NBER WORKING PAPER SERIES

STEREOTYPES

Pedro Bordalo
Nicola Gennaioli
Andrei Shleifer

Working Paper 20106
http://www.nber.org/papers/w20106

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2014

We are grateful to Nick Barberis, Roland Bénabou, Dan Benjamin, Tom Cunningham, Matthew Gentzkow, Emir Kamenica, Larry Katz, David Laibson, Sendhil Mullainathan, Josh Schwartzstein, Jesse Shapiro, Alp Simsek and Neil Thakral for extremely helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Stereotypes
Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer
NBER Working Paper No. 20106
May 2014
JEL No. D01,D03,D83,D84

ABSTRACT

We present a model of stereotypes in which a decision maker assessing a group recalls only that group's most representative or distinctive types relative to other groups. Because stereotypes highlight differences between groups, and neglect likely common types, they are especially inaccurate when groups are similar. In this case, stereotypes consist of unlikely, extreme types. When stereotypes are inaccurate, they exhibit a form of base rate neglect. They also imply a form of confirmation bias in light of new information: beliefs over-react to information that confirms the stereotype and ignore information that contradicts it. However, stereotypes can change – or rather, be replaced – if new information changes the group's most distinctive trait. Applied to gender stereotypes, the model provides a unified account of disparate evidence regarding the gender gap in education and in labor markets.

Pedro Bordalo
Department of Economics
Royal Holloway
University of London
Egham Hill, Egham, TW20 0EX
United Kingdom
pedro.bordalo@rhul.ac.uk

Andrei Shleifer
Department of Economics
Harvard University
Littauer Center M-9
Cambridge, MA 02138
and NBER
ashleifer@harvard.edu

Nicola Gennaioli
Department of Finance
Università Bocconi
Via Roentgen 1
20136 Milan, Italy
nicola.gennaioli@unibocconi.it

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 ("male drivers are aggressive"), demographic groups ("Florida residents are elderly"), and activities ("flying is dangerous"). Stereotypes play an important cognitive role. Psychologists define them 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" (Hilton and von Hippel 1996). While stereotypes allow for a quick and intuitive assessment of groups, they may also cause distorted judgment and biased behavior, such as discrimination and inter-group conflict. The nature of stereotypes is not completely understood and there are many open questions: How do stereotypes form? How do they affect beliefs and actions? Why do some stereotypes have a reasonable amount of validity ("men are aggressive drivers"), while others have much less ("Florida residents are elderly ")? How do stereotypes change?

We present a psychologically motivated model in which stereotypes are simplified models of reality, consisting of features or types that automatically come to mind when thinking about a group. Psychologists have proposed several factors that shape which types come to mind and become stereotypes. These include representativeness, likelihood, and availability of types (e.g., due to media coverage). We focus on representativeness. We build on Gennaioli and Shleifer's (GS, 2010) model of the representativeness heuristic, in which a group's representative types are those that most distinguish it from other groups. This approach views the core of stereotyping as drawing differences among groups, and captures Kahneman and Tversky's (1972) notion 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 a relevant reference class."

Formally, we assume that a type t is representative for group G if it is diagnostic of G

relative to a comparison group $-G$, in that the likelihood ratio $\Pr(G|t)/\Pr(-G|t)$ is large. To explore the role of representativeness in the simplest setting, we assume that, due to limited working memory, only the most representative types are recalled and used in judgments, be it for inference or for prediction. The stereotype of G is thus formed by truncating the true probability distribution $\Pr(t|G)$ of group G to its $d \geq 1$ most representative types. Non-representative types are neglected.

Relative to a Bayesian, distortions in beliefs can be drastic, particularly when the types that come to mind are not the most likely ones. To illustrate this logic, consider the formation of the stereotype "Florida residents are elderly". The proportion of elderly people in Florida and in the overall US population is shown in the table below.[1]

age	$0-18$	$19-64$	$65+$
Florida	20.7%	61.1%	18.2%
US	23.5%	62.8%	13.7%

The table shows that the most representative type of a Florida resident is someone over 65, because this age bracket maximizes $\Pr(\text{Florida}|t)/\Pr(\text{US}|t)$. However, and perhaps surprisingly, only about 18% of Florida residents are elderly. The vast majority of Florida residents, nearly as many as in the overall US population, are in the age bracket "19-64", 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 focus on representativeness yields the following insights and predictions:

- Whether stereotypes are accurate (recalling likely types) or inaccurate (recalling unlikely types) depends on the underlying distribution of group types. In particular, because stereotypes highlight the differences between groups, they are especially inaccurate when groups are fairly similar and differ only in the tails. These are the cases in which representativeness and likelihood differ the most. Our theory thus explains why stereotypes are often extremely unlikely, as in the Florida example.

[1] See http://quickfacts.census.gov/qfd/states/12000.html.

- Because stereotyping relies on limited recall of types, not only do stereotypes emphasize differences between groups, they often also minimize variability within groups.

- Stereotypes can exhibit a specific form of neglect of base rates which is distinct from – and yields different predictions than – the standard approach in which the impact of base-rates on Bayesian updates is dampened.

- Stereotypes distort reaction to information. So long as stereotypes do not change, people display a form of confirmation bias in that they over-react to information consistent with stereotypes, and under-react or even ignore information inconsistent with stereotypes. Base-rate neglect and confirmation bias are two sides of the same coin of representativeness based recall.

- Stereotypes change – or rather, are replaced – if sufficient contrary information is received (e.g. observing more women than men studying math), or if an entirely different feature becomes more representative (e.g. observing many Black athletes). A change of stereotypes then leads to a drastic reevaluation of already available data. However, more information does not necessarily lead to a better (more likely) stereotype.

Since Kahneman and Tversky's (1972, 1973) work on heuristics and biases, several studies have formally modeled 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 DM has an incorrect model in mind or incorrectly processes available data. Our approach is instead based on the single assumption that only representative information comes to mind when making judgments. Neglect of some information simplifies the judgment problem in a way that is related to models of categorization (e.g. Mullainathan 2002, Fryer and Jackson 2008). In these models, however, DMs 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 a role of context in shaping assessments. Our emphasis on representative and distinctive features or types is closely related to our previous work on salience (BGS 2012, 2013).

Stereotypes also play a role in models of statistical discrimination (Arrow 1973, Phelps 1972). In these models, stereotypes fill up for the lack of information about agents, but equilibrium stereotypes are accurate on average. Our model can generate similar dynamics, but it additionally emphasises the role of self-stereotypes, namely beliefs about oneself that are influenced by group membership and across-group comparisons, with potentially important economic consequences. In Section 5 we connect our work with a recent literature on the role of beliefs and preferences in gender stereotypes and outcomes (Goldin, Katz and Kuziemko 2006, Niederle and Vesterlund 2011, Bertrand 2011).

In the next section, we introduce the notion of representativeness in the context of categorical (discrete) distributions and describe our model. We explore the forces that shape stereotypes and their accuracy. In Section 3, we describe how stereotypes can cause both under- and over-reaction to new information. Section 4 extends the analysis to continuous distributions. Section 5 applies the model to develop a theory of gender stereotypes. Section 6 extends the model to account for the role of likelihood in recall. Section 7 concludes. Appendix A contains the proofs. In Appendices B and C we consider the cases of unordered types and multidimensional types, respectively.

2 A Model of Representativeness and Stereotypes

2.1 The Model

A decision maker (DM) faces a *prediction* problem, which entails representing the distribution of types t in a group G. The DM may be 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 his future earnings based on his own educational background. The DM solves this problem by forming a simplified representation of G, which relies on recalling from memory only the most representative types of group G relative to an alternative group $-G$.[2]

[2]Our model is concerned with the specific mental operation of *recalling* the conditional distribution of types t given G, which is stored in memory. We do not consider the related operation of inference, namely of determining the probability that G vs $-G$ is true. Gennaioli and Shleifer (GS, 2010) offer a model of representativeness-based inference, in which DMs assess a hypothesis by recalling only its most representative scenarios relative to the alternative hypothesis. In GS (2010), scenario t is more representative of hypothesis G against an alternative $-G$ if, conditional on t, G is more likely to be true than $-G$. This is closely related

Formally, the DM must assess the distribution of a categorical random variable T in a group G, which is a proper subset of the entire population Ω. The random variable T takes values in a type space $\{t_1, \ldots, t_N\}$ that is naturally ordered, with $t_1 < \ldots < t_N$ (and in many examples is assumed to be cardinal). In the examples of the introduction, G is male, or a Florida resident, or firms, while types are assertiveness, age, or stock returns.[3] We denote by $\pi_{t,G}$ the true conditional probability $\Pr(T = t|G)$ of type t in group G and by π_t the true unconditional probability $\Pr(T = t)$ of type t in Ω.

The DM has stored in memory the full conditional distribution $(\pi_{t,G})_{t\in\{t_1,\ldots,t_N\}}$, but he assesses this distribution by recalling only a limited and selected set of types. Recall is limited in that the DM recalls only a subset of $d \in \{1, \ldots, N\}$ types. When $d = 1$, memory limits are so severe that the DM recalls only one type for G. When $d = N$, there are no memory limits, and the DM recalls all possible types for G. Recall is selective in that, for given $d < N$, the recalled types are the most *representative* of group G, in the sense that they are most diagnostic of G relative to other groups in Ω. Following GS (2010), we formalize representativeness as follows.

Definition 1 *The representativeness of type t for group G is defined as $R(t, G) = \Pr(G|T = t)/\Pr(-G|T = t)$, where $-G = \Omega\backslash G$. Bayes' rule implies that representativeness increases in the likelihood ratio:*

$$\frac{\Pr(T = t|G)}{\Pr(T = t| - G)} = \frac{\pi_{t,G}}{\pi_{t,-G}}. \tag{1}$$

A type t is representative of group G if, after observing t, a Bayesian DM is more confident that the type is drawn from G relative to $-G$. Put differently, type t is representative of G if it is diagnostic about G in this sense. This notion captures Kahneman and Tversky's (1972) intuition, whereby a type t is representative of G if it is relatively more likely to occur in G than in $-G$. When thinking about the age distribution of Floridians, our minds find it easy

to our Definition 1 below.

[3]The model applies also to cases in which types are not ordered, representing for instance occupations, or when they are multi-dimensional, capturing a bundle of attributes such as occupation and nationality. We return to these possibilities in Appendices B and C 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).

to retrieve those age brackets that are relatively more common in Florida, as compared to the rest of the US population.

Definition 1 leads to the following property.

Remark 1 *Suppose that $\pi_{t,G} \geq \pi_{t,-G}$. Then, the representativeness of type t for group G:*

i) increases, for given baseline probability $\pi_{t,-G}$, in the difference $(\pi_{t,G} - \pi_{t,-G})$.

ii) decreases, for given difference $\pi_{t,G} - \pi_{t,-G}$, in the baseline probability $\pi_{t,-G}$.

Our DMs are attuned to log differences in probabilities:[4] property i) says that a type is more representative the more likely it is to occur under G than under $-G$, and property ii) captures a form of diminishing sensitivity, whereby a given probability difference is more attended to when it occurs in a relatively unlikely type of a group G. This is because such types can be very diagnostic.

The DM's assessment of the distribution of types over G works as follows.

Definition 2 *Denote by $r \in \{1, \ldots, N\}$ the representativeness ranking of types, and denote by $t(r)$ the r-th most representative type for G. The DM forms his beliefs according to the modified probability distribution:*

$$\pi^{st}_{t(r),G} = \begin{cases} \frac{\pi_{t(r),G}}{\sum_{r'=1}^{d} \pi_{t(r'),G}}, & \text{for } r \in \{1, \ldots, d\}. \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Because representativeness drives recall, the DM's beliefs about G consist of a truncated probability distribution on the d most representative types (ties are resolved randomly). In this way, diagnosticity of types shapes the DM's predictions about G, even though diagnosticity is normatively irrelevant for prediction tasks. We call the distribution $(\pi^{st}_{t(r),G})_{r=1,\ldots,d}$ the stereotype for G (where *st* stands for stereotype), and sometimes refer to the represented types, $\{t_1, \ldots, t_d\}$, as the stereotype for G. These are the types that are at the "top of mind." The $(N - d)$ least representative types are at the back of mind and are neglected by

[4]This feature connects to our previous work on salience, which also builds on Weber's law. In BGS (2012) we postulated that, in a choice among two lotteries, a lottery outcome is more salient when it entails: i) a larger payoff difference (ordering), and ii) a lower payoff level (diminishing sensitivity). Remark 1 shows that here these same properties characterize recall in the domain of probabilistic types.

the DM. These less representative types are not viewed as impossible; they are just assigned zero probability in the DM's current thinking. This formulation allows us to model surprises or reactions to zero probability events, which we come back to in Section 3.

In the extreme case where $d = 1$, the DM recalls only group G's most representative type $t(1)$, which psychologists call the *exemplar*, and assigns it probability $\pi^{st}_{t(1),G} = 1$. In less extreme, and perhaps more realistic cases, $d > 1$ and the stereotype of G includes the exemplar and some less representative types. When thinking about Floridians, people think about not only retired baby boomers, but also college students.

Stereotypes depend on true probabilities. Equation (2) implies that, conditional on coming to mind, the assessed odds ratios of any two types is consistent with the DM's experience and information. Past experience or information about types is stored in the DM's long-term memory and thus, conditional on coming to mind, shapes assessments. Since past experiences or information may vary across individuals, our model allows for individual heterogeneity in stereotypes, driven for instance by culture (see Section 5).

2.2 Discussion of Assumptions

Implementing Definition 1 of representativeness raises two important issues: i) what is the set of types T considered by the DM, and ii) what is the comparison group $-G$.

Any prediction problem specifies the group G and a possibly coarse version of the type space T. Often, the problem itself provides a natural specification of T as well as of the comparison group $-G$. When, as we assume here, types have a natural order (such as income, age, education), T is naturally given by the problem (income, age and years of schooling brackets). Other settings may not automatically prime a natural set of types. For example, suppose a person is asked to guess the typical occupation of a democratic voter. 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).[5]

[5]Strictly speaking, this is also an issue when types are ordered. However, in contrast to non-ordered categories where granularity is of central importance (professor of comparative literature is very different from professor of business administration), in ordered categories distributions are typically smoother, so changing the bracketing has minor effects on estimates. In some settings, such distributions have natural bracketing, such as in educational attainment.

Psychologists have sought for years to construct a theory of natural types (see Rosch 1998). Here we do not make a contribution to this problem.

The second question raised by Definition 1 is that of the comparison group, or equivalently the set of possible groups Ω (given that $-G = \Omega \backslash G$). The comparison group $-G$ 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 = Black Americans, $-G$ =White Americans. Sometimes there are several natural comparison groups and the specification of Ω can influence the stereotype for G. For example, the stereotype of college athletes in the population of all college students might be "below average academic", but in comparison to professional athletes the stereotype might be "not very strong."[6] We do not have a theory of what determines Ω when it is not pinned down by the problem itself (though the specification of G provides natural bounds for Ω, e.g. when G is a social group, Ω is a larger subset of mankind).

At a broader level, in our model stereotypes are simplified mental representations of groups characterized by selective recall of those groups' types. Our emphasis on representativeness implies that a stereotype exaggerates the distinctive traits of the group it represents, consistent with the social psychology of stereotyping (Hilton and Hippel 1996).

Representativeness is not the only psychological force that shapes stereotypes. Decision makers may for instance find it easier to recall types that are sufficiently likely. In Section 6 we formally incorporate in our model a more general mechanism driven by a combination of representativeness and likelihood of types. More generally, Kahneman and Tversky (1972) stress that selective recall is also shaped by availability, broadly understood as the "ease" with which information comes to mind. This may capture likelihood but also aspects such as recency and frequency of exposure, which might be independent of likelihood or representativeness. For instance, in the aftermath of the 9/11 terrorist attacks, a US respondent asked about what Arabs are like might more easily recall terrorists than Bedouins, even when there are vastly more bedouins than terrorists among Arabs, and even though all Bedouins are Arabs, so that Bedouins are more representative of Arabs than terrorists.[7] The extension

[6] As this example suggests, the type space T can in general have multiple dimensions. In Appendix C, we explore stereotype formation in this case, and in particular which dimension becomes stereotypical.

[7] A Gallup poll conducted shortly after the 1993 terrorist bombing of the World Trade Center

in Section 6 can also capture the role of frequency of exposure on recall, but a full model of availability is beyond the scope of this paper (and none is readily available in the Psychology literature). Moreover, because representativeness captures the central property that stereotypes highlight differences among groups, representativeness is a necessary, and often sufficient, mechanism for stereotyping.

Finally, consider the assumption that stereotypical beliefs are truncations of true distributions, Definition 2. This assumption captures the observation that in most group assessments some non-representative types do not come to mind at all, consistent with the robust finding in social psychology that stereotypes minimise within-group variability (Hilton and von Hippel 1996). When the type space is finite, it is more tractable to assume that selective recall operates by discarding a number of types. When t is continuous, as in Section 5, it is more natural to think that representative types come to mind up to a certain total probability mass. The qualitative features of our model also hold under smooth discounting of the probability of less representative types.[8]

2.3 Properties of Stereotypes

Some stereotypes hold a high degree of validity (e.g., men are stronger than women), while others are widely off the mark (e.g., Florida residents are elderly). We now explore what determines the accuracy of stereotypical beliefs in our model.

To evaluate whether a stereotype's distribution $(\pi^{st}_{t,G})_{t=t_1,\ldots,t_N}$ is an accurate representation of the true distribution $(\pi_{t,G})_{t=t_1,\ldots,t_N}$ we use two standard measures. The first is

found that "majorities of Americans said the following terms applied to Arabs: religious (81%), terrorists (59%), violent (58%) and religious fanatics (56%). Related, a recent poll by Pew's Global Attitudes Project found that Westerners view Muslims as fanatical (58% of respondents) and violent (50%), while Muslims view Westerners as selfish (68%), violent (66%) and greedy (64%). Curiously, selfishness and greed are among the traits that Westerners least associate with Muslims. Sources: http://www.gallup.com/poll/4939/Americans-Felt-Uneasy-Toward-Arabs-Even-Before-September.aspx and http://www.pewglobal.org/2011/07/21/muslim-western-tensions-persist/.

[8]Smooth discounting may be useful in those cases where there are only a few relevant types, such as when the DM's assessment naturally divides a group into two broad types (e.g. people on welfare vs people not on welfare). Formally, given a weighting function $\delta(\pi_{t,G}/\pi_{t,-G})$ which increases in the likelihood ratio (i.e., $\delta'(\cdot) > 0$) one can define:

$$\pi^{st}_{t,G} = \frac{\delta(\pi_{t,G}/\pi_{t,-G}) \cdot \pi_{k,G}}{\sum_k \delta(\pi_{k,G}/\pi_{k,-G}) \cdot \pi_{k,G}}$$

In this formulation, the probability of types that have a higher representativeness ratio is inflated.

the quadratic loss function $L = \sum_t (\pi^{st}_{t,G} - \pi_{t,G})^2$, which captures the average discrepancy between the stereotype and the true probability distribution. This measure captures the extent to which stereotypes shift probability mass across types, and holds for both ordered and non ordered types. The second is the difference between the stereotype's mean and the true mean, $L = \left|\sum_t (t \cdot \pi^{st}_{t,G} - t \cdot \pi_{t,G})\right|$. This measure is only valid for cardinal types and captures the distortion induced by the stereotype on the variable of interest.

As measured by the quadratic loss function, stereotype accuracy depends on the link between representativeness and likelihood. Accuracy is high if the stereotype includes the types that are objectively most likely but decreases as the (recalled) representative types become less likely. To see what determines the relationship between representativeness and likelihood, suppose that the distribution of types in the comparison group $-G$ is a monotonic transformation of that in group G. Formally, let $(\pi_{t,G})_{t=t_1,\ldots,t_N}$ be the true conditional distribution in group G, and suppose that the true conditional distribution in $-G$ is defined by $\pi_{t,-G} = \pi^* \cdot \pi^{\alpha}_{t,G}$ for all t, where $\pi^* = 1/\sum_t \pi^{\alpha}_{t,G}$ is a normalizing constant. In this formulation, α controls the relationship between the likelihood ranking of types for groups G and $-G$. This is shown in Figure 1, where for the sake of illustration we assume $(\pi_{t,G})_{t=t_1,\ldots,t_N}$ is unimodal (and approximated by a continuous distribution).

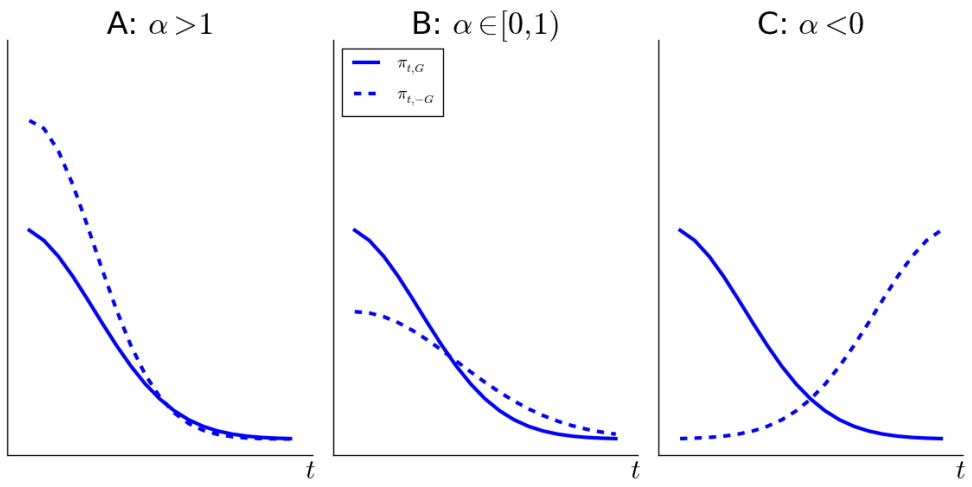

Figure 1: Likely and unlikely stereotypes.

If $\alpha > 0$ (panels A and B), the likelihood ranking of types for groups G and $-G$ coincide:

the two distributions are "similar" and in particular have the same modal type. There are two subcases : if $\alpha \in [0, 1)$ (panel B), then group G is more concentrated around its mode and group $-G$ has fatter tails, while if $\alpha > 1$ (panel A), then group G has fatter tails and $-G$ is more concentrated around its mode. On the other hand, if $\alpha < 0$ (panel C), then the likelihood ranking of types for G is the opposite of that for $-G$.

In this special case, we have that $R(t, G) \propto \pi_{t,G}^{1-\alpha}$ while $R(t, -G) \propto \pi_{t,G}^{\alpha-1}$. That is, the representativeness ranking of types for G is the opposite of that for $-G$. Therefore, the accuracy of stereotypes for groups G and $-G$ critically depends on whether the two distributions have the same likelihood ranking.

Proposition 1 *Let $\pi_{t,-G} = \pi^* \cdot \pi_{t,G}^{\alpha}$ as above. Then:*

i) If $\alpha > 1$, the stereotype for G is its d least likely types, while the stereotype for $-G$ is its d most likely types.

ii) If $\alpha \in [0, 1)$, the stereotype for G is its d most likely types, while the stereotype for $-G$ is its d least likely types.

iii) If $\alpha < 0$, the stereotypes for G and $-G$ are each group's d most likely types.

Proposition 1 describes the conditions under which a group's likely types are selected by representativeness. Broadly speaking, when groups G and $-G$ have the same likelihood ranking ($\alpha > 0$) then one group has an inaccurate stereotype. In Case i), which is described in panel A of Figure 1, the stereotype for G is unlikely but that for $-G$ is likely, because $\pi_{t,G}$ has heavier tails than $\pi_{t,-G}$. In case ii) (panel B), the reverse is true because $\pi_{t,G}$ is more concentrated than $\pi_{t,-G}$ around its mode. Finally, when the groups have the opposite likelihood ranking, as in case iii) (panel C), then the most representative types are also the most likely types for each group.

Proposition 1 captures a sense in which comparing similar groups leads to bad stereotyping. Intuitively, the psychology of representativeness induces the DM to focus on differences among groups and to neglect types that are likely to occur in both groups. When common types are infrequent, the DM neglects some variability across types but still has a reasonably accurate mental representation of each group. However, when groups are similar, the DM fails in two ways: as before, he neglects within group heterogeneity, but he also dispropor-

tionately recalls unlikely types. In this case, stereotypes are very inaccurate because the DM not only perceives within group variability to be too small, but also generalizes to the entire group a trait that may be very infrequent.

This logic may help explain countless negative stereotypes held about social groups, despite the fact that social groups are broadly similar, precisely because such groups tend to differ in unlikely types.[9] Suppose for instance that a decision maker assesses the wealth distribution of Blacks and Whites in the US, considering only two types: poor and not poor, as measured by the US Census Bureau. Because a much higher proportion of blacks are poor than of whites (27.4% vs 9.9% as of 2010), with $d = 1$ the DM stereotypes blacks as poor and whites as not poor. This is the case even though only a minority of each group is poor.[10] In this case, the stereotype of Blacks is inaccurate but that of Whites is accurate.

We now turn to a second measure of stereotype accuracy, namely the discrepancy between the average type in group G and the average type in the stereotype of G (recall that the type space is cardinal). Representativeness implies that stereotypes are often not just unlikely, but also extreme, in the sense of being dominated by types that are at the extremes of the type space. To proceed, we focus on a case where the characterisation of stereotypes is particularly simple, namely where the likelihood ratio $\pi_{t,G}/\pi_{t,-G}$ is monotonic in t. The monotone likelihood ratio property (MLRP) holds to first approximation in many empirical settings (see Section 5) and is also satisfied in many economic models, for instance in canonical agency models. We then have:

Proposition 2 *If MLRP holds, and $d < N$, then:*

i) if the likelihood ratio $\frac{\pi_{t,G}}{\pi_{t,-G}}$ is increasing, the stereotype for G is the right tail of types

[9]The anthropology and psychology literatures documents that extremely negative ethnic stereotypes play an important role in inter-group conflict (Tajfel, 1982). A central finding of this literature is the asymmetry in valence between stereotypes about one's own group (the in-group) and one's stereotype about other groups (out-groups): stereotypes of in-groups tend to be more detailed and more positive than those of out-groups (Hilton and von Hippel 1996). Although we do not predict systematic differences between in-group and out-group stereotypes, differences might emerge if the DM is assumed to have less accurate information about the out-group. Moreover, the in-group/out-group asymmetry does not seem to be universally valid: in many cases, such as in the case of gender self-stereotypes explored in Section 5, the in-group (women) may share with the out-group (men) a negative stereotype about itself and a positive stereotype about the out-group.

[10]Note that, by neglecting the poor White type, the DM can estimate poor Blacks to outnumber poor Whites. In fact, because the White population is over five times larger than the Black population, White poor outnumber Black poor by 2 to 1. We return to the issue of base-rate neglect below.

$\{N-d+1,\ldots,N\}$. *Moreover,*

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

ii) if the likelihood ratio $\frac{\pi_{t,G}}{\pi_{t,-G}}$ *is decreasing, the stereotype for* G *is the left tail of types* $\{1,\ldots,d\}$. *Moreover,*

$$\mathbb{E}^{st}(t|G) < \mathbb{E}(t|G) < \mathbb{E}(t)$$

Intuitively, under MLRP extreme observations are most informative about – and thus representative of – the group they come from. When MLRP holds, representative types are located at the extremes of the distribution.[11] Thus the DM's belief about G are formed by truncating from the original distribution the least representative tail, and focusing on the most representative tail. This leads to three important effects.

First, the DM's mean assessment of group G is shifted in the same direction as the true conditional mean $\mathbb{E}(t|G)$ relative to the unconditional mean $\mathbb{E}(t)$.

Second, because his assessments are biased in the direction of the (extreme) exemplar, the DM's estimate of the mean type is too extreme (e.g. $\mathbb{E}^{st}(t|G) > \mathbb{E}(t|G)$ if the right tail is representative). Indeed, the monotone likelihood ratio property implies a correlation between types and groups: group G is relatively more associated with high (low) types if the ratio is increasing (decreasing). It follows from Proposition 2 that stereotyping induces the DM to overestimate this correlation.

Third, for a large class of distributions, the DM's assessment of the variance $\mathrm{Var}(t|G)$ is dampened as he neglects types in the non-stereotypical tail. In 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.[12]

In our model, stereotypes indeed "provide the greatest differentiation between groups, and [...] show the least within-group variation" (Hilton and von Hippel 1996). The combination of these forces can shed light on several phenomena. When assessing the performance of firms

[11] Note that MLRP holds in all panels of Figure 1, so that stereotypes for both groups are indeed extreme types. As this example illustrates, extreme types need not be unlikely.

[12] Intuitively, this holds as long as the distribution $\pi_{t,G}$ does not have heavy tails. The result is easier to formalise in the continuous case, see Proposition 5.

in a hot sector of the economy, the investor recalls highly successful (and some moderately successful) firms in that sector. However, he neglects the possibility of failures, because failure 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.[13] Similarly, 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.

Proposition 2 thus implies that the DM overreacts to information that assigns people to groups, since such information generates extreme stereotypes.[14] We now show how this logic provides a novel psychological account of some instances of base-rate neglect (Kahneman and Tversky, 1973). 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 result increases the relative likelihood of sickness, and a negative result increases the relative likelihood of health for *any* prior. Formally:

$$\frac{\Pr(+|s)}{\Pr(+|h)} > 1 > \frac{\Pr(-|s)}{\Pr(-|h)}. \tag{3}$$

This condition has clear implications: the representative person who tests positive is sick,

[13] By emphasizing stereotypical outcomes in the valuation of firms' stock, this logic provides a new mechanism for the growth-value puzzle in asset pricing (Lakonishok, Shleifer and Vishny 1994). Because stereotypical outcomes are also extreme, this mechanism is very similar to that described using the model of salience (Bordalo, Gennaioli and Shleifer 2013b). In general, stereotypes and salience produce different results.

[14] In Section 3 we explore in detail how stereotypical beliefs react to a different kind of information, namely information about the distribution of types when groups are *given*.

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 DMs assessment about the group that tested negative is fairly accurate but is severely biased for the group that tested positive.

Our account of base-rate neglect is starkly different from a mechanical underweighting of base-rates in Bayes rule. In the context of the medical test example, such underweighting is modelled by postulating the modified Bayes rule (Grether 1980, Bodoh-Creed, Benjamin and Rabin 2013):

$$\Pr(h\,|G) = \frac{\Pr(G\,|h)\cdot \Pr(h)^{\eta}}{\Pr(G\,|h)\cdot \Pr(h)^{\eta} + \Pr(G\,|s)\cdot \Pr(s)^{\eta}} \tag{4}$$

where $G = +, -$ denotes the result of the test and parameter $\eta \in [0,1]$ modulates the strength of base-rate neglect. When $\eta = 1$, the DM follows Bayes' rule. When $\eta < 1$, the DM dampens the base-rates of h and of s. Equation (4) implies that, upon receiving information, 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 without any information, which cannot happen in our model.[15,16]

3 Stereotypes and Reaction to New Information

Our model can be naturally extended to investigate how stereotypes and beliefs change by the arrival of new information over time. To explore these dynamics, we suppose that at the outset, unlike in Section 2, the decision maker does not have perfect information about the categorical distribution $(\pi_{t,G})_{t=1,\dots,N}$ of the group G of interest, or about the distribution $(\pi_{t,-G})_{t=1,\dots,N}$ of the comparison group $-G$. Instead, the DM has priors over

[15]Updating in the wrong direction occurs when the probability of being healthy is sufficiently high, so that neglecting it reduces the posterior assessment of health for either test outcome.

[16]The mechanical underweigthing of base-rates in (4) accounts for other instances of base-rate neglect not covered by our model, in particular those that arise in inference problems. This includes Kahneman and Tversky 's (1972) lawyers-engineers example as well as Griffin and Tversky's (1992) biased coin example. See Bodoh-Creed, Benjamin and Rabin (2013) for a detailed discussion.

these distributions that are described by the Dirichlet distribution:

$$g\left[\pi_{t,W},\alpha_{t,W}\right]_{t=t_1,\ldots,t_N}=\frac{\Gamma\left(\sum_t\alpha_{t,W}\right)}{\Pi_t\Gamma(\alpha_{t,W})}\cdot\Pi_t\pi_{t,W}^{\alpha_{t,W}-1}\ ,\qquad \text{for } W=G,-G,$$

which are conveniently conjugate to the categorical distributions assumed so far. Parameters $\alpha_G=(\alpha_{t,G})_{t=t_1,\ldots,t_N}$ and $\alpha_{-G}=(\alpha_{t,-G})_{t=t_1,\ldots,t_N}$ pin down the prior expectations of a Bayesian agent:

$$\Pr(T=t|\alpha_W)=\mathbb{E}(\pi_{t,W}|\alpha_W)=\frac{\alpha_{t,W}}{\sum_u\alpha_{u,W}},\qquad \text{for}\quad W=G,-G. \tag{5}$$

In contrast to the Bayesian agent, the stereotype initially held by the DM depends on the probabilities in Equation (5) according to Definition 1. For simplicity, we set $\sum_t\alpha_{t,G}=\sum_t\alpha_{t,-G}$.

Suppose that a sample $n_W=(n_{1,W},\ldots,n_{N,W})$ is observed, where $n_{t,W}$ denotes the observation count in type t and let $\sum_t n_{t,W}$ be the total number of observations for group W. Then, the posterior probability of observing t assessed by a Bayesian agent is

$$\Pr(T=t|\alpha_W,n_W)=\mathbb{E}(\pi_{t,W}|\alpha_W,n_W)=\frac{\alpha_{t,W}+n_{t,W}}{\sum_u(\alpha_{u,W}+n_{u,W})}, \tag{6}$$

which is a weighted average of the prior probability of Equation (5) and the sample proportion $n_{t,W}/n_W$ of type t. As new observations arrive, the probability distribution in group W, and thus stereotypes, are updated according to Equation (6).[17]

Consider how a DM influenced by representativeness updates beliefs. Given Equations (5) and (6), Proposition 3 considers how new information changes the set of types that come to mind, shedding light on when and how stereotypes change. Proposition 4 in turn considers the effect of information on probability assessments for a given set of types included in the stereotype.

Proposition 3 *Suppose that the DM observes the same number of realizations from both groups, formally* $\sum_u n_{u,G}=\sum_u n_{u,-G}=n$. *Then:*

[17]While we assume for simplicity that updating is Bayesian, the representativeness mechanism that links priors to stereotypes can naturally be coupled with a non-Bayesian updating process. Psychologists have documented a tendency to search for information that confirms one's beliefs (Lord, Ross and Lepper 1979, Nickerson 1998). Schwartzstein (2014) proposes a model of biased learning in which information is used to update beliefs only about dimensions that are attended to.

i) If for both groups all observations occur on the same type t that is initially non-representative for G, then this type does not become representative for G. Formally, if $n_{t,G} = n_{t,-G} = n$ for a type t such that $\alpha_{t,G}/\alpha_{t,-G} < 1$, then $\Pr(X = x|\alpha_W, n_G)/\Pr(X = x|\alpha_W, n_{-G}) < 1$ for all n.

ii) If all observations for G occur in a non representative type for G, while those for $-G$ occur in a type that is representative for G, then for a sufficiently large number of observations the stereotype for G changes. Formally, if $n_{t,G} = n$ for a type t such that $\alpha_{t,G}/\alpha_{t,-G} < 1$, while $n_{t',-G} = n$ for a type t' such that $\alpha_{t',G}/\alpha_{t',-G} > 1$, then for n sufficiently large $\Pr(T = t'|\alpha_W, n_G)/\Pr(T = t'|\alpha_W, n_{-G}) < 1 < \Pr(T = t|\alpha_W, n_G)/\Pr(T = t|\alpha_W, n_{-G})$.

The stereotype for a group does not necessarily change if the new observations are contrary to the initial stereotype. For the stereotype of group G to change, the contrary observations must render previously neglected types sufficiently more likely in G, and thus representative, than in the comparison group $-G$.

To see this, consider first case i), in which the data disconfirming G's initial stereotype uniformly accrue in the two groups G and $-G$. In this case, the non-representative type never becomes representative for G despite the fact that the data consistently point to its relevance. Reductions in the overall incidence of crime do not debunk a negative stereotype about a group if a majority of criminals still come from that same group. A process of economic development that improves the livelihoods of all groups in a population does not improve the stereotype of a group that continues to include a disproportionally high share of underdogs. The intuition for this result comes from diminishing sensitivity of the likelihood ratio (Remark 1): types that are highly likely to occur in both groups are ceteris paribus less representative.

Although stereotypes do not change when new information is symmetric across groups, they can change quickly when information is asymmetric. In case ii), the n observations for G occur in a non-representative type t for G, while the n observations for $-G$ occur in a representative type t' for G. In this case, for n sufficiently large, t becomes representative for G while t' becomes unrepresentative for G. One intuitive instance of this process is the asymmetric reduction in the incidence of tail (but highly representative) events in a

group. Reducing crime in certain high-incidence neighborhoods, but not overall, decreases the association between the population of those neighborhoods and crime, debunking the group's crime-based stereotype. The rapid rise of a new commercial class out of an underdog group creates a new stereotype for that group. Some periods of above market performance turns an uninteresting company into a growth stock. The arrival of new information, while beneficial for a rational agent, may render stereotypes less accurate: in the case of the listed company, its recent above average performance may be due to noise. But the investor leaves little room for noise. He looks for causal patterns and quickly jumps to conclusions, even if the informativeness of stereotype-changing information is low. After all, he thinks, above average performance is the distinctive mark of great companies.

We now consider how the initial stereotype for group G (formally, the priors over G and $-G$) affects the way in which the DM processes new information about G. We only consider information concerning G: since the set of types included in the stereotype is assumed to be constant, information about $-G$ is irrelevant.

Proposition 4 *Let $d > 1$. Suppose that one observation about type t is received in group G (formally, $n = n_{t,G} = 1$). Then:*

i) If t belongs to the stereotype of G and its probability is sufficiently low, the DM over-reacts (relative to the Bayesian) in revising upward his assessment of t's probability. Formally, there is a threshold $\nu \in (0, 1/2)$ such that the DM's assessment of t over-reacts if and only if $\alpha_{t,G} / \sum_u a_{u,G} < \nu$.

ii) If t does not belong to the stereotype of G, the DM does not update its probability at all, so he under-reacts relative to the Bayesian DM.

Proposition 4 indicates that stereotypes can both over and under-react to information. In case i), the DM strongly over-reacts to information confirming the stereotype. Intuitively, because the DM neglects non-representative types, he does not fully account the current observation may be due to sampling variability. As a consequence, his beliefs overreact when a type he does attend to is confirmed by the data. If criminal activity is part of a group's stereotype, the DM over-reacts to seeing a criminal from that group and his judgments become even more biased against the group. If a growth company generates surprisingly

positive earnings, investors further upgrade their belief that the stock is a good investment, because they neglect the possibility that an extreme observation may be due to noise.

At the same time, case ii) shows that the DM under-reacts (relative to a Bayesian) to information inconsistent with the stereotype. This is because insofar as the stereotype is unaffected, the probability of a non-stereotypical type is not upgraded, as the type remains neglected in the assessment of the group. Upon observing a highly successful member of a group stereotyped as the underdog, DMs code the occurrence as an "anomaly" and continue to believe that the group at large should be viewed through the lens of the negative stereotype. People can espouse racist views and yet be friendly with individual members of the group they disregard. However, as shown in Proposition 4, non-stereotypical information is often ineffective at changing beliefs even if it swamps the few instances underlying the stereotype.

Putting the two cases together, Proposition 4 implies that the DM exhibits a type of confirmation bias (Lord, Ross and Lepper 1979, Nickerson, 1998). Faced with two observations of different types from group G (formally, $n_{t,G} = n_{t',G} = 1$ and $n = 2$), such that t belongs to the stereotype of G but t' does not, the DM over-reacts to information consistent with the stereotype and ignores information inconsistent with it. In this way, our approach provides a unified mechanism that gives rise to both base-rate neglect and confirmation bias: base-rate neglect arises when representative types are unlikely, while confirmation bias arises when new information does not change representativeness and allows stereotypes to persist. In the context of representativeness-based predictions, these biases are two sides of the same coin.

4 Continuous Distributions

Many distributions of interest in economics can usefully be approximated by continuous probability distributions. Here we extend our model of stereotypes cover the case of continuous distributions. We then use this extension in our application in Section 5.

4.1 Basic Setting

Let T be a continuous variable defined on the support $\overline{T} \subseteq R^k$. Denote by $t \in \overline{T}$ a realization of T which is distributed according to a density function $f(t) : \overline{T} \to \mathbb{R}_+$. Denote by $f(t|G)$ and $f(t|-G)$, the distributions of t in G and $-G$, respectively. In line with Definition 1, we define representativeness as:

Definition 3 *The representativeness of $t \in \overline{T}$ for group G is measured by the ratio of the probability of G and $-G$ at $T = t$, where $-G = \Omega \backslash G$. Using Bayes' rule, this implies that representativeness increases in the likelihood ratio $f(t|G)/f(t|-G)$.*

In the continuous case, the exemplar for G is the realization t that is most informative about G. For one dimensional variables, the exemplar for G is $\sup(\overline{T})$ if the likelihood ratio is monotone increasing, or $\inf(\overline{T})$ if the likelihood ratio is monotone decreasing, just as in Proposition 2.

The DM constructs the stereotype by recalling the most representative values of t until the recalled probability mass is equal to the bounded memory parameter $\delta \in [0, 1]$. When $\delta = 0$, the DM only recalls the most representative type. When $\delta = 1$ the DM recalls the entire support $\overline{T}$ and his beliefs are correct. When δ is between 0 and 1, we are in an intermediate case.

Definition 4 *Given a group G and a threshold $c \in \mathbb{R}$, define the set $\overline{T}_G(c) = \left\{ t \in \overline{T} \mid \frac{f(t|G)}{f(t|-G)} \geq c \right\}$. The DM forms his beliefs using a truncated distribution in $\overline{T}_G(c(\delta))$ where $c(\delta)$ solves:*

$$\int_{t \in \overline{T}(c(\delta))} f(t|G)dt = \delta.$$

The logic is similar to that of Definition 2, with the only difference that now the memory constraint acts on the recalled probability mass and not on the measure of states, which would be problematic to compute when distributions have unbounded support. This feature yields the new implication that changes in the distribution typically change also the support of the stereotype by triggering the DM to recall or forget some states, even when the states' relative representativeness does not change.

4.2 The Normal Case

When $f(t|G)$ and $f(t|-G)$ are univariate normal, the stereotype of G is easy to characterize.

Proposition 5 *In the normal case, the stereotype works as follows:*

i) Suppose $\sigma_G = \sigma_{-G} = \sigma$. Then, if $\mu_G > \mu_{-G}$ the stereotype for G is $\overline{T}_G = [t_G, +\infty)$, where t_G decreases with δ. Moreover, $\mathbb{E}^{st}(t|G) > \mu_G > \mu_{-G} > \mathbb{E}^{st}(t|-G)$.

If instead $\mu_G < \mu_{-G}$, the stereotype for G is $\overline{T}_G = (-\infty, t_G]$, where t_G now increases with δ. Moreover, $\mathbb{E}^{st}(t|G) < \mu_G < \mu_{-G} < \mathbb{E}^{st}(t|-G)$. In both cases, $Var^{st}(t|G) < Var(t|G)$ and $Var^{st}(t|-G) < Var(t|-G)$.

ii) Suppose that $\sigma_G < \sigma_{-G}$. Then, the stereotype for G is $\overline{T}_G = [\underline{t}_G, \bar{t}_G]$ where $\underline{t}_G$ decreases and $\bar{t}_G$ increases with δ. Moreover, $Var^{st}(t|G) < Var(t|G)$.

iii) Suppose that $\sigma_G > \sigma_{-G}$. Then, the stereotype for G is $\overline{T}_G = (-\infty, \underline{t}_G] \cup [\bar{t}_G, +\infty)$ where $\underline{t}_G$ increases and $\bar{t}_S$ decreases with δ. Moreover, $Var^{st}(t|G) > Var(t|G)$.

When the two distributions have the same variance, the stereotype is formed by truncating from the original distribution the least representative tail (as in Section 2.3). In fact, when the mean in G is above the mean in $-G$, the likelihood ratio is monotone increasing and the exemplar for G is $+\infty$; otherwise it is $-\infty$. In both cases, the exemplar is inaccurate because it relies on a highly representative but very low probability realization.

Figure 2 represents the distribution considered by the DM for the high mean group when traits are normally distributed with the same variance across groups. In this example, the true mean μ_G is included in the support, which in turn means that the value of δ is above .5, e.g. $\delta = .7$. Clearly, in this case the assessed mean is above μ_G and the assessed variance is below the true variance σ_G. Both features are due to the fact that the distribution is distorted towards the group exemplar at $+\infty$. Because each distribution is represented by its stereotypical tail, stereotypical thinking underestimates the variance of both distributions, in line with the idea that stereotypes typically show little "within-group variation" (Hilton and von Hippel 1996).[18]

[18]The effects of stereotypical thinking on the perceived mean and variance of distributions described in Proposition 5 hold more generally for all log-concave distributions, which includes normal as well as many other common distributions; see Heckman and Honoré (1990).

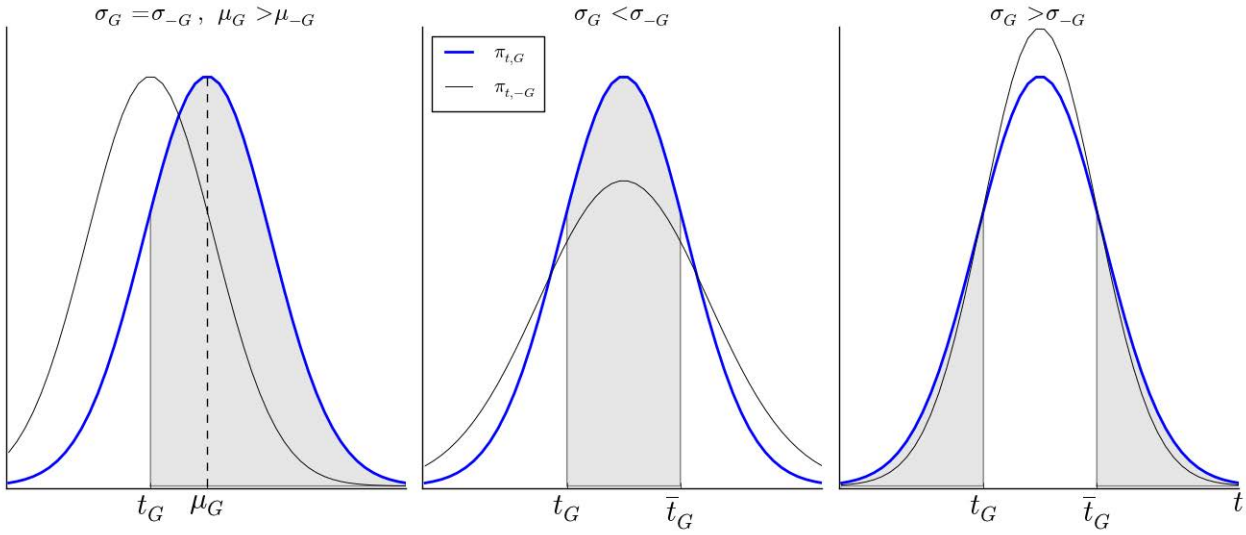

Figure 2: Stereotypes of a Normal distribution as a function of μ_{-G} and σ_{-G}.

Consider now case ii), where the variance of G is lower than that of $-G$. The stereotype consists of an interval around an intermediate exemplar, denoted by $\hat{t}_G$. As in Proposition 1, when the distribution in G is more concentrated than that in $-G$, the exemplar is accurate and captures a relatively frequent, intermediate event. It is however somewhat distorted, because $\hat{t}_G$ lies below the group's true mean μ_G if and only if $\mu_G < \mu_{-G}$. Interestingly, when the mean in the two groups is the same, the low variability group is represented by its correct mean, namely μ_G. Again, because the distinctive feature of group G is being more "average" than group $-G$, its stereotype neglects extreme elements and decreases within group variation.

Finally, consider case iii). Now the variance in G is higher than that in $-G$. As a consequence, both tails are exemplars and the stereotype includes both tails, truncating away an intermediate section of the distribution. This representation increases perceived volatility and thus captures the distinctive trait of G relative to $-G$, which is precisely its higher variability. Stereotyping now induces the DM to recall group G's most extreme elements and to perceive G as more variable than it really is. This is in contrast with the previous cases, and with the common description that stereotypes reduce within-group variability (Hilton and Von Hippel 1996). However, it is consistent with the more basic

intuition that stereotyping highlights the most distinctive features of group G, in this case its extreme elements. As an illustration of this mechanism, when thinking about stock returns, investors may think of positive scenarios where returns are high, or negative scenarios where returns are low, but neglect average returns, which are more typical of safer asset classes. In this respect, case iii) provides a new setting in which to test our model's predictions.

Consider now dynamic updating in this normal case. The DM receives information about the distributions $f(t|G)$ and $f(t|-G)$ over time. In each period k, a sample $(t_{G,k}, t_{-G,k})$ of outcomes is observed, drawn from the two groups. The history of observations up to period K is denoted by the vector $\mathbf{t}^K = (t_{G,k}, t_{-G,k})_{k=1,...,K}$.

Based on $\mathbf{t}^K$, and thus on the conditional distributions $f(t|W, \mathbf{t}^K)$ for $W = G, -G$, the DM updates stereotypes and beliefs. In one tractable case, the $k = 0$ initial distribution $f(t|W)$ is also normal for $W = G, -G$. Formally, suppose that $t_W = \theta_W + \varepsilon_W$ where ε_W is i.i.d. normally distributed with mean 0 and variance v, and θ_W is the group specific mean. Initially, groups are believed to be identical, in the sense that both θ_G and θ_{-G} are normally distributed with mean 0 and variance γ. After observing $(t_{G,1}, t_{-G,1})$, the distribution of θ_W is updated according to Bayesian learning. Updating continues as progressively more observations are learned. Thus, after observing the sample $\mathbf{t}^K$, we have:

$$f(t|W, \mathbf{t}^K) = \mathcal{N}\left(\frac{\gamma \cdot K}{v + \gamma \cdot K} \cdot \frac{\sum t_{W,k}}{K}; v \cdot \frac{v + \gamma \cdot (K+1)}{v + \gamma \cdot K}\right). \tag{7}$$

The posterior mean for group W is an increasing function of the sample mean $\sum t_{W,k}/K$ for the same group. The variance of the posterior declines in sample size K, because the building of progressively more observations reduces the variance of θ_W, in turn reducing the variability of outcomes. However, and importantly, because the same number of observations is received for each group, both groups have the same variance in all periods.

Consider now how learning affects stereotypes. Proposition 5 implies:

Proposition 6 *At time K, the stereotype for group G is equal to $[t_G, +\infty)$ if $\sum t_{G,k} > \sum t_{-G,k}$ and to $(-\infty, t_G]$ if $\sum t_{G,k} < \sum t_{-G,k}$. As a result:*

i) Gradual improvement of the performance of group G does not improve that group's exemplar (and only marginally affects its stereotype) provided $\sum t_{G,k}$ stays below $\sum t_{-G,k}$. In

particular, common improvements in the performance of G and $-G$ (which leave $\sum t_{G,k} - \sum t_{-G,k}$ constant) leave stereotypes unaffected.

ii) Small improvements in the relative performance of G that switch the sign of $\sum t_{G,k} - \sum t_{-G,k}$ have a drastic effect on stereotypes.

Even in the normal case, the process of stereotyping suffers from both under- and over-reaction to information. If new information does not change the ranking between group averages, exemplars do not change and stereotypes only respond marginally. Thus, even if a group gradually increases its average, its stereotype may remain very low. On the other hand, even small pieces of information can cause a strong over-reaction if they reverse the ranking between group averages.

5 Group Identity, Gender Stereotypes, and Attitudes towards Mathematics

Group identity plays a key role in sociologists' thinking on group conflict, discrimination, and cultural values. Akerlof and Kranton (2000) construct an economic model of identity, based on the idea that individual preferences depend on one's group membership.[19] By taking identity as given, this approach does not describe how group identity is formed and evolves with the social context.

We propose a stereotype-based model of identity formation and change, in which social context plays a central role. As in our basic model, decision makers form stereotypes by contrasting the features of different social groups. This approach views self-identity as the DM's stereotype of his own group, consistent with the Oxford Dictionary definition of self-identity as "the recognition of one's potential and qualities as an individual, especially in relation to social context." In fact, according to Turner (1985), when group membership is emphasised, "people come to see themselves more as the interchangeable exemplars of a social category than as unique personalities defined by their differences from others".

[19]For a dynamic perspective on identity, including aspects of self-reputation, see Benabou and Tirole (2011).

We develop a model of gender stereotypes along these lines to shed light on the gender gap in attitudes towards education and mathematics. Goldin, Katz and Kuziemko (2006) document that, since the 1930s, women in the US have lagged behind men in average school grades, but started gaining ground in the 1970s, surpassing men in recent years. A similar pattern holds with respect to college enrollment and graduation, with women initially lagging behind but recently overtaking men. Even in mathematics, a recent analysis of performance at the high-school level in the US shows that women are performing essentially as well as men (Hyde et al, 2008). Despite this improvement in overall school performance, men are still over-represented at the highest performance levels in mathematics, in particular in standardised math tests. A much starker difference arises in the choices of college degree, with women disproportionately choosing humanities and health related degrees and careers (Weinberger 2005). This occurs even though there is a significant wage premium to quantitative skills obtained with EMS (engineering, mathematics and science) degrees.[20]

In light of the small gender differences in mathematics test scores, explanations for this gender gap have turned to the role of factors such as gender specific preferences for different fields of study (Croson and Gneezy 2009) or risk aversion (see Bertrand 2011 for a review). One important hypothesis holds that women are less competitive than men and thus more reluctant to pursue the competitive technical fields. Gneezy, Niederle and Rustichini (2003), Niederle and Vesterlund (2007) and others provide evidence in this direction. Recent work, however, suggests that women's preferences for math and math competitions depends on context, and in particular on women's confidence about their relative performance (Niederle and Vesterlund 2008, Dreber, Essen and Ranehill 2012, Coffman 2014). Our model of stereotypes parsimoniously accounts for these disparate pieces of evidence, and delivers new predictions by offering a psychological foundation for the origins of gender identity.

[20]In 2001, women accounted for 57.4% of bachelor degrees in the US, including 85% of degrees in Health professions, 60% in Biology and Life Sciences, 27% in Computer Science and 20% in Engineering (Livingstone and Wirt, 2004). This occurs even though prospective college students are reasonably informed about this wage premium (Betts, 1996).

5.1 Confidence in Math Ability and Attitudes Toward Competition

We begin by considering gender identities formed with respect to performance on standardised math tests, such as the Scholastic Aptitude Test (SAT) or the National Assessment of Educational Progress (NAEP).[21] The score distributions obtained from either assessment have an inverse-U shape for both men and women. The average math scores are only slightly higher for men than for women (531 vs 499 out of 800 on the SAT, 308 vs 304 out of 500 on the NAEP in 2013). For the SAT, the monotone likelihood ratio property holds over nearly the entire range of scores, with men having a heavier right tail than women. Men are twice as likely to have a perfect SAT math score than women.[22]

Several factors might contribute to the observed gender gaps in math and other disciplines: differences in individual effort, innate ability, or investment by third parties (parents or teachers).[23] We first consider a model in which stereotypes reflect stable differences in observable math skills, as proxied by performance in math tests, to investigate how self identity is formed. In section 5.2, we consider math skills as a product of innate ability and effort, and explore how effort choices and self identity are jointly endogenously determined.

Consider a population that varies in mathematical skill z, as proxied by test scores. There are two groups, M and F (male and female). Skill z is normally distributed in group $G = F, M$ with mean z_G and variance σ^2. To capture the evidence on test scores, we let M have a slightly higher average skill than F, $z_M > z_F$. Given that the two groups have the same variance, MLRP holds and M has a heavier right tail than F.

We make the extreme assumption, which we later relax, that individuals are uninformed about their own skill, but observe the full skill distributions of both groups (e.g. through

[21] SAT is a standardised college admission test, so the population taking it is not representative of the full population. Also, more women take the test (53%) which may bias women's results downwards relative to men's. The NAEP conducts yearly assessment of a representative sample's proficiency in several domains, including mathematics and reading. For SAT scores see http://media.collegeboard.com/digitalServices/pdf/research/SAT-Percentile-Ranks-By-Gender-Ethnicity-2013.pdf. For NAEP scores for 17 year olds in mathematics, see http://nationsreportcard.gov/ltt_2012/age17m.aspx. See Hyde et al (2008), Fryer and Levitt (2009), and Pope and Sydnor (2010) for in-depth empirical analyses of the gender gap in mathematics.

[22] About 1% of men, and 0.45% of women, taking the SAT obtain the perfect score of 800, see http://media.collegeboard.com/digitalServices/pdf/research/SAT-Mathemathics-Percentile-Ranks-2012.pdf

[23] The following analysis is in all ways similar to that of gender gap in other disciplines. We focus on mathematics because it is an important driver of subsequent career choices and outcomes, and as a result has been the object of intense attention in the literature.

grades). Because these distributions satisfy MLRP, the stereotype of the slightly better group M lies in the right tail while that of the slightly worse group F lies in the left tail (Proposition 5). Formally, there exist two thresholds $z_M(\delta)$ and $z_F(\delta)$ such that the stereotypical assessments of groups M and F are:

$$z_M^{st} = \mathbb{E}\left[z | z \geq z_M(\delta), M\right], \qquad z_F^{st} = \mathbb{E}\left[z | z \leq z_F(\delta), F\right]$$

It follows that, for $\delta < 1$, stereotypes exaggerate differences in the assessments of skill.

Lemma 1 *When comparing genders along the dimension of math skill, the gender stereotypes satisfy (assuming $\delta < 1$):*

$$z_M^{st} > z_M > z_F > z_F^{st}. \tag{8}$$

Comparing the performance of the two groups leads to self-stereotyping. Women underestimate their skill in math, stereotyping themselves as being worse than they really are, $z_F^{st} < z_F$. Women stereotype men as disproportionately coming from the right tail, and thus overestimate men's skills, $z_M^{st} > z_M$. Similarly, men overestimate their skill in math, and see women as less able. Perceived group differences in skill are large even if true differences are in fact tiny because stereotypes exaggerate the differences among groups.

Proposition 1 has a number of implications that help clarify the literature.

Prediction 1 *Women are less likely to participate in math tasks than men because they underestimate their own math skill, even in non-competitive settings.*

Suppose that participation in a math task entails a cost c, for instance exerting effort in solving a math test or in studying for an engineering college degree. The benefit of participation is equal to math skill z, and the payoff from non participation is zero. Then, in a mixed-gender environment, a member of group G chooses to participate if and only if:

$$z_G^{st} > c. \tag{9}$$

In a mixed-gender environment men are disproportionally more likely to engage in math

related activities than women. This is consistent with the evidence that, controlling for math grades, women are much less likely to choose engineering, mathematics and science majors in college (Weinberger 2005, Bertrand 2011).

This logic is consistent with the so called "stereotype threat" effect, whereby an individual's performance in a task deteriorates when he is reminded of belonging to a negatively stereotyped group. In our model, making a person's group membership salient invites a cross-group comparison that triggers the self-stereotype, even in non-competitive tasks. Coffman (2014) provides compelling evidence that, in team decision making – a cooperative rather than competitive setting – women are under-confident, and men are over-confident (conditional on measured skill), in answering trivia questions in stereotypical male domains such as Geography and Sports.

The literature offers a range of evidence consistent with the importance of confidence to explain participation in mixed-gender math tasks: controlling for performance, women are less confident than men about their ability in math (Eccles 1998, Niederle and Vesterlund 2007, Buser, Niederle and Oosterbeek 2012), and this difference helps account for educational choices (Buser, Niederle and Oosterbeek 2012) and other outcomes (Stein, 2013). Similarly men participate in mathematics tasks because they are overconfident about their abilities, and so are willing to bear the cost of participation.[24]

Starting with Gneezy, Niederle and Rustichini (2003), a large literature shows that the gender gap is even stronger in tournament-like structures in which rewards go the the top performer in a math task. Women are indeed less willing to compete than men in mixed math contests, and the conventional explanation is that women have an intrinsic distaste for competition. While we do not deny the potentially important role of such gender-specific preferences, our model can explain why women are less willing than men to engage in mixed math contests and also makes the following addditional predictions.

Prediction 2 *Women's willingness to compete is shaped by their self-stereotype in the com-*

[24]Some studies also document that both women and men are overconfident with respect to their actual location in the distribution of task scores (but men are more overconfident than women). A literature in Social Psychology documents that individuals are overconfident in tasks where their absolute performance is good (e.g. driving), and under confident in tasks where their absolute performance is poor (e.g. juggling), see Moore and Cain (2007). We do not address this issue because it requires an analysis of how individual members of a group stereotype themselves relative to other members of the same group.

petitive environment. In particular:

i) Women are unwilling to compete in mixed-gender math tournaments.

ii) Women are willing to compete in single-gender math tournaments.

iii) Women are willing to compete against men in areas that are stereotypically neutral or stereotypically female.

Once again, suppose that to participate in a math tournament a DM must bear a cost c. The participant with the highest skill receives a prize larger than c, the other receives zero. If participants have the same skill, each gets the prize with probability $1/2$.

Case i) captures the stylised fact (Gneezy et al 2003, Niederle and Versterlund 2007) that, controlling for performance, women are less likely to choose a tournament-based compensation scheme than a piece-rate scheme in a mixed-gender math contest. According to (8), when competing against men, women underestimate their own skill, and overestimate their opponent's skill. Thus, they attach a lower probability to winning and are less likely to participate in the tournament than they would be under rational beliefs.[25] Consistent with our model, the evidence shows that after controlling for confidence the gender gap in tournament entry diminishes significantly (Niederle and Vesterlund 2007) or vanishes altogether (Dreber, Essen and Ranehill 2012).[26] Similarly, Buser, Niederle and Oosterbeek's (2012) evidence on the choice of education path in Dutch high schools suggests that the gender gap is significantly, though not entirely, reduced once confidence is taken into account.

Consider now case ii). When competing against other women, women perceive their skill distribution correctly and thus attach probability 0.5 to winning. Women now are as likely to enter a single-sex competition as men. This prediction is confirmed by a range of experimental and field evidence. Gneezy, Niederle and Rustichini (2003) document that

[25]Formally, if stereotypes are sufficiently severe (δ is low enough) so that $z_M(\delta) > z_F(\delta)$, women attach zero probability to the outcome of outperforming a male competitor. In the less extreme case where $z_M(\delta) < z_F(\delta)$, women attach some probability to competing with a man who is less able than them. It is still the case, though, that women are reluctant to compete against men, given that they stereotype the latter as disproportionally coming from the right tail.

[26]Measuring confidence as estimated rank in past tasks, Niederle and Vesterlund (2007) suggest that gender differences in confidence account for only 27% of gender differences in tournament entry. This suggests a smaller role of confidence than do Dreber et al (2012) and Buser et al (2012). One possible reason is that estimation of future performance may be systematically different from estimation of past relative performance or of overall skill, particularly if – due to stereotypical thinking – past positive performance is perceived as simply due to luck. Broader measures of confidence include subjective beliefs about own skill in math (as in Buser et al 2012), and estimates of future performance.

women are as likely as men to enter single-gender math tournaments. Niederle and Vesterlund (2008) show that introducing quota-like (affirmative action) schemes into tournaments boosts female participation. Like Niederle and Vesterlund, we interpret quotas as making the tournament more like a single-sex competition. In explaining the evidence, however, we do not assume that female distaste of competition falls in less mixed environments; instead, we argue the quotas change the self-stereotype of participants by changing the group they compare themselves to. Specifically, in our model quotas exert two effects. First, they reduce the probability that a woman attaches to competing against right-tail men, encouraging participation. Second, by moving the setting towards a single-sex tournament, quotas relieve the stereotype threat for women, improving their confidence. Gneezy et al (2003) find that women's performance on single sex tournaments is significantly better than when competing against men. Their analysis shows this results from higher effort in single-sex tournaments (particularly from women of average skill), in agreement with our prediction on the impact of tournament structure on beliefs about skill.[27] Finally, Booth and Nolen (2009) offer suggestive evidence that women educated in single-sex schools are as competitive as men, even in mathematics.

Consider now case iii). A stereotypically neutral activity is characterized by $z_M = z_F$, while a stereotypically female activity has $z_M < z_F$. Women now (weakly) over-estimate their skill relative to men's, and by reversing the previous argument, are (weakly) more likely to participate in the tournament. A recent but growing body of experimental evidence shows that, in verbal tasks, women are as likely to compete as men (Günther, Ekinci and Schwieren 2010, Grosse and Riener 2010, Kamas and Preston 2012a,b, Dreber, Essen and Ranehill 2012). Shurchkov (2012) shows that women outperform male competitors in verbal tasks, particularly under reduced time pressure. Verbal tasks are seen as weakly stereotypically female,[28] so the evidence that women are as competitive as, but not necessarily more

[27]Evidence from Gneezy et al (2003) also suggests that men increase effort slightly (though not significantly) when competing against women, as compared to single sex tournaments.

[28]The stereotype that women are better than men at verbal tasks is generally perceived to be weaker than the stereotype that men are better at math. Kimura (1999) suggests that women are better at verbal association, but not in verbal fluency. This view is consistent with SAT scores: men are 30 points ahead in math, 5 points ahead in reading and 10 points behind in writing (out of a total of 800 points). Moreover, the monotone likelihood ratio property holds for math scores, with the likelihood ratio (men/women) ranging from 3 at the highest scores down to 0.5 at the lowest observed scores. In comparison, in the writing and

competitive than, men is in line with our prediction. Coffman (2014) shows that women are both better and more overconfident than men in knowledge of Art History and Pop Culture, but the reverse holds for questions about Geography and Sports. In agreement with our model, Art History and Pop Culture are perceived by the subjects as being stereotypical female domains, while Geography and Sports are perceived as stereotypical male domains.

Turning to field evidence, Flory, Leibbrandt and List (2010) show in a natural field experiment that both men and women are less likely to apply for jobs in which the compensation scheme depends on relative performance, but while there is a large gender gap when the job has male connotations (sports news assistant), this gap disappears when these connotations are absent (news assistant). Smith (2013) shows that girls' performance is as resistant to competitive pressure as boys' in the National Spelling Bee competition.

So far we have considered the extreme scenario in which individuals are completely uninformed about their own performance and form stereotypes and inferences based on the population distribution of skill z. Our results extend to the case where individuals observe their own performance (e.g., from test scores) and use that information in forming their self-stereotype. For simplicity we assume that, after observing performance t, the expected skill of a member of group G is $(1-\alpha)z_G + \alpha t$, where α increases in the signal to noise ratio.[29] Thus, even a woman whose performance is above the male average, $t > z_M$, may be stereotyped as not good at math provided $t < z_F + \frac{1-\alpha}{\alpha}(z_M - z_F)$. The reason is that, even if her average performance is good, high level performance from a member of group F is perceived as a fluke while representative performances come from the left tail. By the same token, even men whose performance lies below the average female performance, $t < z_F$, can be stereotyped as good at math provided $t > z_M - \frac{1-\alpha}{\alpha}(z_M - z_F)$. This logic delivers an interesting prediction:

Prediction 4 *The gender gap in attitudes towards mathematics is stronger for individuals of "average" skill, whose performance is located close to the population mean.*

reading sections of distribution of SAT scores are much more similar across genders. However, other results suggest that at a younger age women are far better than men at reading (Pope and Sydnor 2010, Guiso, Monte, Sapienza and Zingales 2008).

[29]The underlying assumption is that own performance t does not affect the group stereotype encoded in z_G^{st}. This simplifies the analysis, but similar results obtain when individual performance shapes the representativeness of skill realisations for that individual.

To see this, note that when individual performance is extreme, it dominates the individual's choices: even though on average women's self-assessment is below that of men, very talented individuals of either gender are likely to participate in math tournaments, and left tail individuals of either gender are not. However, because average women have significantly worse self-stereotypes than average men, this greatly affects their choices to participate. This finding is consistent with both experimental and field evidence. Niederle and Vesterlun (2007) find that too few high skill women, and too many low skill men, enter the competitive tournament. Buser, Niederle and Oosterbeek (2012) document that the gender gap in curriculum choice shows up precisely at the mean: while average men choose highly mathematical curricula, average women choose very humanities-intensive curricula, so that women are over-represented in the latter while men are over-represented in the former.

Our predictions are consistent with the literature on the gender gap in performance in mathematics, but also with the broader literature on identity (Akerlof and Kranton 2000, Bénabou and Tirole 2011), self-categorisation (Turner 1985, Benjamin, Choi, Strickland 2010) and stereotype threat (Steele and Aaaronson 1995).[30] Our account provides a new mechanism that helps link these different concepts. The social context defines the group individuals compare themselves to and thus shapes their self-stereotypes. Applied to performance in mathematics, competition between genders primes each gender to evoke its (self-)stereotypes. The resulting perception of own skill distorts incentives to participate and provide effort.

5.2 Gender Stereotypes in Equilibrium

In our model so far, stereotypes are formed on the true distribution of math skill z, as proxied by performance in tests. Performance in math tests, however, reflects in part innate ability and in part learning effort by individuals.[31] Effort and ability are difficult to measure directly, but we show that - by shaping effort - stereotypes about ability emerge in equilibrium even

[30] According to Turner (1985), feelings of identity need not be permanent but can be primed by social interactions. Related, the literature on stereotype threat holds that when group membership is associated with a negative (positive) stereotype, emphasising it provokes feelings of anxiety (or elation) that affect performance in a way that confirms the stereotype.

[31] There is also a well established role of investment by third-parties, such and families and teachers (Carrell, Page and West 2009). Here we focus on the role of self-stereotypes and decisions about individual effort.

with no underlying ability differences between groups.

Suppose that skill z is given by:

$$z = e + \theta_G + \varepsilon. \tag{10}$$

Here e is individual-level effort, θ_G is the average innate ability in group G, and ε captures the individual effect in innate talent or in the productivity of effort (due, for instance, to non-cognitive skills). The realization of ε is not yet known when the individual chooses his or her effort level e. We assume that ε is distributed normally, with mean 0 and variance σ^2.

The effort choice is determined by long run economic returns from math skills. There is a convex effort cost $c(e)$ and a "mincerian return" π to skills in the market. The expected market wage is $\mathbb{E}[w(z)] = \mathbb{E}\left[e^{\pi z}\right] = e^{\pi \mathbb{E}[z] + \pi^2\sigma^2/2}$.

A fully rational and risk neutral individual chooses effort e to maximize $\mathbb{E}[w(z)] - c(e)$, which yields:

$$c'(e) = \pi e^{\pi(e+\theta_G)+\pi^2\sigma^2/2}. \tag{11}$$

This equation identifies an increasing function $e^*(\cdot)$ such that an individual who perceives his average ability to be θ_G exerts the effort level $e^*(\theta_G)$.[32] Individuals and thus groups with higher innate ability invest more and their observed higher math skills are due to both higher ability and higher effort.

We now turn to stereotypes. As in the rational case, the decision maker trades off the cost of effort $c(e)$ against the return associated with the skill distribution $z(e)$. However, by comparing the observed skill z in the two groups, the DM develops stereotypes about each group's innate abilities. This boils down to individuals forming a stereotype for realisations ε of each group. To see this, suppose for simplicity that every individual in group $G = M, F$ chooses the same effort e_G. As a consequence, skill in group $G = M, F$ is normally distributed with mean $\mathbb{E}[z|G] = e_G + \theta_G$ and variance σ^2. The logic of Section 5.1, then implies the following property.

Lemma 2 *When $e_G + \theta_G > e_{-G} + \theta_{-G}$, the self-stereotype of G truncates the left tail of z, so that $\mathbb{E}^{st}[\varepsilon|G] > 0$. When $e_G + \theta_G < e_{-G} + \theta_{-G}$, the self stereotype of $-G$ truncates the*

[32] We assume that the cost function $c(\cdot)$ is sufficiently concave that the f.o.c. identifies a maximum.

right tail of z, so that $\mathbb{E}^{st}[\varepsilon|G] < 0$. When $e_G + \theta_G = e_{-G} + \theta_{-G}$, assessments are correct and $\mathbb{E}^{st}[\varepsilon|G] = 0$.

When men exhibit higher average skill (10) than women, $e_M + \theta_M > e_F + \theta_F$, the stereotype for a man is to have right tail ability, while that for a woman is to have left tail ability.[33] Formally, $\mathbb{E}^{st}(\varepsilon\,|M) > 0 > \mathbb{E}^{st}(\varepsilon\,|F)$. Intuitively, men are disproportionally common in the right tails of the test scores distribution, which evokes images of men with high innate ability and leads women to stereotype themselves as a low innate ability group. When men have lower average skill – namely $e_M + \theta_M < e_F + \theta_F$ – the reverse is true, in that $\mathbb{E}^{st}(\varepsilon\,|M) < 0 < \mathbb{E}^{st}(\varepsilon\,|F)$. When men and women have the same average skill, all performance levels are equally representative for both groups and assessments are on average correct, in that $\mathbb{E}^{st}(\varepsilon\,|M) = \mathbb{E}^{st}(\varepsilon\,|F) = 0$.

Critically, math skills depend not just on innate ability but also on effort. As in models of statistical discrimination, the fact that stereotypes depend on endogenous effort choices can create self fulfilling stereotypes: a positive stereotype begets higher effort which in turn confirms the stereotype itself.[34] Indeed, a member of group G chooses an optimal effort level equal to $e^*(\theta_G^{st})$, where $\theta_G^{st} = \theta_G + \mathbb{E}^{st}(\varepsilon\,|G)$. An equilibrium thus consists of effort levels (e_F^*, e_M^*) and stereotypes $(\theta_F^{st}, \theta_M^{st})$ such that: i) effort levels are optimal given stereotypes, namely $e_F^* = e^*(\theta_F^{st})$ and $e_M^* = e^*(\theta_M^{st})$, and ii) stereotypes are endogenously confirmed, namely $e_M^* + \theta_M > e_F^* + \theta_F$ if and only if $\theta_M^{st} > \theta_F^{st}$.

In the special case in which the two groups have exactly the same average innate ability, the equilibria are as follows.

[33] We implicitly assume that individuals form stereotypes about their ability based on the performance of their group as a whole, neglecting the individual effort differences. An alternative specification, in which effort differences are not neglected, is as follows: when choosing an effort level e, an individual of group G forms his self-stereotype by comparing his skill distribution $\theta_G + e + \varepsilon$ to the equilibrium skill distribution of group $-G$. This individual stereotypes himself as a right tail individual if and only if he provides sufficiently high effort, $\theta_G + e > \theta_{-G} + e_{-G}^*$. In this model, stereotypes only arise when there are underlying group differences in ability, but stereotyping exaggerates them. An individual of the lower ability group G might consider providing enough effort to match $-G$'s skill level, but because the effort of $-G$ is inflated (all members of $-G$ think they are right tail individuals) it is too costly to do so. Only extreme effort levels are consistent with the stereotypes.

[34] The literature of statistical discrimination is also concerned with how beliefs can be self-fulfilling when effort provision is endogenous. However, that literature is focused on the beliefs of others about the DM, while here we emphasise stereotypical beliefs about the self. Below we expand on the links and differences between stereotypes and statistical discrimination.

Proposition 7 *When $\theta_F = \theta_M = \theta$, there are three possible equilibria:*

i) Women have a negative stereotype, and $e_F^* < e^*(\theta) < e_M^*$.

ii) Men have a negative stereotype, and $e_M^* < e^*(\theta) < e_F^*$.

iii) Stereotypes are correct, and $e_M^* = e_F^* = e^*(\theta)$.

Consider case i). Because women believe they have lower ability, $\theta_F^{st} < \theta_M^{st}$, they also exert less effort at math, $e_F^* < e_M^*$. As a consequence, women indeed perform worse than men at math, confirming the negative female stereotype in equilibrium. A similar logic holds with respect to the positive male stereotype. In case iii), individuals hold correct expectations on their equal innate ability levels and exert the same effort, which are consistent with correct stereotypes. Among these three possible equilibria, only those with incorrect stereotypes are stable.[35]

Proposition 7 shows that stereotypes and the ability distributions influence each other in the long run. Stereotypes can cause prior beliefs concerning the abilities of women and men to become self reinforcing, even though these beliefs may not be based on direct experience. Prior belief that men have greater ability, $\theta_F < \theta_M$, render an equilibrium in which women suffer a negative stereotype more likely. The negative stereotype, in turn, causes even larger observed differences in skill through differential effort choices. This logic implies that prior beliefs about groups' abilities may lead, through effort choices, to outcomes that are unfounded in the groups' true underlying characteristics. A society can hold a negative stereotype about women (and a positive stereotype about men) even if women have on average higher innate ability than men, as long as under those stereotypes men put in sufficient effort to counterbalance women's ability advantage.[36]

The model predicts that societies with greater beliefs of gender disparity exhibit greater gender gap in performance in tasks that are strongly stereotypical of a particular gender. This prediction is confirmed by several studies that link the gender gap in mathematics to measures of gender (in)equality, both at the country level (Guiso et al 2008, Fryer and Levitt

[35]In fact, a small increase in effort provision by a group would displace accurate beliefs, leading to a breakdown of equilibrium iii).

[36]Formally, $(\theta_G^{st} - \theta_{-G}^{st}) \cdot (\theta_G - \theta_{-G}) < 0$ holds if and only if the difference in true average abilities is smaller than the difference in optimal effort provision under the stereotypical abilities. In particular, if $0 < \theta_G - \theta_{-G} < e^*(\theta_{-G}^{st}) - e^*(\theta_G^{st})$ then $\theta_{-G}^{st} > \theta_G^{st}$.

2010) and at a state level in the US (Pope and Sydnor 2010). Consistent with our model, both Guiso et al (2008) and Pope and Sydnor (2010) also find that the negative gender gap in women' performance in mathematics is correlated with a positive gender gap in women' performance in reading tasks, which are more typically associated with women.[37]

The result that endogenous effort might lead to self-fulfilling beliefs about skill also obtains in models of statistical discrimination (Phelps 1972, Arrow 1973, Coate and Loury 1993). That approach focuses on the beliefs of others (including potential employers) about one's ability, which shape returns to effort. In contrast, our approach highlights the role of self identity, which allows us to account even for phenomena in which beliefs held by others play no role.[38,39]

The analysis of stereotype evolution in Section 3 may provide further insights into the dynamics of the gender gap. Goldin et al (2006) document a large increase in schooling and college attendance for both men and women in the US in the middle of the 20th century, but a greater increase for men. Consistent with Proposition 3, greater access to education for both groups did not lead to a change in the stereotype that higher education is a stereotypical male activity. More recently, however, there has been a disproportionate increase in the educational achievement of women. According to Proposition 4, this can eventually reverse the stereotype for educational attainment, so that attending college might become

[37]Equilibria of type ii) can arise in societies where cultural norms assign a larger role to women. Gneezy, Leonard and List (2009) document the case of the Khasi, a matrilineal society in India, in which inheritance and clan membership are transmitted through the female lineage. Among the Khasi, women choose to compete in the authors' experiment at twice the rate of men. This suggests a reversal of roles relative to the previous case, such that women have a positive self-stereotype while men have a negative self-stereotype.

[38]To fully evaluate the role of stereotypes in education efforts and outcomes it is necessary to also consider the stereotypes of prospective employers. In case i) of Proposition 7, both men and women have a negative stereotype of women, so it is rational for an individual woman to cut effort (even if she has an accurate perception of her ability). Moreover, employees who are hired through affirmative action may be stereotyped as less competent, independently of their own ability, reinforcing the negative stereotypes of their group. On the other hand, if there is an improvement of stereotypes, then the effect is a double-whammy. In the case where women's stereotypes become positive the assessment of women's ability increases, causing women to put in more effort and be compensated by higher recognition in the job market. This can lead to a new high effort/high performance equilibrium for this group.

[39]Disentangling stereotypes from statistical discrimination in the market is beyond the scope of this paper. We note, however, that our model makes the strong prediction that removing the effect of negative stereotypes (e.g. by reducing the tournament-like structure of job allocation or using the single-sex tournament) which is related to establishing quotas for women (Niederle and Vesterlund 2008) leads to an *increase* in effort provision (though this may occur under some circumstances in models of statistical discrimination as well, see Coate and Loury 1993).

a stereotypical female activity. In fact, Goldin et al (2006) suggest that the gender gap in overall college education is outright reversed.

More generally, innovations that affect the genders asymmetrically can drive changes in gender stereotypes. Examples of such innovations include social innovations (the feminist movement), technological innovations (home appliances taking over previously stereotypical female activities), and interactions between the two (the contraceptive pill allowing women to better plan and balance work and family).

To conclude, we note that gaps in educational achievement exist not only across genders, but across ethnic and socio-economic groups. Stereotypes might also be instrumental in understanding why, for instance, individuals from poor backgrounds underinvest in education. We return to this point in the conclusion.

6 Likelihood, Availability, and Stereotypes

The assumption that stereotypes are formed based only on representativeness allows for a tractable model of stereotype formation and change. The model yields clear predictions that account for many characteristics of stereotypes described by psychologists (Hilton and Von Hippel, 1996), including the fact that stereotypes emphasise group differences and often minimise within group variation, as a consequence of the central (and previously unexplored) insight that stereotypes are context dependent.

At the same time, as we discussed in Section 2.2, our formulation leads in some instances to extreme predictions and, importantly, it neglects other factors that influence what features come to mind when thinking about a group, such as likelihood and availability.[40] When stereotyping the occupation of a democratic voter, people think about "professor" rather than a "comparative literature professor." While the latter is probably more representative, the former is more likely and thus it comes to mind more easily. When US survey respondents think about Arabs in the aftermath of 9/11, terrorists rather than Bedouins are more likely to come to mind because, even though the latter are more likely and more representative (all

[40] According to Kahneman and Frederick (2005) "the question of why thoughts become accessible – why particular ideas come to mind at particular times – has a long history in psychology and encompasses notions of stimulus salience, associative activation, selective attention, specific training, and priming".

Bedouins are Arabs, but not all terrorists are Arabs), the former are more available because they are more frequently brought up in news sources (see footnote 7).

In this section we show that our model can be easily adapted to account for some effects of likelihood on recall. When we do so, our predictions become less extreme, in the sense that stereotypes become centered around relatively more likely or available types, but the distortions of stereotypes still follow the logic of representativeness, as in our main analysis. This extension can also capture the effects of a crude measure of availability on recall.

Suppose that the ease of recall of a type t for group G is given by:

$$R_k(t,G) = \frac{\pi_{t,G}}{\pi_{t,-G} + k} = \frac{1}{\frac{1}{R(t,G)} + k \cdot \frac{1}{\pi_{t,G}}} \tag{12}$$

where $k \geq 0$ and $R(t,G)$ is representativeness as defined in Definition 1. In Equation (12), the ease of recalling type t increases when that type is more representative, namely when $R(t,G)$ is higher, but also when type t is more likely in group G, namely when $\pi_{t,G}$ is higher. The value of k modulates the relative strength of these two effects: for small k, representativeness drives ease of recall, while for large k likelihood drives recall.[41]

In this new formulation, the stereotype is formed as in Definition 2 except that now what comes to mind are the d types that are easiest to recall. When representative types are also likely, as in case iii) of Proposition 1, recall based on Equation (12) does not change the stereotype for group G. When instead representativeness and likelihood differ for group G, as in cases i) and ii) of Proposition 1, recall driven by $R_k(t,G)$ may yield a different stereotype than a pure representativeness model.

To see how the model can capture some features of availability, note than the term $\pi_{t,G}$ in (12), and also in (1), may be broadly interpreted as capturing the availability, rather than just the frequency, of type t for group G. Formally, in the model of learning of Section 3, we would assume that the estimate of $\pi_{t,G}$ is determined by the share of observations from G that are of type t, even if these observations are not independent. Thus, as the same episodes of terrorism are mentioned repeatedly in the news, their ease of recall is inflated. In

[41] When $k = 0$, we are in a pure representativeness model. As k increases, likelihood becomes progressively more important in shaping recall relative to representativeness. As $k \to \infty$, only likelihood matters for shaping recall and stereotypes.

this approach, availability is related to neglect of the correlation structure of information (as discussed in Section 2.2, understanding the psychology of availability is beyond the scope of this paper).

The concrete implications of Equation (12) are best seen in the case where the type space is continuous, and more specifically when t is normally distributed in groups G and $-G$, with means μ_G, μ_{-G} respectively, and variance σ. In this case, the easiest to recall type t for group G is given by:

$$t_{E,G} = \text{argmin}_t e^{\frac{((t-\mu_G)^2-(t-\mu_G)^2}{2\sigma^2}} + k \cdot e^{\frac{(t-\mu_G)^2}{2\sigma^2}}$$

When $\mu_G > \mu_{-G}$, the easiest to recall type $t_{E,G}$ satisfies:

$$k \cdot (t_{E,G} - \mu_G) \cdot e^{\frac{\left(t_{E,G}-\mu_G\right)^2+2(\mu_G-\mu_{-G})\cdot\left(t_{E,G}-\frac{\mu_G+\mu_{-G}}{2}\right)}{2\sigma^2}} = \mu_G - \mu_{-G} \tag{13}$$

The left hand side of (13) is increasing in $t_{E,G}$, which implies that $t_{E,G}$ is a strictly increasing function of k satisfying $\lim_{k\to\infty} t_{E,G}(k) = \mu_G$ and $\lim_{k\to 0} t_{E,G}(k) = \infty$. In words, the group G with higher mean is stereotyped with an inflated assessment that goes in the direction of the most representative type $t = \infty$. The extent of this inflation increases as k gets smaller. The stereotype for group G in this case is an interval around the easiest to recall type that captures a total probability mass of δ (truncating both tails, but especially the left one). Moreover, as in the case $k = 0$, the stereotype has a lower variance than the true distribution. A corresponding result is obtained if group G has a lower mean than $-G$.

This analysis implies that the basic insights that stereotypes emphasise differences, and lead to base rate carry through to this case, as does the analysis of self-stereotypes.[42] The dynamic updating of stereotypes of Proposition 6 carries over to this formulation as well: as long as the ranking of group means is maintained, group stereotypes (in the sense of which group has a positive or negative stereotype) do not change.

[42]In the extended model given by (12) and (2), the parameters δ and k capture two natural types of bounds on recall: δ determines "how much" comes to mind (which might depend on effort), while k corresponds to the relative weight of likelihood in recall, which may vary across people.

7 Conclusion

We presented a model of stereotypical thinking, in which decision makers making predictions about a group recall only a limited range of the group's types or attributes from memory. Recall is limited but also selective: the recalled 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 types about the group relative to other groups.

Our approach provides a parsimonious and psychologically founded account of how DMs 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. First, 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 any validity. In a dynamic setting, the model explains when DMs under- or over-react to information. In particular, the model accounts for stereotype persistence and stereotype change.

This same logic allows us to describe a number of heuristics and psychological biases, many of which arise in the context of prediction problems. Our model generates both base-rate neglect and confirmation bias (and makes novel predictions for when they occur). To our knowledge, ours is the first model to reconcile these two patterns of behavior, and in fact shows they both arise out of the assumption of representativeness-based recall. The approach can also unify several other biases, such as overconfidence but also – under appropriate extensions not discussed in the paper – polarisation effects.[43]

In a different vein, we show how stereotypes provide fresh insight into the notion of identity, both in terms of distorted beliefs about one own's abilities (self-identity) and about others (discrimination). Applied to gender stereotypes, our model offers a parsimonious account of many findings in a vast literature covering the gender gap in educational attainment, in competitive settings and in labor outcomes. Group membership generates self-stereotypes, which distort incentives to invest in education and to participate in the job market. The same mechanism can also account for a variety of stylised facts on educational attainment across

[43]Polarization arises as a consequence of confirmation bias when DMs have heterogeneous priors. Proposition 3 then implies that a given set of observations can lead different DMs with different stereotypes to each reinforce their own stereotype, and thus update in opposite directions.

socio-economic groups, as described for instance in Banerjee and Duflo (2011). Stereotyping a highly educated person as having a good government job, and a less educated person as being unemployed, generates counter-factual beliefs that returns to education are convex. This discourages students – particularly those who for a variety of reasons do not expect to attain the maximum level of education – from investing even moderately in schooling.

Our model is centrally based on representativeness and it does not capture all the features of stereotypical thinking. However, we argue that it captures perhaps the central feature, namely that when we think of a group, we focus on what is most distinctive about it, and neglect the rest. Many aspects of social preferences might thus be interpreted as stereotypical beliefs: predictable, persistent and yet evolving with circumstances.

References

Akerlof, George and Rachel Kranton, 2000. "Economics and Identity." *Quarterly Journal of Economics* 115 (3): 715 – 753.

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.

Banerjee, Abhijit and Esther Duflo. 2011. *Poor Economics*. PublicAffairs.

Bénabou, Roland and Jean Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets." *Quarterly Journal of Economics* 126 (2): 805 – 855.

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

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

Betts, Julian. 1996. "What Do Students Know about Wages? Evidence from a Survey of Undergraduates." *Journal of Human Resources* 31 (1): 27 – 56.

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

Booth, Alison and Patrick Nolen. 2009. "Gender Differences in Competition: the Role of Single-Sex Education." CEPR Discussion Paper 7214.

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.

Buser, Thomas, Muriel Niederle and Hessel Oosterbeek. 2012. "Gender, Competitiveness and Career Choices." *Quarterly Journal of Economics*, forthcoming.

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.

Coate, Stephen and Glenn Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83(5): 1220 – 1240.

Coffman, Katherine. 2014. "Gender Stereotypes Silence Experts (Even in the Absence of Discrimination)." Mimeo Ohio State University.

Croson, Rachel and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47(2): 448 – 474.

Dreber, Anna, Emma Von Essen and Eva Ranehill. 2012. "In Bloom: Gender Differences in Preferences Among Adolescents." Mimeo Stockholm School of Economics.

Eccles, Jacquelynne. 1998. "Perceived Control and the Development of Academic Motivation: Commentary." *Monographs of the Society for Research in Child Development* 63 (2-3): 221 – 231.

Flory, Jeffrey, Andreas Leibbrandt and John List. 2010. "Do Competitive Work Places Deter Female Workers? A Large-Scale Natural Field Experiment on Gender Differences in Job-Entry Decisions." NBER Working Paper 16546.

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

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.

Gneezy, Uri, Muriel Niederle and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *Quarterly Journal of Economics* 118(3): 1049 – 1074.

Gneezy, Uri, Kenneth Leonard and John List. 2009. "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society." *Econometrica* 77(5): 1637 – 1664.

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.

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

Grosse, Niels and Gerhard Riener. 2010. "Explaining Gender Differences in Competitiveness: Gender-Task Stereotypes." Jena Economic Research Series, No. 2010 – 017.

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

Günther, Christina, Neslihan Arslan Ekinci and Christiane Schwieren. 2010. "Women can't jump? An Experiment on Competitive Attitudes and Stereotype Threat." *Journal of Economic Behavior & Organization* 75(3): 395 – 401.

Heckman, James and Bo Honoré. 1990. "The Empirical Content of the Roy Model," *Econometrica* 58(5): 1121 – 1149.

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

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

Kahneman, Daniel, and Shane Frederick. 2005. "A Model of Heuristic Judgment," in The Cambridge Handbook of Thinking and Reasoning, Keith J. Holyoak and Robert G. Morrison, eds. Cambridge, UK: Cambridge University Press.

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.

Kamas, Linda and Anne Preston. 2012a. "Are Women Really Less Willing to Compete Than Men? Gender Stereotypes, Confidence, and Social Preferences." Mimeo Haverford College.

Kamas, Linda and Anne Preston. 2012b. "The Importance of Being Confident: Gender, Career Choice, and Willingness to Compete." *Journal of Economic Behavior & Organization* 83(1): 82 – 97.

Kimura, Doreen. 1999. *Sex and Cognition.* MIT Press, Cambridge MA.

Lord, Charles, Lee Ross, and Mark Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37(11): 2098 – 2109.

Moore, Don and Daylian Cain. 2007. "Overconfidence and underconfidence: When and Why People Underestimate (and Overestimate) the Competition." *Organizational Behavior and Human Decision Processes* 103(2): 197 – 213.

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

Nickerson, Raymond. 1998. "Confirmation Bias: a Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2(2): 175 – 220.

Niederle, Muriel and Lise Vesterlund, 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (3): 1067 – 1101.

Niederle, Muriel and Lise Vesterlund. 2008. "Gender Differences in Competition." *Negotiation Journal* 24 (4): 447 – 463.

Niederle, Muriel and Lise Vesterlund, 2011. "Gender and Competition" *Annual Review of Economics* 3(1): 601 – 630.

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.

Reinhart, Carmen and Kenneth Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly.* Princeton, NJ: Princeton University Press.

Schwartzstein, Joshua. 2014. "Selective Attention and Learning." *Journal of the European Economic Association*, forthcoming.

Shurchkov, Olga. 2012. "Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints." *Journal of the European Economic Association* 10 (5): 1189 – 1213.

Smith, Jonathan. 2013. "Peers, Pressure, and Performance at the National Spelling Bee." *Journal of Human Resources* 48(2): 265 – 285.

Steele, Claude and Joshua Aronson. 1995. "Stereotype Threat and the Intellectual Test Performance of African Americans." *Journal of Personality and Social Psychology* 69(5): 797 – 811.

Stein, Carolyn. 2013. "Confidence Counts: The Gender Gap at High Levels of Math Achievement." BA Thesis, Harvard College.

Tajfel, Henri. 1982. "Social Psychology of Intergroup Relations." *Annual Review of Psychology* 33 (1): 1 – 39.

Turner, J.C. 1985. "Social Categorization and the Self-Concept: A Social Cognitive Theory of Group Behavior." In Lawler, E. J. ed. *Advances in Group Processes: Theory and Research.* Greenwich, CT: JAI Press.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124 – 1131.

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

Livingston, Andrea and John Wirt. 2004. "The Condition of Education 2004 in Brief", U.S. Department of Education Institute of Education Sciences NCES 2004 – 076.

A Proofs

Remark 1. We first establish that representativeness $R(x, S)$ of a type t for group S increases in the likelihood ratio $Pr(X = x|S)/Pr(X = x| - S)$. From Definition 1, $R(x, S) = Pr(S|X = x)/Pr(-S|X = x)$. Using Bayes' rule, $Pr(S|X = x) = Pr(X = x|S) \cdot Pr(S)/Pr(X = x)$ and similarly for $-G$, leading to

$$R(x, S) = \frac{Pr(X = x|S)}{Pr(X = x| - S)} \cdot \frac{Pr(S)}{Pr - S}$$

Given that $Pr(S)$ and $Pr - S$ do not depend on t, the result follows.

Denote the conditional probability $Pr(X = x|W)$ by $\pi_{t,W}$. Rewrite the likelihood ratio as

$$\frac{\pi_{t,G}}{\pi_{t,-G}} = \frac{(\pi_{t,G} - \pi_{t,-G}) + \pi_{t,-G}}{\pi_{t,-G}}$$

from which result i) follows immediately. Moreover, it is clear that keeping the difference $\pi_{t,G} - \pi_{t,-G}$ fixed, the likelihood ratio decreases with the baseline probability $\pi_{t,-G}$ if and only if the difference is positive. ■

Proposition 1. The likelihood ratio of type t is (up to a normalizing constant) equal to $\pi_{t,G}^{1-\alpha}$. If $\alpha > 1$, then this ratio decreases with the probability $\pi_{t,G}$ of t in G, so that the representativeness ranking and the likelihood ranking of types for G are the opposite. This proves case i). If $\alpha < 1$, this ratio increases with $\pi_{t,G}$, so that the representativeness ranking and the likelihood ranking of types for G coincide. This shows the G part of cases ii) and iii). If $\alpha > 0$ the distributions for G and $-G$ are co-monotonic so that $-G$ is stereotyped by its least likely types, while if $\alpha < 0$ the distributions of G and $-G$ have opposite likelihood rankings of types, so that $-G$ is stereotyped by its most likely types. ■

Proposition 2. Index the types $t \in \{1, \ldots, N\}$ according to the "natural" ordering relation (e.g. type 1 is on the left and type T is on the right). Suppose the likelihood ratio $\pi_{t,G}/\pi_{t,-G}$ is monotonically decreasing in t. Then it follows that the average over G is lower than the average over $-G$, and therefore lower than the unconditional average, $\mathbb{E}(t|G) < \mathbb{E}(t)$. Moreover, the ordering of types by representativeness coincides with the natural ordering of

types, so that the stereotype consists of types 1 through d. By truncating the upper tail, it follows that $\mathbb{E}^{st}(t|G) < \mathbb{E}(t|G)$.

If the the likelihood ratio is monotonically increasing in t, then the ordering of types by representativeness coincides with the inverse of the natural ordering of types, so that the stereotype consists of types $N - d + 1$ through N. By truncating the lower tail, it follows that $\mathbb{E}^{st}(t|G) > \mathbb{E}(t|G) > \mathbb{E}(t)$. ■

Proposition 3. We assume that the same number of observations are received at each stage of the learning process for both groups G and $-G$. This assumption is not restrictive, since only the relative frequency of observations matter. In particular, all probabilities remained unchanged if the sample size of one group is scaled up relative to the sample size of the other. Thus we can set $\sum_{t'} a_{t',G} = \sum_{t'} a_{t',-G} = a$ and $\sum_{t'} n_{t',G} = \sum_{t'} n_{t',-G} = n$.

Representativeness of a type t is now measured by the ratio

$$\frac{Pr(X = x|\alpha_S, \mathbf{n}_S)}{Pr(X = x|\alpha_{-G}, \mathbf{n}_{-G})} = \frac{\alpha_{t,G} + n_{t,G}}{\alpha_{t,-G} + n_{t,-G}}$$

Consider case i) where all observations occur in type t, so that $n_{t,G} = n$ and $n_{t',G} = 0$ for $t' \neq x$, and similarly for $-S$. Then the representativeness of types other than t do not change, while the representativeness of t is $(\alpha_{t,G} + n)/(\alpha_{t,-G} + n_{t,-G})$. This tends to one monotonically as n increases. Therefore, if $a_{t,G}/a_{t,-G} < 1$ then $(a_{t,G} + n)/(a_{t,-G} + n) < 1$ for all n: namely, if t is non-representative to begin with, then no amount of observations of t in population G (when accompanied by observations of t in population $-G$) will make t representative for G.

Consider now case ii), where all observations in G occur in a non-representative type t while all observations in $-G$ occur in a representative (for G) type t'. In that case, the representativeness of t for group G increases as $(a_{t,G}+n)/(a_{t,-G})$, while the representativeness of t' for group G decreases as $(a_{t',G} + n)/(a_{t',-G} + n)$. The result follows. ■

Proposition 4. Consider the case where a single observation of group G occurring in type x does not change the representativeness ranking of types – and thus the stereotype – for G.

If t is in the stereotype of G, then its estimated probability is $a_{t,G}/\sum_{t'=1}^{d} a_{t',G}$, which

is boosted by a factor of $\sum_{t'=1}^{N} a_{t,G} / \sum_{t'=1}^{d} a_{t',G} > 1$, where d is the number of types in the stereotype. Suppose an observation occurs in type t. Its representativeness for G increases, and its assessed probability jumps to $(a_{t,G}+1)/(\sum_{t'=1}^{d} a_{t',G}+1)$. This corresponds to a larger increase of assessed probability than that done by a Bayesian whenever

$$\frac{a_{t,G}+1}{\sum_{t'=1}^{d} a_{t',G}+1} - \frac{a_{t,G}}{\sum_{t'=1}^{d} a_{t',G}} > \frac{a_{t,G}+1}{\sum_{t'=1}^{N} a_{t',G}+1} - \frac{a_{t,G}}{\sum_{t'=1}^{N} a_{t',G}}$$

namely when

$$\frac{a_{t,G}}{\sum_{t'=1}^{N} a_{t,G}} < \frac{\sum_{t'=1}^{d} a_{t,G}}{1+\sum_{t'=1}^{d} a_{t,G}+\sum_{t'=1}^{N} a_{t,G}} < \frac{1}{2}$$

The intuition is that the stereotype ignores some observations, it is as though the probability is being updated over a smaller sample size. Therefore, as long as the prior of t (in the stereotype) is not too large, the DM boosts it more than the Bayesian.

If t is not in the stereotype, then – given that the stereotype does not change – it does not become representative. Its assessed probability stays at zero, so the decision maker under-reacts to this observation relative to a Bayesian. ■

Proposition 5. Let ρ_{μ,σ^2} denote the probability density of $\mathcal{N}(\mu,\sigma^2)$, namely $\rho(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$. The exemplar $\hat{t}_G$ of $G \equiv \mathcal{N}(\mu_G,\sigma_G^2)$ relative to $-G \equiv \mathcal{N}(\mu_{-G},\sigma_{-G}^2)$ satisfies $\hat{t}_E = \text{argmax}_t \frac{\rho_{\mu_G,\sigma_G^2}}{\rho_{\mu_{-G},\sigma_{-G}^2}}$ where

$$\frac{\rho_{\mu_G,\sigma_G^2}}{\rho_{\mu_{-G},\sigma_{-G}^2}} = \frac{\sigma_{-G}}{\sigma_G} \cdot \exp\left\{-t^2\left(\frac{1}{2\sigma_G^2}-\frac{1}{2\sigma_{-G}^2}\right)+t\left(\frac{\mu_G}{\sigma_G^2}-\frac{\mu_{-G}}{\sigma_{-G}^2}\right)-\left(\frac{\mu_G^2}{2\sigma_G^2}-\frac{\mu_{-G}^2}{2\sigma_{-G}^2}\right)\right\}$$

When $\sigma_G < \sigma_{-G}$, the function above has a single maximum in t, namely that which maximizes the parabola in the exponent, $\hat{t}_E = \frac{\frac{\mu_G}{\sigma_G^2}-\frac{\mu_{-G}}{\sigma_{-G}^2}}{\frac{1}{\sigma_G^2}-\frac{1}{\sigma_{-G}^2}}$ from which the result follows.

When $\sigma_G > \sigma_{-G}$, the function above is grows without bounds with $|t|$, so that $\hat{t}_G \in \{-\infty,+\infty\}$.

When $\sigma_G = \sigma_{-G} = \sigma$, the exemplar $\hat{t}_G$ of $G \equiv \mathcal{N}(\mu_G,\sigma^2)$ relative to $-G \equiv \mathcal{N}(\mu_{-G},\sigma^2)$ satisfies

$$\hat{t}_G = \text{argmax}_t e^{-\frac{\mu_G^2-\mu_{-G}^2}{2\sigma^2}} \cdot e^{\frac{t}{2\sigma^2}(\mu_G-\mu_{-G})}$$

so that $\hat{t}_G = -\infty$ if $\mu_G < \mu_{-G}$ and $\hat{t}_G = +\infty$ otherwise. If $\mu_G < \mu_{-G}$ all values of t are equally representative. ■

Proposition 6. Since the variances of the sample populations G and $-G$ are equal, the stereotypes are fully determined by the sample means. From Proposition 5, if $\sum_t t_{G,k} > \sum_t t_{-G,k}$, then the sample mean of G is larger than that of $-G$, so that its exemplar is $\hat{t}_G = +\infty$. If instead $\sum_t t_{G,k} < \sum_t t_{-G,k}$, the exemplar of G is $\hat{t}_G = -\infty$. Cases i) and ii) follow directly from this. ■

Lemma 1. This follows from Proposition 5, and from the fact that a left (right) truncated normal distribution satisfies the MLRP relative to the original distribution, where the likelihood ratio is increasing (decreasing). ■

Lemma 2. The proof is identical to that of Lemma 1. ■

Proposition 7. Consider case i). Women believe they have lower ability, $\theta_F^{st} < \theta_M^{st}$. The first order condition (11) implies that the optimal effort $e_F^* = e^*(\theta_F^{st})$ for any individual woman to provide is below that of men. As a result, average skills for women $\theta + e_F^*$ are lower than average skills of men, thus confirming the stereotype. Symmetrically, men believe they have higher ability, so it is individually optimal for a man to put a level of effort e_M^* which is higher than that of women, thus achieving higher skill levels and again confirming the stereotype. Case ii) is the reverse.

Consider now case iii). If both men and women believe they have the same ability θ, then it is individually optimal to provide the (rational) level of effort, which leads both groups to have the same level of skills, thus confirming the stereotype. However, this equilibrium is not stable: playing off the equilibrium path (due to heterogeneous beliefs, or mistakes) destroys this equilibrium. ■

B Unordered Types

In many settings, decision makers must assess groups in terms of their distributions over unordered type spaces. For instance, one may be interested in the distribution of occupations, or of political views, or of beliefs of different social groups. Our model applies directly to these settings, provided the type space is specified, or at least implied, by the problem at hand. While there is no notion of "extreme" types in unordered type spaces, the central insight about how representativeness and likelihood combine to determine stereotype accuracy continues to hold (Proposition 1): when groups are very similar, representative differences tend to be relatively unlikely, while when groups are different representative differences tend to be likely, and thus generate more accurate stereotypes.

To illustrate this logic in the context of unordered types, consider the formation of the stereotypes "Republicans are creationists" and "Democrats believe in Evolution". In May 2012, Gallup conducted a public opinion poll assessing the beliefs about Evolution of members of the two main parties in the US. The results on the beliefs of Republicans and Democrats, largely unchanged in the three decades over which such polls have been conducted, are presented below:[44]

	Creationism	*Evolution*	*Evolution guided by God*
Republicans	58%	5%	31%
Democrats	41%	19%	32%

The table shows that being a creationist is the distinguishing feature of the Republicans, not only because most Republicans are creationist but also because more Republicans are creationists than Democrats. In this sense, stereotyping a Republican as a creationist yields a fairly accurate assessment. Formally, $t = Creationism$ maximizes not only $\Pr(Republicans|t)/\Pr(Democrats|t)$ but also $\Pr(t|Republicans)$.

On the other hand, the distinguishing feature of the Democrats is to believe in the "standard" Darwinian Evolution of humans, a belief four times more prevalent than it

[44]The three options were described as "God created Humans in present form in the last 10,000 years", "Humans evolved, God has no part in process" and "Humans evolved, God guided the process". See http://www.gallup.com/poll/155003/Hold-Creationist-View-Human-Origins.aspx for details.

is among Republicans. However, and perhaps surprisingly, only 19% of Democrats believe in Evolution. Most of them believe either in creationism (41%) or in Evolution guided by God (32%), just like Republicans do. Formally, $t = \textit{Evolution}$ maximizes $\Pr(\textit{Democrats}|t)/\Pr(\textit{Republicans}|t)$ but not $\Pr(t|\textit{Democrats})$. Evolution is not the most likely belief of Democrats, but rather the belief that occurs with the highest relative frequency. As a consequence, a stereotype-based prediction that a Democrat would believe in the standard evolutionary account of human origins, and would not believe in Creationism, is a bad prediction.[45]

C Multidimensional Types

In the real world, types are often multidimensional. Members of social groups vary in their occupation, education, religion, income and other dimensions. Firms differ in their sector, location and management style. The state of the economy includes GDP growth, interest rates, and inflation. While multiple dimensions are subsumed in our previous analysis, in which each of the N types may consist of a unique specification of a possibly large set of attributes, for many groups stereotypes are formed along specific dimensions. Thus, some social groups are stereotyped by their occupations ("immigrants work in menial jobs"), others by their political views ("the young are liberal"), still others by their religious customs ("Buddhists meditate").[46] How are these dimensions selected?

Our model of representativeness provides a parsimonious perspective on this issue: the stereotype for group G will be organized around the dimension along which G is most

[45] Another example in this spirit is as follows. Suppose the DM must assess the time usage of Americans and Europeans. For the sake of simplicity, we consider only two types, namely $T = \{$time spent on work, time spent on vacation$\}$. The Americans work 49 weeks per year, so the conditional distribution of work versus vacation time is $\{0.94, 0.06\}$. In contrast, the Europeans work 47 weeks per year, with work habits $\{0.9, 0.1\}$. In both cases, work is by far the most likely activity. However, because the Americans' work habits are more concentrated around their modal activity, the stereotypical American activity is work. Because Europeans have fatter vacation tails, their stereotypical activity is enjoying the dolce vita. This stereotype is inaccurate, precisely because the vast majority of time spent by Europeans is at work. Still, due to its higher representativeness, vacationing is the distinctive mark of Europeans, which renders the image of holidays highly available when thinking of that group.

[46] As alluded to in Section 3.3, stereotypes may vary depending on circumstances according to changes in the comparison group. Walking in a deserted neighborhood may evoke a crime-based stereotype, while watching a sport event may evoke an athleticism-based stereotype for the same ethnic group.

different from $-G$. To see what this means, consider an example in which social groups in the US are described in terms of educational attainment (share of group members with higher education degree) and demand for social services (share of group members on welfare). Suppose that 35% of the white population has a college degree and 2% are on welfare, while 21% of the black population has a college degree and 10% are on welfare.[47] In terms of representativeness, the black population differs the most from the white population along the welfare dimension, not along the educational attainment dimension. This follows from the diminishing sensitivity of representativeness (Remark 1): even though the difference in educational attainment is larger (79% of blacks versus 65% of whites without college degrees), the most distinguishing feature of the black population is its higher relative demand for welfare (10% versus 2%). In this sense, to be formalized below, the stereotype for the black population is to be on welfare, despite the fact that only a small minority is on welfare (and even if the higher share of blacks on welfare were partially driven by their lower rates of college graduation). Conversely, the stereotype for whites is their higher share of graduates, not the fact that fewer are on welfare, even though a minority of whites go to college and a majority of whites are not on welfare. This is both because relatively more whites go to college and because most blacks are also not on welfare. The example shows that when groups are characterized by multidimensional types, they can be stereotyped along different dimensions. In particular, due to diminishing sensitivity, both groups can be stereotyped with unlikely types.

We now formalize the intuition described in this example. Suppose that the original random variable t is the product two categorical variables Y and Z, where $Y \in \{1, \ldots, N_Y\}$ and $Z \in \{1, \ldots, N_Z\}$, where $N_Y, N_Z > 1$. In the previous notation, $N = N_Y \times N_Z$ is the number of types. Types are indexed by realizations (y, z) of the two variables. According to Definition 1, the representativeness of type (y, z) for G is then defined by $\Pr(y, z|S)/\Pr(y, z| - S)$. In this setup, a stereotype consists of the d most representative realizations (y, z) of the two variables. To make progress, consider the special case in which the representativeness of a realization of Z does not depend on that of Y, formally

[47] Data from the National Center for Education Statistics (http://nces.ed.gov/programs/digest/d12/tables/dt12_008.asp) and from Statistic Brain (http://www.statisticbrain.com/welfare-statistics/).

$\Pr(z|y,S)/\Pr(z|y,-S) = \Pr(z|S)/\Pr(z|-S)$ for all z and all y (an assumption implicit in the previous example). The representativeness of type (y,z) is then an increasing function of:

$$\frac{\Pr(z|S)}{\Pr(z|-S)} \cdot \frac{\Pr(y|S)}{\Pr(y|-S)}. \tag{14}$$

The representativeness of (y,z) is simply the product of the representativeness of y and z considered independently. This condition holds, for instance, when being uneducated (low y) is predictive of lower income (low z), but this correlation may be independent of group identity and so acts uniformly across groups. Under this assumption, if one group has a higher share of poor members, that must be because it has also a higher share of uneducated members.

When equation (14) holds, the organization of a stereotype is pinned down by comparing the variation in representativeness along the two dimensions Y and Z. Denote by y_r the r-th most representative type of Y, when representativeness for type y is defined by $\Pr(y|S)/\Pr(y|-S)$. If y_1 is much more representative than y_2, then type (y_1,z) is more representative than (y_2,z') for any z and z'. In this case, the stereotype intuitively becomes "lexicographic," in the sense that it allows for little variation in types of the highly representative dimension Y and for much more variation in types of the less representative dimension Z. Specifically, the first N_Z types that come to mind are combinations of y_1 with all possible realizations of Z. The result below characterizes the cases in which this lexicographic ranking arises.

Proposition 8 *When (14) holds, the stereotype is lexicographic in dimension Y if:*

$$\min_r \left[\frac{\Pr(y_r|S)}{\Pr(y_r|-S)} \Big/ \frac{\Pr(y_{r+1}|S)}{\Pr(y_{r+1}|-S)} \right] > \left[\frac{\Pr(z_1|S)}{\Pr(z_1|-S)} \Big/ \frac{\Pr(z_{N_Z}|S)}{\Pr(z_{N_Z}|-S)} \right], \tag{15}$$

where v_r denotes the r-th most representative realization $v = y, z$, when representativeness for v is defined in isolation, formally $\Pr(v_r|S)/\Pr(v_r|-S)$. In particular, the stereotype is lexicographic in Y if Z is uninformative, $\Pr(z|S) = \Pr(z|-S)$ for all z.

Equation (15) identifies a stark condition for the stereotype to be lexicographic, namely that the maximum percentage variation in the likelihood ratio along Z is lower than the

minimum variation along Y. Not only the ranking of Y types by representativeness matters, but also how large an increase in representativeness is obtained by recalling y_1 rather than y_2, and so on. In particular, the stereotype is lexicographic in Y when the non-diagnostic dimension Z is undistinguishable across groups. When comparing Americans and Europeans, stereotypes do not focus on particular age groups, in the sense that the stereotypical European or American can be of a wide range of ages.

More importantly, however, Proposition 2 says that stereotypes can be organized along a given dimension Y if each type along Y is sufficiently more representative than the next. Remark 1 implies that representativeness of types becomes more extreme when the most representative types are unlikely. This suggests, as in the previous example on the demand for welfare, that Equation (15) tends to select bad stereotypes.

C.1 Proofs

Proposition 8. Following the assumptions of the proposition, write $Pr(z|y,S) = Pr(z|S) \cdot \phi(x,y)$ and $Pr(z|y,-S) = Pr(z|-S) \cdot \phi(x,y)$. We now describe the most extreme way that the stereotype may be organised along dimension Y, in which all variation along dimension Z is taken into account, namely $d_Z = |N_Z|$ (maximal) and $d_Y = d/|N_Z|$ (minimal). The representativeness of type (y,z) is given by

$$\frac{Pr(z|y,S)}{Pr(z|y,-S)} \cdot \frac{Pr(y|S)}{Pr(y|-S)} = \frac{Pr(z|S)}{Pr(z|-S)} \cdot \frac{Pr(y|S)}{Pr(y|-S)}$$

Because the representativeness of type (y,z) increases in the representativeness of y keeping z fixed (and vice versa), it is useful to consider the ranking of (unconditional) types $y \in Y$ and $z \in Z$. Let y_i (resp. z_i) denote the i-th most representative type in Y (resp. Z). Then, intuitively, the stereotype organises around Y if the variation in representativeness along the entire Z dimension is smaller than the variation in representativeness between any two types in Y. Formally, the representativeness ranking is lexicography if and only if

$$\frac{Pr(z_1|S)}{Pr(z_1|-S)} \Big/ \frac{Pr(z_{|C_Z|}|S)}{Pr(z_{|C_Z|}|-S)} < \min_r \frac{Pr(y_r|S)}{Pr(y_r|-S)} \Big/ \frac{Pr(y_{r+1}|S)}{Pr(y_{r+1}|-S)} .$$

■