INTRODUCTION

In what follows, I review a body of research from social psychology which suggests that individual performance on academic assessments is affected by the salience of that individual’s membership in stereotyped social groups. Given that individuals presumably know the social groups to which they belong, and given that economic theory generally assumes that individuals make consistent, optimal choices over all available information, this poses an interesting economic problem: why should the salience of our known social categories affect our decisions? Social psychologists have proposed anxiety as the driver of these results, but this is both unsatisfactory and inconsistent with economic assumptions about individual motivation and choice. To better address the problem, I turn to Akerlof and Kranton’s (2000) model of identity and its effects on economic outcomes. I justify the notion that individuals regularly and rationally choose to perform below their capability, and then explore research from cognitive psychology and neuroscience which has firmly established that humans have a limited attentive capacity, particularly when reasoning and comprehending (two tasks central to decision making). Since economic theory has typically assumed that individuals optimize over all known information, previous economic models have failed to account for effects of human cognitive limitations on the decision making process. By incorporating these limitations into a modified version of the Akerlof and Kranton (2000) model, human decision making can more accurately be analyzed. The result is a generalizable model which can account for an individual’s under- or over-performance (relative to the control condition in which her social categories are not made salient) according to the salience of the stereotyped social categories to which she belongs.
Within the field of social psychology it has become widely accepted that there is a psychological phenomenon which relates *the salience* of an individual’s membership in a particular social group with his performance in domains for which there is a widespread belief (such as a stereotype) that members of that group typically perform better or worse than the societal norm.

Often called *stereotype threat* when referring to a negative impact on performance which follows from negative stereotypes, it has been claimed that the phenomenon is one of the most significant contributors to the wide performance gaps between students of differing race (Steele and Aronson (1995); Shih, Pittinsky, & Ambady (1999)), students of differing gender (Shih et al (1999); Inzlicht and Ben-Zeev (2000)), and children of differing social status (Hoff and Pandey (2006)), after controlling for socioeconomic status and past scores on standardized tests (where applicable). At the same time, the less referred-to *stereotype lift* is purported to further aggravate performance gaps by promoting the above-average performance of members of groups associated with positive performance stereotypes, or even of members of groups which are merely *not* associated with negative stereotypes when there exist strong negative stereotypes about the performance of certain out-groups within a particular domain (Shih et al (1999); Walton and Cohen (2003); Hoff and Pandey (2006)).

In all of these studies, researchers found that much (though not all) of the performance gap disappeared when the participants’ particular social groups were not made salient to them (by priming) before they were evaluated. For example, after controlling for socioeconomic status and past SAT scores, Steele and Aronson (1995) found that Black Americans performed significantly worse than did White Americans on a portion of the verbal reasoning section of the Graduate Records Examination (GRE) when they were asked to indicate their race beforehand, but that much of this performance gap disappeared when race was not primed. Anxiety surrounding the negative stereotypes surrounding the academic ability of Black Americans was proposed as the underlying cause. Shih et al (1999) found that Asian women scored lower than average on quantitative tests when they were first given a questionnaire which asked questions priming their gender, while they scored higher than average when first given a questionnaire which asked questions priming their race. The negative stereotypes about females’ quantitative abilities and the positive ones about Asians’ academic abilities were proposed as the underlying causes of these performance inconsistencies. Inzlicht and Ben-Zeev (2000) determined that situational cues alone may be enough to make an individual’s status salient to her, thus affecting her performance. They found that women writing a portion of the quantitative reasoning section of the GRE in a room with two males scored twelve percentage points lower than did women who wrote in a
room with two females, other things equal. Anxiety about the negative stereotypes surrounding women's mathematical abilities was again proposed as the underlying cause.

Each of these studies focused on stereotypes based on physical differentiation (eg. race and gender), which presumably meant that the social categories to which the subjects belonged was publicly salient through simply observing them—though such apparent public salience had no effect on performance if status was not made privately salient: test subjects were only affected when they were immediately aware of their own social categorizations. Hoff and Pandey (2006) departed from this approach by measuring the performance of Indian children of different castes. They primed private cognizance of each child’s own caste beforehand by individually confirming his caste when he arrived at the experiment site. In the control condition they did not publicly announce the children’s castes, and there is presumably there is no way for the children to have visually identified members of different castes. Asking their subjects to complete puzzles on a piece-rate basis, they found that lower caste children performed just as well as upper caste children when castes were not publicly announced or when they were publicly announced only around children of the same caste. However, when caste was publicly announced in groups of mixed caste children, lower caste children performed worse and upper caste children performed better.

I shall henceforth refer to the collection of the aforementioned papers as ‘the ST literature’. The empirical work found in these and other papers on stereotype threat is generally excellent, but the proposed explanations for the results seem lacking. The stereotype threat phenomenon is usually explained by psychologists as being the result of a self-confirming anxiety, faced by members of negatively stereotyped groups, that they may perform poorly and thus confirm the stereotypes about their group. According to many social psychologists, the more salient an individual’s group identity is to her, the greater her anxiety, and the worse she performs. Yet such a narrow explanation does not seem fully consistent with the opposing stereotype lift effect, since the proposed mechanism of stereotype threat—anxiety that one will confirm negative stereotypes about one’s social group—is clearly not a factor in the improved performance caused by stereotype lift (which may be why Shih et al (1999) stopped short of proposing a mechanism for stereotypes to impact performance). Moreover, this hypothesis offers no reasons to think individuals are indeed concerned that their marginal performances might so conclusively confirm a stereotype; nor, for that matter, does it satisfactorily explain why individuals would be so concerned about the purported validity of stereotypes. Finally, since people presumably know the social groups to which they belong, one is left wondering why the salience of
these social groups should affect their behaviour. An alternative explanation seems necessary—one which (1) satisfactorily explains why the social categories to which an individual belongs should affect her performance—either positively or negatively—and which (2) explains why the salience of those social categories should matter given that self-aware individuals presumably know the social categories to which they belong.

Within the realm of rational choice theory in economics, an individual is presumed to act so as to maximize his utility net of the costs and subject to various constraints and his own exogenously determined preferences. His choices/actions are thus assumed to always be those he ‘believes’ will yield his best possible outcome, all things considered. In a seminal paper, Akerlof and Kranton (2000) adapted the standard utility function to take into account an individual $j$’s own actions, the actions of others, and $j$’s identity. Identity is in turn a function of $j$’s actions and the actions of others, the social categories to which $j$ belongs, $j$’s exogenously determined personal characteristics, and the exogenously determined prescribed norms (including stereotypes) for individuals belonging to any given social category. Some social categories, such as vocation, may be chosen by $j$, while others, such as race, are beyond $j$’s control. $j$’s actions are identity decreasing if they represent a departure from the prescribed norms given the social categories of which his identity is comprised, and his utility decreases if he suffers such an “identity loss”. $j$’s actions may also represent a threat to the identities of others, who may (or may not) then respond with actions which decrease $j$’s utility. The brilliant result was that $j$’s utility can increase, decrease, or be unaffected by his choice of any particular action, taking into account the action’s direct impact on utility, the indirect impact through the identity function, and its influence on others’ actions which impact $j$. Consistent with rational choice theory, $j$ then chooses his actions to maximize his utility net of the costs and subject to his constraints.  

As an example, consider a white man whose personal characteristics include having skin of a light pigmentation, male genitalia, and being a homo sapien. He and others then place him into the social categories ‘Caucasoid’, ‘Man’, and ‘Human’, among others, and he faces the prescribed norms associated with those categories. If the prescriptions for ‘Human’ include eating only ‘human’ food, yet he enjoys eating ‘dog’ food, then he suffers an identity loss associated with his humanity. If he is happier eating dog food and suffering the associated identity loss than he would be if he only ate human food,

---

1 Akerlof and Kranton (2000) note—and it bears repeating—that they speak of individual ‘choice’ “...advisedly. We do not presume one way or another that people are aware of their own motivations, as in standard utility theory which is agnostic as to whether an individual shopper is aware or not of the reasons for her choices.”
he will want to eat dog food. However, other people may witness this departure from the prescribed norms and ridicule him for it. If he still prefers to eat the dog food and suffer the consequent identity loss and ridicule, then he will eat dog food; if he is worse off eating dog food and suffering the consequent identity loss and ridicule than he would be if he ate only human food, he will not eat dog food (or he will do so only in private).

Akerlof and Kranton (2000) offer a number of less ridiculous examples of “identity-related behaviour”. A non-exhaustive list includes acts of self-mutilation, perceived challenges to “manhood” stemming from insults, challenges to gender-roles in the workplace, and political action motivated by a sense of belonging. More generally and relevantly, their model explicitly explains why individuals care about the social categories to which they belong, and why an individual’s performance (viewed as the direct consequence of his actions) might ever be beneath his capabilities: an individual will only perform as well as he must to maximize his utility net of the costs and subject to his constraints. Even if the direct costs of performing well are relatively low, doing so may be such a departure from the prescribed norms for individuals in his social categories that he would be better off performing poorly. When this is the case he will ‘choose’ to perform poorly, and it will be a rational, utility-maximizing decision. The model does not discount the effects of ability on outcomes, but rather explains why, ceteris paribus, an individual belonging to a social category for which performance is negatively stereotyped may perform worse than an individual who does not belong to that social category. For example, a black English student may find herself in, among others, the social categories ‘Negroid’, ‘Female’, ‘English Student’, and ‘Human’, and subject to the prescribed norms. If the prescribed norms for ‘Negroid’, ‘Female’, and ‘English Student’ all suggest poor performance on math tests, then to do well on a math test would give her not only the benefits associated with doing well, but also an identity loss (and any consequent effects from the responses of others). If the net effect on her utility of doing well is negative, she will optimize by doing poorly.

For this model to explain positive performance effects in the same way, it is necessary to assume that, even before considering identity, individuals generally do not perform to the full measure of their abilities. More generally, an individual’s performance will almost always be such that he could do both better and worse. This hardly seems a strong assumption: individuals choose their levels of performance to equate the marginal benefits with the marginal costs. To the extent that higher performance is more costly, an optimizing individual should choose a level of performance less than that of which she is capable. Assuming she knows her own production functions, a rational individual should choose a level
of performance *below peak* which equates the marginal benefit of higher performance with the marginal cost (in the same way as firms choose output to equate marginal revenue with marginal cost, rather than attempting to maximize revenue). Optimal results are thus defined by marginal considerations.

What about when academic performance is output, as in the stereotype lift/threat literature described earlier? Are there costs associated with higher performance levels which would make an individual want to perform below his ability? It seems obvious that costly activities such as studying will improve performance, but these costs (of studying) are obviously incurred long before an individual sits down to write a test, and should thus be viewed as sunk once writing commences. A more relevant question is whether there are costs associated with achieving higher exam scores which are incurred while the exams are being written. Many studies, among them Sundre and Kitsantas (2004), Wise and Kong (2005) and Wise and DeMars (2010), conclude that students perform more poorly on low-stakes tests when their motivation is low. It seems plausible to interpret these results as suggestive of a cost associated with higher performance (otherwise, so long as there is any positive motivation to perform, students should still be motivated to perform as well as possible). Wolf, Smith and Birnbaum (1995) explicitly explored this line of reasoning, and concluded that mental fatigue caused by taxing questions did indeed factor into performance.

It seems reasonable to conclude that optimizing individuals generally perform below the levels which their abilities allow. If this is true, then the model above can easily be adapted to explain positive performance effects associated with an individual’s social categories. In the same way that departures from prescribed norms result in an identity loss, behaviour which conforms to prescribed norms now results in an “identity gain”. Consider an Asian male economist whose personal characteristics result in him being placed into the social categories ‘Asian’, ‘Man’, and ‘Economist’. If the prescribed norms for these categories all suggest above-average performance on math tests, then doing well on a math test will positively influence his utility both directly through the utility function and indirectly through an identity gain. Thus the marginal benefit of doing well is higher than without the identity gain. Since the marginal cost of doing well is the same for otherwise identical individuals, our Asian male economist will optimize by choosing a higher level of performance than would an otherwise identical individual who did not belong to one or more of the positively stereotyped social categories.
As powerfully as it explains why an individual’s social categories may prompt her to choose to perform more poorly than she is able, or to perform better than an otherwise identical individual, this model does not yet satisfactorily explain the results from the ST literature. The model thus far suggests that an individual will choose her level of performance to maximize her own utility, but it also suggests that she will be consistent in those choices. There is, as yet, no mechanism to allow the salience of an individual’s social categories to affect his performance; he simply considers everything he knows and makes a decision. Yet the studies conducted by the ST literature authors suggest this is not what is occurring. The test subjects in these studies chose different performance levels contingent on their social categories being primed; but surely the women, blacks, and Asians involved in these studies were aware that they were women, or black, or Asian. Why, then, should their performances have been affected by the priming of their statuses? Why should they not always perform as well or as poorly as their social categories and the associated costs and benefits dictate?

Within the fields of cognitive psychology and neuroscience is a general consensus that humans face limits on our cognitive capacities, particularly as relates to our ability to attend to and process information. Miller (1956) famously claimed “My problem is that I have been persecuted by an integer,” suggesting that humans have a working memory capacity limit of approximately seven “chunks” of information (though the limit on “bits” of information which can exist within each chunk is substantially greater). Unique bits of information initially present as individual chunks until and unless the individual processing the information groups two or more bits into higher-order chunks. For example, an individual may struggle to remember a sequence of ten un-chunked individual digits. However, once she groups them into the three standard chunks which comprise a North American phone-number, XXX-XXX-XXXX, she will have an easier time recalling each chunk, as well as associated information such as the person she will reach if she dials the number. In the classic example within the literature, a beginner chess player may struggle to remember the individual movements which can be made with each piece and the moves available to his opponent, each of which is an individual bit of information un-chunked from any other. An expert, on the other hand, can consider entire sequences of moves, or ‘tactics’, which can be used to attack or defend against her opponent. The expert is then said to have chunked together the legal movements of each piece with the sequences of moves which comprise a tactic. Having done so,

2 Taking as given the standard assumptions of complete and transitive preferences.
the expert can recall, recognize, and consider many more of the options available to her than can the beginner.

Many authors agree with Miller that there is a limit on working memory capacity, though many of them suggest this limit is closer to four or five than seven, and that it varies according to the type of information being considered (see Simon (1974); Baddeley, Thomson, and Buchanan (1975) Cowan (1995),(2000); and Klingberg (2009)). Within the realm of economics, the notion of limited attention and its effect on decision making has been explored by Nobel laureate economist and psychologist Daniel Kahneman (1973), and more recently by Gifford (1997) and Hirshleifer and Teoh (2003) in the context of partially attentive investors.

The general theory is that working memory allows individuals to temporarily store and manipulate information in cognitive tasks such as reasoning and comprehension. Attention is the cognitive focus which must be given to information for it to enter the working memory. Researchers tend to consider attention as being divisible into different types; for example, Klingberg (2009) differentiates between controlled attention—that type of attention which an individual might use to force herself to read something, and of which, if her thoughts wander elsewhere, she is said to have lost control, and stimulus-driven attention—that type of attention “which is involuntarily attracted to an unexpected event in a person’s immediate environment—such as [a] coffee cup hitting the ground.”

Though Cowan (1995) and Klingberg (2009) refer to working memory and attention as technically distinct concepts which intersect each other, for simplicity I will refer to the combined cognitive process as ‘attention’. Moreover, while the semantic distinction between controlled attention and stimulus-driven attention may seem consistent with common conceptions of the ways in which our attention is directed, I believe it is ultimately irrelevant. Take as given that humans have a limited attentive capacity; it seems clear enough that I have no control over which things to which I attend with so-called stimulus-driven attention, but it is much less clear how I might have control, or choice, over those things α to which I attend with controlled attention. Since my limited attention limits my choices to those things to which I am attending, if I can choose α then I must choose it from a set of options β to which I am attending. If that set β is comprised only of things which have been exogenously brought to

---

3 For the sake of relevance and brevity I will not consider the distinct processes and functions of long term memory.
4 Klingberg (2009) also discusses arousal as a unique form of attention.
my controlled attention as options, then I am only choosing from choices over which I had no control, which is essentially the same case as with stimulus-driven attention. On the other hand, if that set of options $\beta$ from which I choose those things $\alpha$ to which I attend is *itself* comprised of a set of options $\gamma$ to which I have chosen to attend, then I again face the question of whether *that* set of options $\gamma$ is exogenously brought to my controlled attention—in which case I still seem to lack control over those things to which I attend—or whether it is comprised of a yet another set of options $\delta$ which I have chosen to attend, and so on. It seems clear that to maintain the notion of controlled attention sets me off on an infinite regress in which it is ultimately impossible for me to distinguish those things to which I choose to attend from those things which ultimately attract my attention apart from my own volition. Thus the distinction between controlled attention and stimulus-driven attention, if it truly exists, is irrelevant for practical purposes, and attention can be treated as something over which I have no ultimate control. Put another way, the things to which I ultimately attend are exogenously determined.

THE MODEL

As mentioned earlier, a model which satisfactorily explains such results as those the ST literature must convincingly explain why an individual’s performance should be affected by her social categories, and why her performance could change with the salience of those social categories.

I propose a modified version of the Akerlof and Kranton (2000) utility function which is consistent both with the above discussion regarding the possibility of a utility-maximizing individual performing more poorly than he is able, and with the research and above discussion on human attention and its limits. I model utility as a function of identity, own actions which can affect identity, own actions which do not affect identity, and exogenous shocks such as a loud noise, an attractive person, or a rainstorm. Identity is modeled as follows: I propose that attention is binary—objects are either in our working memory, or they are not—and introduce it as a vector of dummy variables. I also propose a vector of functions of own actions which can affect identity, an individual $j$’s personal characteristics, the social categories to which $j$ belongs, and the prescribed norms for members of those social categories. Each of these functions incorporates numerous ‘bits’ of related information, so each can be thought of as a ‘chunk’ of information as described earlier. I assume that identity related information is sufficiently chunked such that attending to any individual identity bit involves attending to all related bits (for example, paying attention to one’s own social categories allows one to also attend to the prescribed norms for those categories without ‘using up’ additional working memory capacity). I include a vector of
continuous variables representing the (exogenously determined) nominal importance to \( j \)'s identity of each social category function, and assume that this information is chunked with the social category function. The identity function then interacts the attention dummies with the social category chunks (the nominal importance of particular social category functions and the social category functions themselves). Own actions can affect multiple social category functions or none at all. People often do things without being very aware of the fact they are doing them, so it is possible for people to do things which would be identity affecting if they were attending to certain social category chunks without drawing their attention to those chunks. For analytic simplicity I do not consider the effects of the actions of others, and I assume that information is costless—that is, were \( j \)'s attention not limited, he would operate with complete information.\(^5\) I begin with an example of the simplest case, and then follow with a formal representation of the general model.

Imagine a simple case where an individual \( j \)'s identity involves personal characteristics which place him in only two social categories, ‘Night Club Bouncer’ and ‘Gay Man’. Imagine \( j \) faces the option of committing only two identity affecting actions which, for simplicity, he either does or does not do: he can intimidate people; and he can speak with a lisp. He also faces the option of committing an undefined number of actions which do not affect his identity, as well as an undefined number of exogenous shocks. All of these things compete for his attention, which is limited and focused by exogenous factors. Imagine the prescribed norms for ‘Night Club Bouncer’ include intimidating people, while the prescribed norms for ‘Gay Man’ include speaking with a lisp.

\( j \) faces a number of scenarios concordant with his attention being (exogenously) drawn to (a) neither social category chunk, (b) to the Night Club Bouncer chunk only, (c) to the Gay Man chunk only, or (d) to both chunks.

(a) If \( j \) attends to neither, his actions will not affect his identity and he will choose whether or not to commit them based purely on their net ‘direct’ effect on his utility (this is the classic utility function without identity or attention). His optimal choice can change depending on his circumstances. For example, if he is with friends who he does not wish to intimidate, the direct effect on utility of intimidating may be negative; if he is working and if he will lose his

\(^5\) This unrealistic assumption is not necessary, but I make it simply to show that this is not a problem of costly information. Even if information were costless, cognitive limits would still prevent individuals from considering all of it when making decisions.
job if he does not intimidate people, then the direct effect on utility of not intimidating may be negative. Similarly, circumstance-relative direct effects may exist for speaking with a lisp. For simplicity, assume that the direct effects on utility of each action are unaffected by his choice of the other action. He will make his choices accordingly.

(b) If $j$ attends to the Night Club Bouncer chunk only, he will choose whether or not to intimidate people based on both the direct net effect on his utility and the indirect effect through his identity. Similarly, he will choose whether or not to speak with a lisp based on the combined direct and indirect effects on utility. The effects on his identity of intimidating (or not) and speaking with a lisp (or not) are cumulative but non-complementary. If speaking with a lisp does not affect the Night Club Bouncer part of his identity, then he considers only the direct effects on utility of speaking with a lisp. If he does not intimidate people, he suffers an identity loss. As in (a), the direct effect on utility of intimidating people could be either positive or negative depending on his circumstances. If the direct effect is positive, he will clearly intimidate people (because failing to do so will decrease his utility both directly and indirectly). If the direct effect is negative, however, his behaviour will depend on the size of the indirect effect: if Night Club Bouncer is a very important part of his identity and if failing to intimidate results in a large enough identity loss, then he will intimidate even when he would not do so were he not attending to the Night Club Bouncer chunk. On the other hand, if Night Club Bouncer is not a very important part of his identity, or if intimidating is not a heavily weighted norm in his Night Club Bouncer social category function, then choosing to not intimidate will not result in a large enough identity loss to counteract the positive direct effect. In this case, the direct effect dominates the identity loss and he will not intimidate.

(c) Similarly to (b), if $j$ attends the Gay Man chunk only, he will choose whether to speak with a lisp or not based on both the direct net effect on his utility and the indirect effect through his identity. As with speaking with a lisp in (b), he will choose whether to intimidate or not based on the combined direct and indirect effects on utility. If he does not speak with a lisp, he suffers an identity loss. If the direct effect on his utility of speaking with a lisp is positive, then he will clearly speak with a lisp (since failing to do so would decrease utility both directly and indirectly). If the direct effect on his utility is negative, then his behaviour
depends on the overall effect on his utility: if Gay Man is a very important part of his identity and if not speaking with a lisp results in a large enough identity loss, then he will speak with a lisp even when he would not do so were he not attending to the Gay Man chunk.

Conversely, if Gay Man is not a very important part of his identity, or if speaking with a lisp is not a heavily weighted norm in his Gay Man social category function, then failing to speak with a lisp will not result in large enough identity loss to make a difference. In this case, the direct effect again dominates the identity loss and he will not speak with a lisp.

(d) The most interesting case occurs when $j$ attends both the Night Club Bouncer chunk and the Gay Man chunk. To begin, consider the case where the direct effects on his utility of intimidating and speaking with a lisp are both positive. He must thus consider the indirect effects through his identity. His behaviour will depend, as usual, on the net total effect on utility of speaking with a lisp. Now, though, the nominal importance to $j$’s identity of each social category function matters. Since he is attending to multiple social category chunks, each one becomes (weakly) less proportionally important to his identity. The importance to his identity of each social category function now depends on the nominal importance of each as a proportion of the aggregate importance of all the attended social category functions (this proportion was one when only one chunk was attended). Each action can have a positive, a negative, or a neutral effect on identity. Speaking with a lisp positively affects utility through the Gay Man social category function, but it may have positive or negative effects on utility through the Night Club Bouncer social category function. If it has positive effects, then he will clearly speak with a lisp (since doing so increases utility directly and indirectly through two identity lifts, though less-so through each than it did when only one chunk was attended).

If, however, speaking with a lisp has negative effects on utility through the Night Club Bouncer social category function, things are more complicated. For simplicity, assume speaking with a lisp and being intimidating each significantly impact his social category functions in opposite directions (that is, the prescribed norms for Night Club Bouncer are to be intimidating and to not speak with a lisp, while those for Gay Man are to speak with a lisp and to not be intimidating). Also assume that being intimidating affects Night Club Bouncer more than does speaking with a lisp, while speaking with a lisp affects Gay Man more than does being intimidating. If Gay Man is proportionally more important to his identity than
Night Club Bouncer, then (since speaking with a lisp has a positive direct effect on utility in this case) he will still choose to speak with a lisp. If, however, the proportional importance of Night Club Bouncer is both higher than that of Gay Man and large enough to allow the negative effect of speaking with a lisp on Night Club Bouncer to outweigh the positive effect of doing so on Gay Man, then speaking with a lisp will yield an identity loss. Whether or not he speaks with a lisp will then be determined by whether the effect on his utility of this identity loss is larger or smaller than the positive direct effect on his utility of speaking with a lisp.

An identical analysis applies to his choice of being intimidating. The analysis can also be applied in predictable and similar ways for the cases where the direct effects on his utility of both actions are negative and where the direct effect of one action is positive while the other is negative. The only other result worth mentioning, when two chunks are attended, occurs when the direct effect on utility of an action has a different sign than the indirect effect. When only one chunk was attended, the nominal importance to identity of the associated social category function did not matter (since the proportional importance to identity was one). When two chunks are attended, though, the proportional importance to identity of each social category function is decreased. It is entirely possible that actions which have negative (positive) direct effects on utility but positive (negative) indirect effects through identity, and which may have been engaged in (avoided) when only one chunk was attended (if the direct effect was smaller (larger) than the indirect effect) may now be avoided (engaged in) because of the decreased proportional importance to identity of the social category function. That is, actions which may impact j’s identity enough to affect his behavioural choices when he is attending to only one social category chunk may have too little an impact on his identity to affect his behavioural choices when he is attending to both chunks.

A formalized and more general expression of the model now follows. The utility function I propose is

\[ U_j = U_j(o_j, a_j, I_j, u_j). \]

The utility of person j depends on his identity \( I_j \), a vector of his actions \( a_j \) which affect his identity, a vector of his actions \( o_j \) which do not affect his identity, and on a vector of exogenous shocks \( u_j \) which
affect \(j\). Both \(o_{gj}\) and \(a_{hj}\) are arguments which affect \(U_j\) at some known cost to \(j\) determined by \(U_j(\cdot)\). \(g \in [1, \infty)\), and \(h \in [1, \infty)\). All \(o_{gj} \geq 0\) are continuous variables representing the amount of any particular \(o_{gj}\) engaged in by \(j\), and all \(a_{hj} \geq 0\) are continuous variables representing the amount of any particular \(a_{hj}\) engaged in by \(j\).

\(I_j\) can be represented as

\[
I_j = I_j(\lambda_j \pi_j \theta_j(a_j, c_j, \epsilon_j, P)).
\]

The identity of person \(j\) depends on a vector of social category functions \(\theta_j\) which interact \(j\)'s assigned social categories \(c_j\) with \(j\)'s identity affecting actions \(a_j\), own characteristics \(\epsilon_j\), and the social prescriptions \(P\) of appropriate behaviour for members of any given social category such that

\[
P(c_j) - f_j(a_j, \epsilon_j) = \theta_j \leq 0, \text{ and }
\]

\[
\frac{\partial f_j}{\partial a_j} \geq 0,
\]

where \(j\)'s actions are always defined such that engaging in any particular \(a_{hj}\) runs contrary to \(P(c_j); j\) doing something that conforms with \(P(c_j)\) thus involves engaging in less \(a_{hj}\). Then

\[
\frac{\partial \theta_j}{\partial a_{hj}} \leq 0, \text{ and }
\]

\[
\frac{\partial I_j}{\partial \theta_j} \geq 0.
\]

\(I_j\) also depends on the (also exogenously determined) nominal importance \(\pi_{ij}\) of any \(\theta_{ij}\) to \(I_j\), and on a vector of dummy variables \(\lambda_j\) such that

\[
\lambda_{ij} = \begin{cases} 1 & \text{if } j \text{ is attending to a particular chunk of information } \theta_{ij}, \\ 0 & \text{otherwise.} \end{cases}
\]

That is, attention given to any particular chunk \(\pi_{ij}\theta_{ij}\) is binary. I assume that \(\lambda_{ij}\) is exogenously determined (see discussion above), and that

\[
0 \leq \sum_{i=1}^{s} \lambda_{ij} = m_j \leq n_j,
\]

where \(m_j\) is the number of identity related chunks to which \(j\) is attending during any given choice of \(a_{hj}\), \(n_j\) is the maximum number of chunks to which \(j\) can attend (the limit on working memory capacity,
which as discussed is likely around four or five, and almost certainly no more than seven), $n_j - m_j$ is the number of chunks attended to by $j$ which are not related to his identity (such as those involving $a_j$ or $u_j$), and $z$ is the total number of $j$'s identity related chunks $\pi_{ij}\theta_{ij}$. I also assume that $j$ chunks together all bits of information concerning $c_{ij}, e_{ij}, a_{hj}$, and $P(c_{ij})$ in the function $\theta_{ij}$.

$$0 \leq \frac{\pi_{ij}}{\sum_{i=1}^{z} \pi_{ij}} \leq 1 \text{ subject to } \sum_{i=1}^{z} \pi_{ij} > 0$$

is the proportional importance of any $\theta_{ij}$ to $l_j$, and

$$0 \leq \frac{\lambda_{ij} \pi_{ij}}{\sum_{i=1}^{z} \lambda_{ij} \pi_{ij}} \leq 1 \text{ subject to } \sum_{i=1}^{z} \lambda_{ij} \pi_{ij} > 0$$

is the proportional importance to $l_j$ of any particular $\theta_{ij}$ for any given choice of $a_{hj}$. As long those $\theta_{ij}$ already attended to do not change, I can also say

$$0 \geq \frac{\frac{\lambda_{ij} \pi_{ij}}{\sum_{i=1}^{z} \lambda_{ij} \pi_{ij}}}{\Delta \theta_{ij}} \bigg|_{\lambda_{ij}=1}$$

That is, as the number of $\theta_{ij}$ to which $j$ attends increases, the proportional importance to $l_j$ of any particular $\theta_{ij}$ for any given choice of $a_{hj}$ is weakly decreasing, assuming those $\theta_{ij}$ initially attended to are still attended to. As a proof of this, consider a symmetric world in which all $\theta_{ij}$ are equally important to $l_j$ (all $\pi_{ij}$ are equal and positive). If (11) were not true, then an increase in $\sum_{i=1}^{z} \lambda_{ij} \pi_{ij}$, keeping $\lambda_{ij} \pi_{ij}$ constant, could lead to an increase in $\frac{\lambda_{ij} \pi_{ij}}{\sum_{i=1}^{z} \lambda_{ij} \pi_{ij}}$, which is clearly false; therefore (11) is true. This can be extended to a non-symmetrical world as long as those $\theta_{ij}$ initially attended are still attended as $m_j$ increases.

Given (2) and (7), I can state

$$\frac{\partial l_j}{\partial \theta_{ij}} \begin{cases} = 0 & \text{if } \lambda_{ij} = 0, \\ \geq 0 & \text{otherwise.} \end{cases}$$

Given (2), (5), (6), (10), I define

$$\frac{\partial u_j}{\partial l_j} \geq 0.$$
Since the sign of $\frac{\partial u_j}{\partial a_{hj}} \geq 0$ depends on $j$'s preferences and costs, I can say that $j$ chooses $a_{hj} = 0$ when

$$\frac{\partial u_j}{\partial a_{hj}} \leq \frac{\partial u_j \partial l_j \partial \theta_j}{\partial l_j \partial \theta_j \partial a_{hj}}$$

and that $j$ chooses $a_{hj} > 0$ when

$$\frac{\partial u_j}{\partial a_{hj}} \geq \frac{\partial u_j \partial l_j \partial \theta_j}{\partial l_j \partial \theta_j \partial a_{hj}}$$

This last case clearly only occurs if $\frac{\partial u_j}{\partial a_{hj}} > 0$. When this is the case, $a_{hj} > 0$ occurs for sure if $\lambda_{ij} = 0$ (or if $m_j = 0$) for all $\theta_{ij}$ where $\frac{\partial \theta_{ij}}{\partial a_{hj}} < 0$, and it occurs with greater likelihood as $\pi_{ij}$ approaches 0 and as $\frac{\partial \theta_{ij}}{\partial a_{hj}}$ approaches 0 for all $\theta_{ij}$ where $\lambda_{ij} = 1$ and $\frac{\partial \theta_{ij}}{\partial a_{hj}} < 0$. Since $\lambda_{ij} = 0$ for most $\theta_{ij}$ during any given choice of $a_{hj}$, it will usually (though not always) be the case that $a_{hj} > 0$ when $\frac{\partial u_j}{\partial a_{hj}} > 0$. In such cases, any exogenous occurrence which then sets $\lambda_{ij} = 1$ for any $\theta_{ij}$ where $\frac{\partial \theta_{ij}}{\partial a_{hj}} < 0$ will make $j$ want to reduce $a_{hj}$ in order to increase $\theta_{ij}$ (toward 0). When $P(c_{ij})$ is consistent with high level performance, this scenario will prompt $j$ to maximize his utility by performing better. When $P(c_{ij})$ is consistent with lower level performance, this scenario will prompt $j$ to maximize his utility by performing worse.

**DISCUSSION**

The model satisfies the criterion I established as necessary to explain the phenomenon found in the ST literature: it explains why an individual’s performance might be affected by her social categories, and also why her performance might change with the salience of those social categories. To the extent that black students take poor performance on tests as a prescribed norm which is a non-trivial part of being black, and to the extent that they consider being black an important part of their identities, then whenever they are paying attention to their ‘black’-ness it is the case that high-level performance would have to provide a substantial positive direct impact on their utility to make them want to do well on such tests. When this positive impact is too small, they will rationally choose to do more poorly. The same applies to women when writing math tests. On the other hand, to the extent that Asian students take high-level performance on quantitative tests as a prescribed norm which is non-trivially a part of
being Asian, and to the extent that being Asian is an important part of their identities, then whenever they attend to their ‘Asian’-ness they face increased payoffs to doing well on these tests.

It is important to recognize that, as always, general trends by no means apply to any specific individual. The importance of conforming to any prescribed norm for any individual’s sense of belonging to a social category, as well as the importance of belonging to that social category for any individual’s sense of identity, are both individual-specific and determined outside the model. There is no reason to believe that either prescribed norms or the importance of such things to individuals are immutable; members of negatively stereotyped social categories need not be forever stuck in a poor-performance trap. It is also important to note that, as usual, conclusions should only be drawn ceteris paribus; a great equating of prescribed norms and importance levels would not equate performance results if members of certain social categories are actually under-equipped to perform well. Nonetheless, the model reconciles apparently inconsistent and detrimental behaviour with the classic economic assumptions of individual rationality, and complete and transitive preferences. The only qualification is that such conditions hold only for each set of attended options; when attention shifts, so do payoffs and optimal choices.

The model is general enough to be extended to non-academic performance. Possible applications include the workplace performance of individuals with different job titles, the labour force and consumption decisions of individuals in different income brackets, and the dietary and exercise decisions of individuals of different sizes. Extensions could add Akerlof and Kranton’s (2000) effect of the actions of others on own utility, or could look at the effects on decision making of the $n_j - m_j$ non-identity-related chunks attended by individuals (particularly the $o_j$). In this way, the model may explain such things as the effectiveness of relatively uninformative but brief advertisements and political platforms—for example, a person may well choose to vote for a political candidate whose campaign promises are easily attended, even if the complex platform of a different candidate is more broadly consistent with the body of values held by the voter.

CONCLUSION

People belonging to social groups beset by negative stereotypes tend to do more poorly on tests when their membership in those groups is primed than do control groups; people belonging to positively stereotyped groups tend to perform better when their membership in those groups is primed. The explanation offered by social psychologists seems lacking from an economic perspective, but economists
have thus far not built a model which satisfactorily accounts for all of the dynamics at play. Using a modified version of Akerlof and Kranton’s (2000) model of identity’s effects on economic outcomes, I applied insights on human attentive limits from other academic disciplines to develop a model which explains the phenomenon while maintaining the general assumptions of rational choice theory. All else equal, when individuals who belong to stereotyped social categories have their attention drawn to their membership in those categories, they will be more likely to act in ways which confirm those stereotypes because it is utility increasing to do so. Dividing their attention by making salient their membership in additional social categories will decrease the effect on utility of any one category, and can thus change optimal behavioural choices. These results explain the phenomenon which motivated the model, and can be further applied to other economic problems involving individual choice.

REFERENCES


Miller, George A. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review 63, no. 2*, March 1956: 81-97.


