Self-Esteem and Rational Self-Handicapping (Draft) *

Rachel Mannahan †

December 4, 2021

Abstract

In order to protect their self-esteem, agents may create or seek out performance-inhibiting obstacles to avoid negative feedback about their ability. This behavior allows agents to attribute failures to self-imposed obstacles rather than to a lack of competence. Psychologists refer to this phenomenon as self-handicapping. In this paper, I model rational agents with preferences for self-esteem. This allows me to provide sufficient conditions under which these self-esteem concerned agents engage in rational self-handicapping behavior. I then consider the effects of self-esteem concerned agents in two policy relevant applications: education and tournaments. In education, designing an exam with noisier questions results in more studying by self-handicappers without discouraging other students. In tournaments, policies aimed at encouraging a particular group to show up to job interviews can result in strategic adjustment by the targeted group leading to their lower overall attendance.

†I am incredibly thankful to Lina Andersson, Pierpaolo Battigalli, Andreas Blume, Inga Deimen, Martin Dufwenberg, Amanda Friedenberg, Senran Lin, Kelli Marquardt, Ernesto Rivera, Tim Roberson, Jin Sohn, and seminar participants at the University of Arizona for their helpful comments.
1 Introduction

Self-esteem is an evaluation of one’s own traits that can impact preferences. It develops at an early age\(^1\) and it is influenced by an individual’s successes and failures. In particular, successes often increase self-esteem, while failures often decrease self-esteem.\(^2\) For example, earning a high test grade may increase self-esteem, whereas failing the test may decrease self-esteem. Because low self-esteem is generally considered to be harmful to a person’s health and well-being, people are driven by a desire to protect and maintain self-esteem.\(^3\) Thus, when self-esteem impacts a decision-maker’s preferences, they may seek to take actions which increase self-esteem, even if these actions are sometimes costly in other domains.

One way in which a decision maker may increase self-esteem is through self-handicapping, a term coined by Berglas and Jones (1978), which occurs when someone self-inflicts performance impediments. These impediments, which reduce the quality of the performance, distort the signal a self-handicapper receives from a poor performance. For example, a student may not study for an exam so that her resulting score is less informative. Rhodewalt, Saltzman, and Wittmer (1984) find evidence that athletes self-handicap by reducing or skipping practice for a competition. Berglas and Jones (1978) find evidence of participants in an intelligence test self-handicapping by choosing performance inhibiting drugs. While self-handicapping behavior may temporarily protect one’s self-esteem, this behavior may negatively impact long-run outcomes. For example, if a student does not study for an exam, she may temporarily feel better after she fails; however, the consequences of not studying may result in worse future performance in the subject matter.

To explain why decision makers sometimes engage in these potentially harmful self-handicapping behaviors, psychologists have hypothesized a link between self-handicapping and self-esteem maintenance.\(^4\) In this paper, I utilize the mathematical framework of psychological game theory to

---

\(^1\)Harter (1993) discusses the mental health consequences of low self-esteem in youths and the importance of early interventions.

\(^2\)Bongers, Dijkstra, and Spears (2009), Crocker, Sommers, and Luhtanen (2002), and Heatherton and Polivy (1991) all discuss the relationship between success, failure, and self-esteem.

\(^3\)Baumeister and Leary (2000), Crocker and Park (2004), and Tesser (2000) describe the self-esteem maintenance motive and its mechanisms.

\(^4\)Baumeister and Leary (2000), Jones and Berglas (1978), and Tice (1991) discuss the relationship between self-esteem and self-handicapping.
establish a formal link between self-esteem maintenance and self-handicapping behavior.\textsuperscript{5} Formalizing this link allows for a better understanding of when preferences for self-esteem maintenance lead to self-handicapping behavior (and when they do not), and for a more precise understanding of how different incentives alter self-handicapping behavior. This is relevant for analyzing the impact of different policies on a population in which some agents have preferences for self-esteem maintenance.

I begin by establishing the belief-dependent utility of an agent who has a preference for self-esteem maintenance. To capture an agent’s utility for self-esteem, I assume that each agent is endowed with a type, which represents her ability. The agent knows the underlying distribution of types, but does not know her actual type. The agent can take actions in a game that provide or conceal information about her type. This information allows the agent to update her belief about her type and, in turn, impacts the agent’s utility.

The self-esteem concerned agent’s utility function has two components: a material payoff component and a self-esteem component. The agent’s utility is increasing in both of these components. The material component can be conceived of as the utility of an outcome (e.g., a high grade, success, a win, a payment, etc.). The self-esteem component takes into account the agent’s belief about her underlying ability level.\textsuperscript{6}

Next, I define self-handicapping behavior and use this notion to establish sufficient conditions on self-esteem preferences that result in self-handicapping. Definition 2.1 will specify that one strategy is self-handicapping relative to another if it results in a lower material payoff and induces a distribution over ex post expected types which second-order stochastically dominates the distribution induced by the other strategy. This allows me to explore the extent to which self-esteem preferences are conducive to self-handicapping behavior. I show that when the utility of an agent’s self-esteem payoff is strictly concave, the self-handicapping strategy can be rational.\textsuperscript{7}

\textsuperscript{5}Psychological game theory, first developed by Geanakoplos, Pearce, and Stacchetti (1989) and extended by Battigalli and Dufwenberg (2009) and Battigalli, Corrao, and Dufwenberg (2019), allows agents in games to have utilities which depend on their beliefs. See Battigalli and Dufwenberg (2020) for an overview of work involving psychological game theory, in which they mention its many applications, including self-image concerns.

\textsuperscript{6}This is directly consistent with the view Brown and Dutton (1995), in which agents care about their beliefs about their attributes. Notice that the agents care about the true, underlying type only in that it impacts her beliefs.

\textsuperscript{7}Note that this relies on the availability of a self-handicapping strategy.
To understand potential adverse consequences of self-handicapping behavior and potential policy implications, I analyze two applications that are policy relevant. First, I analyze the optimal test to administer to a heterogeneous population of students, some of whom are concerned with self-esteem. I show that increasing the proportion of a test grade assigned to noisy questions increases the amount of studying on the exam, even without altering the material incentives to study. Second, I analyze a two-player job tournament which is biased against one candidate (e.g., there is a targeted policy that only favors one of the candidates). I show that the candidate who is favored is less likely to apply, and the candidate who is disadvantaged is more likely to apply.

This paper is the first to analyze self-esteem in the frame of psychological games, and the first formalize self-handicapping behavior to clarify the link to self-esteem maintenance. This paper draws upon several foundational works in psychology on self-esteem, self-esteem maintenance, and self-handicapping.\footnote{See Baumeister and Leary (2000) for an overview of the vast literature on self-esteem. See Tesser (2000) for an overview of self-esteem maintenance and maintenance mechanisms. Berglas and Jones (1978) and Jones and Berglas (1978) are the two seminal papers on self-handicapping. The recent book “Self-Handicapping: The Paradox That Isn’t” by Higgins, Snyder, and Berglas (2013) provides and in-depth view of self-handicapping behavior.}

In economics, the most closely related work is that of Köszegi (2006), who uses Caplin and Leahy’s (2001) psychological expected utility to model an agent with “ego-utility.” In an initial learning stage, the agent can observe any number of free signals about their ability. Then, the agent may choose an ambitious or unambitious asset in two subsequent financial decisions. Notably, the ambitious option is more likely to materialize when the agent has higher ability. As a result of their desire to maximize ego-utility, the agent tends to become overconfident. One key difference between Köszegi’s work and this paper is that Köszegi’s analysis is limited to individual decisions in a specific context. Another crucial difference is that this paper is focused on the self-esteem and self-handicapping relationship, which is not the primary focus of Köszegi. Furthermore, my paper is related to a handful of others in economics.\footnote{Similar in theme are papers by Bénabou and Tirole (2002), Cowen and Glazer (2007), Ishida (2012). All of these papers explore some notion of self-esteem or information avoidance, although the technical approaches and the exact focus all deviate substantially from this paper.}

Section 2 rolls out a two-stage game form during which beliefs are formed, and defines the agent’s ex post utility. It also includes examples that demonstrate self-esteem utility in both
individual decisions and games. Section 3 defines self-handicapping behavior. Section 4 focuses on optimal exam design for student populations with self-esteem concerns. Section 5 analyzes biased tournaments in populations with self-esteem concerns. Section 6 concludes.

2 Model

There is a finite set of players $I$. Each player $i \in I$ has a compact set of types $T_i \subseteq \mathbb{R}$. Write $T_{-i} = \prod_{j \in I \setminus \{i\}} T_j$ and $T = \prod_{j \in I} T_j$. A type profile $\tau = (\tau_i : i \in I)$ is drawn from a full-support distribution $\mu \in \Delta(T)$. Players do not observe this draw, so they do not know their own type or the types of the other players.

The players interact in a two-stage game form. The first stage involves simultaneous-moves by each player, where player $i$ chooses a strategy from a finite set $S_i$. Write $S_{-i} = \prod_{j \in I \setminus \{i\}} S_j$ and $S = \prod_{j \in I} S_j$. The second stage involves only a move by chance ($c$). Chance’s set of actions depends on the actions chosen in the first stage, but not the type profile. Write $A_c(s)$ for the finite, nonempty set of chance’s choices when the players chose $s$ in the first stage. The set $A_c(s)$ may be a singleton; this corresponds to the case where chance does not move after $s$ is played in the first stage.

Player $i$’s strategy set is $S_i$. I will make use of $S_c = \prod_{s} A_c(s)$, which I refer to as chance’s strategy set. While this strategy set $S_c$ does not depend on $\tau$, the probability that chance chooses a particular strategy may depend on $\tau$. That is, there is a profile of distributions $q_c = (q_c(\cdot \mid \tau) : \tau \in T)$, where each $q_c(\cdot \mid \tau)$ may assign probability 0 to some strategies in $S_c$. The ex ante distribution of $T \times S_c$ is $\nu \in \Delta(T \times S_c)$ where $\nu(E \times F) = \int_E q_c(F \mid \tau) d\mu(\tau)$ for $E \subseteq T$ and $F \subseteq S_c$ measurable.

The two-stage game form induces a set of terminal nodes $Z$. I take these terminal nodes to only describe the strategies of the players and chance, but not to directly include information about the type profile. So, there is a function $\zeta : S \times S_c \rightarrow Z$ that maps the strategy profiles of players and chance into the induced terminal node. Write $\hat{\zeta}[s_i] : T \times S_{-i} \times S_c \rightarrow T \times Z$ for the mapping.
with \( \hat{\zeta}[s_i](\tau, s_{-i}, s_c) = (\tau, \zeta(s_i, s_{-i}, s_c)) \).

Material payoffs are determined both by the terminal node and the type profile. Write \( \pi_i : Z \times T \rightarrow \mathbb{R} \) for the material payoff function. I assume that \( \pi_i(z, \tau_i, \tau_{-i}) \) is weakly increasing in \( \tau_i \). That is, for any given \( (z, \tau_{-i}) \), \( i \) receives a weakly higher material payoff if \( \tau_i \) is higher.

There is a player-specific partition \( \mathcal{P}_i \) of \( Z \times T \) called the set of terminal information sets for \( i \). Let \( \mathcal{P}_i[z, \tau] \) denote the element of the partition that contains \( (z, \tau) \). These terminal information sets satisfy the following properties:

(i) For each \( (z, \tau) \in \mathcal{P}_i[z', \tau'] \), there is some \( s_i \in S_i \) such that \( \zeta^{-1}(z) \subseteq \{s_i\} \times S_{-i} \times S_c \) and \( \zeta^{-1}(z') \subseteq \{s_i\} \times S_{-i} \times S_c \).

(ii) If \( (z, \tau) \in \mathcal{P}_i[z', \tau'] \), then \( \pi_i(z, \tau) = \pi_i(z', \tau') \).

The first requirement says that player \( i \) knows the strategy \( s_i \) that she chose. That is, if \( (z, \tau) \) and \( (z', \tau') \) are in the same terminal information set, then they both resulted from \( i \) choosing the same strategy \( s_i \). In this case, say \( \mathcal{P}_i[z, \tau] \) is induced by \( s_i \). The second requirement says that player \( i \) knows her material payoff. (She may or may not have additional information.) That is, if \( (z, \tau) \) and \( (z', \tau') \) are in the same terminal information set, then they both induce the same material payoffs. So there is a function \( \Pi_i : \mathcal{P}_i \rightarrow \mathbb{R} \) mapping each terminal information set into a single payoff, i.e., \( \Pi_i(\mathcal{P}_i[z, \tau]) = \pi_i(z', \tau') \) for each \( (z', \tau') \in \mathcal{P}_i[z, \tau] \).

In order to describe a player’s self-esteem, I first need to describe her ex post belief about her type. This belief will depend, in part, on her beliefs about the strategy profiles of other players. Ex ante, player \( i \)'s initial belief about \( S_{-i} \) is some \( \alpha_i \in \Delta(S_{-i}) \). Therefore, \( (\nu \times \alpha_i) \in \Delta(T \times S_c \times S_{-i}) \) describes player \( i \)'s initial belief about \( T \times S_c \times S_{-i} \).

At the end of the game, players learn information which causes them to update their initial beliefs about \( T \times Z \). Say \( \mathcal{P}_i[z, \tau] \) is consistent with \( \alpha_i \) if there exists some \( (s_i, s_{-i}, s_c, \tau') \) such that \( \alpha_i(s_{-i}) > 0 \), \( q_c(s_c \mid \tau) > 0 \), and \( (\zeta(s_i, s_{-i}, s_c), \tau') \in \mathcal{P}_i[z, \tau] \). (Note \( \mu(\tau') > 0 \) by assumption.)

**Definition 2.1.** Call \( \beta_i : \mathcal{P}_i \rightarrow \Delta(Z \times T) \) a terminal belief mapping induced by \( \alpha_i \in \Delta(S_{-i}) \) if the following hold:
(i) For each \( p_i \in \mathcal{P}_i \), \( \beta_i(p_i)(p_i) = 1 \).

(ii) If \( p_i \) is consistent with \( \alpha_i \) and induced by \( s_i \), then for each \( E \subseteq Z \times T \) measurable,

\[
\beta_i(p_i)(E) = \frac{(\nu \times \alpha_i)(\zeta[s_i]^{-1}(E \cap p_i))}{(\nu \times \alpha_i)(\zeta[s_i]^{-1}(p_i))}.
\]

Say \( \beta_i \) is a **terminal belief mapping** if there exists some \( \alpha_i \) such that \( \beta_i \) is a terminal belief mapping induced by \( \alpha_i \). Let \( \mathcal{B}_i \) be the set of all terminal belief mappings.

The first condition states that \( \beta_i \) assigns probability 1 to the information set that is reached by player \( i \). The second condition states that, when player \( i \)'s terminal information set \( p_i \) is consistent with \( \alpha_i \) and arises from \( i \) playing \( s_i \), player \( i \)'s terminal belief at the terminal information set \( p_i \) is the conditional belief induced by \( i \)'s prior \( (\nu \times \alpha_i) \). For a given terminal belief mapping \( \beta_i \) and terminal information set \( p_i \), write \( E_{\beta_i(p_i)}[\tilde{\tau}_i] \) for \( i \)'s ex post expectation about her type at \( p_i \).

Several factors may contribute to a higher ex post expected type. First, higher material payoffs may lead \( i \) to a higher ex post expected type because \( i \)'s material payoff is increasing in her own type. Second, player \( i \) may have a higher ex post expected type after observing the realization of chance's move since the distribution of chance’s strategy may depend on \( \tau \). Third, player \( i \) may learn direct information about the path of play or the type profile, which may increase her ex post expected type.

### 2.1 Self-Esteem Utility

A player’s utility at the end of the game consists of two components: the material payoffs she receives from playing the game and her ex post self-esteem.

Self-esteem will be captured by a function that is strictly increasing in a player’s ex post expectation of her own type. Hence, this function \( f : \mathbb{R} \to \mathbb{R} \) is a strictly increasing function. This is consistent with the view that as one’s assessment of their own ability improves, so does their self-esteem. Furthermore, I assume \( f \) is strictly concave, which I later show is sufficient for self-handicapping behavior when agents are sufficiently concerned with self-esteem.\(^{11}\)

\(^{11}\)See example 3 for a discussion of the link between self-handicapping behavior and risk aversion.
Given a terminal belief mapping $\beta_i$, $i$’s **self-esteem** at $p_i \in P_i$ is $f(\mathbb{E}_{\beta_i}(\tau_i))$. Notice, a higher ex post expected type gives player $i$ (weakly) higher self-esteem. With this, player $i$’s utility is given by a function $u_i : B_i \times P_i \rightarrow \mathbb{R}$ so that

$$u_i(\beta_i, p_i) = \Pi_i(p_i) + \theta_i f(\mathbb{E}_{\beta_i}(\tau_i)),$$

where $\theta_i \geq 0$ is a player-specific exogenous parameter that captures $i$’s sensitivity to self-esteem. The higher $\theta_i$, the more weight player $i$ places on her self-esteem relative to her material payoffs. Notice that, if $\theta_i = 0$, player $i$ only cares about her material payoffs.

The following example will illustrate how to apply self-esteem utility to an individual decision problem.

**Example 1: Self-Esteem Utility in an Individual Decision Problem**

A student may have a high ability level, corresponding to $\tau_i = 1$, or a low ability level, corresponding to $\tau_i = 0$. Each occur with probability $\frac{1}{2}$. She must decide whether to “study” ($S$) for an exam or “not study” ($N$) for an exam. The student may have a high ability level, corresponding to $\tau_i = 1$, or a low ability level, corresponding to $\tau_i = 0$. If she chooses $N$, she receives a grade of 0 on the test regardless of her type. If she chooses $S$, she receives a 100 if $\tau_i = 1$ and a 50 if $\tau_i = 0$.

Refer to Figure 2.1. At the end of the game, the student only learns her grade. Her material payoff is equivalent to her grade. There are three terminal information sets, written $[N, 0]$, $[S, 50]$, and $[S, 100]$ for convenience. In Figure 2.1, the dashed line represents that $[N, 0]$ contains two elements.\(^{12}\)

Note two features of this example. First, the set of chance’s actions following any strategy profile $s$, $A_c(s)$, is a singleton. Second, because there is only one agent, $i$’s beliefs about the strategies of other players ($\alpha_i$) does not enter the agent’s utility.

Notice that the student reaches the terminal information set $[S, 100]$ only if $\tau_i = 1$. This is reflected by $\beta_i$, which assigns probability 1 to $\{(S, \tau = 1)\}$. As a result, her ex post expected type will be 1. That is, $\mathbb{E}_{\beta_i([S,100])}(\tau_i) = 1$. Thus, her utility is $u_i(\beta_i, [S, 100]) = 100 + \theta_i f(1)$.

\(^{12}\)Notice, the terminal information set I call $[N, 0]$ contains $(\tau_i = 0, N)$ and $(\tau_i = 1, N)$.
Similarly, if the student studies and receives a 50 on the exam, she infers that $\tau_i = 0$ and her utility is $u_i(\beta_i, [S, 50]) = 50 + \theta_i f(0)$. If the student reaches $[N, 0]$, she learns no new information so her terminal belief is the same as her prior belief. Thus, her utility is $u_i(\beta_i, [N, 0]) = 0 + \theta_i f(0.5)$. Notice that $i$’s utility is constant on terminal information sets. If the student has self-esteem concerns, she may choose $N$ to protect her self-esteem.

The next example demonstrates how including other players influences self-esteem utility.

**Example 2: Self-Esteem Utility in Games**

Two students are completing a group project. Before the project deadline, they must simultaneously decide whether to exert effort ($E$) or not exert effort ($N$). As before, if a student chooses $N$, her score is 0; if she exerts effort, her score is 100 (50) if $\tau_i = 1$ ($\tau_i = 0$). However, now, the student does not observe her score. Rather, she observes the average of her score and her partner’s score. Figure 2.2 depicts the game-form. To understand the payoffs, suppose student 2 exerts effort and observes a 50. This could have occurred because she received a 100 and her partner received a 0, because she received a 0 and her partner received a 100, or because both received 50. Once again, let $[s_i, x]$ denote the terminal information set induced by $i$’s strategy $s_i$ at which $i$ receives $x$. Her terminal information set in this scenario is $[E, 50]$. The cells associated with this terminal information set are highlighted in Figure 2.2.

Because there is more than one student, beliefs about the strategies of others, $\alpha_i$, will be necessary to determine self-esteem utility. This $\alpha_i$ will eventually correspond to equilibrium. For now, take $\alpha_i$ to be arbitrary.
\[(\tau_1, \tau_2) = (0, 0)\]

\[
\begin{array}{c|cc}
  & E & N \\
\hline
1 & 50 & 25 \\
N & 25 & 0 \\
\end{array}
\]

\[(\tau_1, \tau_2) = (0, 1)\]

\[
\begin{array}{c|cc}
  & E & N \\
\hline
1 & 75 & 25 \\
N & 50 & 0 \\
\end{array}
\]

\[(\tau_1, \tau_2) = (1, 0)\]

\[
\begin{array}{c|cc}
  & E & N \\
\hline
1 & 75 & 50 \\
N & 25 & 0 \\
\end{array}
\]

\[(\tau_1, \tau_2) = (1, 1)\]

\[
\begin{array}{c|cc}
  & E & N \\
\hline
1 & 100 & 50 \\
N & 50 & 0 \\
\end{array}
\]

Figure 2.2: Group Project Grades

Notice that player 2’s terminal information set \([E, 50]\) could have been reached if player 1 chose \(E\) or \(N\). As such, any belief \(\alpha_2\) is consistent with player 2’s terminal information set \([E, 50]\). If player 1 chose \(E\), player 2 must have ability \(\tau_2 = 0\). If player 1 chose \(N\), player 1 must have ability \(\tau_2 = 1\). So, for any \(\alpha_2\) that has full support, \(\beta_2([E, 50])\) assigns probability \(\alpha_2(E)\) to \((\zeta(E, E), \tau_2 = 0)\) and \(1 - \alpha_2(E)\) to \((\zeta(N, E), \tau_2 = 1)\). (Notice, this is also true if \(\alpha_2\) does not have full-support, i.e., if \(\alpha_2(E)\) is 0 or 1.) Thus, player 2’s utility is

\[
u_2(\beta_2, [E, 50]) = 50 + \theta_2[\alpha_2(E)f(0) + (1 - \alpha_2(E))f(1)].
\]

This is decreasing in \(\alpha_2(E)\). So, when player 2 is more pessimistic about their partner’s effort, they are more optimistic about their own ability.

### 3 Self Handicapping

In the this section, I first discuss an example that illustrates the key components for defining self-handicapping behavior. Then, I formally define self-handicapping behavior and discuss sufficient conditions for said behavior.

**Example 3: Self-Handicapping Beliefs**

Ann can have one of two types, \(\tau_a = 1\) or \(\tau_a = 0\), each with equal probability. Ann moves first
and has three available strategies: $s^1_a$, $s^2_a$, and $s^3_a$. Chance moves after Ann. For each of Ann’s strategies $s_a$, chance has 3 available actions: Report 0, Report 1, and Report ∅. The action Report 0 reports to Ann that her type is $\tau_a = 0$. The action Report 1 reports to Ann that her type is $\tau_a = 1$. The action Report ∅ reports no information about Ann’s type. The probabilities with which chance takes each action depends on Ann’s type. Figure 3.1 describes these probabilities. (The probabilities in Figure 3.1 induce a mixed strategy $q_c$.)

\[
\begin{array}{c|cc}
\tau_a = 0 & \tau_a = 1 \\
\hline
R0 & 0.02 & 0 \\
R1 & 0 & 0.02 \\
R∅ & 0.98 & 0.98 \\
\end{array}
\quad
\begin{array}{c|cc}
\tau_a = 0 & \tau_a = 1 \\
\hline
R0 & 0 & 0 \\
R1 & 0.99 & 0.01 \\
R∅ & 0.01 & 0.99 \\
\end{array}
\quad
\begin{array}{c|cc}
\tau_a = 0 & \tau_a = 1 \\
\hline
R0 & 0.99 & 0.01 \\
R1 & 0.01 & 0.99 \\
R∅ & 0 & 0 \\
\end{array}
\]

Figure 3.1: Chance Probabilities

Ann observes the report (if there is one). Write $[s_a, x]$ for the terminal information set induced by $s_a$, at which Ann observes report $x$. For example, if Ann chooses $s^1_a$ and observes a report of 1, she has reached $[s^1_a, 1]$. Ann’s material payoff from $s^1_a$ is 1, from $s^2_a$ is 2, and from $s^3_a$ is 3.

Each strategy of Ann induces an interim distribution of ex post expected types. For example, if Ann chooses $s^1_a$, in the interim stage - i.e., after choosing $s^1_a$ but before observing the report - there is a 98% chance she will observe Report ∅. In this case, Ann gains no additional information about her type from observing the report, so her ex post expected type will be the same as her ex ante expected type. Additionally, Ann is aware that there is a 1% chance she will receive Report 0, and a 1% chance she will receive Report 1. In these cases, she knows her type with certainty.

Notice then, each strategy induces a distribution over ex post expected types. See Figure 3.2.

Intuitively, if Ann is a self-handicapper, she will prefer $s^1_a$ and $s^3_a$ over $s^2_a$, because $s^2_a$ gives her a high probability of finding out she is certainly type 0. However, comparing $s^1_a$ and $s^3_a$ is not as intuitive. Consider the least preferred possibility that could result from each strategy. Under $s^1_a$, the worst possibility is that Ann certainly has a type of 0. Under $s^3_a$, the worse possibility is that

---

13Note $s_c = \{R0, R1, R∅\}^3$ and $q_c(\cdot \mid \tau) \in \Delta(\{R0, R1, R∅\}^3)$. Figure 3.1 reflects the marginals. I can write $\text{marg}_{A_c(s)} q_c(a_c \mid \tau)$ to reflect the probability chance chooses action $a_c \in A_c(s)$ conditional on the type profile being $\tau$ and the players choosing the strategy profile $s$. For example, $\text{marg}_{A_c(s^1_a)} q_c(\text{Report1} \mid \tau_a = 0) = 0.01$. Notice that in this example, the probability that chance chooses each action is independent of Ann’s type.
Ann *almost* certainly has a type of 0. However, the worst-case in $s^1_a$ will occur with much lower probability than the worst-case in $s^3_a$. How Ann would rank these is not as clear. In fact, it is easy to imagine that some self-handicappers prefer $s^1_a$, while others prefer $s^3_a$. Therefore, self-handicapping alone need not provide a definitive preference over these two strategies.

Second-order stochastic dominance provides a ranking of these strategies that is equivalent to the intuition about which is best for the self-handicapper.\textsuperscript{14} Second-order stochastic dominance will rank higher those strategies for which the induced interim distribution (of ex post expected types) is unambiguously less risky. That is, for two strategies with the same induced interim mean, the strategy with lower variance is preferred. This preferred strategy will reduce the probability of higher ex post expected types in order to also reduce the probability of lower ex post expected types, thereby mitigating some of the risk of believing one has a low type. This is consistent with the idea that self-handicappers disprefer believing unfavorable information about themselves and would be willing to sacrifice potentially good news to avoid this information.

Self-handicapping behavior also requires material suboptimality. Thus, the only candidates for a self-handicapping strategy must be those in which Ann is guaranteeing herself a relatively worse expected material payoff. Although Ann prefers both $s^1_a$ and $s^3_a$ to $s^2_a$, only $s^1_a$ is self-handicapping relative to $s^2_a$ because it results in a lower material payoff.

**Defining Self-Handicapping**

When Ann determines which strategy she will play, she considers the ex post expectations about her own type that could result from each of her strategies. I think of the interim stage as the stage

\textsuperscript{14}Notice that first-order stochastic dominance does not provide a ranking over these strategies because the ex ante mean of these distributions is the same for each of Ann’s strategies.
after Ann chooses her own strategy, but has not learned the strategy of other players or chance.

Write $\Phi_i(\cdot | s_i, \alpha_i)$ for player $i$’s interim cumulative distribution function of ex post expectations of $\tau_i$ given that $i$ plays $s_i$ and $i$ believes $\alpha_i$.

Write $\mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i | s_i]$ for player $i$’s interim expected material payoff given strategy $s_i$. So

$$\mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i | s_i] = \sum_{p_i \in P_i} \Pi_i(p_i)(\nu \times \alpha_i)((s_{-i}, s_c, \tau) : (\zeta(s_i, s_{-i}, s_c), \tau) \in p_i).$$

Write $\mathbb{E}_{(\nu \times \alpha_i)}[u_i | s_i]$ for player $i$’s interim expected utility when she plays strategy $s_i$.

**Definition 3.1.** Fix $\alpha_i$. Strategy $s'_i$ is $\alpha_i$-self-handicapping relative to $s''_i$ if:

i. (Material Suboptimality) $\mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i | s'_i] < \mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i | s''_i]$ and

ii. (Self-Esteem Maintenance) $\Phi_i(\cdot | s'_i, \alpha_i, q_c)$ second-order stochastically dominates $\Phi_i(\cdot | s''_i, \alpha_i, q_c)$.

To say that $s'_i$ is $\alpha_i$-self-handicapping relative to $s''_i$ requires that $s'_i$ results in a lower expected material payoff than $s''_i$ and in an unambiguously less risky distribution of ex post expected types. Notice that for a fixed $\alpha_i$, the self-handicapping relation is a strict partial order.

**Definition 3.2.** Strategy $s'_i$ is $\alpha_i$-self-handicapping if there exist some $s''_i$ such that $s'_i$ is $\alpha_i$-self-handicapping relative to $s''_i$.

**Definition 3.3.** Let $\sigma = (\sigma_i : i \in I)$ be an equilibrium profile of mixed strategies. Say $\sigma$ induces self-handicapping behavior if there exists an $i$ and some $s_i$ in the support of $\sigma_i$ so that $s_i$ is $\sigma_{-i}$-self-handicapping.

**Theorem 3.1.** Suppose $s'_i$ is $\alpha_i$-self-handicapping relative to $s''_i$. Then there exists some $\bar{\theta} > 0$ so that $\mathbb{E}_{(\nu \times \alpha_i)}[u_i | s'_i] > \mathbb{E}_{(\nu \times \alpha_i)}[u_i | s''_i]$ if and only if $\theta_i > \bar{\theta}$.

**Proof.** Suppose $s'_i$ is $\alpha_i$-self-handicapping relative to $s''_i$. Define

$$\bar{\theta} = \frac{\mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i | s''_i] - \mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i | s'_i]}{\mathbb{E}_{(\nu \times \alpha_i)}[f | s'_i] - \mathbb{E}_{(\nu \times \alpha_i)}[f | s''_i]}.$$ (1)
Note, the numerator of Equation (1) is strictly positive if and only if condition (i) of Definition 3.1 holds. The denominator is strictly positive if and only if condition (ii) of Definition 3.1 holds. (See Wolfstetter (1999). This makes use of the fact that $f$ is strictly concave.) Since $s'_i$ is $\alpha_i$-self-handicapping relative to $s''_i$, $\bar{\theta} > 0$.

Now observe that $\theta_i > \bar{\theta}$ if and only if

$$E_{(\nu \times \alpha_i)}[\Pi_i \mid s'_i] + \theta_i E_{(\nu \times \alpha_i)}[f \mid s'_i] > E_{(\nu \times \alpha_i)}[\Pi_i \mid s''_i] + \theta_i E_{(\nu \times \alpha_i)}[f \mid s''_i].$$

Equivalently, $\theta_i > \bar{\theta}$ if and only if

$$E_{(\nu \times \alpha_i)}[u_i \mid s'_i] > E_{(\nu \times \alpha_i)}[u_i \mid s''_i].$$

Theorem 3.1 shows that strict concavity of $f$ is sufficient for an $\alpha_i$-self-handicapping strategy to increase an agent’s utility. This implies that self-handicapping behavior can be rational. However, if $s'_i$ is $\alpha_i$-self-handicapping relative to $s''_i$, neither is necessarily a best response. For instance, there may be some $s'''_i$ that gives $i$ both a higher material payoff and higher self-esteem. (In Example 3, Ann’s strategy $s^3_o$ has this property.)

While strict concavity is sufficient for rational self-handicapping, it is not necessary. In the Appendix, I provide an example in which self-esteem utility that is not strictly concave but results in self-handicapping behavior and I mention potential drawbacks of such a formulation of self-esteem.

4 Education

According to Verschueren, Marcoen, and Schoefs (1996), children as young as 5 years old possess a sense of self. For these children, a positive self-view is positively correlated with higher competence, behavioral adjustment, and social acceptance. As a result, establishing good study habits and a positive relationship with school in early childhood can have a strong impact on how children develop into adults. Furthermore, falling behind at any phase of education can make catching up in subsequent years more difficult. Therefore, educators and educational institutions may wish to foster an environment that encourages students to exert effort so that they reap the benefits of
higher levels of educational attainment and better educational outcomes.\textsuperscript{15}

Avoidance of studying may prove detrimental to students’ long-term development, resulting in lower levels of human capital. Society may label those who do not exert effort in school as “lazy;” however, these students may instead be engaging in self-handicapping behavior. Though these self-handicappers may appear indistinguishable from with the “lazy” group, they will respond differently to certain incentives. By making use of these incentives, we may be able to circumvent a cycle of falling behind in self-handicapping prone students. In this section, I discuss simple testing interventions aimed at increasing the work ethic of a self-handicapping student (by way of increasing their utility from exerting effort).

I consider the decision of a student choosing how to prepare for an exam with three different kinds of questions. Questions differ on whether they distinguish between student types or not. If the student is sufficiently concerned with self-esteem, she may not study for the test, guaranteeing a failing grade. However, some distributions of questions may be more likely to elicit studying from a student who is inclined to self-handicap. These exams, which blur the lines between luck and ability, increase the number of students who study without altering the average score for all students who study. This way, the effect on studying is entirely due to exam restructuring, rather than inflating grades with additional points.

4.1 Setup

A student can be of type \( \tau = 0 \) or \( \tau = 1 \). The probability of \( \tau = 0 \) is \( q \in (0, 1) \). The student chooses whether to “study” (S) for an exam or “not study” (N) for an exam.\textsuperscript{16} The exam consists of three kinds of questions. Easy questions make up \( x_e \geq 0 \) points of the exam, hard questions make up \( x_h \geq 0 \) points, and noisy questions make up \( x_n \geq 0 \) points. There are \( N_e \in \mathbb{N} \cup \{0\} \) easy questions, \( N_h \in \mathbb{N} \cup \{0\} \) hard questions, and \( N_n \in \mathbb{N} \cup \{0\} \) noisy questions on the exam. Points for each type of question are divided evenly amongst all questions of that type. For example, each easy question is worth \( \frac{x_e}{N_e} \) points.

\textsuperscript{15}We may choose to think of higher levels of educational attainment (or alternatively, more successful educational outcomes) as one form of increased human capital.

\textsuperscript{16}Jones and Berglas (1978) consider the example of whether or not to get a good night’s rest before the exam.
Whether the student gets each question correct or incorrect depends on her type and her action. If she chooses not to study, she answers all questions incorrectly. If she chooses to study, both types answer the easy questions correctly, but only type $\tau = 1$ answers hard questions correctly. If she chooses to study, type $\tau = 1$ answers each noisy question correct (C) with probability $r_1 \in (0, 1)$, and type $\tau = 0$ answers each noisy question correct (C) with probability $r_0 \in (0, 1)$. I assume that $r_1 \geq r_0$, reflecting that $\tau = 1$ are more likely to answer noisy questions correctly than $\tau = 0$. These chance probabilities are summarized in Figure 4.1. Notice that chance independently determines whether each noisy question is correct.

If the student gets a question correct, she earns points for that question. If she gets a question incorrect, she earns no points for the question. The student’s material payoff is her grade on the exam, i.e., the sum of the points she earns on each question. Thus, the student’s maximum possible score is $x_e + x_h + x_n$, and her minimum possible score is 0. The student observes her grade, but does not observe her performance on any individual question. Notice then that not studying for the exam is a self-handicapping strategy for the student.

I restrict attention to a class of exams with two properties. First, the maximum number of points on the exam is $T$, i.e., $x_e + x_h + x_n = T$. Second, the number of points assigned to easy questions is fixed at $\bar{x}_e \leq T$.

I focus on how changing the proportion of points assigned to each type of question affects the decision of self-esteem concerned agents without changing their material payoff. That way, the results only reflect how changing the exam structure effects the decision of self-esteem concerned agents. Thus, I restrict attention to exams for which the interim expected material payoff conditional on studying is the same for all possible point distributions. This implies that, for any exam

\[ x_e \]
I consider, $\mathbb{E}_{(\nu \times \alpha_i)}[\Pi_t | S] = x_e + (1 - q_t)x_h + (q_t \cdot r_0 + (1 - q_t) \cdot r_1)x_n$.

4.2 The Set of Possible Exams

Given the restrictions, the set of all possible exams can be broken into two general categories. In the first, there are terminal information sets following $S$ that are not singletons. That is, when the student observes her score, she is uncertain about the extent to which it comes from her ability level, luck, or both. In the second, the student can perfectly infer her type at the end of the game if she studies; that is, the terminal information sets following $S$ are singletons.

All of the exams in Category 1 result in some terminal information sets following $S$ that are not singletons. The following conditions characterize all such exams:

1. $x_h = 0$ or

2. $x_n \geq x_h \geq 0$, $N_n \geq 1$, and $x_n / N_n$ is a factor$^{18}$ of $x_h$.

These conditions ensure that the points the student earns due to noisy questions obscure the points the student earns due to hard questions. In Lemma 7.1 in the Appendix, I prove that any exam that satisfies one of these two conditions must have a non-singleton terminal information set reached with positive probability. In the first condition, because the student can perfectly infer her type if she answers hard questions correctly, there must at least as many noisy points available, so that upon seeing her score, a low ability student believes she could have earned those points because she has high ability.$^{19}$ The second condition ensures that, for a fixed number of noisy questions, the number of points assigned to noisy questions is such that it is possible to mistake noisy points for hard points. Figure 4.2 depicts one possible Category 1 exam, in which $N_n = 1$ and $x_n > x_h = 0$.

I will illustrate these conditions using several examples. First, suppose $N_n = 0$. Then the exam only asks easy and hard questions. In this case, if $x_h > 0$, the student can perfectly infer her type. Thus, in order to have some non-singletons, condition 1 must be satisfied.

Second, suppose $N_n = 1$. Then there is one noisy question. In this case, the number of points

---

$^{18}$By factor, I mean there exists some number $W \in \mathbb{N}$ such that $\frac{x_n}{N_n} \cdot W = x_h$.

$^{19}$This condition is trivially satisfied if the exam assigns zero points to hard questions.
for noisy questions must equal the number of points for hard questions, or else the student can back out whether or not they answered the hard questions correctly.

Finally, suppose $N_n = 2$, $x_n = 6$, and $x_h = 5$. Notice that the first part of condition 2 is satisfied. However, there are 3 points per noisy question. That means, whether the student answers 0, 1, or 2 noisy questions correct (corresponding to 0, 3, or 6 points from noisy questions), the student will still be able to determine whether they earned the 5 points from hard questions, failing the second part of condition 2. However, if $x_h = 3$, as long as the student answers at least one noisy question correctly, she will be uncertain whether she earned that 3 points from a noisy question or from the hard questions, satisfying the second part of condition 2.

One notable exam that falls under Category 1 is an exam that only consists of easy questions ($\bar{x}_e = T$). Whenever $\bar{x}_e < T$, increasing the number of points assigned to noisy questions must be traded off with decreasing the number of points assigned to hard questions. Doing so makes obscuring the hard questions more likely.

Category 2 includes any distribution of points for which the student can perfectly infer her type at the end of the game if she studies. This includes any exam that fails either of the two conditions required for Category 1. Notice, this includes Example 1 where $\bar{x}_e = 50$, $x_h = 50$, and $x_n = 0$. Figure 4.3 depicts the terminal information sets of a Category 2 exam with $N_n = 1$ and $x_n > 0$.

For all exams in Category 2, the student, upon observing her grade, can determine whether
she answered the hard questions correctly. Therefore, she will be certain of her type. Notable examples of Category 2 exams include exams with only hard questions, exams with only easy and hard questions, and exams that assign the most points to hard questions.

Notice that the exam designer faces a tradeoff between the number of points assigned to noisy questions and hard questions. Because $\bar{x}_e$ is fixed, the number of points assigned to noisy and hard questions must sum to $T - x_e$. As a result, increasing the number of points assigned to hard questions decreases the number of points assigned to noisy questions. Thus, increasing the number of points assigned to hard questions makes satisfying condition 2 above more difficult (since $x_n \geq x_h$).

Furthermore, increasing the number of points assigned to hard questions also reduces the probability that a student’s score obscures her type. For example, suppose $x_h = 6$, $x_n = 12$, and $N_n = 2$, satisfying condition (2). Notice that the student only needs to answer one noisy question correctly in order to end up in a non-singleton terminal information set. If $x_h$ increases to 9, the $x_n$ must decrease to 9. Now, the student must answer both noisy questions correctly in order to end up in a non-singleton terminal information set. Therefore, by increasing the number of hard points, the probability that a student ends up in a non-singleton information set decreases.
4.3 Optimal Exam Version

Suppose a teacher is interested in choosing the exam version that maximizes the number of students who study. The teacher may or may not have a choice of $r_0$ or $r_1$ (the probabilities with which noisy questions are answered correctly for each type) - these result in two different optimization problems. First, I consider the optimal exam when the teacher has no choice of $r_0$ and $r_1$. For an exam with $r_0 \neq r_1$ the optimal exam is one that only asks easy questions ($\bar{x}_e = T, x_h = 0, x_n = 0$).\footnote{This is true even if there are students in the population with convex self-esteem preferences. These students will be psychologically indifferent between studying and not studying, so they will study due to higher material payoffs.} However, on this exam, the teacher is entirely rewarding participation, and will be unable to distinguish between high and low ability students. Despite the fact that the maximum number of students will study, this may be undesirable if the teacher would like to incentivize studying while still distinguishing between high and low ability students. Notably, every exam in Category 1 results in strictly less self-handicapping than every exam in Category 2. See Proposition 7.1 in the Appendix. Therefore, the teacher may choose to administer an exam that includes a mix of all three questions types, but choose the distribution of points to satisfy the Category 1 conditions and weight hard questions relatively less.

Second, if the teacher has a choice of $r_0$ and $r_1$, she will increase the number of students who study by setting $r_0 = r_1$. This way, the noisy questions are purely based on luck, making them essentially equivalent to easy questions. When $r_0 = r_1$, any exam with $x_h = 0$ will maximize the number of students who study.\footnote{This is also true in the case where the teacher does not have a choice, but $r_0 = r_1$.}

4.4 Policy Implications

Students who choose not to study (either because it is costly or the outside option is too appealing) may seem indistinguishable from those with self-esteem concerns. However, the policy intervention for these students will be different. Taking measures to reduce costs of studying - i.e., parents forcing their kids to stay at the library on exam night - will not have the same impact on the students who are concerned with self-esteem, since they may still receive negative information about their ability. That is why, in the absence of costs, such students still avoid studying. Reformatting the exam,
however, would cause some of these students to change their behavior. Additionally, changing the exam in this way will not negatively impact any students in the class who are solely motivated to earn the highest grade possible. This is because the expected test score conditional on studying is preserved across all exams.

The results of the section suggest that, perhaps, teachers may provide more opportunities for participation scores, which is equivalent to setting $\bar{x}_e = T$. Furthermore, teachers may impose harsh time constraints or ask questions from supplementary course material to make questions noisier, which is like increasing $x_n$. Finally, they may choose coarser rubrics, which is like decreasing $x_h$.\(^{22}\)

Another way of helping self-handicapping students is to change their beliefs about their production function of effort and ability. Notice that if the student believes that the exam is more ability-based, rather than effort-based then she is less inclined to study.\(^{23}\) This is consistent with the psychological literature on locus of control, which suggests that students who believe a task is more effort-based spend more time on activities and find those activities more enjoyable. (See Lefcourt (2014).)

5 Tournaments

An agent is considering whether to submit a costly application to a job. Whether or not the prospective applicant applies depends on the probability she assigns to getting hired. This probability depends, in part, on whether she is more qualified than other applicants. When the cost of applying is zero, a materially motivated applicant will apply to the job.

This is no longer the case when the prospective applicant is motivated by self-esteem. Whether there are other applicants for the job changes both the probability that the agent is hired, and the information that the agent gleans from being selected (or not). Finding out that she was not selected for the position may reduce her self-esteem by so much that, in anticipation, she may avoid applying altogether - even if applying is otherwise costless.

\(^{22}\)For example, Pass/Fail may be preferred to letter grades, which may be preferred to assigning exact points.

\(^{23}\)The optimal exam (all participation) is entirely effort based - the student’s grade only depends on her action and not on her type.
The size and direction of the effect of a hiring decision on an agent’s self-esteem depends on the relative attractiveness of the applications. For example, one applicant may possess a more desirable feature than another. This feature may be a job-related qualification, or some quality that is unrelated to the position (such as the applicant’s gender). Whether these features move the hiring decision towards or away from an applicant impacts the informativeness of the hiring decision. This, in turn, affects the agent’s self-esteem when she applies.

5.1 Setup

There are two players, $i=1,2$. Each has a type $\tau_i$ drawn from a uniform distribution on $[0,1]$. Each $\tau_i$ is drawn independently. The players simultaneously choose to Apply (A) or to Not Apply (N). The winner depends on the players’ strategies, types, and a known bias against one of the players. If player $i$ wins, her material payoff is 1; if she loses, her material payoff is 0. I abstract away from applications costs. Thus, the only reason why either player would choose not to apply in equilibrium is due to self-esteem concerns.

Call the bias $b$ and let $b \in (-1,1)$. The winner of the tournament is determined as follows: if player $i$ chooses N then $i$ loses. If only one player applies, that player wins. If both players apply, 1 wins if $\tau_1 > \tau_2 + b$, and 2 wins if $\tau_2 + b > \tau_1$. This is the sense in which $b$ is a bias against one of the players. (If $b > 0$, it is a bias against 1, and if $b < 0$, it is a bias against 2.)

5.2 Equilibrium Analysis and Key Results

One important and perhaps counterintuitive consequence of adding bias to the tournament is that players are less likely to apply if the tournament is biased in their favor. To show this result, I begin by characterizing the pure strategy equilibria in the case where $b = 0$. All proofs in this section are located in the Appendix.

---

24 All of the results in this section still hold if applying to the job is materially costly.

25 As one can see from the results, there will also exist a mixed strategy equilibrium. However, one must be careful in specifying the terminal information. For example, if the players only observe the hiring decision (but not the outcome of the mix), this will lead to a different equilibrium than when players observe the outcome of the opponent’s mixed strategy. In the pure strategy case, this does not effect the equilibria, as players have correct beliefs about the strategies of others.
Proposition 5.1. Suppose $b = 0$. There exists some $\hat{\theta}$ such that the pure strategy equilibria when $b = 0$ are as follows:

1. When $\theta_1 \geq \hat{\theta}$, $(N, A)$ is an equilibrium.

2. When $\theta_2 \geq \hat{\theta}$, $(A, N)$ is an equilibrium.

3. When $\hat{\theta} \geq \theta_1$ and $\hat{\theta} \geq \theta_2$, $(A, A)$ is an equilibrium.

Proposition 5.1 can be visualized in Figure 5.1.

When $b = 0$ and at least one $\theta_i > \hat{\theta}$, only one of the players chooses to apply for the job in equilibrium. By not applying when her opponent applies, the prospective applicant reduces her material payoff (she does not get the job for sure), but she maintains her self-esteem (she avoids observing a lower ex post expected type from losing). The prospective applicant who applies also gets no information from being hired. Thus, this player will apply even if she is sufficiently motivated by self-esteem concerns.

When both $\theta$s are sufficiently low, both players apply to the job in equilibrium. In this case, neither player is highly concerned with maintaining their self-esteem. The potential upside of a

---

26When the prospective applicant does not apply, she is engaging in $\alpha_i$-self-handicapping behavior. Here, the $\alpha_i$ is that which is determined in equilibrium.
higher the material payoff from the tournament is enough to outweigh self-esteem concerns for both of these players.

I now compare the case where there is no bias \((b = 0)\) to that where there is bias in favor of \(2 (b > 0)\). Proposition 5.2 characterizes the equilibria in this case, and states that the threshold sensitivity above which players choose not to apply is lower for player 1 and higher for player 2. Intuitively, player 1 receives a more favorable signal both when she wins (she beat an opponent with an advantage) and when she loses (she lost to an opponent with an advantage). Player 2 experiences the opposite effect.

**Proposition 5.2.** Suppose \(b > 0\). There exists some \(\bar{\theta}_2 > \hat{\theta} > \bar{\theta}_1\) such that the following are equilibria:

1. When \(\theta_1 \geq \bar{\theta}_1\), \((N, A)\) is an equilibrium.
2. When \(\theta_2 \geq \bar{\theta}_2\), \((A, N)\) is an equilibrium.
3. When \(\bar{\theta}_1 \geq \theta_1\) and \(\bar{\theta}_2 \geq \theta_2\), \((A, A)\) is an equilibrium.

To apply in equilibrium, player 1 must now be less sensitive to self-esteem than when \(b = 0\). Furthermore, player 2 must now be more sensitive than when \(b = 0\) to self-esteem. Notice that the case where \(b > 0\) from player 2’s perspective is equivalent to the case where \(b < 0\) from player 1’s perspective. The contrast between both cases seen in Figure 5.2.\(^{27}\)

If we think of an agent being drawn from a population with heterogeneity in \(\theta_i\), then Proposition 5.2 implies that if the tournament favors a player’s opponent, then she is more likely to apply in equilibrium. The intuition behind this result is as follows: when the applicant applies to a job where she knows she will be discriminated against, the prospective applicant (player 1) gets a relatively weak signal when she does not get the job, but a strong signal when she does get the job. In order to beat out another candidate who was favored, she must have had high ability. However, if she does not get the job, this could be entirely due to the discrimination.

On the other hand, if the tournament favors the applicant, she is less likely to apply in equilibrium. Unlike before, when player 1 wins she receives a weak signal and when she loses she receives

\(^{27}\)I denote the thresholds for \(b < 0\) by \(\hat{\theta}_1\) and \(\hat{\theta}_2\).
Figure 5.2: Change in Thresholds due to Nonzero Bias

a strong signal. When she is hired, this might only be due to discrimination and not her ability. However, when she loses in spite of the advantage, she must have had low ability.

5.3 Implications

In the case of no bias, prospective job applicants who are concerned with self-esteem may not apply to a job, even if applying is costless. Such results are supported by a growing body of empirical evidence in the tournament literature. In this literature, a well-documented result is that women shy away from competition. Specifically, laboratory results suggest that women are much less likely to opt-in to a tournament than men, even when the tournament involves completing gender-neutral task. One potential explanation is that women are more concerned with self-esteem than men, and therefore choose not to enter a tournament.

In the cases where $b \neq 0$, the bias term $b$ can be interpreted as discrimination by the employer, an ability gap between the players, or a recruitment policy. Qualification/ability gap between the players. Given the discrimination interpretation, the results of this section suggest that discrimination may affect the pool of candidates who apply for a job in a counterintuitive way: the party who is discriminated against may be more inclined apply and the more favorable candidates may opt not to apply if all parties prefer to maintain their self-esteem.

Opting-in to a tournament is analogous to applying for the job.
The results of this section are equivalent if $\tau_2$ is instead drawn from a uniform distribution on $[b, 1+b]$. This can be though of as an ability gap between players 1 and 2. Given this interpretation of $b$, the results suggest that competitors concerned with self-esteem may prefer to enter a tournament against an expert rather than an amateur. Furthermore, job seekers may prefer to apply to jobs for which they are underqualified rather than overqualified. The experimental work by Niederle and Vesterlund (2007) documents that the women who opt-out of tournaments often have higher ability levels. Self-esteem concerns are one plausible explanation for this empirical phenomenon.

Finally, given the recruitment interpretation of $b$, the results of this section suggest that policies targeting a specific group with the intent of encouraging higher application rate may lead to strategic adjustment by that group that could, in part, undermine the policy goal. For instance, if a potential applicant pool has self-esteem concerns, a policy aimed at encouraging female applicants may result in a larger pool of male applicants. These effects may compound if there is both an ability gap and a recruitment policy, resulting in a large pool of low ability male applicants when the desired employee may be a high ability female.

6 Conclusion

Self-esteem preferences may have an nontrivial impact on typical behavior of the people with these preferences. Agents who wish to maintain their self-esteem may resort to self-handicapping in situations where not doing so may be materially beneficial. Exploring the link between self-esteem preferences and self-handicapping behavior is critical for understanding how to mitigate self-handicapping.

To explore this link, I formalize the utility of agents with concerns for maintaining their self-esteem. Then, I define self-handicapping behavior. I show that strict concavity of an agent’s self-esteem utility is sufficient for self-handicapping behavior to be rational. However, self-handicapping is not necessarily a best response. This is due to the nature of the self-handicapping relation, which is a strict partial order.

To understand the potential policy implications for a population with self-esteem preferences,
I explore two applications. In the first application, I show that noisier exam questions can increase the number of students who study for an exam. In the second application, I show that policies with the goal of increasing participation from some group may, in fact, have the opposite impact on participation. Fully understanding self-esteem preferences allows for further analysis of important behaviors and their impact on policies. This paper demonstrates the importance of that understanding, and provides several key insights into the mechanism underlying self-handicapping behavior. While I explore two applications of self-handicapping, there are many possibilities for future exploration of empirically relevant applications using the formal notions established in this paper.
References


7 Appendix

7.1 Self-Handicapping without Strict Concavity

In this example, Ann has self-esteem utility

\[ f(\mathbb{E}_{\beta_a(p_a)}[\tau_a]) = \min\{[\mathbb{E}_{\beta_a(p_a)}[\tau_a] - 0.5], 0\} \]

at terminal information set \( p_a \). Notice that \( f \) is not strictly concave. None the less, this results in self-handicapping behavior. Notice that Ann receives negative self-esteem when her expected type is below 0.5, and 0 self-esteem when her expected type is above 0.5. This utility function can be interpreted as Ann suffering whenever she finds out she is below average.

Suppose Ann has two equally probable types, \( \tau_a = 0 \) and \( \tau_a = 1 \). First Ann chooses between \( s^1_a \) and \( s^2_a \). Following each of Ann’s strategies, chance can choose one of three actions: Report 0, Report 1, or Report \( \emptyset \). Chance’s actions can be interpreted in the same way as Example 3, but with different probabilities. Figure 7.1 depicts these probabilities.

<table>
<thead>
<tr>
<th></th>
<th>( \tau_a = 0 )</th>
<th>( \tau_a = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>R1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>R\emptyset</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 7.1: Chance Probabilities

If Ann chooses \( s^1_a \) (respectively, \( s^2_a \)), she receives a material payoff of 1 (respectively, 0). Ann’s strategy \( s^2_a \) is \( \alpha_i \)-self-handicapping relative to strategy \( s^1_a \). Clearly, \( s^2_a \) has a lower material payoff than \( s^1_a \). Furthermore, \( s^2_a \) results in an interim distribution that places equal probability on an ex post expected type of 0 or 1. On the other hand, \( s^1_a \) results in an interim distribution that places probability 1 on an ex post expected type of 0.5.

If Ann chooses \( s^1_a \), there is \( \frac{1}{2} \) probability with which Ann realizes she is a type 1, in which case Ann’s utility is 1. There is also a \( \frac{1}{2} \) probability with which Ann realizes she is a type 0, in which case Ann’s utility is \( 1 - 0.5\theta_a \). Thus, Ann’s interim expected utility given strategy \( s^1_a \) is

\[ \mathbb{E}_{(\nu \times \alpha_i)}[u_a | s^1_a] = 1 - 0.25\theta_a \]

Ann’s interim expected utility from strategy \( s^2_a \) is \( \mathbb{E}_{(\nu \times \alpha_i)}[u_a | s^2_a] = 0 \). When \( \theta_a \geq 2 \), Ann prefers the self-handicapping strategy \( s^2_a \). Even though \( f \) is not strictly concave, self-handicapping behavior still maximizes Ann’s utility.
Notice that in this example, Ann has a reference point of 0.5. In general, self-handicapping behavior will not occur for any game with any reference point in this form of self-esteem utility. Furthermore, an appropriate reference point depends on the context of the game, and may not generalize well.

### 7.2 Education Proofs

**Lemma 7.1.** If an exam satisfies

1. \( x_h = 0 \) or

2. \( x_n \geq x_h \geq 0 \), \( N_n \geq 1 \), and \( \frac{x_n}{N_n} \) is a factor of \( x_h \),

then there exists some terminal information set that is a non-singleton following \( S \). If neither condition is satisfied, then all information sets that follow \( S \) are singletons.

**Proof.** First, suppose only condition 1 is satisfied. Then there is some positive probability with which both high and low types who study earn the same score on the exam (consider, for example, the case where both types answer all noisy questions incorrectly). Then there is at least one terminal information set that occurs with positive probability that is a non-singleton.

Next, suppose only condition 2 is satisfied. There is some positive probability with which a high type who studies answers all noisy question correctly, resulting in a final score of \( \bar{x}_e + x_h \). Notice, a low type will never answer hard questions correctly. However, because \( \frac{x_n}{N_n} \) is a factor of \( x_h \), and \( x_n \geq x_h \), there is some positive probability with which a low type answers enough noisy questions correctly such that their final score is \( \bar{x}_e + x_h \). This is one terminal information set that occurs with positive probability that is a non-singleton.

Finally, suppose neither condition is satisfied. Because \( x_h > 0 \), high types always earn a positive number points from hard questions that low types do not. The minimum score that can be earned by a high type is then \( \bar{x}_e + x_h \). The maximum score for a low type is \( \bar{x}_e + x_n \). If \( x_n < x_h \), then \( \bar{x}_e + x_n < \bar{x}_e + x_h \), so there is no non-singleton terminal information set that can be reached. If \( x_n > x_h \) but \( \frac{x_n}{N_n} \) is not a factor of \( x_h \), then there is no number of noisy questions such that, if a
low type answers all correctly, they can earn the same final score as a high type. Thus, there is no non-singleton terminal information set that can be reached.

**Proposition 7.1.** All Category 1 exams (non-singleton) result in less self-handicapping than all Category 2 (singleton) exams.

*Proof.* Because $E_{(\nu, \alpha_i)}[\Pi_i]$ is constant across all exams, the only difference between the expected utility of exams in each Category is the student’s expected self-esteem utility. Notice that for all Category 2 exams, the student perfectly infers her type. For all Category 1 exams, the student remains less informed. Thus, her interim distribution over ex post expected types for any Category 1 exam will second-order stochastically dominate her interim distribution over ex post expected types for any Category 2 exam, resulting in a higher expected self-esteem utility, and thus a higher expected utility. Therefore, students with some $\theta_i$’s will study on a Category 1 exam, who would not study on a Category 2 exam, resulting in less self-handicapping. 

**7.3 Tournament Proofs**

Let $X$ and $Y$ be independent random variables with absolutely continuous cdfs $G_X$ and $G_Y$. Write $g_X$ and $g_Y$ for the associated pdfs. Write $g_{X|X<Y}$ for the pdf of $X$ conditional on $X < Y$, and similarly write $g_{X|X>Y}$ for the pdf of $X$ conditional on $X > Y$.

**Lemma 7.2.** $g_{X|X<Y}(x) = \frac{(1-G_Y(x))g_X(x)}{\int_{-\infty}^{\infty} G_X(y)g_Y(y)dy}$ and $g_{X|X>Y}(x) = \frac{G_Y(x)g_X(x)}{1-\int_{-\infty}^{\infty} G_X(y)g_Y(y)dy}$.

*Proof.* First, notice that,

$$g_{X|X>Y}(x) = \frac{g_{X>Y|X}(x)g_X(x)}{g_{X>Y}(x)}.$$ 

The term

$$g_{X>Y|X}(x) = g_{Y<X|X}(x) = G_Y(x).$$

By independence, the denominator

$$g_{X>Y}(x) = 1 - g_{X<Y}(x) = 1 - \int_{-\infty}^{\infty} G_X(y)f_Y(y)dy.$$
Thus,

\[ g_{X|X>Y}(x) = \frac{G_Y(x)g_X(x)}{1 - \int_{-\infty}^{\infty} G_X(y)f_Y(y)dy}. \]

Similarly,

\[ g_{X|X<Y}(x) = \frac{g_{X<Y|X}(x)g_X(x)}{g_{X<Y}(x)}. \]

Notice that the term \( g_{X<Y|X}(x) = 1 - g_{X>Y}(x) \) and the term \( g_{X<Y}(x) = 1 - g_{X>Y}(x) \). Then,

\[ g_{X|X<Y}(x) = \frac{(1 - G_Y(x))g_X(x)}{\int_{-\infty}^{\infty} G_X(y)f_Y(y)dy}. \]

Lemma 7.3. \( E[\tau_i > \tau_{-i} + b] \) and \( E[\tau_i < \tau_{-i} + b] \) are increasing in \( b \).

Proof. Let \( \tau_i = X \), which is uniform on \([0, 1]\). Let \( \tau_{-i} + b = Y \), which is uniform on \([b, 1 + b]\). If \( i \) chooses \( A \) and loses the tournament, then

\[
E_{\beta_i[A,1]}[\tau_i] = E[\tau_i \mid \tau_i < \tau_{-i} + b] = E[X \mid X < Y].
\]

Applying Lemma 7.2,

\[
E[X \mid X < Y] = \int_0^1 x \frac{g_X(x)(1 - G_Y(x))}{\int_b^{1+b} G_X(y)g_Y(y)dy} dx = \frac{1 + 3b}{3 + 6b},
\]

which is increasing in \( b \).

Similarly, if \( i \) chooses \( A \) and wins the tournament, then

\[
E_{\beta_i[A,1]}[\tau_i] = E[\tau_i \mid \tau_i > \tau_{-i} + b] = E[X \mid X > Y].
\]

Applying Lemma 7.2,

\[
E[X \mid X > Y] = \int_b^1 x \frac{g_X(x)G_Y(x)}{1 - \int_b^{1+b} G_X(y)g_Y(y)dy} dx = \frac{\frac{b^3 - b}{2(1-b)} + \frac{1-b^3}{3(1-b)}}{\frac{1-b}{2}},
\]
which is increasing in $b$. □

**Proof of Proposition 5.1.** In equilibrium, $i$’s belief about $j \neq i$’s strategy will be correct. Thus, I will first consider the case when $j$ plays $N$. Notice that when player $i$ chooses $N$, she always loses and learns no new information about her type. When player $i$ chooses $A$, she always wins and learns no new information about her type. Thus, her self-esteem payoffs from $(A, N)$ and $(N, N)$ are the same. Therefore, player $i$ will play $A$ because it maximizes her material payoff.

Now I consider the case in which player $j$ plays $A$. If player $i$ plays $N$, player $i$ will lose the tournament and gain no new information about her type. Then,

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid N] = 0 + \theta_i f\left(\frac{1}{2}\right).$$

If player $i$ plays $A$, player $i$ will either win, and learn that her type was higher than $j$’s, or lose and learn that her type was lower than $j$’s. Therefore,

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A] = Pr(\tau_i > \tau_j)u_i(\beta_i, [A, 1]) + Pr(\tau_i < \tau_j)u_i(\beta_i, [A, 0]).$$

If player $i$ plays $A$ and wins, her ex post expected type (from Lemma 7.3) is

$$\mathbb{E}_{\beta_i, [A, 1]}[\tau_i] = \mathbb{E}[\tau_i \mid \tau_i > \tau_j] = \frac{2}{3}.$$

Similarly, if player $i$ plays $A$ and loses, her ex post expected type (from Lemma 7.3) is $\frac{1}{3}$. So,

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A] = \frac{1}{2}(1 + \theta_i f\left(\frac{2}{3}\right)) + \frac{1}{2}(0 + \theta_i f\left(\frac{1}{3}\right)).$$

Player $i$ then prefers to play $N$ when $\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid N] \geq \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A]$, which is true when

$$\theta_i \geq \frac{1}{2(f(\frac{1}{3})-\frac{1}{2}(f(\frac{1}{3})+f(\frac{1}{3})))}. \quad \text{Player } i \text{ plays } A \text{ when the inequality is reversed.}$$

**Proof of Proposition 5.2.** From Proposition 5.1, there exists a $\hat{\theta}$ for which $\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid N, b = 0] \geq \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b = 0]$. Now consider $b > 0$ and let $\alpha_i$ assign probability 1 to player $j \neq i$ playing $A$. }
Because $b < 1$,

$$\mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i \mid A, b > 0] > \mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i \mid N, b > 0].$$

By Jensen’s inequality, player 1’s interim cdf over ex post expected types induced by strategy $N$ second-order stochastically dominates the interim cdf induced by strategy $A$ when $b > 0$. Thus, by Definition 3.1, $N$ is $\alpha_i$-self-handicapping relative to $A$ (where $\alpha_i$ assigns probability 1 to player $j \neq i$ choosing $A$). Therefore, by Theorem 3.1, there exists some $\hat{\theta}_1$ for which

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid N, b > 0] \geq \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b > 0].$$

Similarly, there exists some $\hat{\theta}_2$ for which

$$\mathbb{E}_{(\nu \times \alpha_2)}[u_2 \mid N, b > 0] \geq \mathbb{E}_{(\nu \times \alpha_2)}[u_2 \mid A, b > 0].$$

Now notice that $\mathbb{E}_{(\nu \times \alpha_i)}[\Pi_i \mid A]$ is decreasing in $b$. From Lemma 7.3, $\mathbb{E}[\tau_i \mid \tau_i > \tau_{-i} + b]$ and $\mathbb{E}[\tau_i \mid \tau_i < \tau_{-i} + b]$ are increasing in $b$. Therefore, there exists some $\hat{\theta}_1^*$ such that, for all $\theta_i > \hat{\theta}_1^*$,

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b > 0] > \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b = 0].$$

Let $\hat{\theta} \geq \hat{\theta}_1^*$. Then

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid N, b = 0] = \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid N, b \neq 0].$$

Therefore, for all $\theta_i \geq \hat{\theta}$,

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid N, b = 0] \geq \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b > 0] \geq \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b = 0].$$

Thus $\hat{\theta} \geq \hat{\theta}_1$.

Similarly, notice that there exists some $\bar{\theta}_2^*$ such that, for all $\theta_i \geq \bar{\theta}_2^*$,

$$\mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b = 0] > \mathbb{E}_{(\nu \times \alpha_i)}[u_i \mid A, b > 0].$$
Let $\hat{\theta} \geq \bar{\theta}_2^*$. Then for all $\theta_i \geq \bar{\theta}_2$,

$$E_{(\nu \times \alpha_i)}[u_i \mid N, b = 0] \geq E_{(\nu \times \alpha_i)}[u_i \mid A, b = 0] \geq E_{(\nu \times \alpha_i)}[u_i \mid A, b < 0].$$

Thus $\hat{\theta} \geq \bar{\theta}_2$. \qed