Jackson under grants SES-9986190 and SES-0316493.

"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. The distinguished social psychologist 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." Indeed, there is a long tradition in social psychology that treats certain biases such as stereotyping and prejudice as inevitable consequences of categorization (for example, see Allport (1954), Hamilton (1981), Tajfel (1969), or Fiske (1998) for a recent review). 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), but their potential has yet to be realized in many other social sciences (e.g., economics and sociology). This is due, in part, to the lack of formal models linking categorization to social decision making.

This paper builds a model of cognition centered on the basic principle that humans process information with the aid of categories; providing a link between categorization and social decision making. 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 central idea is that the number of categories is limited, and so the decision maker is forced to group heterogeneous experiences in the same category. 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.

Our main focus is on "optimal" categorizations, which we define to be ones that minimize the sum across categories of within category variation. We show that under some mild conditions this is equivalent to categorizing in a way that maximizes expected utility. An optimal solution to this problem necessarily lumps less frequent types of experiences into categories that end up being more heterogeneous. An important implication 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. We establish this in a series of results that partly characterize an optimal categorization, and we show how the categorization of objects depends on relative frequencies and some measure of distance between

objects.

We think of our contribution in two parts. The first is developing a model of how experiences are sorted into categories and how categorization biases decision making. The second is developing implications of this in specific social contexts. For instance, in a labor example, minorities will not be as finely sorted based on their investments in human capital. 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 are certainly not the first to provide a model of categorization. There is a rich literature 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). The novelty of our analysis is that it 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.

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

## 2 Three Examples

For pedagogical purposes, we begin with a simple example that illustrates some basic ideas of categorization.

**Example 1** *A Simple Labor Market Example*

Consider a population of employers and a population of workers. The population of workers consists of 90 percent "white" workers and 10 percent "black" workers. Thus, the black workers are the "minority" group. Workers come in two human capital levels: high and low. So, overall, workers come in four flavors: black-high, black-low, white-high, and white-low. 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? 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 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,·)'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, as illustrated in Table 1.

| Categories | Mean | Difference | Total Variation |
|---|---|---|---|
| (0,1)<br>(0,0)<br>{(1,0); (1,1)} | (0,1)<br>(0,0)<br>(1,.5) | 0<br>0<br>90*(0,.5) | 45 |
| (1,0)<br>(0,0)<br>{(0,1); (1,1)} | (1,0)<br>(0,0)<br>(.9,1) | 0<br>0<br>5*(.9,0); 45*(.1,0) | 9 |
| (1,1)<br>(0,1)<br>{(0,0); (1,0)} | (1,1)<br>(0,1)<br>(.9,0) | 0<br>0<br>5*(.9,0); 45*(.1,0) | 9 |
| (1,1)<br>(0,0)<br>{(0,1); (1,0)} | (1,1)<br>(0,0)<br>(.9,.1) | 0<br>0<br>5*(.9,.9); 45*(.1,.1) | 18 |
| (1,0)<br>(0,1)<br>{(0,0); (1,1)} | (1,0)<br>(0,1)<br>(.9,.9) | 0<br>0<br>5*(.9,.9); 45*(.1,.1) | 18 |
| (1,1)<br>(1,0)<br>{(0,0); (0,1)} | (1,1)<br>(1,0)<br>(0,.5) | 0<br>0<br>10*(0,.5) | 5 |

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 $\frac{1}{2}$. 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. [1]

Once we couple this with the observation that prototypes are important in forming expectations, discrimination results. 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. This 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.

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.[2] The discrimination that emerges from our model is not malevolent, nor is it derived from some primitive preference or taste for one's own race.

[1] This is consistent with the experimental evidence in social psychology and cognitive neuroscience that agents tend to categorize others by race (Brewer, 1988; Bruner, 1957; Fiske and Neuberg, 1990).

[2] 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)). There are many papers in the literatures on discrimination that have followed the seminal contributions of Becker (1957) and Arrow (1973). See Fryer (2002a) for a recent review of theoretical models of discrimination in the economics literature.

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.

**Some Evidence on Coarse Sorting of Minorities from Audits on Resumés**

A small literature using audit studies involving resumés (Jowell and Prescott-Clarke 1970, Hubbick and Carter 1980, Brown and Gay 1985, Bertrand and Mullainathan 2003) provides evidence for a coarser sorting of blacks by employers, which closely mirrors the above examples. These studies send resumes of fictitious applicants to potential employers. The main difference between the two resumés is that on one resumé the applicant has a distinctively black name and on the other the applicant has a traditionally white name. Such studies repeatedly have found that resumés with white names are substantially more likely to lead to job interviews than the identical resumés with distinctively black names. In terms of our example, the name represents the first attribute – broken down as white or black – and the rest of the content on a resumé corresponds to the second attribute, human capital. Further, Bertrand and Mullainathan (2003) show that the gap in call back rates between blacks and whites are larger among higher levels of skill and education. This is the precise prediction that comes out of our example and model.

**A Remark on Endogenizing Human Capital**

The 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".[3] 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.[4]

[3] It is not clear that one benefits from being over-estimated. There are two reasons. First, one may be thought to be overqualified for a particular job. Second, there are cost to being assigned to a job that is above one's level of expertise.

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

The idea of a feedback from discrimination to human capital investment is well-developed in Lundberg and Startz (1983), who build off of discrimination due to racially biased tests (a type of bias analyzed by Phelps (1972)). As such feedback effects are well understood, we will contain our analysis to the cognitive model.

**Example 2** *Social Interactions and Expected Utility in the Labor Market Example*

There are two key pieces missing from Example 1. First, why do people keep track of race at all? Second, why does the decision maker want to categorize in a manner that minimizes the total variation? A straightforward extension of the above example provides answers to both of these questions.

Suppose that there are two types of interaction in an employer's life: "social" and "economic." For social interactions, correctly assessing a person's race or culture is important, while in economic interactions the human capital attribute is the most relevant. Consider the following thought experiment. Suppose the employer has past objects categorized as in the first example. The employer then meets a new object in either a social or economic setting. The probability of it being a social setting is denoted $p_s$ and the probability of it being an economic setting is $p_e$.[5] The employer then matches the new object most closely to a category. This might happen in any of a number of ways, which are all equivalent for the purposes of this example. The employer might match this object to the category which contains the most objects of this type, or to the category whose vector of average characteristics is closest to this object. Once the object is matched to a category, the employer's prediction of what to expect is made based on the average experience from that category in the past. The payoff to the employer from the interaction depends on how closely the employer's prediction matches the actual object, weighted by some factor which captures the marginal impact of a correct versus incorrect prediction on the employer's utility.

Let us make this more concrete by revisiting Example 1 in some detail. Let $V_s$ be the marginal utility of a correct (versus incorrect) prediction in social setting, and let $V_e$ be the corresponding marginal utility in an economic setting. If an object has attributes $(0, 1)$ and is matched to a category with average attribute $(\frac{1}{2}, 1)$, then the prediction will be have an error in distance $\frac{1}{2}$ in the social dimension and 0 in the economic dimension. In that case, the payoff would be $\frac{1}{2}V_s$ if it turns out that this is a social interaction and would be $V_e$ if it turns out that this is an economic interaction. Based on this, we can develop the following expected utilities for the two most pertinent

[5] We don't need to have these sum to one, as it may be that some settings are both social and economic.

categorizations from Example 1.

First, consider the expected payoff to the "race-based" categorization (i.e., the categorization in which the employer assigns whites to different categories than blacks, and then subdivides whites into different categories by human capital). The expected utility is

$$p_s V_s 100 + p_e V_e (90 + 10(.5)) = p_s V_s 100 + p_e V_e 95.$$

Next, consider the expected payoff to the "human capital" categorization (i.e. the categorization in which employer assigns high human capital types to different categories than low human capital, and then subdivides one of the two human capital levels into different categories by race). The expected utility is

$$p_s V_s (50 + 45(.9) + 5(.1)) + p_e V_e 100 = p_s V_s 91 + p_e V_e 100.$$

Thus, sorting by human capital is better than the sorting by race if and only if $p_e V_e > 1.8 p_s V_s$. Therefore, the expected economic payoff needs to dominate the expected social payoff by a factor of almost two in a situation with numbers as in this example, before it becomes worthwhile to sort primarily based on the economic attribute.

We are now poised to answer the questions posited at the beginning of this example. First, categorization and memory are used for many tasks. If keeping track of race is useful in one venue of one's life, it can have spillovers to other venues.[6] Second, if $p_s V_s$ is similar to $p_e V_e$, then minimizing the variance is equivalent to maximizing expected utility.

Let us make one last remark about the example. There are profits to be had if one employer is able to overcome their categorical bias while others do not.[7] The question is whether there are sufficiently many employers who overcome such cognitive bias to give incentives to minority group members to make efficient investments in education and human capital. This point mirrors that developed by Becker (1957) in a model with tastes for race. If there are frictions in the market, for instance any search costs in finding employment, having some unbiased employers around might not be sufficient to induce efficient investment in human capital by minorities. A few cognitively biased employers could tilt hiring in favor of majority group members.

[6] There is substantial experimental evidence that individuals tend to keep track of other's race. See, for instance, Hart et. al., (2000) and Phelps et. al., (2000), though there is evidence that this may be context dependent (see Wheeler and Fiske (2002)).

[7] 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.

**Example 3** *Beyond Cognition: A Marketing Example*

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.[8]

Consider an advertiser who will produce $n$ different advertisements. A consumer is represented by a list of attributes, possibly including demographic information, tastes, consumption patterns, television watching behavior, price sensitivity, etc. Suppose the advertiser organizes the consumers into $n$ categories to minimize the total of within category variation. 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.

Consider a situation where $n$ is three and there are also three different attributes. Thus, there are 8 different types of consumers. Suppose that there exist 400 consumers with attributes (1,1,1); 400 with (1,1,0); 4 with (1,0,1); 4 with (1,0,0); 100 with (0,1,1); 100 with (0,1,0); 1 with (0,0,1); and 1 with (0,0,0), as in Table 2. The first attribute can be interpreted as gender, the second as human capital, and the third as race. So, there are 400 male, high human capital, of race 1, etc.

| | 1 | | 0 | |
|---|---|---|---|---|
| | 1 | 0 | 1 | 0 |
| 1 | 400 | 400 | 100 | 100 |
| 0 | 4 | 4 | 1 | 1 |

In this situation, it is straightforward to see that the advertiser will categorize as follows. Place (1,1,1)'s and (1,0,1)'s in one category; (1,1,0)'s and (1,0,0)'s in the second category; and all $(0,\cdot,\cdot)$'s in the third category.

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

There are several interesting things to note about this categorization. First, race is ignored in the categorization. That is, assignments to categories would be the same if that attribute were eliminated from the example completely. This is due to the very small number of consumers that have race attribute equal to zero. Advertisers don't find it useful, with limited resources, to distinguish consumers based on this attribute. If instead, the advertiser had five categories, the optimal sorting would involve some differentiation based on race. This previews a discussion of the endogeneity of the number of categories, which is something that we return to in Section 7.

Second, the interpretation of the model is quite different here from the cognitive discussion in the previous examples. Here, the advertiser not only sees all of the attributes; but might also be fully cognizant of them and able to understand and process them. It is the limitation in available advertisements that leads to the categorization. This is in contrast to the earlier examples where the potential employer observed all attributes, but because of limited cognitive abilities stored and processed the information in a boundedly rational fashion.

# 3 A Model of Categorization

## A. The Basic Building Blocks

### Categories

$C = \{C_1, \ldots, C_n\}$ is a finite set of categories. These categories can be thought of as "file folders" in our decision maker's brain that will be useful for the storing of information. 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.[9] Effectively, this is a bounded rationality story in which there are both costs to storing details of past interactions and delays in activating stored information based on how finely it is stored. We take the number of categories as given and discuss endogenizing this number in Section 6.

### Objects

[9] 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.

$O$ 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.[10] 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 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.[11]

Let $\theta : O \to \{0,1\}^m$ denote the function, written as $(\theta_1(o), \ldots, \theta_m(o))$, which describes the attributes that each object has.[12] 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

[10] In the psychology literature the term attributes often refers to the association of a given category with a series of different possible behaviors or other characteristics (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.

[11] Simply extend the range of the $\theta$ 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.

[12] Of course, the range of $\theta$ can be easily extended to the continuum $[0,1]$.

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.[13]

**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 $f : O \to C$ 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 (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 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.[14]

[13] 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 (see, for instance, Fiske (1993)).

[14] In our model, we are careful to use the term "prototype" for the representative of a category, rather than the term "stereotype". There is evidence that individuals can identify a "stereotype" for a given vector of attributes that will be common to other individuals, even without having that as their own belief. So, a stereotype might be thought of as knowing something about other people's categorizations and prototypes. See Hilton and Von Hippell (1996).

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.

**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 actions with 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 $\theta$. When confronted with object $o$, the decision maker's mind calls up some category $f(o)$. This might be done by comparing the given object's attributes to the prototypes of different categories until a closest match is found. There is substantial experimental evidence that when faced with an object or person, a person's brain "automatically" activates a category that, according to some metric, best matches the given object (and at times context) in question.[15] Then the expected utility of taking action $a$ when faced with object $o$ is

$$EU(a, o) = U(a, \overline{\theta}(f(o))). \tag{3}$$

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.[16]

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

[16] One can find many alternative methods for making predictions for a given categorization. An alternative to what we propose is an adaptation of case-based decision theory, developed by Gilboa and Schmeidler (2001), to the

**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. [17] 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 ($\ell_1$ norm). 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. 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. Unless indicated otherwise, we stick with the city-block metric as it simplifies the analysis.

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

$$Var(O) = \sum_{o \in O} d\left(\theta(o), \overline{\theta}(O)\right), \tag{4}$$

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

$$Var(f, O) = \sum_{C_i \in C} Var(\{o : f(o) \in C_i\}). \tag{5}$$

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.

[17] 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 focus on the "end" categorization under the presumption that it has been constructed in some approximately efficient manner.

An *optimal categorization function* relative to $O$ is a categorization $f^*$ that minimizes $Var(f, O)$[18].

**When is Expected Utility Maximization Equivalent to Variance Minimization?**

Let us comment on why minimizing the variance is a sensible objective, and how this relates to expected utility maximization.

A possible goal of the decision maker is to use their categorization to form accurate predictions for expected future interactions. A best guess at the distribution of future interactions is based on the frequency of past interactions. Similar to Example 2, let an attribute $k$ be relevant for a decision with probability $p_k$ and let $V_k$ denote the marginal utility benefit of a correct (versus incorrect) prediction. A categorization $f$, which maximizes expected utility is then a solution to the problem

$$\max_f \sum_o \sum_k p_k V_k \left[1 - d\left(\theta_k\left(o\right), \overline{\theta}_k\left(f\left(o\right)\right)\right)\right]$$

Without loss of generality, let us code the attributes so that the $p_k V_k$ are similar across $k$. This can be done by coding attributes a number of times that is proportional to this weighted utility. For instance, if $p_k V_k$ is twice as high for one attribute versus another, then we can include two entries of this attribute in our vectors for each coding of the other attribute. Once this is done, then the above maximization problem is equivalent to the minimization problem

$$\min_f \sum_o \sum_k \left| d\left(\theta_k\left(o\right), \overline{\theta}_k\left(f\left(o\right)\right)\right)\right|,$$

which is precisely our objective function.

# 4 A Categorization Theorem

We assume throughout that $2^m > n$, so that there are fewer categories than types. This rules out the degenerate case where each type of 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. The question of which types end up grouped together has important implications, as we have already seen in the examples in Section 2. In those examples, we saw that it was the smallest groups of types that were categorized together. We can now develop this idea more generally, and show how the categorization model operates.

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

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. The technical results on optimal categorization proceed in four parts. First, we prove that objects of the same type are always put in the same category. Second, we provide a proposition that states necessary and sufficient conditions for sorting four homogeneous groups of objects into three categories. Third, we state a theorem that provides a characterization of optimal categorizations. Our last result provides sufficient conditions for complete segregation of a minority group under a categorization.

Let us begin with a lemma that is important to the analysis.

**Lemma 1** *Under an optimal categorization, objects of the same type are assigned to the same category. That is, if* $\theta(o) = \theta(o')$*, then* $f_d^*(o) = f_d^*(o')$*.*

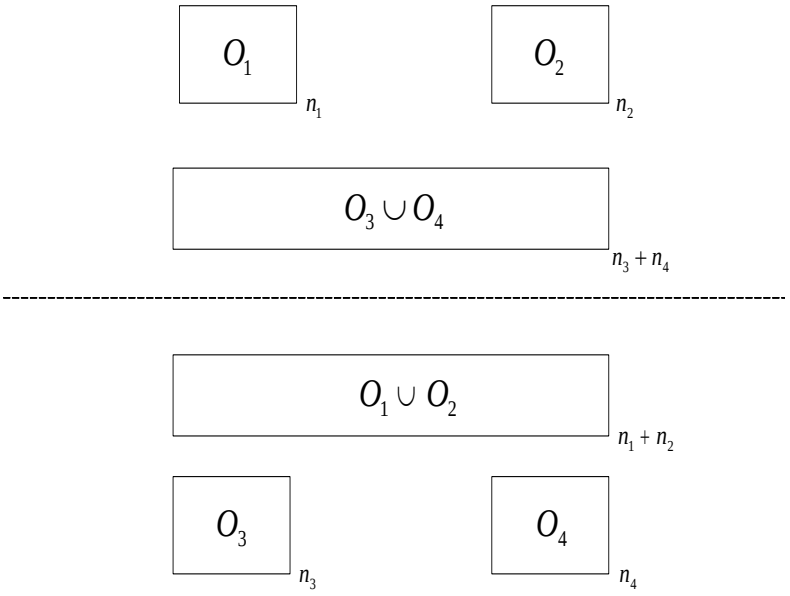
**Proof.** See Appendix ■

On the surface, this lemma may seem obvious, but it takes a bit of proof. The difficulty is that grouping objects of the same type together can lead some categories to be quite full and others much less so. In cases where all categories contain objects of a variety of types, one could imagine that equalizing size might help with variance. Lemma 1 shows that this is not the case, and so one can work with blocks of objects that are all of the same type. This makes the forthcoming analysis much easier.

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 $\theta(o) \neq \theta(o')$.

Consider four groups of objects $O_1$, $O_2$, $O_3$, and $O_4$ with corresponding cardinalities $n_1$, $n_2$, $n_3$, and $n_4$. Suppose that we have categorized things so that $O_1$ and $O_2$ have their own category and $O_3$ and $O_4$ are grouped together. When would we do better by re-categorizing so that we split up $O_3$ and $O_4$ and instead put $O_1$ and $O_2$ together? The answer is given in the following proposition.

Let $h_{ij}$ be the number of attributes on which two (possibly heterogeneous) groups objects $O_i$ and $O_j$ differ at all (which may even be 0).




**Proposition 1** *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 $n_4$.*

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

*if and only if*

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

**Proof.** See Appendix ■

To paraphrase Proposition 1, suppose that we have four 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 $\frac{n_i n_j h_{ij}}{n_i+n_j}$, are the best ones to put together.

Rather than have $O_1$ and $O_2$ assigned to their own categories and $O_3$ and $O_4$ lumped together, it is better to split up $O_3$ and $O_4$, 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 $\frac{1}{n_1}+\frac{1}{n_2}$ ,
- $O_1$ and $O_2$ are fairly similar in their attributes ( $h_{12}$ is small), and
- $O_3$ and $O_4$ are relatively large and so it is optimal to assign them to their own categories (so $\frac{1}{n_3}+\frac{1}{n_4}$ is small).

Let us return to Example 1. 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 splitting of low types into $O_3$ and $O_4$ being according to the other attribute, human capital. So, $O_3$ is white-high with $n_3 = 45$ and $O_4$ is white-low with $n_4 = 45$. Here, since $h_{34} = h_{12} = 1$ and $\frac{1}{5} + \frac{1}{5} > \frac{1}{45} + \frac{1}{45}$, then it is optimal to categorize the blacks into one category and separate the whites into two.

Proposition 1 is useful and transparent, in part, because we are dealing explicitly with homogeneous groups of objects. Our main result, which is stated next, relaxes this assumption and provides a fuller characterization of optimal categorization.

For an arbitrary (possibly heterogeneous) group of objects $O_i$, let $n_k^{i+}$ be the number of objects in $O_i$ with $\theta_k(o) = 1$, and let $n_k^{i-}$ be the number of objects in $O_i$ with $\theta_k(o) = 0$.

**Theorem 1** *Consider groups of objects $O_1$, $O_2$, ..., $O_J$, and suppose we are considering combining two of these groups. The two groups that when combined lead to the lowest total variation summed across these groups are any two that lead to the smallest factor:*

$$\frac{\sum_k (n_k^{i+} n_k^{j-} - n_k^{i-} n_k^{j+})^2}{n_i n_j (n_i + n_j)}. \tag{6}$$

**Proof.** See the Appendix. ■

Our results provide some feeling for which groups of objects it makes the most sense to lump together in the same category. This stops far short of providing a full description of what an optimal categorization looks like. This is a hard problem (in the language of computer science, an NP-hard problem). One starting point in terms of an algorithm for choosing an algorithm would be to use Theorem 1 iteratively. Start by assigning groups to different categories until one is faced with more groups than categories. Group together the two that produce the smallest variation based on the groups currently faced, and continue in this manner.

It is not clear how optimal this would be, or whether more efficient algorithms exist. It certainly will be history dependent, which provides another consideration for cognitive discrimination.

## 5 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 develop-

ment 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 $\theta_k(o) = 0$.[19] 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 $(0,\cdot,\cdot)$ 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

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

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

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

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.

**Segregation 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).*

**Proof.** See Appendix ■

Corollary 1 provides sufficient conditions for a complete segregation of minority objects from majority ones, and more coarse sorting of the minority group. The very strong conclusions of this result require very 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. 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.

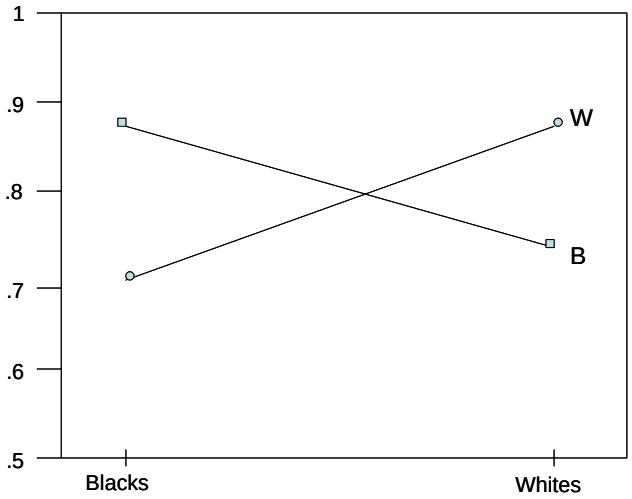
## 6 Evidence on Racial Differences in Facial Recognition

There is an impressive literature on racial and ethnic differences in facial recognition.[20] The terms "cross-race recognition deficit," "cross-race effect," and "own-race bias" all describe the frequently observed performance deficit of one ethnic group in recognizing faces of another ethnic group compared with faces of one's own group (see Sporer 2001 for a detailed review). In other words, "they all look alike to me" is a reasonable caricature of how members of one group categorize another.

Own-race bias in the recognition of facial stimuli is observed when two factors, ethnic group of participant and ethnic group of stimulus face, interact significantly in the expected direction. Ideally, the interaction can be presented graphically as a complete crossover. Namely, both ethnic groups recognize members of their own better than members of other ethnic groups. Figure 2, extracted from Devine and Malpass 1985, is one such illustration.[21]

[20] We thank Andrew Postlewaite for suggesting this line of inquiry.

[21] Meissner and Brigham (2001) provide a detailed meta-analytic review of the last thirty years of literature investigating the own-race bias in facial recognition. They review 39 research articles involving the responses of over 5,000 subjects. There are a few studies that fail to find a cross-race effect. The overwhelming consensus among social psychologists, however, is that these effects not only exist, but are quite large (Meissner and Brigham 2001).




This data is consistent with predictions of our model to the extent that individuals interact more frequently with members of their own racial group. Our model of categorization predicts that individuals with more *inter*-group contact will be better at distinguishing subtle features about other groups than individuals with less inter-group contact. There is substantial evidence in this regard (see Sporer 2001, Table 2). As Meissner and Brigham (2001) report:

> "Several studies demonstrate that adolescents and children living in integrated neighborhoods better recognize novel other-race faces than did those living in segregated neighborhoods.[22]"

An interesting study testing the relationship between contact with other groups and facial recognition is Li, Dunning, and Malpass (1998). They demonstrate that white "basketball fans" were superior to white "basketball novices" in recognizing black faces. The idea is that basketball fans watch the National Basketball Association games on a regular basis, which provides frequent exposure to black faces, given that a sizeable majority of the players are black. Participants were black and white men and women. They were presented with black and white faces on a video monitor. The subjects were informed that they would be tested on their ability to recognize the faces viewed. Performance of basketball fans who were white was indistinguishable from blacks in their ability to recognize black faces, whereas the white subjects who were not basketball fans performed at a significantly worse level, while there is no such difference in the ability to recognize white faces! In recognizing white faces, there was no difference between basketball fans and novices.

[22] There have been several studies that replicate this finding (Cross et. al. 1971, Feinman and Entwisle 1976).

While this evidence does not *prove* that our model of categorization is correct, it is certainly consistent with its predictions. And, the robustness of the conclusions of such studies across subjects and experimental designs over the years makes it hard to dismiss.

# 7 Some Further Remarks

### Toward a Theory of 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 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. The model we have put forth in the previous sections provides a tool for the study of identity, as we can view *identity as self-categorization*.

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.[23] 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.

### Culture and Identity

Consider a stylized community. Each agent decides whether to invest their time in learning the local language and traditions or computer programming. Investments in computer programming are valued in the global labor market, whereas, the local knowledge 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 local knowledge 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

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

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 local knowledge (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.[24]

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 a low level of local knowledge may imply that you are likely to leave the community. Hence, the community will not cooperate with agents who invest too much in computer programming or too little in local knowledge. 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.[25] 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, 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

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

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

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.

**Endogeneity of Categories**

We have treated the set of categories as a given. We know from the developmental 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".[26] 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.[27] 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 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 laboratory evidence 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

[26] 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.

[27] 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.

tested to see whether such types of objects have more biased predictions associated with them. We referred to the large literature on facial recognition in Section 6, which provides substantial evidence consistent with our model, but more tests are feasible.

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. While the facial literature convincingly documents the relationship between interracial contact and facial recognition – they do not tests whether a lack of interaction results in individuals reverting to the mean. A way of testing this effect is compare human capital investment decisions by ethnic minorities 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. Further, whereas a taste-based model would predict that larger numbers of blacks would result in more discrimination – the categorization model has the opposite prediction. Categorical discrimination can also 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.

**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.[28] 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 another new view of a role model: teaching the decision makers (e.g., employers) and not just the

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

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.

**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 (forthcoming) 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.[29]

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

[29] 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."

(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.

## 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.

[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] Bertrand, M. and S. Mullainathan, 2003. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," NBER working paper number 9873.

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

[18] 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.

[19] Brown, C. and P. Gay (1985) *Racial Discrimination 17 Years after the Act*, London: Policies Studies Institute.

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

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

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

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

[24] 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.

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

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

[27] 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.

[28] 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.

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

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

[31] Fryer, R., and Levitt, S. forthcoming. "Understanding the Black-White Test Score Gap in the First Two Years of School." *Review of Economics and Statistics*

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

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

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

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

[36] 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.

[37] 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.

[38] 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.

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

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

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

Hubbick, J. and S. Carter 1980. *Half a Chance? A Report on Job Discrimination against Young Black Males in Nottingham*. London: Commission for Racial Equality.

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

[43] 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.

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

[45] Jowell, R. and Prescott-Clarke, P. 1970. "Racial Discrimination and White-Collar Workers in Britain," *Race*, vol. 11, pp 397-417.

[46] 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.

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

[48] Li, J., Dunning, C., and Malpass, R. 1998. "Cross-racial Identification Among European-Americans: Basketball Fandom and the Contact Hypothesis. Working paper.

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

[50] Lundberg, S. and Startz, R. 1983. Private Discrimination and Social Intervention in Competitive Labor Markets. *American Economic Review*, 73, 340-347

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

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

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

[54] Meissner, C. and Brigham, J. 2001. "Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review." *Psychology, Public Policy, and Law*, 7 (1), 3-35.

[55] 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.

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

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

[58] 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.

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

[60] 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.

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

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

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

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

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

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

[67] Sporer, S. 2001. "Recognizing Faces of Other Ethnic Groups: An Integration of Theories," *Psychology, Public Policy, and Law*, 7 (1), 36-97.

[68] 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.

[69] 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.

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

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

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

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

# 8 Appendix: Proofs

The following lemmas are useful in the proof of Theorems **??**, 1, 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 2** *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 2:**

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 3** *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 3:** By Lemma 2,

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

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. ∎

**Proof of Lemma 1:** 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 $\delta$, by Lemma 2 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 $\delta$ [30] is (after simplifying some terms)

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

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

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

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

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

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

Given Lemma 1, we can think of the categorization of objects in terms of which types ($\theta$'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.

[30] Even though $\delta$ 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 $\delta$ is at one of the endpoints.

**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.[31]

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 **??** and 1 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 2.

**Proof of Theorem 1:** It is sufficient to show that

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

holds if and only if

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

Lemma 3 implies that this boils down to showing that

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

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

**Proof of Proposition 1:** This follows directly from Theorem 1 noting that for $h_{12}$ of the $k$'s that

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

[31] 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.

and the for remaining $k$'s this is 0. Similarly for groups $O_3$ and $O_4$. (6) 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 1 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 $\theta^1$ and $\theta^2$, and the majority type be $\theta^3$, 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 **??**, 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. ∎