Cognition 266 (2026) 106288

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

A signaling theory of self-handicapping

Yang Xiang [a,*], Samuel J. Gershman [a,b,c,1], Tobias Gerstenberg [d,1]

[a] Department of Psychology, Harvard University, Cambridge, MA 02138, United States of America
[b] Center for Brain Science, Harvard University, Cambridge, MA 02138, United States of America
[c] Center for Brains, Minds, and Machines, MIT, Cambridge, MA 02139, United States of America
[d] Department of Psychology, Stanford University, Stanford, CA 94305, United States of America

ARTICLE INFO

Dataset link: https://github.com/yyyxiang/self-handicapping

Keywords:
Self-handicapping
Signaling
Competence
Bayesian inference
Attribution
Theory of mind

ABSTRACT

People use various strategies to bolster the perception of their competence. One strategy is *self-handicapping*, by which people deliberately impede their performance in order to protect or enhance perceived competence. Despite much prior research, it is unclear why, when, and how self-handicapping occurs. We develop a formal theory that chooses the optimal degree of self-handicapping based on its anticipated performance and signaling effects. We test the theory's predictions in two experiments ($N = 400$), showing that self-handicapping occurs more often when it is unlikely to affect the outcome and when it increases the perceived competence in the eyes of a naive observer. With sophisticated observers (who consider whether a person chooses to self-handicap), self-handicapping is less effective when followed by failure. We show that the theory also explains the findings of several past studies. By offering a systematic explanation of self-handicapping, the theory lays the groundwork for developing effective interventions.

1. Introduction

One of the most important attributes of people is their competence—the ability to perform well in various aspects of life (Anderson, 1968; Festinger, 1954; White, 1959). People use a variety of strategies to bolster others' perception of their competence (Bradley, 1978; Snyder & Higgins, 1988; Tesser, 1988; Wills, 1981). One strategy is *self-handicapping*, where a person deliberately impedes their performance to protect perceived competence in case of failure, or enhance it in case of success (Berglas & Jones, 1978). For example, a student might procrastinate before an exam and then use tiredness as an excuse for poor performance, rather than lack of ability.

Much work has documented self-handicapping (Beck et al., 2000; Berglas & Jones, 1978; Ferrari & Tice, 2000; Greenberg, 1985; Higgins & Harris, 1988; Rhodewalt & Davison, 1986; Smith et al., 2009; Thompson & Richardson, 2001; Tice, 1991; Tice & Baumeister, 1990; Tucker et al., 1981). In the context of academic learning, self-handicapping can be harmful. It decreases performance over time (Gadbois & Sturgeon, 2011; Nurmi et al., 1995; Schwinger et al., 2014; Urdan, 2004; Urdan et al., 1998; Zuckerman et al., 1998), lowers well-being, self-esteem, academic and competence satisfaction, and intrinsic motivation (Eronen et al., 1998; Zuckerman & Tsai, 2005). Therefore, it is important to better understand self-handicapping to design effective interventions.

Despite much prior research, self-handicapping remains poorly understood from a theoretical standpoint. Prior work has described different situations in which people self-handicap and observations of how this behavior is perceived. However, these theories have largely been verbal descriptions and do not specify the cognitive processes that underlie the behavior. As a result, they cannot easily explain conflicting empirical observations of why people self-handicap in some situations but not others (Self, 1990), or why observers sometimes view self-handicappers as more competent (Luginbuhl & Palmer, 1991) and sometimes as less (Rhodewalt et al., 1995). Past work has mainly explained the phenomenon as the self-handicapper signaling their competence by affecting the observer's causal attributions (e.g., Arkin & Baumgardner, 1985; Arkin & Oleson, 1998; Berglas & Jones, 1978; Rhodewalt & Tragakis, 2002), often using Kelley's (1973) discounting principle (where the role of a particular cause is reduced if other plausible causes are also present) and augmentation principle (where the role of a particular cause is amplified if an event occurs despite the presence of known constraints, costs, or risks). For example, failing the exam is attributed to tiredness instead of lacking competence, and succeeding despite having been tired leads to increased perceptions of competence. Existing theories have also identified situational factors (Elliot et al., 2006; Schwinger et al., 2014; Self, 1990; Shepperd & Arkin, 1989) and personality traits (Bobo et al., 2013; Prapavessis & Grove, 1998; Ross

* Corresponding author.
E-mail address: yyx@g.harvard.edu (Y. Xiang).
[1] Equal senior authors.

https://doi.org/10.1016/j.cognition.2025.106288
Received 1 April 2025; Received in revised form 19 June 2025; Accepted 8 August 2025

et al., 2002; Tice & Baumeister, 1990) that predict self-handicapping behaviors. However, none of these theories explains precisely what conditions would motivate an actor to carry out this gambit, when this gambit might backfire, and how different factors are weighed against each other. Thus, a formal mathematical theory of self-handicapping is needed to understand the computational principles driving the behavior and its interpretation by others. Such a theory not only explains the relevant phenomena, but also allows us to derive novel predictions—for example, that self-handicapping may emerge not only when actors anticipate failure, but also when they are confident enough to risk a harder task to enhance their perceived competence.

To understand the actor's behavior, it is important to highlight the complexity of their decision. The actor chooses whether to self-handicap before the task begins—before knowing whether they will succeed or fail. This distinguishes self-handicapping from classic "explaining away" scenarios, where an observer updates beliefs about one latent cause (e.g., competence) after observing an outcome (e.g., success or failure), while conditioning on the presence of another (e.g., a handicap). Classical attribution theories can explain how observers weigh dispositional and situational causes (e.g., Gilbert & Malone, 1995; Jones & Davis, 1965; Jones & Harris, 1967; Kelley, 1973; Walker et al., 2015), such as whether failure was due to low ability or to a situational handicap. However, in the case of self-handicapping, the actor's choice is not a fixed explanatory variable—it directly alters the probability of success. As a result, observers must reason not just about which latent cause best explains the outcome, but also about how the actor's anticipatory choice shaped the outcome in the first place. To interpret such behavior, observers must consider how the actor anticipated being evaluated, which requires a form of nested social reasoning that goes beyond standard causal attribution.

Here, we develop a signaling theory of self-handicapping that predicts when these attributional principles apply and how they influence the behavior. The theory explains why, when, and how self-handicapping occurs. The theory involves a naive observer, an actor, and a sophisticated observer. The naive observer evaluates the actor's competence based on their outcome and handicap. The actor seeks to impress the naive observer through strategic self-handicapping. The sophisticated observer considers the actor's decision whether to self-handicap and evaluates the actor's competence accordingly. This distinction between naive and sophisticated observers follows past work showing that observers who were previously self-handicapping actors think differently about the tactics of other actors (Smith & Strube, 1991).

The signaling theory builds on recent progress in Theory of Mind modeling, in which people reason about others' beliefs and desires recursively. Recursive Theory of Mind models have been applied across several domains, including pedagogical reasoning (Gweon, 2021; Shafto et al., 2014), rational speech act (Beller & Gerstenberg, 2025; Goodman & Frank, 2016), collaborative decision making (Xiang et al., 2023), and social reasoning under uncertainty, such as deception and skepticism (Alon et al., 2023). By modeling how a sophisticated observer interprets the actor's behavior through reasoning about how the actor anticipated being judged, the signaling theory extends this approach to a new domain: competence inference where motives are strategic, the context is evaluative, and choices affect not only others' beliefs but also task outcomes themselves, in contrast to most prior work that assumes cooperative or communicative goals.

We tested the theory in two experiments ($N = 400$) that use a game show setting where actors' competence is judged by observers. Actors can choose to self-handicap. Participants played both actors and observers in different phases. We manipulated the level of observer sophistication by having participants play the role of an observer twice: once before they played the actor role, when they were "naive", and again afterward, when they were "sophisticated" and could think through how an actor might behave. Consistent with the theoretical predictions, we found that: (a) Participants were more likely to self-handicap when they were either very incompetent or very competent, but not when they were just good enough for the task; and (b) self-handicapping when the actor failed increased naive observers' evaluations, but less so for sophisticated observers. We additionally show that the theory captures several results from earlier self-handicapping studies.

2. Theory

As illustrated in Fig. 1, the theory involves: (a) a *naive observer* who evaluates an actor's competence; (b) an *actor* who seeks to impress the naive observer through strategic self-handicapping; and (c) a *sophisticated observer* who sees through the actor's intent and evaluates the actor's competence knowing that they chose whether to handicap. Suppose an actor with competence c performs a task of difficulty d, achieving outcome $s \in \{0, 1\}$, where $s = 0$ indicates failure and $s = 1$ success. The actor can choose to self-handicap, which creates an impediment such that only a proportion $\gamma \in (0, 1]$ of their competence is applied to the task (e.g., taking a performance-inhibiting drug as in Berglas and Jones (1978)). $\gamma = 1$ means that the actor did not self-handicap at all—thus preserving full ability, whereas $\gamma = 0$ means that the actor is completely unable to carry out a task (e.g., paralyzed).

2.1. Naive observer

The naive observer evaluates the actor's competence based on the outcome s and the handicap factor γ, using Bayesian inference:

$$P(c|s, \gamma) \propto P(s|c, \gamma)P(c), \tag{1}$$

where $P(c)$ is the naive observer's prior over competence, which we assume to be uniform, and $P(s|c, \gamma)$ is the likelihood of success or failure given c and γ.

The likelihood of success given c and γ follows a logistic function:

$$P(s = 1|c, \gamma) = \frac{1}{1 + e^{-k((\gamma c - d) - b)}}, \tag{2}$$

where k controls the steepness of the curve and b adjusts the position of the sigmoid midpoint. γ controls what proportion of an actor's competence is used to perform the task. The probability of success thus depends on the difference between this fractional competence and the task difficulty d.

2.2. Actor

We assume that the actor has two goals: (1) maximizing their perceived competence, and (2) succeeding at the task. We formalize the first goal as follows:

$$\mathbb{E}[\hat{c}|c, \gamma] = \sum_s P(s|c, \gamma)\mathbb{E}[c|s, \gamma], \tag{3}$$

where $\hat{c} = \mathbb{E}[c|s, \gamma] = \int_c P(c|s, \gamma)c\, dc$ is the naive observer's perception of the actor's competence after observing s and γ. This expected competence $\hat{c}$ reflects a general expectation of what competence levels are more likely to satisfy the observed variables (γ and outcome s). Thus it is independent of the actor's true competence. The actor takes advantage of the way $\hat{c}$ is computed—knowing their true competence, they maximize the observer's perception of their competence by strategically choosing γ values that bring about favorable perceptions, while keeping in mind how their γ choice might affect the outcome (as weighted by the probability of each outcome occurring).

The second goal is maximizing the likelihood of success:

$$\mathbb{E}[s|c, \gamma] = \sum_s P(s|c, \gamma)s = P(s = 1|c, \gamma) \tag{4}$$

The two goals are then combined by a weight parameter $w \in [0, 1]$ that controls the relative weight the actor places on maximizing perceived competence (the first goal) versus performance (the second

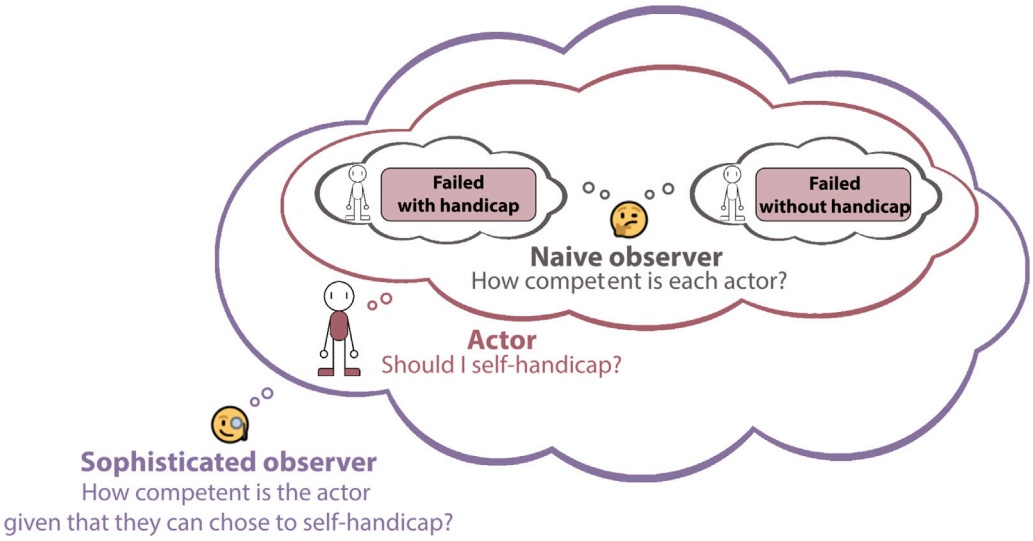

Fig. 1. An illustration of the signaling theory of self-handicapping. The naive observer evaluates the actor's competence based on whether they were handicapped (γ) and whether they succeed or fail (s). The actor decides whether to self-handicap by considering what the naive observer would infer about their competence. The sophisticated observer evaluates the actor's competence knowing that they decided whether to self-handicap.

goal). We use $Q_c(\gamma)$ to denote the value of choosing γ when competence is c:

$$Q_c(\gamma) = w\mathbb{E}[\hat{c}|c,\gamma] + (1-w)\mathbb{E}[s|c,\gamma]. \tag{5}$$

We assume that the actor chooses γ to optimize the choice value $Q_c(\gamma)$. To allow for some stochasticity in choice behavior, we assume a softmax choice probability:

$$P(\gamma|c) \propto \exp[\tau Q_c(\gamma)], \tag{6}$$

where $\tau \geq 0$ is an inverse temperature parameter that controls choice stochasticity by scaling the choice values. Smaller τ produces more stochasticity.

2.3. Sophisticated observer

A sophisticated, "mentalizing" observer considers the actor's decision process and recognizes that γ provides information about c even before s is observed. Therefore, for a sophisticated observer, $P(c)$ in Eq. (1) is replaced with $P(c|\gamma)$:

$$P(c|s,\gamma) \propto P(s|c,\gamma)P(c|\gamma), \tag{7}$$

where $P(c|\gamma)$ is computed by incorporating information gained from the actor's choice of γ:

$$P(c|\gamma) \propto P(\gamma|c)P(c). \tag{8}$$

Note that in order to compute $P(\gamma|c)$, it is necessary for the sophisticated observer to infer the actor's choice value, which in turn requires the actor to infer the observer's beliefs, leading to an infinite recursion. In practice, we cut off this recursion after one step. The theory does not require the sophisticated observer to know w, as it can be marginalized over. For simplicity, we assume w is observed in the experiments.

2.4. Alternative models

We compare the theory to two alternative models inspired by past work. Each model defines an action value function $Q_c(\gamma)$, which represents the value of choosing a particular γ given the actor's competence c. In both models, α is a scaling parameter that controls sensitivity to deviations from the optimal γ, and β is a bias term. The values are then passed through a softmax function (see Eq. (6)) to yield choice probabilities.

The first alternative is an *ambiguity-seeking model* where the actor avoids revealing their true competence. This model captures a behavioral tendency drawn from classic self-deception theories: the avoidance of situations that would lead to a clear and potentially threatening inference about one's true ability (Covington, 1992; Quattrone & Tversky, 1984). In this model, the action value is defined as:

$$Q_c(\gamma) = -\alpha(\gamma-\beta)^2 \tag{9}$$

Notice that this equation is not a function of the actor's competence c. In other words, the actor always prefers to self-handicap (i.e., prefers smaller γ), regardless of their competence.

The second alternative is a *hide-incompetence model* where the actor conceals their competence when it is low, but reveals their competence when it is high (Urdan et al., 1998). In this model, the action value is given by:

$$Q_c(\gamma) = -\alpha(\gamma-\tilde{c}-\beta)^2, \tag{10}$$

where $\tilde{c} \in [0,1]$ represents the actor's competence c normalized to match the scale of γ. Here, the actor prefers smaller γ when competence is low, and prefers larger γ when competence is high.

Each of these alternative models has three free parameters: α, β, and inverse temperature parameter τ for the softmax function. Thus, the number of free parameters matches that of the signaling theory.

3. Experiments 1 and 2

In both experiments, participants first played the role of a naive observer who evaluated actors' competence without knowing that actors could choose whether to self-handicap. Then, they played the role of actors who decide whether to self-handicap based on their competence. Finally, participants played the role of a sophisticated observer who re-evaluated actors' competence knowing that they could choose whether to self-handicap. We manipulated the actors' goals across the two experiments; actors in Experiment 1 aimed to maximize their perceived competence, whereas actors in Experiment 2 aimed to succeed.

3.1. Method

3.1.1. Participants

We recruited 200 participants for Experiment 1 (78 Female, 120 Male, 1 Non-binary, 1 Other; mean age 44 years, range 23–77 years)

and 200 participants for Experiment 2 (87 Female, 112 Male, 1 Non-binary; mean age 44 years, range 21–76 years) via Amazon's Mechanical Turk platform (MTurk). This sample size was selected based on a power analysis on a pilot study of Experiment 1 with 29 participants, which revealed that at least 176 participants are required to detect an effect of evaluation change in the 'Fail10' condition with 90% power. We decided to be conservative and collect 200 participants for each experiment. This decision was preregistered. We did not exclude any participant or observation. Participants received $4 for completing the experiment. The experiments were approved by the Harvard Institutional Review Board and preregistered at https://aspredicted.org/f4h3-f4xv.pdf.

3.1.2. Procedure

In each experiment, 200 participants read vignettes about "Hidden Genius", a game show where actors answered general knowledge questions and judges evaluated their competence. Actors were assigned 20 questions. They could choose to self-handicap and only be evaluated on a random subset of 10 questions—but they have to make this decision before seeing the questions and will not know whether they passed until after they submit all 20 answers. Importantly, this setup allows us to deconfound self-handicapping and putting in less effort. Even if the actor chooses to self-handicap, they still need to answer all 20 questions. Passing required giving at least 8 correct answers, regardless of the number of questions evaluated. However, the judges did not know the exact scores; they only knew whether each actor was evaluated on 10 or 20 questions, and whether they passed (i.e., whether the 10 or 20 answers contained at least 8 correct responses).

It is worth pointing out that this task is fundamentally different from decisions about revealing or concealing information. Regardless of the actor's choice, the judges always observe both the evaluation condition (e.g., whether the performance is assessed on all 20 answers or a subset of 10) and the outcome (pass or fail). Thus, the actor cannot manipulate what information is available to the observer. Importantly, the actor's choice directly alters the difficulty of the task and must be made in advance—before knowing the outcome. Thus, the actor must weigh whether introducing a handicap will ultimately benefit or harm their perceived competence—knowing that it also changes their own chances of success. This structure mirrors classic self-handicapping examples, such as intentionally getting insufficient sleep before an exam (Berglas & Jones, 1978), where an individual imposes a real obstacle on themselves prior to performance.

Each experiment consisted of three blocks, illustrated in Fig. 2. In the first block (Fig. 2(a)), participants played naive judges who thought that actors *could not* choose how many of their answers were evaluated. Participants evaluated four combinations of outcomes and handicaps: An actor could pass with 20 answers evaluated ("Pass20") or 10 ("Pass10"), or fail with 20 answers evaluated ("Fail20") or 10 ("Fail10"). We showed participants four actors with different results on the same screen and participants answered the question "How competent is each contestant?" on a sliding scale that ranged from "Not competent at all" (coded as 0) to "Extremely competent" (coded as 100). The slider selection button was hidden until the slider was clicked.

In the second block (Fig. 2(b)), participants played the role of 11 actors whose average accuracy in practice tests ranged from 0% to 100% (in steps of 10%). The actors were presented in random order, and participants answered the question "How many answers should this contestant choose to be evaluated on?" on a sliding scale ranging from "Definitely 10 answers" (coded as 100% probability of self-handicapping) to "Definitely 20 answers" (coded as 0% probability of self-handicapping), with the middle being "Unsure" (coded as 50% probability of self-handicapping). The actors' goal differed across two experiments. In Experiment 1, the actors' goal was to maximize their competence evaluations. In Experiment 2, the actors' goal was to maximize their chances of succeeding.

Table 1
Summary table of parameter values for modeling new and past experiments.

	c	γ	d	k	b	τ	w
New experiments							
Experiment 1	[0,20]	{0.5,1}	8	1.2	−0.8	0.3	1
Experiment 2	[0,20]	{0.5,1}	8	1.2	−0.8	2.5	0
Past experiments							
Luginbuhl and Palmer (1991)	[0,100]	{0.8,1}	{95,75,55}	0.3	−3	–	–
Tice (1991)	[0,10]	{0.6,1}	3	2	0	15	1
Rhodewalt et al. (1995)	[0,10]	{0.5,1}	5	1	0	0.5	1

In the third block (Fig. 2(c)), participants played sophisticated judges who knew that actors could choose – while the actors thought the judges did not know – and re-evaluated the four actors' competence from the first block. We showed participants four actors with different results on the same screen and participants answered the question "How competent is each contestant?" on a sliding scale that ranged from "Not competent at all" (coded as 0) to "Extremely competent" (coded as 100). The sliders were initialized at the responses from the first block and participants were able to update their evaluations by dragging the selection button. They could also choose to keep the same evaluation by clicking the selection button. To remind participants of their responses in the actor block, we showed them a summary table of their responses next to the response sliders. Additionally, right before this block, we asked participants three reflection questions about what number of answers a more competent, averaged-skilled, or less competent actor should choose based on their previous responses.

Participants completed a few comprehension check questions about the game show after the instructions and right before each block. They were allowed to proceed only after correctly answering all of the questions. They were directed back to the relevant set of instructions each time they failed a comprehension check. A complete list of task instructions and comprehension check questions is included in the Supplement.

3.1.3. Model fitting

Table 1 summarizes the parameter values for modeling the experiments. Task difficulty $d = 8$, handicap factor $\gamma \in \{0.5, 1\}$, and competence $c \in [0, 20]$ (defined as the number of correct answers with the actor's full capacity) were prescribed by the task instructions. Note that the discrete set of γ values was constrained by the experimental setup; this assumption is not required by the theory. We assumed a uniform prior over competence $P(c)$ to avoid making assumptions about participants' prior beliefs about how competent the actors were in general. The model had three free parameters. Two of them were the logistic growth rate k and the x-value of the sigmoid midpoint b in Eq. (2), which we assumed to be shared across both experiments. The last free parameter was the inverse temperature parameter in Eq. (6) that controls the stochasticity of γ selection, which we fit to each experiment because the choice values in the two experiments were on different scales. These three parameters were fit to participant-averaged data in the actor block using the Nelder–Mead optimization algorithm. Loss was measured as the total sum of squared error across two experiments. Data from the other two blocks were compared to the model but not used to fit the model.

3.2. Results

3.2.1. Self-handicapping increases naive observers' evaluations

We first analyzed data from the naive observer block (Block 1). Fig. 3 shows that participants rated actors who passed the test as more competent, and the "Pass10" actor as the most competent (M = 83.64, SD = 14.50 in Experiment 1, M = 82.87, SD = 15.67 in Experiment 2), followed by the "Pass20" actor (M = 74.15, SD = 15.70 in Experiment 1, M = 80.80, SD = 17.42 in Experiment 2).

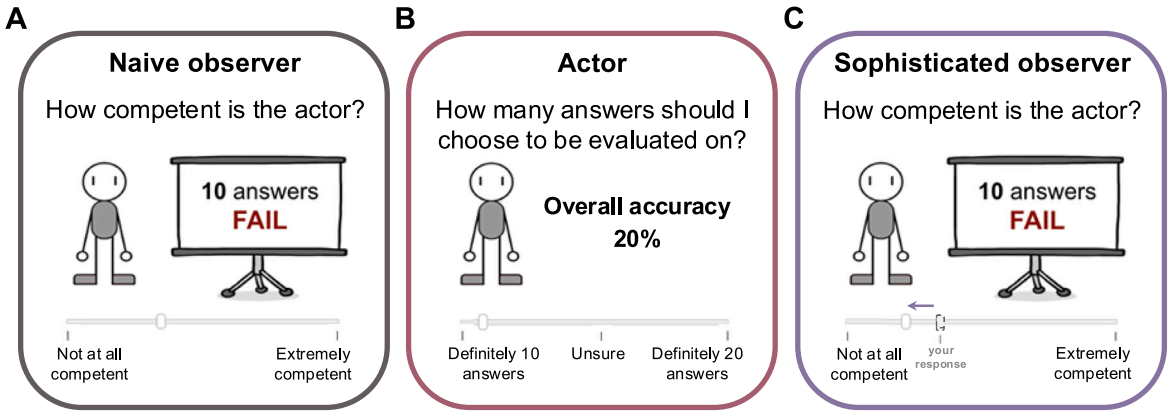

Fig. 2. An illustration of the experiments. (A) In Block 1, participants played the role of naive observers and evaluated the actors' competence based on the results. (B) In Block 2, participants played the role of actors with different competencies and decided whether to be evaluated on 10 or 20 answers. In Experiment 1, their goal was to maximize perceived competence. In Experiment 2, their goal was to maximize chances of succeeding. (C) In Block 3, participants played the role of sophisticated observers and adjusted their previous evaluations. The "your response" indicates the participant's response in Block 1.

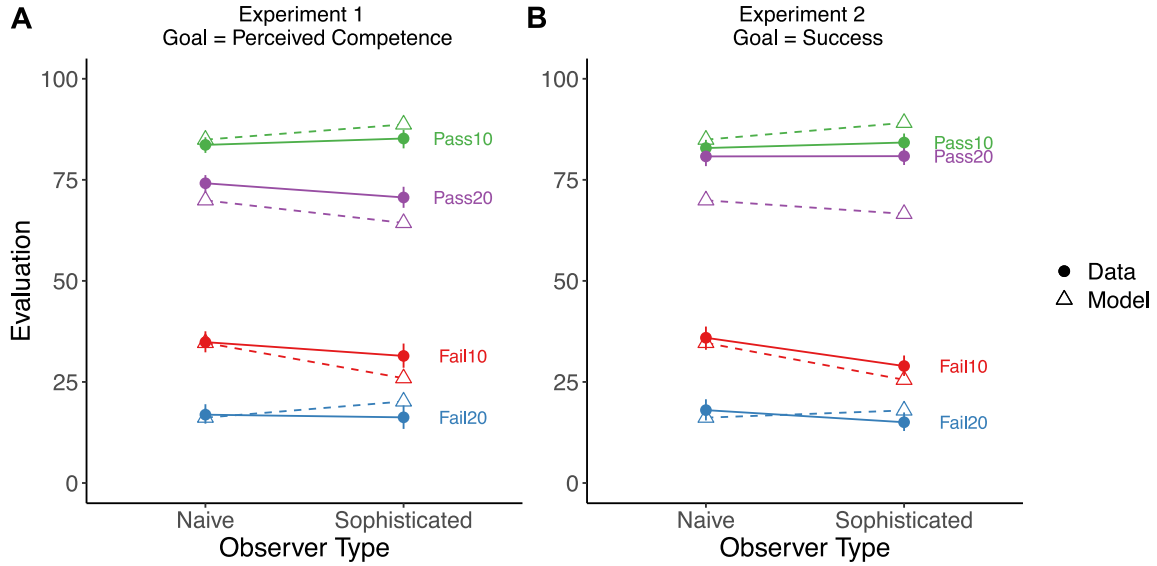

Fig. 3. Observer evaluations of actors' competence. (A) In Experiment 1, the actors' goal was to maximize perceived competence. (B) In Experiment 2, the actors' goal was to maximize chances of succeeding. In both experiments, naive observers rated actors who passed with 10 answered evaluated ("Pass10") more competent than actors who passed with 20 answers evaluated ("Pass20"), followed by actors who failed with 10 answers evaluated ("Fail10") and 20 ("Fail20"). Sophisticated observers rated "Fail10" actors less competent than naive observers. These patterns were captured by the model. Error bars indicate bootstrapped 95% confidence intervals.

Among actors who failed, the "Fail10" actor (M = 34.85, SD = 19.20 in Experiment 1, M = 35.93, SD = 20.12 in Experiment 2) was rated as more competent than the "Fail20" actor (M = 16.9, SD = 17.29 in Experiment 1, M = 18.03, SD = 19.00 in Experiment 2). In other words, actors are considered more competent if they pass compared to if they fail, and when the outcome is the same, actors with handicaps are perceived as more competent. The model captures this pattern (see Fig. 3).

We further confirmed this finding with a Bayesian mixed-effects regression predicting participants' competence evaluations in the naive observer block with the outcome (pass or fail), the number of answers the actor was evaluated on (10 or 20), and intercept, along with random intercept and slopes for each regressor grouped by participants. The results are summarized in Table 2. For both experiments, we found a credible positive effect of outcome, indicating that participants rated actors who passed as more competent than actors who failed. We also found a credible negative effect of the number of answers evaluated in both experiments, meaning that actors who self-handicapped received higher evaluations than actors who did not. In summary, when the outcome is held the same, self-handicapping increases competence evaluations in the eyes of a naive observer who does not think the handicap is strategic.

Note that the model predictions looked similar between the two experiments. This is because the naive observers infer the actors' competence with the same information (outcome and handicap), without thinking about the actors' goals. Participants, on the other hand, evaluated the "Pass20" actor as more competent when their goal was to maximize chances of succeeding (Experiment 2), compared to when their goal was to maximize the naive observers' perceptions of their competence (Experiment 1). This pattern was not captured by the model. The alternative models have the same number of free parameters, but none of the parameters control the success likelihood function (Eq. (2)), hence there are no fitted parameters for k and b. However, the alternative models differ from the signaling theory in how the actor makes decisions – not how observers perceive the actors – therefore their predictions would be similar.

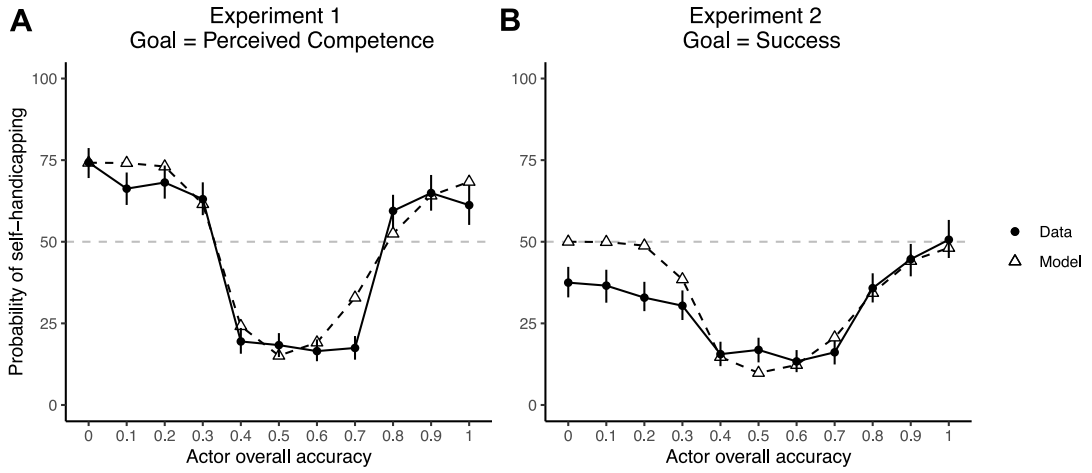

Fig. 4. Actors' probability of handicapping as a function of accuracy. Passing required at least 8 correct responses, regardless of whether 10 or 20 answers were evaluated. (A) In Experiment 1, actors were more likely to self-handicap when they were very incompetent or very competent, but less likely to do so when they were just competent enough for the task. (B) In Experiment 2, actors overall preferred not to self-handicap, although the probability of self-handicapping was slightly higher when the actors were very incompetent or very competent. The model captures these patterns. Error bars indicate bootstrapped 95% confidence intervals.

Table 2
Estimates of a Bayesian mixed effects regression that was fit for the following model: `Naive observer evaluation ~ 1 + Outcome + Answers + (1 + Outcome + Answers | Participant)`. `Outcome` was coded as 0 = 'fail', and 1 = 'pass'. `Answers` was coded as 0 = '10 answers', and 1 = '20 answers'. 'Estimate' shows the mean of the posterior distribution, 'Est. Error' the standard deviation of the posterior distribution, and '95% CrI' the 95% credible interval.

	Estimate	Est. Error	95% CrI
Experiment 1			
(Intercept)	59.25	0.88	[57.52, 60.93]
Outcome	37.51	1.03	[35.50, 39.52]
Answers	−13.72	1.22	[−16.07, −11.39]
Experiment 2			
(Intercept)	59.38	0.88	[57.67, 61.07]
Outcome	38.81	1.19	[36.49, 41.11]
Answers	−9.95	1.14	[−12.18, −7.69]

Table 3
Estimates of a Bayesian mixed effects regression that was fit for the following model: `Probability of self-handicapping ~ 1 + Accuracy² + Accuracy + (1 + Accuracy² + Accuracy | Participant)`. 'Estimate' shows the mean of the posterior distribution, 'Est. Error' the standard deviation of the posterior distribution, and '95% CrI' the 95% credible interval.

	Estimate	Est. Error	95% CrI
Experiment 1			
(Intercept)	86.50	2.82	[80.99, 91.97]
Accuracy^2	198.15	11.07	[176.45, 219.91]
Accuracy	−215.54	10.89	[−236.79, −194.34]
Experiment 2			
(Intercept)	44.63	2.42	[39.96, 49.41]
Accuracy^2	120.33	9.54	[102.16, 139.55]
Accuracy	−113.54	8.77	[−131.10, −96.94]

3.2.2. Actors self-handicap when they are very incompetent or very competent

In the actor block (Block 2), participants decided whether to self-handicap based on their competence and their goal. In Experiment 1, the actors' goal was to achieve the highest competence evaluations. Based on the naive observer's inferences, actors should choose not to self-handicap when the handicap would affect the outcome—that is, if the actors would pass without self-handicapping but fail after self-handicapping. On the other hand, due to the effect of the handicap, the actors should choose to self-handicap when the handicap would not affect the outcome—that is, if the actors would or would not have at least 8 correct responses regardless of whether they were evaluated on 10 or 20 of their answers. This was indeed what we saw in the data from the actor block. Fig. 4(a) shows that participants were more likely to self-handicap when they were very incompetent and could not pass even with their full ability (accuracy between 0% and 30%) or when they were very competent and could still pass even with the handicap (accuracy between 80% and 100%), but not when they were moderately competent and self-handicapping could affect the outcome (accuracy between 40% and 70%). To test for this non-monotonic effect, we fit a Bayesian mixed-effects model predicting participants' probability of self-handicapping with the actors' quadratic accuracy, accuracy, and intercept, with random intercept and slopes for each regressor grouped by participants. As predicted, we found a credible positive effect for the quadratic term (see Table 3).

In Experiment 2, the actors' goal was to maximize their chances of succeeding. Therefore, we predicted that participants should avoid self-handicapping because the handicap would hinder their chances of passing. As predicted, participants were less likely to self-handicap for almost all competence levels except for when accuracy was 100% and the probability of self-handicapping was around chance (M = 50.62%, SD = 41.15%; Fig. 4(b)). This provided a stark contrast to Experiment 1, where participants were more likely to self-handicap, particularly when an actor's accuracy was low. Notably, the curve was still U-shaped (credible quadratic effect; see Table 3) presumably because, when the actors were very competent or very incompetent, both actions (self-handicapping or not) produced similar expected values and similar choice probability according to Eq. (6), therefore it mattered less whether the actors self-handicapped. As shown in Fig. 4, the model captured the patterns in both experiments.

By contrast, both alternative models predicted almost the same γ across different actor accuracies (see Figures S1 and S2 in the Supplement). The magnitude shift from Experiment 1 to Experiment 2 was due to the inverse temperature parameter being fitted to participants' data (see Tables S1 and S2 for the fitted parameter values). Neither of them was able to capture the U-shaped curves observed in the behavioral data.

3.2.3. Self-handicapping is less effective with sophisticated observers when the actor fails

How is self-handicapping perceived by observers who know that actors were able to choose whether to self-handicap? Sophisticated

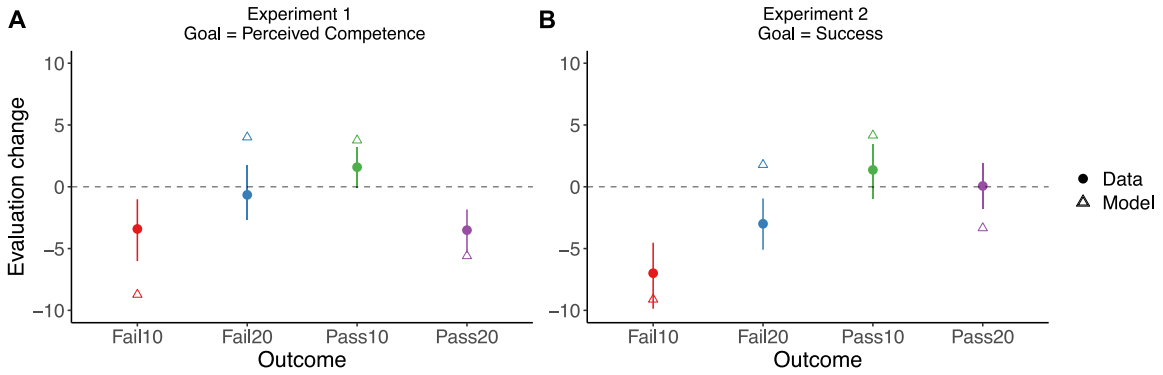

Fig. 5. Changes in observer evaluations of actors' competence (sophisticated minus naive). Notably, in both Experiment 1 (A) and Experiment 2 (B), the "Fail10" actor was rated less competent by sophisticated observers. The model captures this pattern. Error bars indicate bootstrapped 95% confidence intervals.

observers recognize that the actors' choices provide information about their competence even before observing the outcome. By thinking about the actors' decision, they realize that, in both Experiments 1 and 2, an actor who chooses to self-handicap is more likely to be either very incompetent or very competent, whereas an actor who chooses not to self-handicap is more likely somewhere in between the two extremes (see Fig. 4). This realization should in turn affect sophisticated observers' perception of the actors' competence; in particular, for actors who failed, self-handicapping is more likely a strategy used to discount the failure and they should be perceived as less competent than previously thought by a naive observer.

As predicted, Fig. 5 shows that sophisticated observers in both experiments provided lower evaluations of the "Fail10" actor's competence than the naive observers ($M = -3.42$, $SD = 18.21$ in Experiment 1, $M = -7.00$, $SD = 18.22$ in Experiment 2). This effect was confirmed by a Bayesian mixed-effects model regressing evaluations of the "Fail10" actor's competence on observer type (naive or sophisticated) and intercept, with random intercept grouped by participants[2]. We found a credible negative effect of observer type in both experiments (see Table 4), meaning that sophisticated observers' evaluations of actors who self-handicapped and failed were lower than naive observers'. Additionally, we conducted a divergence analysis on observers' belief change, following Alon et al. (2023). This analysis was done with model predictions only, because we did not collect data on participants' belief distributions. Specifically, we computed the Kullback–Leibler (KL) divergence between naive and sophisticated observers' posterior distributions over each actor's competence $D_{\mathrm{KL}}\left(P_{\mathrm{naive}}(c|s,\gamma) \parallel P_{\mathrm{sophisticated}}(c|s,\gamma)\right)$. Because KL-divergence is an asymmetric measure, we also computed $D_{\mathrm{KL}}\left(P_{\mathrm{sophisticated}}(c|s,\gamma) \parallel P_{\mathrm{naive}}(c|s,\gamma)\right)$ and averaged across the two directed divergences. Larger divergence indicates larger belief change. As shown in Fig. 6, the observers had much higher KL-divergence towards the "Fail10" actor than the other actors, suggesting that the "Fail10" actor raised much suspicion among sophisticated observers who knew that the actor could choose whether to self-handicap. We include the regression outputs for the other three actors in the Supplement.

Remember that the alternative models do not have parameters for the success likelihood function (Eq. (2)), therefore there are no fitted parameters for k and b. However, regardless of the values of k and b, $P(c|\gamma)$ would be a uniform distribution because the actor would choose the same γ regardless of their competence. As a result, sophisticated observers would not gain any information from knowing that the actor intentionally chose whether to self-handicap. This means they would not change their evaluations, in contrast to what we see in the behavioral data.

[2] Due to convergence issues, here we deviated slightly from our preregistration and did not include the random slope for observer type grouped by participants.

Table 4
Estimates of a Bayesian mixed effects regression that was fit for the following model: `Evaluation of Fail10 actor ~ 1 + Observer type + (1 | Participant)`. `Observer type` was coded as 0 = 'naive', and 1 = 'sophisticated'. 'Estimate' shows the mean of the posterior distribution, 'Est. Error' the standard deviation of the posterior distribution, and '95% CrI' the 95% credible interval.

	Estimate	Est. Error	95% CrI
Experiment 1			
(Intercept)	34.85	1.48	[31.99, 37.72]
Observer type	−3.41	1.30	[−6.01, −0.79]
Experiment 2			
(Intercept)	35.96	1.40	[33.17, 38.72]
Observer type	−7.00	1.36	[−9.67, −4.31]

4. Modeling past experiments

In this section, we demonstrate that the signaling theory of self-handicapping generalizes to past work, which used different setups and methods. We emphasize the theory's ability to qualitatively capture the patterns found in past work, predicting changes in patterns across experimental conditions, rather than evaluating its quantitative fit. Because these papers had very different setups and measurements, we had to make assumptions in order to generate theoretical predictions, which we spell out below. Another issue we had to deal with was lack of data—the studies only presented participants' mean responses. We therefore hand-tuned the parameters to bring the predictions quantitatively closer to the data. Note that information from the papers is too sparse to strongly constrain the parameter values, but we know that the parameter values have to be in a certain range to produce certain outcomes. For example, to achieve 75 points out of 100 as in Luginbuhl and Palmer (1991), γ has to be relatively large; by way of illustration, it would be unlikely for even the most competent person ($c = 100$) to get at least 75 points if their effective competence is only 20 ($\gamma = 0.2$). However, the same qualitative patterns arise for a broad range of the possible parameter values.

4.1. Method

4.1.1. Study selection

We reviewed the 168 papers cited in Török et al. (2018), the latest comprehensive survey of the self-handicapping literature. We selected studies that: (a) investigated self-handicapping empirically; (b) involved actual instances or descriptions of self-handicapping rather than questionnaires that measured the tendency to self-handicap; and (c) provided the necessary data for the model to generate predictions. This meant that studies where participants were actors needed to include a direct measure of competence or a related proxy, individually or by competence group, and a categorical outcome that

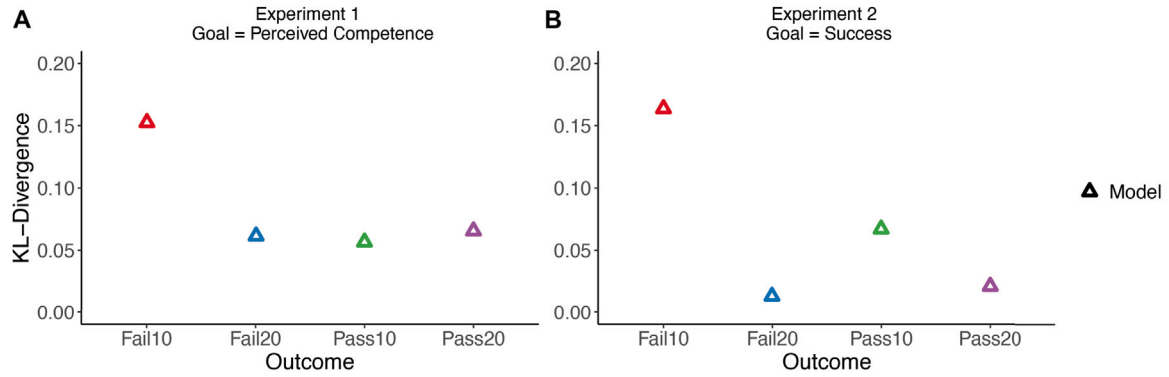

Fig. 6. KL-divergence between naive and sophisticated observers' posterior distributions over actors' competence. In both Experiment 1 (A) and Experiment 2 (B), the "Fail10" actor raised the most suspicion from the sophisticated observer who knew that the actor could choose to self-handicap.

reflected actual competence rather than fabricated feedback (e.g., non-contingent success on unsolvable tasks). Studies where participants were observers needed to report competence judgments for each combination of outcome and handicap condition, instead of separately by each factor (e.g., an average competence judgment of all actors who failed, whether or not they self-handicapped). This procedure yielded four studies: Luginbuhl and Palmer (1991), Rhodewalt et al. (1995), Tice (1991) and Tice and Baumeister (1990). We modeled the first three, but excluded (Tice & Baumeister, 1990) because it was an earlier version of Tice (1991) that manipulated fewer variables and had less information to constrain the model parameters.

4.1.2. Model fitting

Table 1 summarizes the parameter values for modeling the studies. When possible, we used the ranges and values specified in the papers. For parameters we could not infer from the papers, we hand-tuned them in ranges that would make the experimental conditions possible. We assumed a uniform prior competence distribution for consistency. For Luginbuhl and Palmer (1991), we set $c \in [0, 100]$ as the points an actor could get with their full capacity, task difficulty $d = 95$, 75, and 55 as indicated in the paper, and $\gamma \in \{0.8, 1\}$ so that the corresponding outcomes were possible based on the participants' evaluations. For the modeling of Tice (1991), the empirical handicap factor was calculated by dividing the raw data (number of seconds practiced and level of distraction from a tape) by the maximum response in each study. Since the average was 0.8 and it was a weighted combination of the different values in the handicap factor space, we set $\gamma \in \{0.6, 1\}$ so that it aligned with the empirical data. We also set $c \in [0, 10]$ for simplicity. Because it was unclear how competent participants were, we had to make some assumptions about this and the task difficulty. We assumed that the ground truth was $c = 2$ for incompetent participants and $c = 9$ for highly competent participants. Task difficulty d was set to 3 so that it was too hard for incompetent participants and easy for highly competent participants even with the handicap. w was set to 1 according to the traditional definition of self-handicapping, meaning that the actors' only goal was to maximize evaluations. For Rhodewalt et al. (1995), since the original competence range was [−5, 5], we increased all the ratings by 5 to ensure $c \geq 0$. Thus, $c \in [0, 10]$. The paper labeled the middle of the scale as "can't tell (the competence)", so we used that as the task difficulty (i.e., $d = 5$). The excuse statements such as "I can hardly keep my eyes open" suggested that the handicap heavily affected the actors, thus we set $\gamma \in \{0.5, 1\}$ so that it was hard for self-handicappers to succeed. This study did not tell observers the outcome, so the dependent variable was the expectation of competence ($\mathbb{E}[c]$ or $\mathbb{E}[c|\gamma]$, depending on the condition). Again, w was set to 1. For all three studies, the two free parameters for the logistic function (k and b) and the inverse temperature parameter were hand-tuned to bring model predictions quantitatively closer to the data.

4.2. Results

The results of the three studies fit with our experimental results: Luginbuhl and Palmer (1991) captures how naive observers perceive actors' competence; Tice (1991) captures how actors make strategic choices; and Rhodewalt et al. (1995) captures how sophisticated observers view the actors.

4.2.1. Naive observers' evaluations

In two experiments, Luginbuhl and Palmer (1991) showed participants videotapes of an actor the night before an exam; a friend repeatedly asked the actor to go to a movie that night, with the actor repeatedly saying no and reiterating that he had to study. In the self-handicapping condition, participants additionally watched a second clip showing that, after the friend asked once more, the actor agreed to go. The actor's reluctance to go to the movie makes it plausible that participants did not consider the actor's eventual decision to go to the movie as a strategic choice aimed at maximizing competence evaluations; thus this study captures how a naive observer would perceive an actor who did not choose whether to self-handicap.

As indicated in Luginbuhl and Palmer (1991), we translated "receiving Grade A/C/F" to a success rule of getting at least 95/75/55 points out of 100. We used the "predicted future test score" as a proxy for competence. The aggregated data from both experiments and model predictions are shown in Fig. 7(a). Both participants and the model evaluated the actor who received higher grades as more competent, and evaluated the self-handicapper as slightly more competent than the non-self-handicapper when they received the same grade.

4.2.2. Actors' strategy

In two experiments, Tice (1991) studied how trait self-esteem (which we treat as a proxy of competence, an implicit assumption made by Tice, 1991) affected how much participants self-handicapped (note that only Study 1 and Study 2 in Tice, 1991 were actual instances of self-handicapping; the remaining studies measured self-handicapping tendencies with questionnaires). The response variables were number of seconds practiced before an important evaluation (Study 1) and level of distraction from a tape during a test (Study 2), which we divided by the maximum response in each study to convert them to a percentage. Each study manipulated whether failure or success was meaningful. "Failure meaningful" meant that only failure would reveal information about participants' competence, and "success meaningful" meant that only success would reveal information about participants' competence. This was modeled by equating $\mathbb{E}[c|s = 1, \gamma]$ with the prior $\mathbb{E}[c]$ in the "failure meaningful" condition and $\mathbb{E}[c|s = 0, \gamma] = \mathbb{E}[c]$ in the "success meaningful" condition. In other words, no information about competence is gained if the outcome was not meaningful.

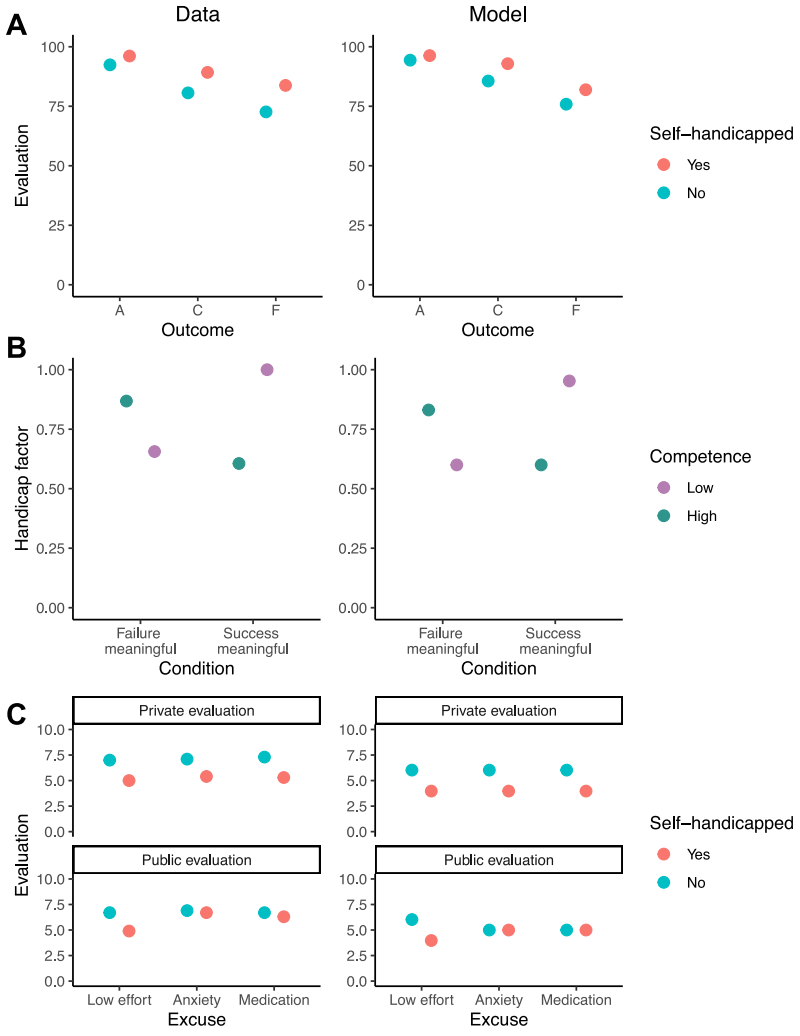

Fig. 7. Modeling results of previous studies. Only participants-averaged responses were available. For illustration purposes, data were aggregated across experiments. (A) Experiment 1 and Experiment 2 in Luginbuhl and Palmer (1991), where participants evaluated the competence of actors who self-handicapped or not and received a grade of either A, C, or F, corresponding to at least 95, 75, or 55 points. (B) Study 1 and Study 2 in Tice (1991), where participants decided how much to self-handicap (amount of practice before important evaluation and level of distraction from a tape during a test) when failure or success was meaningful. (C) Rhodewalt et al. (1995), where participants evaluated the competence of actors who self-handicapped because of low effort, anxiety, or medication.

We computed the predicted handicap factors and compared them to how much participants actually handicapped themselves. Aggregated data from the two studies in Fig. 7(b) show that how much participants handicapped depended on which outcome was meaningful; low competence led to more handicapping to avoid failure when failure was meaningful, whereas high competence led to more handicapping to enhance success when success was meaningful. The theory captures this pattern.

4.2.3. Sophisticated observers' evaluations

Rhodewalt et al. (1995) asked participants to evaluate the competence of actors based on their cartoon captions on a scale of −5 to 5. To apply the theory, we shifted the ratings to a 0–10 scale. Participants listened to a 3-min audiotape allegedly from a non-self-handicapping actor and a self-handicapping actor. The non-self-handicapping actor expressed 10 generic statements about the task (e.g., "These are interesting"), whereas the self-handicapping actor expressed 7 generic statements plus 3 condition-specific statements, either about not trying hard (low effort condition), or performance being affected by anxiety (anxiety condition), or medication making them very sleepy (medication condition). Rhodewalt et al. (1995) asked participants to listen to the tape and categorize the statements by content, indicating how many times each actor mentioned anxiety, medication, etc. Participants also knew that the actor knew that their statements were being recorded and would be listened to. These features of the setup – participants carefully thinking about the actors' excuses and knowing that the actor knows that the excuses would be considered – gave us confidence that participants were thinking about the possibility of actors using strategic excuses to maximize their competence perceptions; thus this study captures how a sophisticated observer would evaluate an actor's competence.

The study additionally manipulated whether the evaluation was public. In the *private* evaluation condition, participants were told that the actors would never see their evaluations. In the *public* evaluation condition, participants were told that they would meet the actors in person and explain their evaluation to the actors. We assumed that participants more carefully considered the actors' excuses in the public evaluation condition since the judgments mattered more.

The validity of the excuses depended on their perceived controllability. As Rhodewalt et al. (1995) themselves pointed out, an actor may not be able to control their anxiety or medication, but they can control how much effort they exert. Following this assumption, the model

predicted that participants were naive observers in the public evaluation condition when actors' gave less controllable excuses (anxiety and medication) and maintained their prior belief about self-handicappers' competence, which was the mean of a uniform distribution. By contrast, in all other situations, participants were sophisticated observers—updating their belief about the self-handicappers' competence from their choice of whether to self-handicap as in Eq. (8). Even though theoretically, knowing that someone intentionally self-handicapped by itself reveals that they could be either very incompetent or very competent, we inferred from the experiment – which labeled average competence ($c = 5$) as "can't tell (the competence)" – that participants were evaluating the actors relative to a task difficulty of 5. The excuse statements such as "I can hardly keep my eyes open" suggested that the handicap heavily affected the actors. Since a severe handicap would make it unlikely for even the most competent actor to succeed at a task that is difficult for an average person, participants were more likely to infer that the self-handicapping actors were at the lower end of the spectrum. As shown in Fig. 7(c), both participants and the model gave lower evaluations to the self-handicapper in the private evaluation condition and in the public evaluation condition when the excuses were more controllable (low effort), whereas the two actors were judged similarly in the public evaluation condition when the excuses were less controllable (anxiety and medication).

5. General discussion

We presented a signaling theory of self-handicapping that explains when people self-handicap and how that is perceived by naive and sophisticated observers. We showed that this theory generates predictions in line with behavioral data and is capable of explaining existing results in the literature. Specifically, we found that self-handicappers were perceived as more competent than non-self-handicappers for the same outcome and, relatedly, actors were more likely to self-handicap when the handicap would not affect the outcome. However, sophisticated observers who reasoned about actors' decision whether to self-handicap perceived them as less competent when they failed the task, compared to naive observers. These patterns were predicted by the theory.

Previous accounts of self-handicapping have largely been informal, explaining the phenomenon as actors capitalizing on discounting and augmentation principles, without specifying exactly when these principles apply and how they determine the actors' behavior and the observers' inferences. As the first formal, normative account of self-handicapping, the signaling theory formalizes why these intuitions emerge, when they hold, and how they interact: it makes precise predictions about how an actor's competence affects their self-handicapping choices and how the self-handicapping behavior affects observers' perception of the actors' competence. For example, the model predicts non-monotonic patterns of self-handicapping and evaluation—patterns that are not obvious from existing informal theories, but emerge naturally from the interaction between competence, outcome likelihood, and social inference. Moreover, the theory does not rely on assumptions typically made in past work, such as assuming that there are trait-level differences in self-handicapping, or that self-handicapping coincides with reducing effort. We showed that the theory explained when and why people self-handicap when only situational factors were considered and when self-handicapping did not reduce effort.

The theory draws a distinction between naive and sophisticated observers, which potentially reconciles a debate in the literature on the effectiveness of self-handicapping. Past work has been inconclusive regarding how observers perceive self-handicappers' competence. While some studies showed that self-handicapping increases perception of competence (Luginbuhl & Palmer, 1991), some showed the opposite (Rhodewalt et al., 1995). Self (1990) pointed out that, in order for self-handicapping to be effective, the handicapper should not appear to desire factors that hinder their performance, and that self-handicapping faces the disapproval of perceivers who detect the use of this strategy (see also Hirt et al., 2003). The signaling theory formalizes these ideas and provides an intuitive explanation for when and why self-handicapping is effective in managing perceived competence. The differences in naive and sophisticated observers' perceptions might be numerically small in our experiments, but the signaling theory is able to capture larger differences, as evidenced by the modeling results of previous studies.

More broadly, our work adds to a growing literature on observer-aware behaviors. Recent work in AI and cognitive modeling highlights how agents that are aware of being observed adapt their behavior to influence others' inferences (Miura & Zilberstein, 2024). Notably, the ability to act strategically to manipulate others' perception – and to reason about others doing the same – has also been documented in Western Scrub Jays (Clayton et al., 2007). An open question is to what extent this type of strategic reasoning is specific to human adults, and how it develops. The recursive nature of the task may place demands on social-cognitive abilities that emerge only later in development. Future work could explore whether children, adolescents, or even nonhuman primates engage in similar forms of anticipatory impression management, and under what conditions they fail to infer the motives behind others' self-handicapping decisions.

The theory can be extended to deal with additional sources of uncertainty. For simplicity, we assumed that actors had perfect knowledge about their competence. While past work has shown that people can form representations of their own competence given limited amounts of data (Leonard et al., 2020; Nicholls & Miller, 1984), they might still be somewhat uncertain, and this may affect when they use these strategies (the inflection points in Fig. 4(a)). We also told observers how exactly the handicap affected the actor (i.e., the handicap factor γ), which might not always be observable, and could be confounded with effort, which has been modeled in a similar way in previous work (Xiang, Landy et al., 2023; Xiang et al., 2025, 2023, 2024). An open question is how self-handicapping relates to the ability-effort trade-off (i.e., for a given level of success, greater attribution to ability would occur when effort is lower Heider, 1958).

We assumed that the actor only had two possible goals (or some weighted combination of the two)—maximizing competence perception (which we tested in Experiment 1) and maximizing chances of success (tested in Experiment 2). People may have other goals, such as deceiving oneself in order to maintain a positive self-image (Quattrone & Tversky, 1984), or obtaining reliable diagnostic feedback about one's competence (Festinger, 1954). Past work has also shown that, while self-handicapping helps with competence evaluations, it can make actors seem less reliable or favorable (Hirt et al., 2003; Levesque et al., 2001; Luginbuhl & Palmer, 1991). How people make these trade-offs should be studied in future work.

The current work empirically tested how self-handicapping behaviors are perceived by third-party observers—how observers infer competence based on an actor's self-handicapping decisions and outcomes. This perspective is relatively underexplored, as most prior work has examined self-handicapping from the actor's perspective. The current work also explored the actor's behavior in the critical pre-performance stage—deciding whether to introduce a handicap that makes success less likely, where the "self" was engaged at the motivational level. It remains an important task for future work to examine self-handicapping from a truly first-person perspective, where participants generate their own performance that is evaluated by others.

Finally, our findings have several implications for designing better interventions against academic self-handicapping. We showed that situational factors alone can explain why students might choose to self-handicap, without having to postulate individual differences in the tendency to self-handicap. We also showed that, even when the goal is to maximize chances of success (Experiment 2)—instead of observers' impressions of competence (Experiment 1)—self-handicapping is more likely to occur in certain situations than others. When the goal was to maximize chances of success, participants overall preferred not to

self-handicap. However, it is possible that people would nonetheless choose do so in certain situations. For example, when self-handicapping reduces effort (e.g., practicing less) or produces pleasurable feelings (e.g., using substances and drinking alcohol), students may be more willing to self-handicap. Inspired by our findings, one possible solution is for teachers to provide students with tasks that are just right for them. As shown in Fig. 4, regardless of the actor's goal, self-handicapping is least likely when the task is close to the actor's competence and would thus affect the outcome. Figuring out the right task difficulty would require teachers to reason about the students' competence and assign tasks accordingly (Chen et al., 2024; Shafto et al., 2014). Teachers may also shift the focus from vertical comparisons (i.e., comparing the performance of different students within a single period) to horizontal comparisons (i.e., comparing each student's performances over time). Past work has shown that evaluative threat induces self-handicapping (Hirt et al., 2000; Snyder et al., 1985; Stone, 2002; Suhr & Wei, 2013; Tandler et al., 2014). This change would emphasize the goal of learning (Xiang & Gershman, 2025) rather than appearing competent or succeeding at a task. The present work provides a theoretical and empirical basis for new interventions that target self-handicapping, helping people to realize their full potential.

CRediT authorship contribution statement

Yang Xiang: Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Samuel J. Gershman:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Tobias Gerstenberg:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Acknowledgments

We thank Thomas Icard, Arnav Verma, Adani Abutto, and Yiqiao Wang for helpful discussions. This research was funded by National Science Foundation, United States of America (DRL-2024462) to S.J.G.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106288.

Data availability

All data, code, and materials are publicly available at https://github.com/yyyxiang/self-handicapping.

References

Alon, N., Schulz, L., Rosenschein, J. S., & Dayan, P. (2023). A (dis-) information theory of revealed and unrevealed preferences: Emerging deception and skepticism via theory of mind. *Open Mind*, *7*, 608–624.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*(3), 272.

Arkin, R. M., & Baumgardner, A. H. (1985). Attribution. Basic issues and applications. In *Self-handicapping*. Orlando, FL: Academic Press.

Arkin, R. M., & Oleson, K. C. (1998). Attribution and social interaction: The legacy of Edward E. Jones. In *Self-handicapping*. Washington, DC: American Psychological Association.

Beck, B. L., Koons, S. R., & Milgrim, D. L. (2000). Correlates and consequences of behavioral procrastination: The effects of academic procrastination, self-consciousness, self-esteem and self-handicapping. *Journal of Social Behavior and Personality*, *15*(5), 3.

Beller, A., & Gerstenberg, T. (2025). Causation, meaning, and communication. *Psychological Review*.

Berglas, S., & Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology*, *36*(4), 405.

Bobo, J. L., Whitaker, K. C., & Strunk, K. K. (2013). Personality and student self-handicapping: A cross-validated regression approach. *Personality and Individual Differences*, *55*(5), 619–621.

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology*, *36*(1), 56.

Chen, A. M., Palacci, A., Vélez, N., Hawkins, R. D., & Gershman, S. J. (2024). A hierarchical Bayesian model of adaptive teaching. *Cognitive Science*, *48*(7), Article e13477.

Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *362*(1480), 507–522.

Covington, M. V. (1992). *Making the grade: A self-worth perspective on motivation and school reform*, Cambridge University Press.

Elliot, A. J., Cury, F., Fryer, J. W., & Huguet, P. (2006). Achievement goals, self-handicapping, and performance attainment: A mediational analysis. *Journal of Sport and Exercise Psychology*, *28*(3), 344–361.

Eronen, S., Nurmi, J.-E., & Salmela-Aro, K. (1998). Optimistic, defensive-pessimistic, impulsive and self-handicapping strategies in university environments. *Learning and Instruction*, *8*(2), 159–177.

Ferrari, J. R., & Tice, D. M. (2000). Procrastination as a self-handicap for men and women: A task-avoidance strategy in a laboratory setting. *Journal of Research in Personality*, *34*(1), 73–83.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140.

Gadbois, S. A., & Sturgeon, R. D. (2011). Academic self-handicapping: Relationships with learning specific and general self-perceptions and academic performance over time. *British Journal of Educational Psychology*, *81*(2), 207–222.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Greenberg, J. (1985). Unattainable goal choice as a self-handicapping strategy. *Journal of Applied Social Psychology*, *15*(2), 140–152.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Heider, F. (1958). *The psychology of interpersonal relations*, New YorkL Wiley.

Higgins, R. L., & Harris, R. N. (1988). Strategic "alcohol" use: Drinking to self-handicap. *Journal of Social and Clinical Psychology*, *6*(2), 191–202.

Hirt, E. R., McCrea, S. M., & Boris, H. I. (2003). "I know you self-handicapped last exam": Gender differences in reactions to self-handicapping. *Journal of Personality and Social Psychology*, *84*(1), 177.

Hirt, E. R., McCrea, S. M., & Kimble, C. E. (2000). Public self-focus and sex differences in behavioral self-handicapping: Does increasing self-threat still make it "just a man's game?". *Personality and Social Psychology Bulletin*, *26*(9), 1131–1141.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. *vol. 2*, In *Advances in experimental social psychology* (pp. 219–266). Elsevier.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107.

Leonard, J. A., Sandler, J., Nerenberg, A., Rubio, A., Schulz, L., & Mackey, A. (2020). Preschoolers are sensitive to their performance over time. In *CogSci*.

Levesque, M. J., Lowe, C. A., & Mendenhall, C. (2001). Self-handicapping as a method of self-presentation: An analysis of costs and benefits. *Current Research in Social Psychology*, *6*(15), 1–13.

Luginbuhl, J., & Palmer, R. (1991). Impression management aspects of self-handicapping: Positive and negative effects. *Personality and Social Psychology Bulletin*, *17*(6), 655–662.

Miura, S., & Zilberstein, S. (2024). Observer-aware planning with implicit and explicit communication. In *Proceedings of the 23rd international conference on autonomous agents and multiagent systems* (pp. 1409–1417).

Nicholls, J. G., & Miller, A. T. (1984). Reasoning about the ability of self and others: A developmental study. *Child Development*, 1990–1999.

Nurmi, J.-E., Onatsu, T., & Haavisto, T. (1995). Underachievers' cognitive and behavioural strategies-self-handicapping at school. *Contemporary Educational Psychology*, *20*(2), 188–200.

Prapavessis, H., & Grove, J. R. (1998). Self-handicapping and self-esteem. *Journal of Applied Sport Psychology*, *10*(2), 175–184.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, *46*(2), 237.

Rhodewalt, F., & Davison, J., Jr. (1986). Self-handicapping and subsequent performance: Role of outcome valence and attributional certainty. *Basic and Applied Social Psychology*, *7*(4), 307–322.

Rhodewalt, F., Sanbonmatsu, D. M., Tschanz, B., Feick, D. L., & Waller, A. (1995). Self-handicapping and interpersonal trade-offs: The effects of claimed self-handicaps on observers' performance evaluations and feedback. *Personality and Social Psychology Bulletin*, *21*(10), 1042–1050.

Rhodewalt, F., & Tragakis, M. W. (2002). Self-handicapping and school: Academic self-concept and self-protective behavior. In *Improving academic achievement* (pp. 109–134). Elsevier.

Ross, S. R., Canada, K. E., & Rausch, M. K. (2002). Self-handicapping and the five factor model of personality: Mediation between neuroticism and conscientiousness. *Personality and Individual Differences*, *32*(7), 1173–1184.

Schwinger, M., Wirthwein, L., Lemmer, G., & Steinmayr, R. (2014). Academic self-handicapping and achievement: A meta-analysis. *Journal of Educational Psychology*, *106*(3), 744.

Self, E. A. (1990). Situational influences on self-handicapping. In *Self-handicapping: the paradox that isn't.* Plenum Press.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shepperd, J. A., & Arkin, R. M. (1989). Determinants of self-handicapping: Task importance and the effects of preexisting handicaps on self-generated handicaps. *Personality and Social Psychology Bulletin*, *15*(1), 101–112.

Smith, J. L., Hardy, T., & Arkin, R. (2009). When practice doesn't make perfect: Effort expenditure as an active behavioral self-handicapping strategy. *Journal of Research in Personality*, *43*(1), 95–98.

Smith, D. S., & Strube, M. J. (1991). Self-protective tendencies as moderators of self-handicapping impressions. *Basic and Applied Social Psychology*, *12*(1), 63–80.

Snyder, C. R., & Higgins, R. L. (1988). Excuses: Their effective role in the negotiation of reality. *Psychological Bulletin*, *104*(1), 23.

Snyder, C., et al. (1985). Adler's psychology (of use) today: Personal history of traumatic life events as a self-handicapping strategy. *Journal of Personality and Social Psychology*, *48*(6), 1512.

Stone, J. (2002). Battling doubt by avoiding practice: The effects of stereotype threat on self-handicapping in white athletes. *Personality and Social Psychology Bulletin*, *28*(12), 1667–1678.

Suhr, J., & Wei, C. (2013). Symptoms as an excuse: Attention deficit/hyperactivity disorder symptom reporting as an excuse for cognitive test performance in the context of evaluative threat. *Journal of Social and Clinical Psychology*, *32*(7), 753–769.

Tandler, S., Schwinger, M., Kaminski, K., & Stiensmeier-Pelster, J. (2014). Self-affirmation buffers claimed self-handicapping? A test of contextual and individual moderators. *Psychology*, *2014*.

Tesser, A. (1988). *Advances in experimental social psychology, Toward a self-evaluation maintenance model of social behavior*. Academic Press.

Thompson, T., & Richardson, A. (2001). Self-handicapping status, claimed self-handicaps and reduced practice effort following success and failure feedback. *British Journal of Educational Psychology*, *71*(1), 151–170.

Tice, D. M. (1991). Esteem protection or enhancement? Self-handicapping motives and attributions differ by trait self-esteem. *Journal of Personality and Social Psychology*, *60*(5), 711.

Tice, D. M., & Baumeister, R. F. (1990). Self-esteem, self-handicapping, and self-presentation: The strategy of inadequate practice. *Journal of Personality*, *58*(2), 443–464.

Török, L., Szabó, Z. P., & Tóth, L. (2018). A critical review of the literature on academic self-handicapping: Theory, manifestations, prevention and measurement. *Social Psychology of Education*, *21*, 1175–1202.

Tucker, J. A., Vuchinich, R. E., & Sobell, M. B. (1981). Alcohol consumption as a self-handicapping strategy. *Journal of Abnormal Psychology*, *90*(3), 220.

Urdan, T. (2004). Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *Journal of Educational Psychology*, *96*(2), 251.

Urdan, T., Midgley, C., & Anderman, E. M. (1998). The role of classroom goal structure in students' use of self-handicapping strategies. *American Educational Research Journal*, *35*(1), 101–122.

Walker, D. W., Smith, K. A., & Vul, E. (2015). The" Fundamental Attribution Error" is rational in an uncertain world. *vol. 37*, In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychological Review*, *66*(5), 297.

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin*, *90*(2), 245.

Xiang, Y., & Gershman, S. (2025). Modeling intrinsic motivation as reflective planning. *vol. 47*, In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, Article 105609.

Xiang, Y., Landy, J., Cushman, F., Vélez, N., & Gershman, S. J. (2025). People reward others based on their willingness to exert effort. *Journal of Experimental Social Psychology*, *116*, Article 104699.

Xiang, Y., Vélez, N., & Gershman, S. J. (2023). Collaborative decision making is grounded in representations of other people's competence and effort. *Journal of Experimental Psychology: General*, *152*(6), 1565.

Xiang, Y., Vélez, N., & Gershman, S. J. (2024). Optimizing competence in the service of collaboration. *Cognitive Psychology*, *150*, Article 101653.

Zuckerman, M., Kieffer, S. C., & Knee, C. R. (1998). Consequences of self-handicapping: Effects on coping, academic performance, and adjustment. *Journal of Personality and Social Psychology*, *74*(6), 1619.

Zuckerman, M., & Tsai, F.-F. (2005). Costs of self-handicapping. *Journal of Personality*, *73*(2), 411–442.