STEREOTYPES

Pedro Bordalo

Katherine Coffman

Nicola Gennaioli

Andrei Shleifer*

February 15, 2016

Abstract

We present a model of stereotypes based on Kahneman and Tversky's representativeness heuristic. A decision maker assesses a target group by overweighting its representative types, defined as the types that occur more frequently in that group than in a baseline reference group. Stereotypes formed in this way contain a "kernel of truth": they are rooted in true differences between groups. Because stereotypes focus on differences, they cause belief distortions, particularly when groups are similar. Stereotypes are also context dependent: beliefs about a group depend on the characteristics of the reference group. In line with our predictions, beliefs in the lab about abstract groups and beliefs in the field about political groups are context dependent and distorted in the direction of representative types. JEL: D03, D83, D84. C91.

*We are grateful to Nick Barberis, Roland Bénabou, Dan Benjamin, Tom Cunningham, Matthew Gentzkow, Emir Kamenica, Larry Katz, Eliana La Ferrara, David Laibson, Sendhil Mullainathan, Josh Schwartzstein, Jesse Shapiro, Alp Simsek, Neil Thakral, and four anonymous referees for extremely helpful comments, and Jesse Graham, Jonathan Haidt, and Brian Nosek for sharing the Moral Foundations Questionnaire data. We thank the Initiative on Foundations of Human Behavior for support of this research, and Maik Wehmeyer, Laura Freitag, and Aaron Englander for research assistance. Corresponding Author: Andrei Shleifer, email: ashleifer@harvard.edu, tel: 6174955046, Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138.

1 Introduction

The Oxford English Dictionary defines a stereotype as a "widely held but fixed and oversimplified image or idea of a particular type of person or thing". Stereotypes are ubiquitous. Among other things, they cover racial groups ("Asians are good at math"), political groups ("Republicans are rich"), genders ("Women are bad at math"), demographic groups ("Florida residents are elderly"), and situations ("Tel-Aviv is dangerous"). As these and other examples illustrate, some stereotypes are roughly accurate ("the Dutch are tall"), while others much less so ("Irish are red-headed"; only 10% are). Moreover, stereotypes change: in the US, Jews were stereotyped as religious and uneducated at the beginning of the 20th century, and as high achievers at the beginning of the 21st (Madon et. al., 2001).

Social science has produced three broad approaches to stereotypes. The economic approach of Phelps (1972) and Arrow (1973) sees stereotypes as a manifestation of statistical discrimination: rational formation of beliefs about a group member in terms of the aggregate distribution of group traits. Statistical discrimination may impact actual group characteristics in equilibrium (Arrow 1973), but even so stereotypes are based on rational expectations.[1] As such, these models do not address the central problem that stereotypes are often inaccurate. The vast majority of Florida residents are not elderly, the vast majority of the Irish are not red-headed, and Tel-Aviv is really pretty safe.

The sociological approach to stereotyping pertains only to social groups. It views stereotypes as fundamentally incorrect and derogatory generalizations of group traits, reflective of the stereotyper's underlying prejudices (Adorno et al. 1950) or other internal motivations (Schneider 2004). Social groups that have been historically mistreated, such as racial and ethnic minorities, continue to suffer through bad stereotyping, perhaps because the groups in power want to perpetuate false beliefs about them (Steele 2010, Glaeser 2005). The stereotypes against blacks are thus rooted in the history of slavery and continuing discrimination. This approach might be relevant in some important instances, but it leaves a lot out. While some stereotypes are inaccurate, many are quite fair ("Dutch are tall," "Swedes

[1]More recent work explores under what conditions stereotypes are self-fulfilling. Assuming that freely available information is used correctly, minorities can invest in visible signals of quality that offset preconceptions (Lundberg and Startz 1983). Glover et al (2015) present evidence on self-fulfilling aspects of stereotypes in labor markets.

are blond.") Moreover, many stereotypes are flattering to the group in question rather than pejorative ("Asians are good at math"). Finally, stereotypes change, so they are at least in part responsive to reality rather than entirely rooted in the past (Madon et. al., 2001).

The third approach to stereotypes – and the one we follow – is the "social cognition approach", rooted in social psychology (Schneider 2004). This approach gained ground in the 1980's and views social stereotypes as special cases of cognitive schemas or theories (Schneider, Hastorf, and Ellsworth 1979). These theories are intuitive generalizations that individuals routinely use in their everyday life, and entail savings on cognitive resources. Hilton and Hippel (1996) define stereotypes as "mental representations of real differences between groups [...] allowing easier and more efficient processing of information. Stereotypes are selective, however, in that they are localized around group features that are the most distinctive, that provide the greatest differentiation between groups, and that show the least within-group variation." A related "kernel-of-truth hypothesis" holds that stereotypes are based on some empirical reality; as such, they are useful, but may entail exaggerations (Judd and Park 1993).

We show that this approach to stereotypes is intimately related to another idea from psychology: the use of heuristics in probability judgments (Kahneman and Tversky 1972). Just as heuristics simplify the assessment of complex probabilistic hypotheses, they also simplify the representation of heterogeneous groups, sometimes causing errors in judgment. We formally explore this idea by modelling stereotype formation as a consequence of Kahneman and Tversky's representativeness heuristic. Tversky and Kahneman (1983) write that "an attribute is representative of a class if it is very diagnostic; that is, the relative frequency of this attribute is much higher in that class than in the relevant reference class." Following Gennaioli and Shleifer (GS 2010), we assume that a type t is representative for group G relative to a comparison group $-G$ if - in line with the Tversky and Kahneman definition - it scores high on the likelihood ratio:

$$\frac{\Pr(t|G)}{\Pr(t|-G)}. \tag{1}$$

The most representative types come to mind first, and so are overweighted in judg-

ments. Predictions about G are then made under a distorted distribution, or stereotype, that overweights representative types. Our results obtain with minimal assumptions on such overweighting. We describe a number of weighting specifications and explore their properties.

To illustrate the logic of the model, consider the stereotype "Florida residents are elderly". The proportion of elderly people in Florida and in the overall US population is shown below.[2]

age	$0-19$	$20-44$	$45-64$	$65+$
Florida	24.0%	31.7%	27.0%	17.4%
US	26.9%	33.6%	26.4%	13.1%

The table shows that the age distributions in Florida and in the rest of the US are very similar. Yet, someone over 65 is highly representative of a Florida resident, because this age bracket maximizes the likelihood ratio $\Pr(t|\text{Florida})/\Pr(t|\text{US})$. When thinking about the age of Floridians, then, the "65+" type immediately comes to mind because in this age bracket Florida is most different from the rest of the US, in the precise sense of representativeness. Representativeness-based recall induces an observer to overweight the "65+" type in his assessment of the average age of Floridians.

This example also illustrates how stereotypes can be inaccurate. Indeed, and perhaps surprisingly, only about 17% of Florida residents are elderly. The largest share of Florida residents, nearly as many as in the overall US population, are in the age bracket "19-44", which maximizes $\Pr(t|\text{Florida})$. Being elderly is not the most likely age bracket for Florida residents, but rather the age bracket that occurs with the highest *relative* frequency. A stereotype-based prediction that a Florida resident is elderly has very little validity.

The same logic of representativeness suggests that the reason people stereotype the Irish as red-headed is that red hair is more common among the Irish than among other groups, even though it is not that common in absolute terms. The reason people stereotype Republicans as wealthy is that the wealthy are more common among Republicans than Democrats.[3] In both cases, the representation entails judgment errors: people overestimate the proportion of red-haired among the Irish, or of the wealthy among the Republicans.

[2]Data from the 2010 US Census, see http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_DP_DPDP1&src=pt.

[3]See www.nytimes.com/packages/pdf/politics/20041107_px_ELECTORATE.xls.

We find that representativeness often generates fairly accurate stereotypes but sometimes causes stereotypes to be inaccurate, particularly when groups have similar distributions that differ most in unlikely types. More generally, our model highlights two critical properties:

- Stereotypes amplify systematic differences between groups, even if these differences are in reality very small. When groups differ by a shift in means, stereotyping exaggerates differences in means, and when groups differ by an increase in variance, stereotyping exaggerates the differences in variances. In these cases (but not always), representativeness yields stereotypes that contain a "kernel of truth", in the sense that they differentiate groups along existing and highly diagnostic characteristics, exactly as Hilton, Hippel and Schneider define them.

- Stereotypes are context dependent. The assessment of a given target group depends on the reference group to which it is compared.

In line with the social cognition approach to stereotypes, a significant body of psychological research on beliefs about gender, race, age, and political groups finds that stereotypes broadly reflect reality but also display biases. Social psychologists have explored the extent to which stereotypes exaggerate real differences, thus possessing a "kernel of truth". Evidence on exaggeration varies by domain. For race and gender, many studies have reported roughly accurate beliefs, while others have found underestimation or exaggeration of true differences (Jussim et al 2015).[4] For age and political stereotypes, evidence points more strongly toward exaggeration.[5] Schneider (2004) summarizes the existing empirical work on stereotype accuracy as follows: "the best we can do by way of general summary is to say that some stereotypes held by some people for some groups are sometimes accurate". Our empirical investigation explores the connection between stereotype accuracy and representativeness, which we measure using the previously defined likelihood ratio.

[4]For evidence of roughly accurate beliefs for race, see, for example, Ashton and Esses (1999) on beliefs of academic ability, Kaplowitz et al (2003) on income, poverty rates, and out-of-wedlock births, and Wolsko, Park, Judd, and Wittenbrink (2000) on a array of positive and negative attributes. Ryan (1996) and Gilens (1996) find evidence of exaggeration of race stereotypes on personal attributes and on poverty rates respectively. Some studies have found roughly accurate beliefs on gender (Briton and Hall 1995, McCauley, Thangavelu, and Rozin 1988, and Diekman, Eagly, and Kulesa 2002), while others find evidence of exaggeration (Martin 1987, Beyer 1999).

[5]See Chan et al (2014) on age. We discuss evidence on political stereotypes in Section 4.

We first assess the role of representativeness and context dependence in the lab. We construct a group of mundane objects, G, and present it to participants next to a comparison group, $-G$. In our baseline condition, the comparison group is chosen so that no type is particularly representative of group G. In our treatment, we change the comparison group, $-G$, while leaving the target group, G, unchanged. The new comparison group gives rise to highly representative types within G. In line with the key prediction of our model, participants in the treatment condition shift their assessment of G toward the new representative types.

We next test the model using two data sets on political preferences, and beliefs about political preferences, in the U.S. Here, groups are political constituencies (Democrats and Republicans) and types are their positions on a number of issues. Holding fixed the groups and the set of types, and varying the groups' true distributions across issues, the data allow us to test whether beliefs about political preferences are shaped by representativeness. We first show that such beliefs depart from the truth by exaggerating (mean) differences, as per the kernel of truth logic. We then explore how the extent of exaggeration varies across issues. Consistent with the model, we find that beliefs systematically depart from rational expectations and that the degree to which they exaggerate true differences is a function of representativeness. While representativeness is not the only heuristic that shapes recall (availability, driven by recency or frequency of exposure, also plays a role), it explains the fact that, in the data, stated beliefs indeed exaggerate differences among groups.

Since Kahneman and Tversky's (1972, 1973) work on heuristics and biases, several studies have formally modelled heuristics about probabilistic judgments and incorporated them into economic models. Work on the confirmation bias (Rabin and Schrag 1999) and on probabilistic extrapolation (Grether 1980, Barberis, Shleifer, and Vishny 1998, Rabin 2002, Rabin and Vayanos 2010, Benjamin, Rabin and Raymond 2011) assumes that the decision maker has an incorrect model in mind or incorrectly processes available data. Our approach is instead based on the assumption that representative information comes foremost to mind when making judgments. The mental operation that lies at the heart of our model – generating a prediction for the distribution of types in a group, based on data stored in memory – also captures base-rate neglect and overreaction to diagnostic information. The

underweighting or neglect of information in our model simplifies judgment problems in a way related to models of categorization (Mullainathan 2002, Fryer and Jackson 2008). In these models, however, decision makers use coarse categories organized according to likelihood, not representativeness. This approach generates imprecision but does not create a systematic bias for overestimating unlikely events, nor does it allow for context dependent beliefs. In our empirical analysis of political beliefs, we explicitly compare the predictions of representativeness-based recall to those of likelihood based models and find that the evidence supports the former.

In modeling representativeness we follow the specification of GS (2010), but investigate a new set of questions. GS (2010) examine how representativeness distorts the assessed probabilities of alternative hypotheses, but not how the probability of a given hypothesis or group is distributed across its constituent elements. In the context of the current setting, GS (2010) ask how imperfect recall affects the assessed probability that a randomly drawn member from a universe Ω belongs to group G. The current paper, in contrast, asks which type t we expect to draw *once we know* that we are facing group G. GS (2010) show how representativeness generates biased probabilistic assessments such as conjunction and disjunction fallacies. The current paper deals with perhaps a broader and more ubiquitous problem of stereotype formation, extensively studied by other social scientists but largely neglected by economists.

Section 2 describes our model. In Section 3 we examine the properties of stereotypes, including the forces that shape stereotype accuracy, and illustrate these properties with examples. In Section 4 we bring the model to the data, performing a lab experiment and analyzing existing surveys of political beliefs. Section 5 concludes. The Online Appendix presents proofs, a number of extensions of the model, and additional results for the experiments and field evidence.

2 A Model of Representativeness and Stereotypes

2.1 The Model

A decision maker (DM) faces a *prediction* problem, such as assessing the ability of a job candidate coming from a certain ethnic group, the future performance of a firm belonging to a certain sector, or future earnings based on own gender.

Formally, there is a set of types of interest T and an overall population Ω, of which group G is a subset. The set of types T can be unordered (e.g., occupations) or ordered (and typically cardinal, e.g., earnings levels). When T is ordered, we write $T = \{t_1, \ldots, t_T\}$ with $t_1 < t_2 < ... < t_T$.[6] There is a probability or frequency distribution $\pi \in \Delta(T \times \Omega)$, that induces a conditional distribution $\Pr(T = t \,|G)$ when restricted to G.[7] In what follows, we denote by $\pi_{t,G} = \Pr(T = t \,|G)$ the probability of type t in group G and by π_G the vector $(\pi_{t,G})_{t \in T}$ containing the conditional distribution.

The DM's goal is to assess the distribution of the types of interest in a particular group G. While the DM has stored in memory the full distribution, he retrieves from memory a distorted version of π_G that overweights the probability of those types that are most representative of G relative to a comparison group $-G$. Generically, $-G$ is a group in Ω that is distinct from G, namely $-G \subseteq \Omega \backslash G$, although in some cases it can coincide with the complement of G.

According to Tversky and Kahneman (1983), a type t is representative of G if it is relatively more likely to occur in G than in $-G$. Definitions 1 and 2 formalize this representativeness-based recall, following GS (2010).

Definition 1 *The representativeness of type t for group G given comparison group $-G$ is*

[6] For simplicity, we also use T to denote the number of types $|T|$. The model applies also to cases in which types: i) are multi-dimensional, capturing a bundle of attributes such as occupation and nationality, or ii) are continuous. We consider these cases in Appendices C and D respectively. Also, G may represent any category of interest, such as the historical performance of a firm or industry, actions available to a decision maker (T = set of payoffs, G = occupations), or categories in the natural world (T = ability to fly, G = birds).

[7] In many applications each individual in Ω is characterized by a deterministic type (e.g. age, hair color, etc). As a result, $\pi(t, \omega) = 1/|\Omega|$. For instance, each Floridian has a single age type (at the finest temporal resolution). When instead types are stochastic, such as when estimating future earnings of a person or a firm, each individual is described by a non-degenerate distribution.

defined as the likelihood ratio:

$$R(t, G, -G) \equiv \frac{\pi_{t,G}}{\pi_{t,-G}}. \tag{2}$$

Definition 1 implies that DMs are attuned to log differences in probabilities: representativeness depends on the percentage probability increase of a type from $-G$ to G. This captures a form of diminishing sensitivity, whereby, for a fixed probability difference, a type is more likely to be overweighed if it is unlikely in the comparison group.[8] Thus, the representative age of a Floridian is 65+ because people in this age bracket are more common in Florida than in the rest of the US. Statistically, representative types are also diagnostic of the target group G. Indeed, the higher is $R(t, G, -G)$, the more confident is a Bayesian DM observing t that t belongs to G rather than to $-G$.[9]

The ease of recall of highly representative types affects judgments because more easily recalled types are overweighted. We model distorted recall as follows. Denote by $\mathbf{R}(t, G, -G) \equiv (\pi_{t,G}/\pi_{t,-G})_{t \in T}$ the vector of representativeness of all types in G. We then have:

Definition 2 *The DM attaches to each type $t \in T$ in group G a distorted probability:*

$$\pi^{st}_{t,G} = \pi_{t,G} \frac{h_t(\mathbf{R}(t, G, -G))}{\sum_{s \in T} \pi_{s,G} h_s(\mathbf{R}(t, G, -G))}, \tag{3}$$

where $h_t : R^T_+ \to \mathbb{R}_+$ is a weighting function such that:

1) The weight h_t is a symmetric function of the representativeness of types $s \neq t$. Formally, $h_t = h\left(\frac{\pi_{t,G}}{\pi_{t,-G}}; \left(\frac{\pi_{s,G}}{\pi_{s,-G}}\right)_{s \in T \setminus \{t\}}\right)$ where $h : R_+ \times R^{T-1}_+ \to \mathbb{R}_+$ is a function that is invariant to a permutation of the last $T-1$ arguments.

2) Weighing of a type increases in own representativeness and decreases in the representativeness of other types. Formally, the function $h(\cdot)$ is weakly increasing in its first argument,

[8] Our definition of representativeness links to Weber's law of sensory perception, see Section 2.2. It also links to our previous work on salience, in which we postulated that log differences in payoffs determine the attention to lottery payoffs, Bordalo, Gennaioli, and Shleifer (2012) and to goods' attributes (Bordalo, Gennaioli, and Shleifer 2013). Equation (2) establishes the same principle for the domain of probabilities.

[9] This insight led Tenenbaum and Griffiths (2001) to define representativeness as individuals' sense, as intuitive Bayesians, of updating in reaction to data. Their definition, like ours, is in terms of the likelihood ratio. However, Tenenbaum and Griffiths interpret representativeness as a mechanism that affects intuitive judgments of similarity, rather than beliefs (e.g. it accounts well for lab evidence where subjects are asked to rank types in terms of representativeness, or of strength of association with a group.) Accordingly, they do not consider the possibility of systematically distorted, and context dependent, beliefs.

and weakly decreasing in the other $T-1$ arguments.

We call the distribution $(\pi^{st}_{t,G})_{t\in T}$ the stereotype for G. If a type t is objectively more likely, namely $\pi_{t,G}$ is higher, then the stereotype attaches higher probability to it. By property 1), distortions are due exclusively to the the fact that a type is more or less representative than the others. In particular, if all types are equally representative, the DM equally weighs all of them at $h(1)$ and holds rational expectations about G. If instead the representativeness of different types differs, property 2) implies that the stereotype ceteris paribus overweights the probability of more representative types.

Most of the results we explore in this paper hold for a general weighing function $h_t(\cdot)$. Specific functional forms capture added assumptions about the psychology of representativeness-based recall, and are useful in applications. We outline a few specifications and their properties.

- Rank-based stereotypes: the ranking of the representativeness of different types shapes distortions. Denote by $r(t) \in \{1, ..., T\}$ the representativeness ranking of type t. When $r(t) = 1$ type t is the most representative one (potentially with ties). We can specify two ways in which a type's representativeness ranking distorts its probability.

  - Rank-based truncation: the DM only recalls the types that have representativeness ranking of at most d, namely $\{t \in T \,|r(t) \leq d\}$. Zero probability is attached to the remaining types.[10] Denote by $I(r(t) \leq d)$ an indicator function taking value 1 if the representativeness ranking of t is at most d. Then, the weighting function is $h_t = I(r(t) \leq d)$ so that:

$$\pi^{st}_{t,G} = \pi_{t,G} \frac{I(r(t) \leq d)}{\sum_{s\in T} \pi_{s,G} I(r(s) \leq d)},$$

    which is the true conditional probability within recalled types. This assumption

[10] These neglected types are not viewed as impossible; they are just assigned zero probability in the DM's current thinking. This formulation allows us to model surprise and reactions to unforeseen contingencies, which have proved useful ingredients in modeling probabilistic judgments (GS 2010) as well as neglect of risk in financial crises (Gennaioli, Shleifer, and Vishny 2012).

is used in Gennaioli and Shleifer (2010).[11]

– Rank-based discounting: The DM discounts by a constant factor $\delta \in [0,1]$ the odds of type t relative to its immediate predecessors in the representativeness ranking. Lower δ implies stronger discounting of less representative types. Formally, the weighting function is $h_t = \delta^{r(t)}$, so that:

$$\pi^{st}_{t,G} = \pi_{t,G} \frac{\delta^{r(t)}}{\sum_{s \in T} \pi_{s,G} \delta^{r(s)}}.$$

- Representativness based discounting: All else equal, the weight attached by the DM to type t increases continuously with its representativeness. One convenient formulation is $h_t = \left(\pi_{t,G}/\pi_{t,-G}\right)^{\theta}$ so that:

$$\pi^{st}_{t,G} = \pi_{t,G} \frac{\left(\pi_{t,G}/\pi_{t,-G}\right)^{\theta}}{\sum_{s \in T} \pi_{s,G} \left(\pi_{s,G}/\pi_{s,-G}\right)^{\theta}},$$

where $\theta \geq 0$ captures the extent to which representativeness distorts beliefs. This formulation is particularly convenient when dealing with continuous distribution of the exponential or power classes.

These functional forms all embody the main idea of our model that the stereotype overweights the probability of more representative types. Rank-based truncation captures a central manifestation of limited memory: forgetting unrepresentative types. Smoother discounting (based on ranking or on representativeness) may be more appropriate when the type space is small, and smooth discounting can be more tractable in certain settings.

Section 3 characterizes the general properties of stereotypes. In particular, it shows their ability to account for social psychologists's "kernel of truth" hypothesis under the general weighting function of Definition 2. To bring the model to the data in Section 4.2, we derive linear approximations of stereotypical beliefs by assuming that the weighting function is

[11] Specifically, in GS (2010) the assessed probability that a certain hypothesis G is true is equal to:

$$\Pr(G) = \frac{\sum_t \pi_{t,G} I(r(t) \leq d)}{\sum_t \pi_{t,G} I(r(t) \leq d) + \sum_t \pi_{t,\Omega/G} I(r(t) \leq d)}$$

which increases in the ratio between the total probability mass recalled for G and that recalled for $-G = \Omega \backslash G$.

differentiable with respect to a type's representativeness. This assumption excludes rank-based weighting but allows for many possibilities.

2.2 Discussion of Assumptions

Before moving to the formal analysis, we discuss some properties as well as limitations of our approach. Representativeness-based recall, the idea that individuals recall distinctive group types, can be viewed as an instance of what Kahneman and Tversky call "attribute substitution". When dealing with the difficult question "what is the distribution of hair color among the Irish?", people intuitively answer to the simpler question "which hair color distinguishes the Irish people?". Critically, as discussed by Kahneman and Tversky, attribute substitution does not occur because people misunderstand the original question, or mechanically confuse the assessment of $\Pr(t\,|G)$ with that of $\Pr(G\,|t)$. Rather, it occurs because the distinctive or representative types immediately come to mind, and individuals anchor their overall probability judgment to it. As a consequence, subjects do not only make mistakes in judging the probability that a Floridian is over 65. They also give too high an answer to the question "what is the average age of a Floridian?"[12]

One interesting question is whether the process of stereotyping we describe is optimal in some sense. Focusing mental representations on a few types can be justified by the costs of thinking or retrieval. This approach, however, is not enough to explain why individuals should focus on representative rather than likely types. We do not formally analyze the optimality of representativeness here, but mention some relevant considerations from cognitive psychology. Kahneman and Tversky (1979) stress the similarity between many perceptual and cognitive operations. For instance, the highlight of contrast – a key principle of visual perception – is invoked to justify the Prospect Theory assumption that the carriers of utility

[12]Indeed, in many cases mere confusion of $\Pr(t\,|G)$ with $\Pr(G\,|t)$ would not yield the phenomenon of stereotyping. In the Irish hair color example, the probability of being Irish conditional on having red hair is:

$$\Pr(Irish\,|red) = \frac{\Pr(red\,|Irish)\Pr(Irish)}{\Pr(red\,|Irish)\Pr(Irish) + \Pr(red\,|Non\ Irish)\Pr(Non\ Irish)}.$$

This probability is clearly very small, given that the Irish population is a tiny fraction of the world population. Confusion of $\Pr(t\,|G)$ with $\Pr(G\,|t)$ would in this case lead to an understatement of the probability of the red haired Irish.

are changes relative to a reference point. The same logic applies to our model, in which stereotypes precisely highlight the contrast between groups.

In visual perception, assessing properties such as brightness, color, size, or distance to an object by comparing them to other proximate objects has been shown to be optimal in the presence of multiplicative background noise (Kersten et al. 2004, Cunningham 2013). Our formulation of representativeness is related to the same idea, in the sense that individuals estimating properties of one group stress differences from another group. In a noisy world in which attention is limited, this process may optimally allow for swift reactions to changes in group characteristics, even if errors are sometimes made.[13] Exploring this idea formally is an interesting avenue for future work.

Consider now some limitations of our model. First, representativeness is not the only heuristic that shapes recall. Decision makers may for instance find it easier to recall types that are sufficiently likely. Another potentially important mechanism is availability, understood by Kahneman and Tversky (1972) as the "ease" with which information comes to mind (because of actual frequency or repetition). In Online Appendix E we present a truncation-based recall mechanism in which distortions are driven by a combination of representativeness and likelihood of types (which is equivalent to relaxing property 1 in Definition 2). This model can offer a useful starting point to capture availability as well, even though a full model of availability is beyond the scope of this paper. Even in this more general setting, the influence of representativeness on recall is the driving force of stereotypes that, in line with the social psychology perspective, are based on underlying differences among groups. As we show in Section 4, this feature is critical in accounting for the evidence.

The second set of model-related issues concerns how to specify the elements of Definition 1 in applications: group G, the type space T, and the reference group $-G$. Take the specification of the group G and of the type space T. Often, the problem itself provides a natural specification of these features. This is the case in the empirically important class of "closed end" questions, such as those used in surveys, which provide respondents with

[13] To give a simple example, suppose that – as in the case of Proposition 3 – the variance in the environment increases, in the sense that extreme tail events become more likely. Then, a likelihood-based stereotype would detect no change while a representativeness based stereotype would focus on the heightened probability of the tails. In particular, an asymmetric increase in tail probabilities that shifts the mean would be detected by a representativeness-based stereotype, even if the distribution's mode does not change.

a set of alternatives, as in the data we use in Section 4. More generally, the problem solved by the decision maker – such as evaluating the resume of a job applicant coming from a certain ethnic group – primes a group, a dimension of interest, and a set of types (e.g., the applicant's qualification or skill levels). When types have a natural order, such as income, age, or education, the granularity of T is also naturally given by the problem (income, age, and years of schooling brackets). When the set of types is not specified by the problem, decision makers spontaneously generate one.[14] It would be useful to have a model of which dimensions and types come to mind, particularly for more open ended problems. Psychologists have sought to construct a theory of natural types and dimensions (Rosch 1998). We do not make a contribution to this problem, but note that in many problems of interest in economics the dimension as well as the set of types is naturally given. Furthermore, in our model details of the type space can be important under rank-based truncation, but they matter less under smooth discounting.

Consider finally the role of the comparison group $-G$. This group captures the context in which a stereotype is formed and, again, is often implied by the problem: when $G =$ Floridians, $-G =$Rest of US population; when $G =$ African Americans, $-G =$White Americans. A distinctive prediction of our model, confirmed by our experiments in Section 4.1, is that the stereotype for a given group G depends on the comparison group $-G$.[15] When $-G$ is not pinned down by the problem itself, to derive testable predictions from representativeness, we set $-G = \Omega \backslash G$ where Ω is the natural population over which the unconditional distribution of types is measured.

[14]For example, suppose a person is asked to guess the typical occupation of a democratic voter in an "open ended" format (without being provided with a set of alternatives). Here the level of granularity at which types are defined is not obvious (e.g. teacher vs a university teacher vs a professor of comparative literature).

[15]Some empirical papers have taken a similar approach, exogenously varying the natural comparison group through priming. Benjamin, Choi, and Strickland (2009) show that priming racial or ethnic identity can impact the risk preferences of participants. Chen et al (2014) find that Asian students cooperate less with outgroup members when primed with their ethnic identity rather than their university identity. Shih et al (1999) show that Asian-American women self-stereotype themselves as better or worse in math, with corresponding impact on performance, when their ethnicity or gender, respectively, is primed. Shih et al (2006) replicate this effect using a verbal task, documenting that Asian-American women performed better when their gender rather than their ethnicity was primed. While the generalizability and replicability of priming has been doubted (Klein et al 2014), this body of evidence is consistent with context dependence.

3 Properties of Stereotypes

We now study stereotypical beliefs and their accuracy. To illustrate the role of representativeness, we first ask to what extent the most representative type is a good fit for the group, namely whether it is modal. Next, we assess the accuracy of the entire stereotypical distribution. To do so, we focus on a cardinal types and compare the stereotype's mean and variance to the true ones.

3.1 Likely vs Unlikely Exemplars

The most representative type for a group is the one that agents most easily recall and associate with the group itself, for instance a red-haired Irishman or a 65+ year old Floridian. Social psychologists call this type the exemplar of the group. Accordingly, we define:

Definition 3 *A type t^* is an exemplar for G given comparison group $-G$ when:*

$$t^* \in \arg\max_t \frac{\pi_{t,G}}{\pi_{t,-G}}.$$

Under any specification of the weighting function h_t in Definition 2, overweighting (weakly) increases as we move toward more representative types, so the exemplar is also the type whose probability is overweighted the most.[16] By analyzing the exemplar, then, we can gauge whether representativeness induces the DM to overweight a likely type (as it happens standard models of categorical thinking) or an unlikely type. When overweighting occurs in unlikely and extreme types, the biases of stereotypes can be particularly severe.[17]

Equation (2) yields the following characterization.

Proposition 1 *Suppose the conditional distributions π_G and π_{-G} are not identical. Consider two extreme cases:*

[16] Consider the function $h(\cdot)$ from Definition 2. When applied to more more representative types, the first argument of the function increases, while one of the other $T-1$ arguments decreases. As a result the weighting factor h_t (and thus overweighing $\pi^{st}_{t,G}/\pi_{t,G}$) increases as well.

[17] In the rank-based truncation model, the frequency of the exemplar provides a measure of stereotype accuracy. By accuracy, we mean the extent to which the stereotype minimizes the distance $\sum_t(\pi^{st}_{t,G}-\pi_{t,G})^2$. When $d=1$ and only one type is recalled (there are no ties), accuracy is maximized if the exemplar is the most likely type and minimized if the exemplar is the least likely type.

i) If for all t, $t' \in T$ we have that $\pi_{t,G} > \pi_{t',G}$ if and only if $\pi_{t,-G} > \pi_{t',-G}$, then the modal type is not an exemplar for at least one group.

ii) If for all t, $t' \in T$ we have that $\pi_{t,G} > \pi_{t',G}$ if and only if $\pi_{t,-G} < \pi_{t',-G}$, then for each group the modal type is the exemplar.

Case i) says that when groups have similar distributions, in the sense of having the same likelihood ranking, the most representative type is unlikely for at least one group, potentially for both. Representativeness draws the DM's attention to group differences, neglecting the fact that the groups are similar, and have the same mode. This mechanism generates inaccurate stereotypes and is illustrated by the Florida example. This result holds under any measure of representativeness that differentiates the two groups (for instance, the difference $\pi_{t,G} - \pi_{t,-G}$).

Case ii) says that the most representative type tends to be likely for both groups when the distributions are very different. In this case, groups differ the most around their modes, so representativeness and likelihood coincide. Thinking of Swedes as "blond haired" and Europeans as "dark haired" is accurate precisely because these are majority traits of the Swedish and European populations, respectively. In these cases, stereotyping yields fairly reliable models. Of course, there is still some inaccuracy. Even in the case of likely exemplars, judgment errors can be significant. For instance, voters in some U.S. states are perceived as "blue" or "red" because a majority of the population indeed votes Democrat or Republican. In reality, even in "blue" states, far from everyone votes Democrat. In the 2012 Presidential election, vote shares of either candidate in most states ranged from 40% to 60%.[18]

When DMs strongly overweight representative types, the most severe biases occur when those types are unlikely and extreme. This is true both under rank based truncations and under smooth discounting functions (see Section 3.2). Ethnic stereotypes based on crime or terrorism exhibit this error: they neglect the fact that by far the most common types in all groups are honest and peaceful.

[18]See https://en.wikipedia.org/wiki/United_States_presidential_election,_2012, section on votes by electoral college.

3.2 Stereotypical Moments

We now characterize how the first two moments of a distribution are distorted by the process of stereotyping. To do so, we must restrict our analysis to cardinal, ordered types. The following results hold for any weighting function $h_t(\cdot)$ satisfying Definition 2. We consider two canonical cases that prove useful in illustrating the predictions of the model.

In the first case, groups G and $-G$ are such that the likelihood ratio $\pi_{t,G}/\pi_{t,-G}$ is monotonic in t. The monotone likelihood ratio property (MLRP) holds to a first approximation in many empirical settings and is also assumed in many economic models, such as standard agency models.[19] If $\pi_{t,G}/\pi_{t,-G}$ is monotonically increasing (decreasing) in t, then group G is associated with higher (lower) values of t relative to the comparison group $-G$. Formally:

Proposition 2 *Suppose that MLRP holds, and assume w.l.o.g. that the likelihood ratio* $\frac{\pi_{t,G}}{\pi_{t,-G}}$ *is strictly increasing in* t*. Then, for any weighting function* $h_t(\cdot)$ *that is not constant in the relevant range:*

$$\mathbb{E}^{st}(t|G) > \mathbb{E}(t|G) > \mathbb{E}(t\,|-G) > \mathbb{E}^{st}(t|-G).$$

Under MLRP, the most representative part of the distribution for G is the right tail if $\pi_{t,G}/\pi_{t,-G}$ increases in t or the left tail if $\pi_{t,G}/\pi_{t,-G}$ decreases in t. The representative tail is then overweighted while the non-representative tail is underweighted. As a consequence, the assessed mean $\mathbb{E}^{st}(t|G)$ is too extreme in the direction of the representative tail.

Critically, in line with the social cognition perspective, the stereotype contains a kernel of truth: the DM overestimates the mean of G if this group has a higher mean than the comparison group, namely $\mathbb{E}\,(t\,|G) > \mathbb{E}\,(t\,|-G)$ and conversely if $\mathbb{E}\,(t\,|G) < \mathbb{E}\,(t\,|-G)$. The DM exaggerates this true difference because he inflates the association of G with its most representative types.[20] For instance, when judging an asset manager who performs well,

[19] Examples include the Binomial and the Poisson families of distributions with different parameters. The characterisation of distributions satisfying MLRP is easier in the case of continuous distributions, see Online Appendix D: two distributions $f(x)$, $f(x-\theta)$ that differ only in their mean satisfy MLRP if and only if the distribution $f(x)$ is log-concave. Examples include the Exponential and Normal distributions. To the extent that discrete distributions sufficiently approximate these distributions (as the Poisson distribution $Pois(\lambda)$ approximates the Normal distribution $N(\lambda,\lambda)$ for large λ), they will also satisfy MLRP.

[20] Depending on the distribution and the weighing function, the DM's assessment of the variance $\mathrm{Var}(t|G)$ may also be dampened relative to the truth. This is often true under the truncation weighing function. In

we tend to over-emphasize skill relative to luck because higher skill levels are relatively more associated with higher performance. This occurs even if for both skilled and unskilled managers high performance is mostly due to luck.

In the second case for which we charaterize the stereotypical distributions, groups G and $-G$ have the same mean $\mathbb{E}(t\,|G) = \mathbb{E}(t\,|-G) = \mathbb{E}(t)$ but differ in their variance. We abstract from skewness and higher moments by considering distributions $(\pi_{t,G})_{t\in T}$ and $(\pi_{t,-G})_{t\in T}$ that share the same support and are both symmetric around the median/mean $\mathbb{E}(t)$.

Proposition 3 *Suppose that in G more extreme types are relatively more frequent than in $-G$. Formally, the likelihood ratio $\frac{\pi_{t,G}}{\pi_{t,-G}}$ is U-shaped in t around $\mathbb{E}(t)$. Then, for any weighting function $h_t(\cdot)$ that is not constant in the relevant range, stereotypical beliefs satisfy:*

$$\begin{aligned} Var^{st}(t|G) &> Var(t|G) > Var(t|-G) > Var^{st}(t|-G), \\ \mathbb{E}^{st}(t|G) &= \mathbb{E}^{st}(t|-G) = \mathbb{E}(t). \end{aligned}$$

When group G has a higher relative prevalence of extreme types, its representative types are located at both extremes of the distribution. The DM's beliefs about G are then formed by overweighting both tails while underweighting the unrepresentative middle. The overweighting of G's tails causes the assessment of its variance $Var^{st}(t|G)$ to be too high. For example, the skill distribution of immigrants to the US may be perceived as having very fat tails, or even bimodal, with immigrants being perceived as either unskilled or very skilled relative to the native population. The mean of the group, in contrast, is assessed correctly, because the stereotypical distribution remains symmetric around $\mathbb{E}(t)$. As before, the stereotype contains a kernel of truth. It induces the agent to exaggerate the true differences between groups, namely the higher variance of G relative to its counterpart.

We present a number of extensions of the model in the Online Appendix. We first consider multi-dimensional type spaces, and show that stereotypes center around the dimension where groups differ the most, in line with the kernel of truth logic (Appendix C). Multidimensional

this case, stereotyping effectively leads to a form of overconfidence in which the DM both holds extreme views and overestimates the precision of his assessment. That extreme views and overconfidence (in the sense of over precision) go together has been documented in the setting of political ideology, among others (Ortoleva and Snowberg 2015).

stereotypes imply that the dimension we think about is influenced by context dependent. For example, the Irish are stereotyped as red-haired when compared to the European population. However, when compared to the Scots, a more plausible stereotype for the Irish is "Catholic" because religion is the dimension along which Irish and Scots differ the most.

In Appendix D we extend the model to continuous type spaces. Many settings of interest in economics can be usefully described by continuous probability distributions, and we show our model is particularly tractable in this case. In Appendix E, we relax Definition 2 and allow weighting of types to also be influenced by their likelihood. We show that the basic insight that stereotypes contain a kernel of truth carries through to these cases as well.

To summarise, the psychology of representativeness yields stereotypes that are consistent with the social cognition approach in which individuals assess groups by recalling and focusing on distinctive group traits. When there are systematic differences between groups, stereotypes get the direction right, but exaggerate differences.

3.3 Some Examples

A growing body of field and experimental evidence points to a widespread belief that women are worse than men at mathematics (Eccles, Jacobs, and Harold 1990, Guiso, Monte, Sapienza and Zingales 2008, Carrell, Page and West 2010). This belief persists despite the fact that, for decades, women have been gaining ground in average school grades, including mathematics, and have recently surpassed men in overall school performance (Goldin, Katz and Kuziemko 2006, Hyde et al 2008). This belief, shared by both men and women (Reuben, Sapienza and Zingales 2014), may help account, in part, for the gender gap in the choices of high school tracks, of college degrees and of careers, with women disproportionately choosing humanities and health related areas (Weinberger 2005, Buser, Niederle and Oosterbeek 2014) and foregoing significant wage premiums to quantitative skills (Bertrand 2011).

Gender stereotypes in mathematics, particularly beliefs that exaggerate the extent of average differences, are consistent with the predictions of our model. The fact that men are over-representated at the very highest performance levels leads a stereotypical thinker to exaggerate the magnitude of mean differences. Figure I shows the score distributions from the mathematics section of 2013's Scholastic Aptitude Test (SAT), for both men and

women.[21] The distributions are very similar, with average scores being slightly higher for men (531 versus 499 out of 800). However, scores for men have a heavier right tail, with men twice as likely to have a perfect SAT math score than women.[22] In light of such data, the stereotypical male performance in mathematics is high, while the stereotypical female performance is poor. Predictions based on such stereotypes are inaccurate, exaggerating true differences. Consistent with this prediction, experimental evidence shows that both genders underestimate women's ability in simple math tasks, even controlling for past performance (Reuben et al. 2014). Coffman (2014) shows a similar pattern extends to confidence about own ability in other male-typed domains, with women reporting significantly less confidence in gender-incongruent than gender-congruent tasks. Our model suggests that these patterns might come from stereotypes based on gender differences in the right tail of the distribution. While differences in the right tail of the distribution are unlikely to be relevant for most decisions, stereotypical thinking driven by these differences has the potential to impact economically-important decisions, whether through self-stereotyping (i.e., choice of careers or majors as in Buser, Niederle, Osterbeek 2014) or through discrimination (i.e., hiring decisions as in Bohnet, van Geen, and Bazerman 2015).

The logic of exaggerated, yet directionally correct, stereotypes can also shed light on the well documented phenomenon of base rate neglect (Kahneman and Tversky, 1973). Indeed, Proposition 2 implies that the DM overreacts to information that assigns people to groups, precisely because such information generates extreme stereotypes. Consider the classic example in which a medical test for a particular disease with a 5% prevalence has a 90% rate of true positives and a 5% rate of false positives. The test assigns each person to one of two groups, $+$ (positive test) or $-$ (negative test). The DM estimates the frequency of the sick type (s) and the healthy type (h) in each group. The test is informative: a positive

[21] Standardized test performance measures not only innate ability but also effort and investment by third parties, Hyde et al (2008). The mapping of test performance into inferences about innate ability is an issue not addressed by our model.

[22] For 2013 SAT Mathematics scores, see http://media.collegeboard.com/digitalServices/pdf/research/SAT-Percentile-Ranks-By-Gender-Ethnicity-2013.pdf. Results are similar for the National Assessment of Educational Progress (NAEP), which are more representative of the overall population. For 2012 NAEP scores for 17 year olds in mathematics, see http://nationsreportcard.gov/ltt_2012/age17m.aspx. See Hyde et al (2008), Fryer and Levitt (2010), and Pope and Sydnor (2010) for in-depth empirical analyses of the gender gap in mathematics.

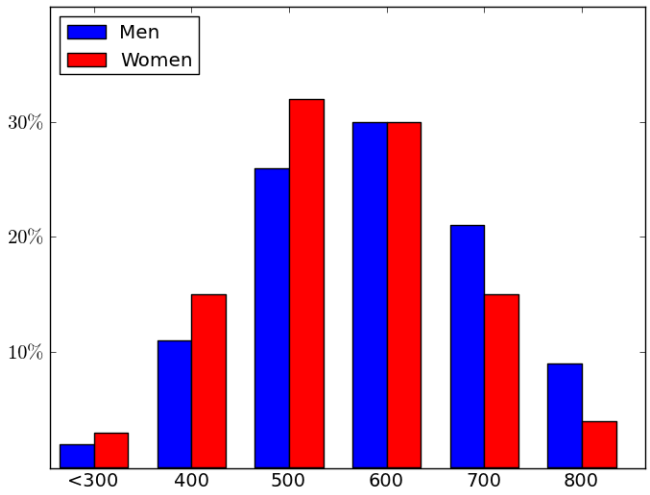


Figure I: SAT Mathematics scores by gender (2013)

result increases the relative likelihood of sickness, and a negative result increases the relative likelihood of health for any prior. Formally:

$$\frac{\Pr(+|s)}{\Pr(+|h)} > 1 > \frac{\Pr(-|s)}{\Pr(-|h)}. \tag{4}$$

This condition has clear implications: the representative person who tests positive is sick, while the representative person who tests negative is healthy. Following Proposition 2, the DM reacts to the test by moving his priors too far in the right direction, generating extreme stereotypes. He greatly boosts his assessment that a positively tested person is sick, but also that a negatively tested person is healthy. Because most people are healthy, the DM's assessment about the group that tested negative is fairly accurate but is severely biased for the group that tested positive. This analysis formalises Tversky and Kahneman's (1983) verbal account of base rate neglect.[23]

[23]Our account is distinct from a mechanical underweighting of base-rates in Bayes rule, as in Grether (1980) and Bodoh-Creed, Benjamin and Rabin (2013). In those models, upon receiving the test results, the DM can update his beliefs in the wrong direction: he can be less confident that a person is healthy after a negative test than under his prior, which cannot happen in our model.

While this prediction of our model seems consistent with introspection, we are not aware of experimental evidence on this point. Griffin and Tversky (1992) present evidence consistent with pure neglect of base rates, but in a significantly different task, namely inferring the bias of a coin from a history of coin flips. Such experiments are hard to compare with the predictions of our model, because subjects are asked to generate distributions of different numbers of coin flips in their minds, which is a much more involved task

The same mechanism may underlie several other instances where decision makers overweight diagnostic information. When assessing an employee's skill level, an employer attributes high performance to high skill, because high performance is the distinctive mark of a talented employee. Because he neglects the possibility that some talented employees perform poorly and that some non-talented ones perform well (perhaps due to stochasticity in the environment), the employer has too much faith in skill, and neglects the role of luck in accounting for the output. Similarly, when assessing the performance of firms in a hot sector of the economy (e.g. $G =$ internet), an investor recalls highly successful firms in that sector ($t =$ return of Google, Amazon, etc.). However, he neglects the possibility of firms being unsuccessful, because lack of success is statistically non-diagnostic, and psychologically non-representative, of a growing sector – even if it is likely. This causes both excessive optimism (in that the expectation of growth is unreasonably high) and overconfidence (in that the variability in earnings growth considered possible is truncated). True, the hot sector may have better growth opportunities on average, but representativeness exaggerates this feature and induces the investor to neglect a significant risk of failure.

4 Evidence on Representativeness and Stereotypes

Testing our model requires evaluating the beliefs individuals hold about a target group against the true distribution of that target group over an attribute space. An ideal data set would consist of naturally defined groups with known distributions over a given space of types, and a corresponding set of beliefs about the distribution of each group over that space of types. This would allow us to test for exaggeration of true differences between groups and to ask whether this exaggeration is well-predicted by overweighting of representative types. To identify a causal role for representativeness or context dependence, we would need exogenous variation in either the comparison group, $-G$, or in the distributions over the type spaces. This is unavailable in existing data sets. Accordingly, we take a two-pronged approach. First, we create a controlled laboratory environment that allows us to induce

than to recall types of a given distribution. Their assessments, then, might be wrong for other reasons. See Bodoh-Creed, Benjamin and Rabin (2013) for a detailed discussion.

the exogenous variation in representative types that we need to test causality, and second, we re-analyze existing field evidence to check for consistency with our predictions. To our knowledge, the prediction that representativeness generates context dependence has not been tested before.

In testing our model, we focus on the two main implications of representativeness-based stereotypes:

- Context dependence: the stereotype of a target group depends on the characteristics of the reference group it is compared to.
- Kernel of truth: stereotypes depend on group characteristics, and – in most (precisely characterized) settings – are slanted toward representative types.

We test the first property with a lab experiment (Section 4.1). We then turn to survey data on beliefs about U.S. political groups (Section 4.2) for an empirical analysis that explores the second property. The survey data is more tightly linked to our interest in social stereotypes. The laboratory experiment, however, allows us to directly test the role of representativeness in generating context dependent beliefs. Online Appendices F and G provide all details, and additional results, for the experiments and field evidence.

4.1 Lab Evidence on Representativeness and Context Dependence

The influence of the representativeness heuristic on recall and on beliefs has been extensively documented in the lab (Kahneman and Tversky 1972, 1983). Our goal here is to consider how representativeness as formalized in Equation (2) gives rise to context dependent beliefs. To our knowledge, the possibility that representativeness may generate context dependence has not been tested before.

To assess this prediction, we perform a controlled laboratory experiment that allows us to isolate representativeness from many confounding factors – historical, sociological, or otherwise – that may affect stereotype formation in the real world. We construct our own groups of ordinary objects, creating a target group, G, and a comparison group, $-G$. We hold the target group G fixed, but explore how participant impressions of it change as we change the comparison group $-G$, and hence representativeness.

We conducted several experiments, in the laboratory as well as on Amazon Mechanical Turk. Each involves a basic three-step design. First, participants are shown the target group and a randomly-assigned comparison group for 15 seconds. In this time, differences between groups can be noticed but the groups' precise compositions cannot be memorized. The second step consists of a few filler questions, which briefly draw the participants' cognitive bandwidth away from their observation. Finally, participants are asked to assess the groups they saw. Participants are incentivized to provide accurate answers.

We randomly assign participants to either the Control or the Representativeness condition. In the Control condition, G and $-G$ have nearly identical distributions, so that all types are similarly representative for each group. In the Representativeness condition, $-G$ is changed in such a way that a certain type becomes very representative for G. Context dependence implies that the assessment of G should now overweight this representative type, even though the distribution of G itself has not changed.

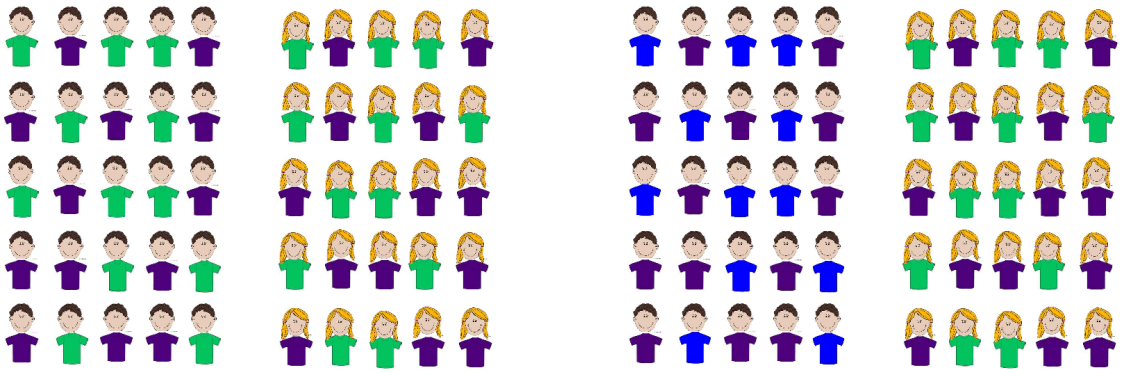
We ran six experiments of this form, with design changes focused on reducing participant confusion and removing confounds. Here, we describe the final, and most refined, version of these experiments. In an attempt to provide a overview of the results while remaining concise, we also provide the results from pooled specifications that use all data collected. In Online Appendix F, we present additional details and report all experiments conducted. We also provide instructions and materials for each experiment and the full data set.

Consider first the experiment illustrated in Figure II. A group of 25 cartoon girls is presented next to a group of 25 cartoon boys in t-shirts of different colors: blue, green, or purple. In the Control condition, Fig.IIa, the groups have identical color distributions (13 purple, 12 green), so no color is representative of either group. The Representativeness condition, Fig.IIb, compares the *same* group of girls with a different group of boys, for whom green shirts are replaced by blue shirts. Now only girls wear green and only boys wear blue. These colors, while still not the most frequent for either group, are now most representative. For each group, girls and boys, participants are asked a number of questions concerning the frequency of T-shirts of different colors worn by that group.

Applying our model, in the control condition the type space is $T = \{\text{green, purple}\}$, and the groups are $G = $ girls, and $-G = $ boys. Given that the color distributions are

identical across groups, both types are equally representative, $\pi_{green,girls}/\pi_{green,boys} = 1 = \pi_{purple,girls}/\pi_{purple,boys}$. As a result, assessment of G should be on average correct, $\pi^{st,control}_{green,girls} = \pi_{green,girls}$ and $\pi^{st,control}_{purple,girls} = \pi_{purple,girls}$ for any weighing function (and the same is true about assessments of $-G$).

In the treatment condition, the distribution of shirt colors remains the same for girls. For boys, green shirts are changed into blue. Thus, the type space changes to $T = \{\text{green, purple, blue}\}$ and the representative color for girls becomes green, $\pi_{green,girls}/\pi_{green,boys} = \infty > 1 = \pi_{purple,girls}/\pi_{purple,boys}$, while that for buys becomes blue. As a result, in the treatment condition subjects should inflate the frequency of green shirts relative to the truth, $\pi^{st,treatment}_{green,girls} > \pi_{green,girls}$ (and the same should happen to assessments of blue shirts for boys). We also expect the assessed frequency of green shirts to go up relative to the control condition, namely $\pi^{st,treatment}_{green,girls} > \pi^{st,control}_{green,girls}$. Critically, the only factor that varies across treatments is the representativeness of the 12-color shirt. Thus, if we see differences across conditions, the causal role of representativeness-based recall in shaping group judgments is clear.[24]



(a) Control Condition (b) Representativeness Condition

Figure II: T-shirts Experiment

We collected data from 301 participants using this T-shirts design.[25] Since the number of green and purple shirts is very similar, we first ask subject the simplest question of which

[24]We vary which colors are used in which roles across participants. Some participants saw this particular color distribution, while others see, for example, green as the modal color, with purple as the diagnostic color for boys in the Rep. condition and blue as the diagnostic color for girls in the Rep. condition. We vary the colors across the roles to avoid confounding the characteristics of any particular color with its diagnosticity.

[25]Throughout our analysis, we exclude any participant who participated in a previous version of the experiment and any participant who self-identified as color blind. In Appendix F, we show that our results are unchanged if we include these additional observations.

shirt color is modal. Next, we ask subjects to assess the share of green and purple shirts.

Consistent with the role of representativeness, participants assigned to the Representativeness condition are 10.5 percentage points more likely to recall the less frequent color, green for girls or blue for boys, as the modal color when it is representative of a group (35% of participants guess the less frequent color is modal in the Control condition, this proportion increases to 46% in the Representativeness condition, p=0.01, estimated from a probit regression reported in Appendix F).

Let us now turn to subjects' estimates of how many T-shirts of each color they saw in each group. In both conditions, the true difference in counts is one (13 purple shirts, 12 green or blue shirts). In the Control condition, participants on average believe they saw 0.54 more purple shirts than green or blue shirts. In the Representativeness condition, participants believe they saw 0.72 fewer purple shirts than green or blue shirts (the across treatment difference is significant with p=0.01 from OLS regression reported in Appendix F).

In total, we collected data for six experiments of this general structure, gathering evidence from more than 1,000 participants. As we describe in Appendix F, while there is substantial variation across experiments, when we pool all data collected we find significant aggregate treatment effects in line with a role of representativeness in judgment. We employ four different unordered types designs, similar to the T-shirts experiment, using six samples (four online and two in the laboratory) with 741 participants. We find effects in the predicted direction for five of the six samples. Using a probit regression that pools all of the data for these unordered type experiments, we find that participants are 9.3 percentage points more likely to guess that the less frequent type is modal when it is representative than when it is not (p=0.002). We also run a family of ordered types experiments (two designs, five samples, 402 participants). Unlike the simpler T-shirts style design, the theoretical predictions for the ordered types designs are more sensitive to the specific assumptions one makes about the weighting function of the decision-maker.[26] We find effects in the predicted direction for three of the fives samples. As we discuss in more detail in Appendix F, the ordered types results vary by platform, with consistently stronger results on Amazon Mechanical

[26]This is in part due to our choice to not study MLRP distributions. We motivate and discuss this design decision in Appendix F.

Turk than in the laboratory samples. Pooling all ordered types experiments, participants are 9.3 percentage points more likely to guess that the group of interest has a greater average than the comparison group when the right tail is representative (p=0.062). Given our simple experimental setting with groups of mundane objects, we interpret our results – a significant and reasonably-sized impact on average beliefs – as an important proof of concept: the presence of representative types biases ex post assessment.

4.2 Empirical Evidence on Political Stereotypes

We examine two data sets on political preferences, and beliefs about political preferences, in the U.S. We investigate the roles of representativeness and context dependence by separately testing for hypotheses that allow us to assess the leading theories of stereotypes.

First, we test whether beliefs are correct or depart systematically from the truth. The statistical discrimination approach builds on the assumption that people hold rational expectations of group traits. Comparing beliefs to the truth allows us to assess the validity of this assumption in our data.

Second, we test if beliefs depart from the truth by exaggerating (mean) differences among groups, as per the kernel of truth hypothesis. This is an implicit test of context dependence, because it implies that beliefs about the target groups are shaped not only by that group's characteristics, but also by those of the reference group.[27]

Third, we test if distortions in beliefs can be accounted for by the overweighting of highly representative types (defined as types that are relatively more frequent in the target relative to the reference group). The second and third tests address the key predictions of our model.

4.2.1 The data

We have two data sets on political preferences and beliefs about political preferences. The first data set, from Graham et al (GNH 2012), contains data from the Moral Foundations Questionnaire. Respondents (1,174 self-identified liberals and 500 self-identified conservatives) answer questions about their position on a subset of 45 issues: 20 moral relevance

[27]Of course, unlike in the laboratory experiment, in this setting we cannot test for context dependence by exogenously varying the reference group.

statements (e.g., "when you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?") and 25 moral judgments (e.g., "indicate the extent to which you would agree or disagree"). For each issue, a randomly determined subset of participants states their own position, another subset states their belief on the position of a "typical liberal", and a third subset states their belief on the position of a "typical conservative". The data thus includes the distribution over positions for both liberal and conservatives, as well as the average believed typical position of liberals and of conservatives, on each of the 45 issues. Each position is elicited on 1 - 6 scale.

The second data set comes from the American National Election Survey (ANES), and contains data from more than 20,000 respondents between 1964 and 2012.[28] The survey covers political issues of the day, such as the optimal amount of government spending and service provision (1984 through 2000), or the proper place of women in society (1972 through 1998). We focus on the 10 issues that ask participants to respond on a multi-point, 1 to 7, scale (rather than just indicate binary agreement or disagreement); each of these 10 issues is asked in multiple years. Participants are asked to provide their own position on the scale and their believed position of the Democratic and Republican party ("Where would you place the Democratic (Republican) party on this scale?"). The data includes, for each issue-year observation, the distribution of participant positions for both self-identified liberals and self-identified conservatives, as well as the distribution of believed typical positions of the Democratic and Republican Parties.

4.2.2 Empirical strategy and results

Our analysis focuses on beliefs about two groups, Conservatives and Liberals. The types are the possible positions for each issue (1, 2, ... , 6, 7). For the GNH data, we interpret beliefs about the "typical" element of a group to coincide with the believed average position in that group. Similarly, for the ANES data we use the believed party positions as a proxy for believed mean of each group.[29] We then take as a benchmark the hypothesis that individuals

[28]This data is publicly available at http://www.electionstudies.org/studypages/anes_timeseries_cdf/anes_timeseries_cdf.htm.

[29]This assumption is consistent with the authors' interpretation of the GNH data (GNH 2012) and with previous studies using ANES (e.g., Westfall et al, WBCJ 2015). Furthermore, to the extent that this assumption holds equally well for most issues within a data set, our focus on across-issue differences should

hold accurate beliefs about each group, and in particular that believed mean position should equal true mean position, at least on average across subjects. The accurate beliefs hypothesis underlies the most common economic model of stereotyping, statistical discrimination.

To assess our representativeness-based model, we perform a regression exercise. To test our model in a linear regression framework, we rely on linear approximations of the weighting function. Our model then yields two regression specifications.

Proposition 4 *Let $G \in \{conservative,\ liberal\}$, and let $h_t \equiv h\left(\pi_{t,G}/\pi_{t,-G}\right)$ be a differentiable and strictly increasing weighting function as in Definition 2. The following hold as a first order approximation around identical distributions $\pi_G/\pi_{-G} = 1$.*

1) Kernel of truth regression. There exists a constant $\kappa > 0$ such that:

$$\mathbb{E}^{st}(t|G) = \mathbb{E}(t|G)\left(1+\kappa\right) - \kappa \cdot \mathbb{E}(t|-G). \tag{5}$$

2) Representativeness regression. Denote $H = \{T-2, \ldots, T\}$ the right tail of types and $R_H^{cons} = \sum_H \pi_{t,cons} / \sum_H \pi_{t,lib}$ as the average representativeness of right tail types for conservatives. Under the further approximation where $\frac{\pi_{t,cons}}{\pi_{t,lib}} \approx R_H^{cons}$ for $t \in H$:

$$\begin{aligned} \mathbb{E}^{st}(t|cons) &= \mathbb{E}(t|cons) + \lambda_{cons}\left(R_H^{cons} - 1\right), & (6) \\ \mathbb{E}^{st}(t|lib) &= \mathbb{E}(t|lib) - \lambda_{lib}\left(R_H^{cons} - 1\right), & (7) \end{aligned}$$

where λ_{cons} and λ_{lib} are positive constants.

The first regression allows us to test for the kernel of truth hypothesis, while the second set of regressions allows us to test for the role of representativeness.

Equation (5) says that respondents in our model inflate the average position of a group, say the conservatives, if and only if the group has a higher average position than the other group, namely the liberals. Formally, $\mathbb{E}^{st}(t|cons) > \mathbb{E}(t|cons)$ if and only if $\mathbb{E}(t|cons) >$

allow us to test the predictions of our model even with an imperfect proxy for beliefs of mean positions. Finally, the data provides some insight into whether subjects are reporting (perceived) modal or mean types. As we show below, the modal type is a poor prediction of stated beliefs, while a distorted mean slanted towards representative types is an accurate prediction of stated beliefs.

$\mathbb{E}(t|lib)$. Because in our measurement scale higher types mean "more conservative", we expect: i) believed conservative average to be higher than the truth, and ii) the extent of overstatement to decrease in the average liberal position $\mathbb{E}(t|lib)$. Conversely, we expect the average liberal position to be lower than the truth, the more so the higher the average conservative position $\mathbb{E}(t|cons)$.

As previously discussed, the basis of these predictions is context dependence: information about the distribution of $-G$ is relevant for the beliefs about G. This context dependence is inconsistent with rational expectations, in which only the group's own means should affect beliefs. We test the hypothesis that the true mean $\mathbb{E}(t|G)$ is a significant predictor of the believed mean $\mathbb{E}^{st}(t|G)$ with a positive sign, while the other group's true mean $\mathbb{E}(t|-G)$ is a predictor of the believed mean with a negative sign.

Equations (6) and (7) say that repondents' assessment bias is shaped by representativeness. When the right tail is more representative for conservatives ($R_H^{cons} - 1$ positive and large), participants should inflate the average conservative position more (higher $\mathbb{E}^{st}(t|cons) - \mathbb{E}(t|cons)$) and deflate the average liberal position more (lower $\mathbb{E}^{st}(t|lib) - \mathbb{E}(t|lib)$). We test the hypothesis that the inflation in conservative positions is positively associated with the representativeness of the right tail for the conservatives, while the inflation in liberal positions is negatively associated with it. Once again, the representativeness of the right tail is computed using the true distribution of positions.

In Equations (6) and (7), in many cases the representative tail is also the most likely one. As a consequence, these tests cannot distinguish a representativeness-based from a likelihood-based model of distorted beliefs. We perform two additional tests. First, we run versions of Equations (6) and (7) in which we control for the likelihood of tails (see Table A4 in Online Appendix G). Second, we compute numerically the predictions of a representativeness-based model of stereotypes and of a likelihood-based model of stereotypes. We then assess which of these two is better able to match the data on beliefs.

4.2.3 Empirical Results

To begin, we illustrate the structure of the data and the nature of our predictions with two simple examples from the GNH data set, focusing on beliefs about conservatives. In Example

1, participants are asked about their agreement with the statement, "It can never be right to kill a human being". In Example 2, participants are asked about the moral relevance of "whether or not someone cared for someone weak or vulnerable". As can be seen in Figure III, in Example 1 the modal position (Strongly Disagree (1)) and most representative positions (Strongly Disagree (1)) coincide for conservatives. In contrast, in Example 2 in Figure III, the most representative types (Slightly Relevant (3), Not at all Relevant (1)) are not most likely for the conservative group. Following Proposition 1, we predict that beliefs will be distorted in the direction of the most representative types. Thus, we expect more exaggeration in Example 2 than in Example 1, since in Example 2 the most representative types (in the left tail) are far from the modal type, while in Example 1, they coincide. This is what we find: the conservative position is exaggerated by only 0.09 positions in Example 1 (true mean 2.99, believed mean 2.90), but by 1.06 positions in Example 2 (true mean 4.21, believed mean 3.15).

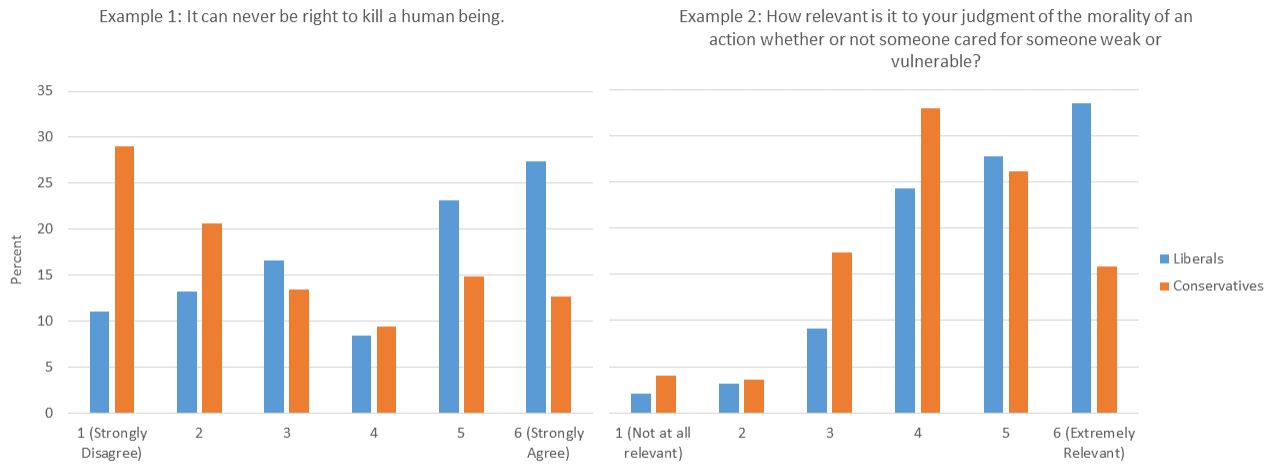


Figure III: Two Examples

In the full data sets, we treat each (issue, year) pair as an observation, and we cluster standard errors at the issue level. For the GNH data, we have 45 observations: 45 issues

each measured in the same year. For the ANES data, we have 66 observations: 10 issues, each measured in multiple years. To begin, we simply document systematic exaggeration in both data sets. This is a primary focus of the original analysis in GNH (2012), and also in WBCJ (2015)'s analysis of the ANES data.[30] Figure IV shows that the believed difference between typical conservative and typical liberal positions is larger than the true difference in mean positions for 109 of the 111 observations. The data for both GNH (purple squares) and ANES (orange triangles) lie above the 45 degree line (dashed).[31] Average exaggeration is 0.62 positions on the scale (0.66 in the GNH data, 0.59 in the ANES data).[32]

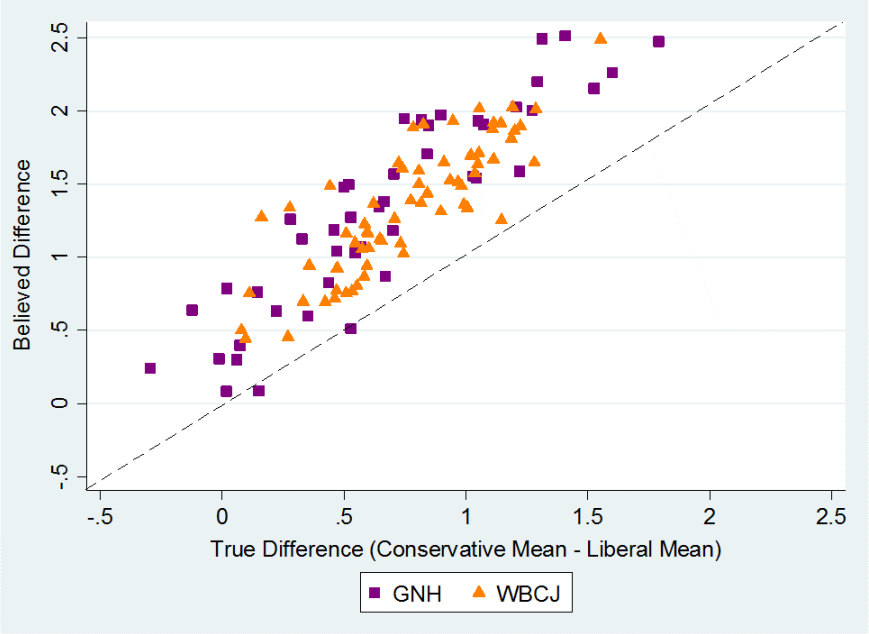


Figure IV: Exaggeration of Differences

The systematic and significant exaggeration of mean differences suggests that the benchmark model of accurate beliefs is missing something important. Indeed, this exaggeration reflects the fact that believed means are typically more extreme than true means. First, note

[30] Chambers and Melnyk (2006) also find evidence of exaggeration of true differences in beliefs about political positions.

[31] For convenience, we recode all issues so that the high end of the scale (6,7) represents the stereotypically more conservative position.

[32] A natural question to ask is how beliefs vary across liberals and conservatives. That is, do beliefs about a group G depend on membership in G versus $-G$. Our model does not speak to this issue. However, in Appendix G, we show that the results we document below hold for both beliefs held by conservatives and beliefs held by liberals; see Tables A9, A10, and A11.

that the Kernel of truth regression (Equation 5) generates exaggeration of mean differences, just as documented in Figure IV:

$$\mathbb{E}^{st}(t|G) - \mathbb{E}^{st}(t|-G) = (1+2\kappa)\cdot[\mathbb{E}(t|G) - \mathbb{E}(t|-G)]$$

To bring this prediction to the data, we regress the believed mean position for each group, $\mathbb{E}^B(t|G)$, on the true mean of the group, $\mathbb{E}(t|G)$, and the true mean of the comparison group, $\mathbb{E}(t|-G)$, across issues. We do this for each group (liberals, conservatives) in each data set separately. We also present pooled specifications which combine the data sets to make a prediction for a given group (liberals, conservatives). In these pooled specifications, we include a dummy variable indicating whether the observation came from the ANES data set. For all specifications, we cluster observations at the issue level.[33] The results from regression are shown in Table I. In every specification, we find that $\mathbb{E}(t|G)$ is a significant predictor of $\mathbb{E}^B(t|G)$ with the predicted positive sign. Crucially, for five of the six specifications, the mean of the comparison group $\mathbb{E}(t|-G)$ is a also significant predictor of $\mathbb{E}^B(t|G)$, with the predicted negative sign.

While these results provide strong evidence of context dependence and are consistent with our model, they do not pin down a role for representativeness of types. Our next test relates the magnitude of representativeness of tail types to the magnitude of belief distortions.

To this end, we implement the regressions in Equations (6, 7). Following Proposition 4, we compute the average representativeness of tail types for conservatives, $R_H^{cons} = \frac{\sum_{t\geq T-2}\pi_{t,cons}}{\sum_{t\geq T-2}\pi_{t,lib}}$. We again test the hypothesis that R_H^{cons} is a significant predictor of $\mathbb{E}^{st}(t|cons)$ with a positive sign, and a predictor of $\mathbb{E}^{st}(t|lib)$ with a negative sign. Table II shows that, conditional on true mean, R_H^{cons} predicts believed mean for each group G as predicted. The first three specifications display the results for predicting beliefs held about conservatives. In both data sets, the average representativeness of tail types for conservatives is a significant, positive predictor of beliefs held about conservatives. The final three specifications display the results for predicting beliefs held about liberals. We find that the average representativeness of tail types for conservatives is a negative predictor of beliefs held about liberals. This effect is

[33] In general, the results presented below for the ANES data are not largely impacted by the decision to cluster at the issue level. Similar results are obtained if the data are not clustered.

Table I: Information about -G Predicts Beliefs about G

OLS Predicting Believed Mean of Group G for Each Issue						
	G = Conservatives			G = Liberals		
	GNH	ANES	Pooled	GNH	ANES	Pooled
True Mean Position of Conservatives for Issue	1.02****	0.98****	0.96****	-0.21****	-0.19	-0.25****
	(0.097)	(0.133)	(0.076)	(0.060)	(0.116)	(0.060)
True Mean Position of Liberals for Issue	-0.35***	-0.86****	-0.58****	0.987****	0.39***	0.73****
	(0.106)	(0.134)	(0.131)	(0.066)	(0.106)	(0.135)
Constant	1.51***	3.35****	2.35****	0.69****	2.58****	1.56****
	(0.195)	(0.269)	(0.279)	(0.122)	(0.249)	(0.270)
R-squared	0.83	0.53	0.66	0.92	0.32	0.68
Obs. (Clusters)	45 (45)	66 (10)	111 (55)	45 (45)	66 (10)	111 (55)

Notes: Std. errors in parentheses, clustered at the issue level. *, **, ***, and **** denote significance at the 10% level, 5%, 1%, and 0.1% level, respectively. In pooled specifications, we include a dummy variable indicating whether the observation came from ANES data set.

significant in the GNH data and marginally significant in the ANES data.

We present several additional results in the Online Appendix G. To further examine the role of representativeness in driving beliefs, we first show the results of Table II carry through when we control for the average likelihood on the tail positions (see Table A4 in Appendix G). Most tellingly, we use the model to predict beliefs across issues and compare those to the data. For simplicity, we use the rank-based truncation specification to predict mean beliefs when stereotypes include the d most representative types, for $d = 1, \ldots, T$. We compare these predictions to those of a model in which beliefs are obtained by restricting the distribution to the d most likely types. Our benchmark for both models is predicting the believed mean from the entire distribution, where $d = T$. We show that the predictions of the rank-based truncation model, with $d = 4$ or 5, compare favorably to both a likelihood-based truncation model (with any d) and the accurate beliefs benchmark. Interestingly, this result suggests that stereotypical beliefs are well approximated by neglecting the least representative types (as opposed to focusing only on the most representative types), and so

Table II: Average Representativeness of Tail Positions Predicts Beliefs

OLS Predicting Believed Mean of Group G for Each Issue						
	G = Conservatives			G = Liberals		
	GNH	ANES	Pooled	GNH	ANES	Pooled
True Mean Position of Group G for Issue	0.78****	0.24**	0.51****	0.72****	0.18***	0.41****
	(0.06)	(0.08)	(0.09)	(0.05)	(0.05)	(0.10)
R_H^{cons} Avg. Rep. of Tail Types for Conservatives	0.19**	0.55**	0.25***	-0.14**	-0.12*	-0.24****
	(0.07)	(0.22)	(0.08)	(0.06)	(0.07)	(0.05)
Constant	1.01****	2.60****	1.84****	0.93****	2.71****	2.01****
	(0.26)	(0.45)	(0.29)	(0.22)	(0.25)	(0.32)
R-squared	0.82	0.48	0.60	0.91	0.31	0.70
Obs. (Clusters)	45 (45)	66 (10)	111 (55)	45 (45)	66 (10)	111 (55)

Notes: Std. errors in parentheses, clustered at the issue level. *, **, ***, and **** denote significance at the 10% level, 5%, 1%, and 0.1% level, respectively. In pooled specifications, we include a dummy variable indicating whether the observation came from ANES data set.

represent a moderate, though systematic, departure from the standard benchmark. Finally, we show that the results we document above hold for both beliefs held by conservatives and beliefs held by liberals; see Tables A9, A10, and A11 in Appendix G.

5 Conclusion

We present a model of stereotypical thinking, in which decision makers making predictions about a group overweight the group's most distinctive types. These overweighted types are not the most likely ones given the DM's data, but rather the most representative ones, in the sense of being the most diagnostic of the group relative to other groups. Representativeness implies that what is most distinctive of a group depends on what group it is compared to. We present experimental evidence that confirms this context dependence in recall-based assessments of groups. Finally, we evaluated the predictions of the model using political data from existing large scale surveys. We find context-dependence to be a key feature of beliefs. Given the richness of the political data, we can go a step further and identify a role

for representativeness in particular. As the representativeness of tail types increases, beliefs of a group are distorted in the direction of that tail.

Our approach provides a parsimonious and psychologically founded account of how decision makers generate simplified representations of reality, from social groups to stock returns, and offers a unified account of disparate pieces of evidence relating to this type of uncertainty. The model captures the central fact that stereotypes highlight the greatest difference between groups, thus explaining why some stereotypes are very accurate, while others lack validity. Still, stereotypes often have a "kernel of truth", when they are based on systematic – even if small – differences between groups. This same logic allows us to describe a number of heuristics and psychological biases, many of which arise in the context of prediction problems. Generically, our model generates overreaction to diagnostic information.

Our model is based on representativeness and does not capture all the features of stereotypical thinking. However, it captures perhaps the central feature: when we think of a group, we focus on what is most distinctive about it, and neglect or underweight the rest.

Royal Holloway, University of London
Ohio State University
Universitá Bocconi and IGIER
Harvard University

References:

Adorno, Theodor, Else Frenkel-Brunswik, Daniel Levinson, and Nevitt Sanford. 1950. *The Authoritarian Personality.* New York, NY: Harper & Row.

Arrow, Kenneth. 1973. The Theory of Discrimination. In O. Ashenfelter and A. Rees, eds. *Discrimination in Labor Markets.* Princeton, N.J.: Princeton University Press: 3 – 33.

Ashton, Michael, and Victoria Esses. 1999. "Stereotype Accuracy: Estimating the Academic Performance of Ethnic Groups." *Personality and Social Psychology Bulletin* 25: 225 - 236.

Bayer, Sylvia. 1999. "The Accuracy of Academic Gender Stereotypes." *Sex Roles* 40 (9): 787 – 813.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny. 1998. "A Model of Investor Sentiment." *Journal of Financial Economics* 49 (3): 307 – 343.

Benjamin, Dan, James Choi, and Joshua Strickland. 2010. "Social Identity and Preferences." *American Economic Review* 100 (4): 1913 – 1928.

Benjamin, Dan, Matthew Rabin, and Collin Raymond. 2015. "A Model of Non-Belief in the Law of Large Numbers." *Journal of the European Economic Association* forthcoming.

Bertrand, Marianne. 2011. "New Perspectives on Gender" in O. Ashenfelter and D. Card eds, *Handbook of Labor Economics* 4 (B): 1543 – 1590.

Bodoh-Creed, Aaron, Dan Benjamin, and Matthew Rabin. 2013. "The Dynamics of Base-Rate Neglect." Mimeo Haas Business School.

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2015. "When Performance Trumps Gender Bias: Joint Versus Separate Evaluation." *Management Science* forthcoming.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *Quarterly Journal of Economics* 127 (3): 1243 – 1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5): 803 – 843.

Briton, Nancy and Judith Hall. 1995. "Beliefs about Female and Male Nonverbal Communication." *Sex Roles* 32 (1/2): 79 - 90.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness and Career Choices." *Quarterly Journal of Economics* 129 (3): 1409 – 1447.

Carrell, Scott, Marianne Page, and James West. 2010. "Sex and Science: How Professor Gender Perpetuates the Gender Gap." *Quarterly Journal of Economics* 125 (3): 1101 – 1144.

Chambers, John, and Darya Melnyk. 2006. "Why Do I Hate Thee? Conflict Misperceptions and Intergroup Mistrust." *Personality and Social Psychology Bulletin* 32: 1295 - 1311.

Chan, Wayne et al. 2012. "Stereotypes of Age Differences in Personality Traits: Universal and Accurate?" *Journal of Personality and Social Psychology* 103 (6): 1050 – 1066.

Chen, Yan, Sherry Xin Liu, Tracy Xiao, and Margaret Shih. 2014. "Which Hat to Wear? Impact of Natural Identities on Coordination and Cooperation." *Games and Economic Behavior* 84: 58 – 86.

Coffman, Katherine. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *Quarterly Journal of Economics* 129 (4): 1625 - 1660.

Cunningham, Tom. 2013. "Comparisons and Choice." Mimeo Stockholm University.

Diekman, Amanda, Alice Eagly, and Patrick Kulesa. 2002. "Accuracy and Bias in the Stereotypes about the Social and Political Attitudes of Women and Men." *Journal of Experimental Social Psychology* 38: 268 - 282.

Fryer, Roland, and Matthew Jackson. 2008. "A Categorical Model of Cognition and Biased Decision-Making." *B.E. Journal of Theoretical Economics* 8 (1): 1 – 42.

Fryer, Roland, and Steven Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal, Applied Economics* 2 (2): 210 – 240.

Gennaioli, Nicola, and Andrei Shleifer. 2010. "What Comes to Mind." *Quarterly Journal of Economics* 125 (4): 1399 – 1433.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. 2012. "Neglected Risks, Financial Innovation, and Financial Fragility." *Journal of Financial Economics* 104 (3): 452 – 468.

Gilens, Martin. 1996. "Race and Poverty in America: Public Misperceptions and the American News Media." *Public Opinion Quarterly* 60 (4): 515 – 541.

Glaeser, Edward. 2005. "The Political Economy of Hatred." *Quarterly Journal of Economics* 120(1): 45 - 86.

Glover Dyland, Pallais Amanda, and Pariente William. 2015. "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores." Mimeo Harvard University.

Goldin, Claudia, Lawrence Katz, and Ilyana Kuziemko. 2006. "The Homecoming of American College Women: The Reversal of the College Gender Gap." *Journal of Economic Perspectives* 20 (4): 133 – 156.

Graham, Jesse, Brian Nosek, and Jonathan Haidt. 2012. "The Moral Stereotypes of Liberals and Conservatives: Exaggeration of Differences across the Political Spectrum." *PLOS One* 7 (12): 1 – 13.

Grether, David. 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *Quarterly Journal of Economics* 95 (3): 537 – 557.

Griffin, Dale, and Amos Tversky. 1992. "The Weighing of Evidence and the Determinants of Confidence." *Cognitive Psychology* 24 (3): 411 – 435.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, Gender, and Math." *Science* 320 (5880): 1164 – 1165.

Hilton, James, and William Von Hippel. 1996. "Stereotypes." *Annual Review of Psychology* 47 (1): 237 – 271.

Hyde, Janet, Sara Lindberg, Marcia Linn, Amy Ellis, and Caroline Williams. 2008. "Gender Similarities Characterize Math Performance." *Science* 321 (5888): 494 – 495.

Judd, Charles, and Bernardette Park. 1993. "Definition and Assessment of Accuracy in Social Stereotypes." *Psychological Review* 100 (1): 109 – 128.

Jussim, Lee, Jarret Crawford, Stephanie Anglin, John Chambers, Sean Stevens, and Florette Cohen. 2015. "Stereotype Accuracy: One of the Largest and Most Replicable Effects in All of Social Psychology " in *The Handbook of Prejudice, Stereotyping, and Discrimination*, Todd Nelson, editor. Mahwah, NJ: Lawrence Erlbaum Publishing.

Kahneman, Daniel, and Amos Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430 – 454.

Kahneman, Daniel, and Amos Tversky. 1973. "On the Psychology of Prediction." *Psychological Review* 80 (4): 237 – 251.

Kaplowitz, Stan, Bradley Fisher, and Clifford Broman. 2003. "How Accurate are Perceptions of Social Statistics about Blacks and Whites?" *Public Opinion Quarterly* 67 (2): 237 - 243.

Kersten, Daniel, Pascal Mamassian, and Alan Yuille. 2004. "Object Perception as Bayesian Inference." *Annual Review of Psychology* 55: 271 – 304.

Klein, Richard et al. 2014. "Investigating Variation in Replicability: A Many Labs Replication Project "*Social Psychology* 45 (3): 142 – 152.

Lundberg, Shelly, and Richard Startz. 1983. "Private Discrimination and Social Intervention in Competitive Labor Markets." *American Economic Review* 73 (3): 340-347.

Madon, Stephanie, Max Guyll, Kathy Aboufadel, Eulices Montiel, Alison Smith, Polly Palumbo, and Lee Jussim. 2001. "Ethnic and National Stereotypes: The Princeton Trilogy Revisited and Revised." *Personality and Social Psychological Bulletin* 27 (8): 996 – 1010.

Martin, Carol. 1987. "A Ratio Measure of Sex Stereotyping." *Journal of Personality and Social Psychology* 52: 489 – 499.

McCauley, Clark, Krishna Thangavelu, and Paul Rozin. 1988. "Sex Stereotyping of Occupations in Relation to Television Representations and Census Facts." *Basic and Applied Social Psychology* 9 (3): 197 - 212.

Mullainathan, Sendhil. 2002. "Thinking through Categories." Mimeo Harvard University.

Ortoleva, Pietro, and Erik Snowberg. 2015. "Overconfidence in Political Behavior." *American Economic Review* 105 (2): 504 – 535.

Phelps, Edmund. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659 – 661.

Pope, Devin, and Justin Sydnor. 2010. "Geographic Variation in the Gender Differences in Test Scores." *Journal of Economic Perspectives* 24(2): 95 – 108.

Rabin, Matthew. 2002. "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics* 117 (3): 775 – 816.

Rabin, Matthew, and Joel Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics* 114 (1): 37 – 82.

Rabin, Matthew, and Dimitri Vayanos. 2010. "The Gambler's and Hot-Hand Fallacies: Theory and Applications" *Review of Economic Studies* 77 (2): 730 – 778.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How Stereotypes Impair Women's Careers in Science." *Proceedings of the National Academy of Sciences* 111 (12): 4403 – 4408.

Rosch, Eleanor. 1973. "Natural Categories." *Cognitive Psychology* 4 (3): 328 – 350.

Ryan, Carey S. 1996. "Accuracy of Black and White College Students' In-Group and Out-Group Stereotypes. *Personality and Social Psychology Bulletin* 22: 1114 - 1127.

Schneider, David. 2004. *The Psychology of Stereotyping.* New York, NY: The Guilford Press.

Schneider, David, Albert Hastorf, and Phoebe Ellsworth. 1979. *Person Perception* (2nd ed.). Reading, MA: Addison-Wesley.

Shih, Margaret, Todd Pittinsky, and Nalini Ambady. 1999. "Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance." *Psychological Science* 10 (1): 80 – 83.

Shih, Margaret, Todd Pittinsky, and Amy Trahan. 2006. "Domain-specific Effects of Stereotypes on Performance." *Self and Identity* 5 (1): 1 – 14.

Steele, Claude. 2010. *Whistling Vivaldi: How Stereotypes Affect Us and What We Can Do.* New York, NY: W. W. Norton & Company.

Tenenbaum, Joshua, and Thomas Griffiths. 2001. "The Rational Basis of Representativeness." *23rd Annual Conference of the Cognitive Science Society* 1036 – 1041.

Tversky, Amos, and Daniel Kahneman. 1983. "Extensional versus Intuitive Reasoning: the Conjunction Fallacy in Probability Judgment." *Psychological Review* 90 (4): 293 – 315.

Weinberger, Catherin. 2005. "Is the Science and Engineering Workforce Drawn from the Far Upper Tail of the Math Ability Distribution?" Mimeo UCSB.

Westfall, Jacob, Leaf Van Boven, John Chambers, and Charles Judd. 2015. “Perceiving Political Polarization in the United States: Party Identity Strength and Attitude Extremity Exacerbate the Perceived Partisan Divide." *Psychological Science* 10 (2): 145 – 158.

Wolsko, Christopher, Bernadette Park, Charles Judd, and Bernd Wittenbrink. 2000. “Framing Interethnic Ideology: Effects of Multicultural and Color-Blind Perspectives on Judgments of Groups and Individuals.” *Journal of Personality and Social Psychology* 78 (4): 635 - 654.