ESSAYS ON REPORTING SYSTEMS, OPERATIONAL DISTORTION, AND BELIEF DISTORTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jeremiah Wayne Bentley

February 2016

ESSAYS ON REPORTING SYSTEMS, OPERATIONAL DISTORTION, AND BELIEF DISTORTION

Jeremiah Wayne Bentley, Ph. D.

Cornell University 2016

This dissertation is a combination of three relatively independent sole-authored papers. The first presents evidence that individuals distort their behaviors and their beliefs in response to performance systems that focus on objective measures of performance. They distort their behaviors and beliefs less when they are allowed to provide unverifiable narrative reports of their performance. I also provide evidence that the reason that they distort their beliefs is to make their beliefs be consistent with their distorted actions. The second essay presents evidence that unselfish, pro-firm actions lead to subsequent operational distortion (i.e. making operational decisions intended to increase a measure of performance rather than the underlying performance that the measure is intended to capture). I also present evidence that these pro-firm actions lead to more operational distortion if the operational distortion occurs in the same division of the firm as the initial pro-firm actions. Both of these essays present the results of experiments using Amazon Mechanical Turk (AMT) workers as participants. My final essay discusses four challenges that arise when conducting research with AMT and how researchers can deal with these challenges.

BIOGRAPHICAL SKETCH

Jeremiah Wayne Bentley completed his PhD at Cornell University in 2015. His primary focus of study was Accounting with minors in Social Psychology and Applied Statistics. He earned his Master of Accountancy degree (Professional Accountancy emphasis) from Brigham Young University in 2010. He earned his Bachelor of Science degree from Brigham Young University in 2010, majoring in Accounting with minors in Economics and Information Technology. Jeremiah is married to Linda Bentley and they have three children: James, Timothy, and Adelaide. He has accepted a position as an assistant professor at the University of Massachusetts – Amherst.

Dedicated to my darling wife Linda, who has supported me through thick and thin, to my children JJ, Timmy, and Adelaide, who always know how to make me smile, and to my parents, who taught me to analyze the world around me.

ACKNOWLEDGMENTS

I thank my dissertation committee members: Rob Bloomfield (Chair), John Bunge (Applied Statistics), Melissa Ferguson (Social Psychology), and Mark Nelson (Accounting) for their valuable guidance and advice. For their helpful comments on chapter two of this dissertation I thank Chris Agoglia, Linda Bentley, Steve Bentley, Amy Donnelly, Jeremy Douthit, Mike Durney, Scott Emett, Matthew Hayes, Steve Kachelmeier, Lisa Koonce, Volker Laux, Zheng Leitter, Bob Libby, Eldar Maksymov, Ken Merkley, Lillian Mills, Joe Pacelli, Jeff Pickerd, Kristi Rennekamp, Todd Thornock, Brian White, and workshop participants at Arizona State University, Cornell University, the University of Central Florida, the University of Massachusetts at Amherst, and the University of Texas at Austin. For their helpful comments on chapter three of this dissertation I thank Linda Bentley, Steve Bentley, Mike Durney, and Scott Emett.

I gratefully acknowledge financial support from the Samuel Curtis Johnson Graduate School of Management at Cornell University and the Deloitte Foundation. Any errors in this dissertation are my own.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND RESEARCH STATEMENT

Broadly speaking, my research examines how accounting and reporting systems lead individuals to distort their beliefs, actions, and communication. In this chapter I briefly discuss how four of my papers (two that are included as part of this dissertation and two that are not part of my dissertation) fit together in this research stream.

Bentley, Bloomfield, Davidai, and Ferguson (2015), which is not included in this dissertation, finds that (1) unstructured meetings allow users to detect reporters' beliefs and (2) a persuasion goal causes reporters to alter their beliefs and thereby become more-effective persuaders. These findings provide two practical suggestions for accounting systems: (1) accounting reports should be accompanied by face-to-face interactions such as in-person performance evaluations, conference calls, and testimonies in order to improve the degree to which users can differentiate between a reporters' cheap talk and sincere beliefs and (2) operational decisions will likely be more biased if decision-makers are also involved in persuasive activities than if decision-makers are not involved in persuasive activities because in an effort to persuade, decision-makers 'drink their own Kool-Aid' thereby distorting their own perception of reality.

A natural follow-up question asks "How can accounting systems be designed to allow reporting while minimizing self-deception?" In chapter two of this dissertation, I explore that question. I begin by replicating prior research that shows that objective performance measures cause operational distortion and surrogation (e.g. Holmstrom and Milgrom 1991; Choi, Hecht,

and Tayler 2012),[1] where surrogation is a specific type of self-deception. I then extend the research to find that individuals distort operations less and surrogate less when they are allowed to provide narrative reports of their performance. Individuals who can explain their actions are less likely to sacrifice true performance in order to increase reported performance and are less likely to 'drink the Kool-Aid.' Thus, while Bentley, et al. (2015b) finds that narrative reporting improves users' decision-making, chapter two of my dissertation presents data suggesting that narrative reporting also improves reporters' decision-making and helps them appropriately weigh all aspects of performance, not just the measured aspects of performance.

Bentley, Bloomfield, Bloomfield, and Lambert (2015), not included in this dissertation, examines how people view the morality of two types of measure management.[2] We find that people generally believe that operational distortion is more acceptable than reporting distortion. We also find that how much an individual values Purity/Sanctity as a moral virtue is a significant predictor of their views on reporting distortion but not their views on operational distortion.

In chapter three of my dissertation I extend this research to investigate how individuals justify operational distortion. I predict and find that individuals are more likely to engage in operational distortion after they have done an out-of-the-ordinary act to benefit the firm than if a coworker performed the out-of-the-ordinary act. Mediation analysis confirms that going

[1] Operational distortion is when individuals make operational decisions with the intent of improving a measure of performance rather than improving the actual performance that the measure is intended to capture. Surrogation is the phenomenon whereby individuals come to believe that an incentivized measure of performance has a greater causal relationship with true performance than it actually has.

[2] Measure management is any action intended to increase a measure of performance rather than the actual performance that the measure is intended to capture. Reporting distortion is when the actions are reporting decisions (e.g. underestimating accruals). Operational distortion is when the actions are operational decisions (e.g. cutting R&D expenses).

the extra mile in an initial task makes participants feel like they deserve a reward and that the feeling of deservingness leads to the operational distortion. I also find that these relationships are stronger when the initial action and the opportunity to distort operations are bracketed together by having them both reported to the same supervisor than when they are bracketed separately by having them reported to different supervisors. I then propose additional research which, if successful, will provide evidence that firms can reduce operational distortion by narrowly bracketing performance.

I continue to be interested in the general topic of deception in performance reporting, especially subtle/ambiguous forms of deception such as operational distortion and self-deception. I expect my future research will continue to look at these types of questions, exploring how reporting systems can lead to or prevent distorted beliefs, actions, and communication.

REFERENCES TO CHAPTER 1

Bentley, J.W., M. Bloomfield, R.J. Bloomfield, and T. Lambert. 2015. The morals of unethical reporting. *Working Paper*

Bentley, J.W., R.J. Bloomfield, S. Davidai, and M.J. Ferguson. 2015. Drinking your own Kool-Aid: The role of beliefs, belief-revision, and meetings in persuasion. *Working Paper*.

Choi, J.W., G.W. Hecht, and W.B. Tayler. 2012. Lost in translation: The effects of incentive compensation on strategy surrogation. *The Accounting Review* 87, (4): 1135-63.

Holmstrom, B., and P. Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*: 24-52.

CHAPTER 2

DECREASING OPERATIONAL DISTORTION AND SURROGATION THROUGH NARRATIVE REPORTING

Abstract:

Prior research shows that agents who are compensated on a single objective measure of performance tend to (1) distort their operational and investment decisions in order to make that measure look better and (2) use that measure as a surrogate measure of true performance even when they are making operational and investing decisions on their own behalf. I predict and find that allowing agents to provide narrative explanations for their actions reduces both operational distortion and surrogation. In my experiment, experienced chess players make bets on 30 in-progress chess games: 10 games on their own behalf, then 10 on behalf of a boss, and finally 10 more on their own behalf. Participants also write an explanation for each bet, describing who they think will win and why.

When they are placing bets on behalf of their boss, all participants are told that their boss will subjectively allocate a bonus among four agents. Half of the participants are told that the boss will see only how their bet aligns with material count, which is an imperfect measure of who has an advantage in the game. The other half are told that their boss will see the participants' unverifiable narrative explanations for each bet in addition to the material count. I find that participants who are permitted to give narrative explanations to their bosses are more likely to make decisions that reflect all dimensions of a chess position, rather than making decisions that favor only the reported measure (material count). They are also less likely to use the reported measure as a surrogate for the true strength of the position when subsequently acting on their own behalf. Finally, I find a moderating effect of participants' preference for consistency, suggesting that surrogation is driven at least partly by a desire to make beliefs be consistent with previous opportunistic actions.

I. INTRODUCTION

> *"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."*
>
> *Campbell's Law (Campbell, 1979, p. 85)*

Managers are all too familiar with Campbell's Law: reporting systems that measure and incentivize performance typically also encourage agents to distort performance (Brown 1990, Holmstrom and Milgrom 1991, see also Prendergast 1999 for a review). In organizations, Campbell's Law arises because management cannot objectively measure and incentivize every aspect of performance. If management decides to measure and incentivize the measurable aspects of performance (e.g. realized returns) and ignore the unmeasurable aspects of performance (e.g. unrealized returns), agents can engage in *operational distortion* (Bloomfield 2014) by making decisions that increase reported performance more than true performance (e.g. by allocating limited resources towards projects that realize returns early but have lower total returns).[1] Furthermore, agents come to believe that the incentivized dimensions of performance have a greater causal relationship with true performance than they actually do, a phenomenon known as *surrogation* (Choi, Hecht, and Tayler 2012, 2013).

In this paper, I predict and find that when agents are allowed to supplement objective performance information with subjective and unverified narrative reporting, they engage in less operational distortion than when they cannot provide narrative reports. Furthermore, agents who distort operations change their beliefs to be consistent with their opportunistic behavior, especially if they have a high preference for consistency. Thus, agents who are

[1] Operational distortion is a generalization of real earnings management. Real earnings management is when managers "manipulate real activities during the year to meet certain *earnings* targets" (Roychowdhury 2006, p. 336; emphasis added). In contrast, operational distortion refers to manipulating real activities in order to add a favorable bias to *any* performance metric.

allowed to provide narrative reports of their performance engage in less operational distortion, which in turn reduces their tendency to use measured performance as a surrogate for true performance, especially if they have a high preference for consistency.

To test my theory, I have 75 experienced chess players solicited from Amazon Mechanical Turk bet on whether White or Black will win in-progress chess games played by computers. Chess is well-suited to testing my theory because supporting a bet requires a multi-dimensional analysis that allows rich narrative reporting to supplement a widely used yet imperfect measure of which side is winning: *material count*. Material count captures one dimension of performance (the numerical value of the pieces remaining for each side), but neglects the strategic dimension of the game: where the pieces are positioned. Advances in computer chess programming allow me to calculate an *engine score* which is a measure of performance that incorporates both the material and the strategic positioning dimensions of the game. In my experiment, participants place bets on 30 games that are paired based on engine score. By having games with the same engine score but with different material counts, I can measure the extent to which participants' bets are biased in favor of material count, as the rational behavior for any level of risk aversion is to place the same bet on the two games.

Participants place the 30 bets in 3 phases. In the pre-agency phase, I get a baseline measure of participants' biases by having them place 10 bets on their own behalf. In the agency phase, I have participants place 10 bets on behalf of a boss. Finally, in the post-agency phase I have participants again place 10 bets on their own behalf (different chess games in each phase, randomized between subjects). For each bet, participants see one chess board and its material count and then place a bet and write an explanation for their bet on a single screen.

My manipulation occurs in the agency phase. I tell all participants that they will place bets on behalf of a boss who will allocate a subjective bonus. The boss always sees the material count and participant's bet for each game, and the boss never sees the chess board or the outcome of the game. This game is analogous to a setting where an employee must choose how to allocate limited resources (e.g. capital) among projects that differ in their short-term and long-term returns, where short-term returns are easily observed and long-term returns are difficult to observe. I manipulate whether or not the boss sees the participant's explanation of the bet. Thus, participants in the narrative reporting condition know that the boss will see their explanation for why they place a particular bet, while participants in the narrative note-taking condition know that their boss will not see their explanation (but they are still required to write one).

As expected, I find no difference in participants' preference for consistency or pre-agency betting between the two conditions, confirming that my random assignment to condition was successful. However, I find that participants who are allowed to provide narrative reports make less-biased decisions when placing bets as an agent and when subsequently placing bets on their own behalf, as compared to participants who are not allowed to provide narrative reports. I also find that participants who distort operations more when acting as an agent also continue to place biased bets when acting on their own behalf, particularly if they have a high preference for consistency. Thus, allowing people to provide narrative reports of their performance is more effective at reducing surrogation for people with a high preference for consistency.

My results contribute to the literatures on operational distortion, surrogation, and cheap talk. First, prior literature has found that measuring performance encourages

operational distortion (e.g.; Brown 1990; Hannan, et al. 2013, 2014; Tafkov 2013). I identify an intervention (narrative reporting) that reduces operational distortion while still allowing managers to measure and incentivize performance. When employees know they will provide narrative performance reports, even if those narrative reports are unverifiable, they concentrate their effort relatively more on improving overall performance than on improving the objective measure of performance. Second, my results add to the literature on surrogation by suggesting that surrogation occurs when people try to make their beliefs be consistent with their opportunistic actions. Choi, et al. (2012, 2013) find that paying for multidimensional performance on the basis of a single objective measure causes subjects to overestimate the importance of that measure even when they are later acting on their own behalf. Choi, et al. (2012) argue that surrogation occurs because incentive compensation increases the salience of the incentivized measures of performance. In contrast, my results show that surrogation is at least partly the result of participants having a desire to make their beliefs be consistent with their opportunistic behavior as agents.

My paper also speaks to the existing cheap talk literature. Narrative reporting can be interpreted as a form of cheap talk. In my setting, participants provide narrative reports that are both subjective and unverifiable. Even if participants choose not to provide false statements, their entire explanation is voluntary and can violate Gricean principles of pragmatic communication, limiting the inferences that principals can draw from the communication (Bloomfield 2012). Thus, participants could choose to exclusively provide information that casts their decisions in a favorable light and omit information that casts their decisions in an unfavorable light. In spite of the cheap nature of the communication, narrative reporting significantly reduces operational distortion and surrogation. Thus, reporting

requirements that allow unverifiable communication may discourage operational distortion and surrogation while reporting requirements that prohibit unverifiable communication may actually encourage operational distortion and surrogation.

The remainder of the paper proceeds as follows. Section 2 presents the background literature and develops my hypotheses. Section 3 describes the experiment. Section 4 presents my primary results, and section 5 presents supplemental analyses. I conclude in Section 6.

II. BACKROUND AND HYPOTHESES

2.1 MULTI-DIMENSIONAL PERFORMANCE EVALUATIONS

A large body of research is based on the premise that principals are unable to perfectly monitor agents' decisions and actions. Models of incentive compensation traditionally assume that agents will take actions to maximize their own welfare rather than the principal's (see e.g., Jensen and Meckling 1976, Holmstrom 1979). When performance has many dimensions, such models predict that agents focus on dimensions that are explicitly incentivized, to the detriment of unincentivized dimensions (e.g. Holmstrom and Milgrom 1991, see also Prendergast 1999 and Bonner and Sprinkle 2002 for reviews of evidence supporting these models). For example, an asset manager may be more likely to invest in short-term projects whose performance is more visible rather than long-term projects that have a higher NPV. Similarly, a salesperson who gets commission based on sales price rather than gross margin may focus more on high-price sales than on high-margin sales.[2] Following Bloomfield (2015), I use the term *operational distortion* to refer to operational decisions (as opposed to reporting decisions) that are intended to increase a measure of performance (e.g. 1-year

[2] A related literature is managerial myopia, see e.g., Kraft, Vashishtha, and Venkatachalam (2014), Bhojraj and Libby (2005), Graham, Harvey, and Rajgopal (2005).

return, sales price) rather than the construct that the measure is supposed to capture (e.g. NPV, gross margin). Operational distortion is a generalization of real earnings management that applies to any performance metric, not just earnings.

Prendergast (1999) argues that firms can reduce operational distortion by using subjective rather than objective performance evaluations. He argues that subjective performance evaluations allow managers to evaluate performance in a more holistic sense than would be possible with ex-ante contracting. While the literature has generally been supportive of Prendergast's argument (see Bol 2008 for a review), subjective performance evaluations are not the perfect cure-all to reduce operational distortion. Managers are prone to overweight salient measures in their subjective performance evaluations (Ittner, Larcker, and Meyer 2003), which may encourage agents to focus on highly visible tasks and salient dimensions of performance (Bol 2008). Thus, subjective performance evaluations are unable to eliminate operational distortion, particularly when some performance measures are not readily available to principals. In the next subsection I discuss how narrative reports may reduce operational distortion.

2.2 NARRATIVE REPORTING AND OPERATIONAL DISTORTION

If agents can supplement imperfect objective performance measures with credible and informative narrative reports, they can make operational decisions that maximize true performance, knowing that they will be able to justify their decisions in their reports. Thus, to the extent that both the principal and the agent believe that narrative reporting is informative and credible, narrative reporting will reduce operational distortion.

However, economic game theory predicts that subjective, narrative reports will be neither informative nor credible as agents will take advantage of subjectivity to provide

opportunistic reports, either by providing false reports of performance (when there are no anti-fraud rules) or by disclosing only the positive dimensions of performance and ignoring other dimensions of performance (when agents cannot explicitly lie but can be selective about disclosures). While this cynical economic analysis suggests that principals should ignore or even disallow subjective reports (see e.g., Rajan and Reichelstein 2009), behavioral research suggests otherwise. For example, Forsythe, et al. (1999) find that reporters are considerably more honest than is predicted by economic theory (see also Evans, et al. 2001). Furthermore, Bentley, et al. (2015) find that lying is much less effective than would be predicted by economic theory and that agents often betray their true beliefs to their own financial detriment. Not only are people unwilling to provide false statements, they are reluctant to withhold damning evidence. Together, these results suggest that narrative reports will be considerably more informative than predicted by traditional economic theory. These results support the hypothesis that reporting environments that allow narrative reports will have less operational distortion than reporting environments that do not allow narrative reports.

H1: Narrative reporting reduces operational distortion.

2.3 NARRATIVE REPORTING AND SURROGATION

Basing pay on imperfect measures not only causes managers to distort operations, but also causes them to distort their beliefs about the importance of the measured dimension of performance. Choi, et al. (2012) use the predictive validity framework (Libby, et al. 2002, Libby 1981) as a way to think about performance measurement systems. Managers have an end goal in mind and have beliefs about what actions will cause that end goal to occur. This hypothesized cause-and-effect relationship is sometimes referred to as a strategy map (Kaplan and Norton 2000), value driver map (Ittner and Larcker 2003), or causal chain of performance

(Tayler 2010). However, these causal relationships are only theoretical in that neither the end goal itself nor the actions can be perfectly measured (i.e., they are conceptual-level constructs). Instead, accounting systems provide imperfect proxy measures for conceptual constructs in much the same way that shadows are imperfect representations of reality (Bloomfield 2014).

Choi, et al. (2012) coin the term *surrogation* to mean the phenomenon when managers "fail to fully appreciate the fact that measures are merely representations of the strategic constructs, and act as though the measures are the constructs of interest" (p. 1135). Choi, et al. (2012, 2013) find that when an explicit incentive compensation system rewards one measure of performance but ignores other dimensions of performance, agents surrogate: they come to believe that the incentivized measure of performance is more important than it truly is. Choi, et al. (2012) motivate their predictions using attribute substitution theory, which states that individuals rely on an easily-accessible heuristic (i.e., maximize the incentivized measure) when making complex decisions (i.e., maximizing firm performance). Choi, et al. argue that incentive compensation increases the accessibility of the incentivized measure, which encourages attribute substitution.

Choi, et al. (2012, 2013) test for surrogation by creating a setting in which participants first perform a task as agents and then perform a similar task on their own behalf (i.e., without a principal-agent relationship). My design employs a similar technique, which allows me to measure the degree to which agents distort operations and the degree to which they surrogate. I first have participants make decisions as an agent and then have them make decisions on their own behalf, with the latter set of decisions allowing me to measure the extent to which a participant has surrogated. I predict that participants will surrogate less if, when acting as

agents, they can provide a narrative report of their performance. When participants provide a narrative report, they are able to focus on unmeasured dimensions of performance, reducing the salience of the numeric measure and increasing the salience of other dimensions of performance.

H2: Narrative reporting reduces surrogation.

In the next subsection, I go into greater detail examining the causal mechanisms that could drive the effect of narrative reporting on surrogation.

2.4 SURROGATION AND PREFERENCE FOR CONSISTENCY

When an individual improves his/her own incentive compensation through operational distortion, the individual has taken actions that differ from what he/she would do absent the information asymmetry that allowed operational distortion to be effective. This difference leads to an inconsistency between the actions the individual took on behalf of the firm and the actions that the individual believes would have benefited the firm the most.[3] Research in psychology has found that some people are uncomfortable when their behavior contradicts their beliefs, when they appear inconsistent to others, or when their friends act in an unpredictable fashion. These people have a high preference for consistency: they try to "do things in the same way" and "make an effort to appear consistent to others" (Cialdini, Trost, and Newsom 1995 p. 328). The preference for consistency literature suggests that agents who have a high preference for consistency will feel discomfort after they take actions that contradict their beliefs. They can resolve this discomfort by changing their beliefs. Thus, I predict that agents who distort operations alter their causal model of success in order to make

[3] I acknowledge that some people may not see these two ideas as inconsistent, which would work against my theory. However, Bentley, Bloomfield, and Lambert (2015) find that most Americans agree or strongly agree that operational distortion is unacceptable.

the actions they took (which a third-party would describe as operational distortion) morally acceptable in their own minds. H3 formally lays out this prediction, proposing a link between operational distortion and surrogation.

H3: Operational distortion is positively associated with surrogation.

H3 is a more powerful test of my theory than H2. My theory predicts that narrative reporting reduces surrogation because it reduces operational distortion. Thus, H2 is a simultaneous test of two links in my causal chain, while H3 directly tests for a link between operational distortion and surrogation.

Finally, the effect of narrative reporting will be stronger for participants who have a high preference for consistency. Individuals with a low preference for consistency can take actions that conflict with their beliefs without feeling discomfort, while individuals with a high preference for consistency will feel great discomfort when their beliefs and actions conflict. Thus, H4 predicts that narrative reporting will have a larger effect for participants with a high preference for consistency than for participants with a low preference for consistency. As discussed previously, my theory argues that the effect of narrative reporting on surrogation occurs through operational distortion. Thus, H5 predicts that there will be a stronger link between operational distortion and surrogation for participants who have a greater preference for consistency.

H4: Narrative reporting reduces surrogation more for participants who have more preference for consistency.

H5: Operational distortion is more positively associated with surrogation for participants who have more preference for consistency.

To the extent that surrogation is driven by preferences for consistency, support for H5 should be stronger than support for H4, because greater surrogation for those who prefer

consistency is predicted only among the subset of participants who engage in substantial operational distortion.

III. METHODS

3.1 EXPERIMENT OVERVIEW

The experiment is designed as a 3 (phase) x 2 (reporting condition in the agency phase) mixed design with preference for consistency as a measured covariate. I have participants make decisions in three phases so that I can measure the degree to which they overweight a verifiable measure of performance when acting on their own behalf before the reporting manipulation (pre-agency phase), when acting as an agent (agency phase), and when acting on their own behalf after the experience as an agent (post-agency phase). Throughout the entire task participants write explanations for their decisions. However, at the beginning of the agency phase I tell participants that they will work for another participant (their boss), who will allocate a subjective bonus among four participants. I manipulate whether the boss sees the participants' explanations in addition to the verifiable measure when he makes the allocation decisions. Participants who know that their boss will see their explanation are in the narrative reporting condition. Participants who know that their boss will not see their explanation are in the narrative note-taking condition. At the end of the experiment I measure participants' preference for consistency (PFC), which prior research has shown to be a stable individual trait (Cialdini, Trost, and Newsom 1995; Nail, et al. 2001). The phases and task are described in greater detail in section 3.2. The reporting condition manipulation and preference for consistency measure are described in section 3.3.

3.2 CHESS TASK

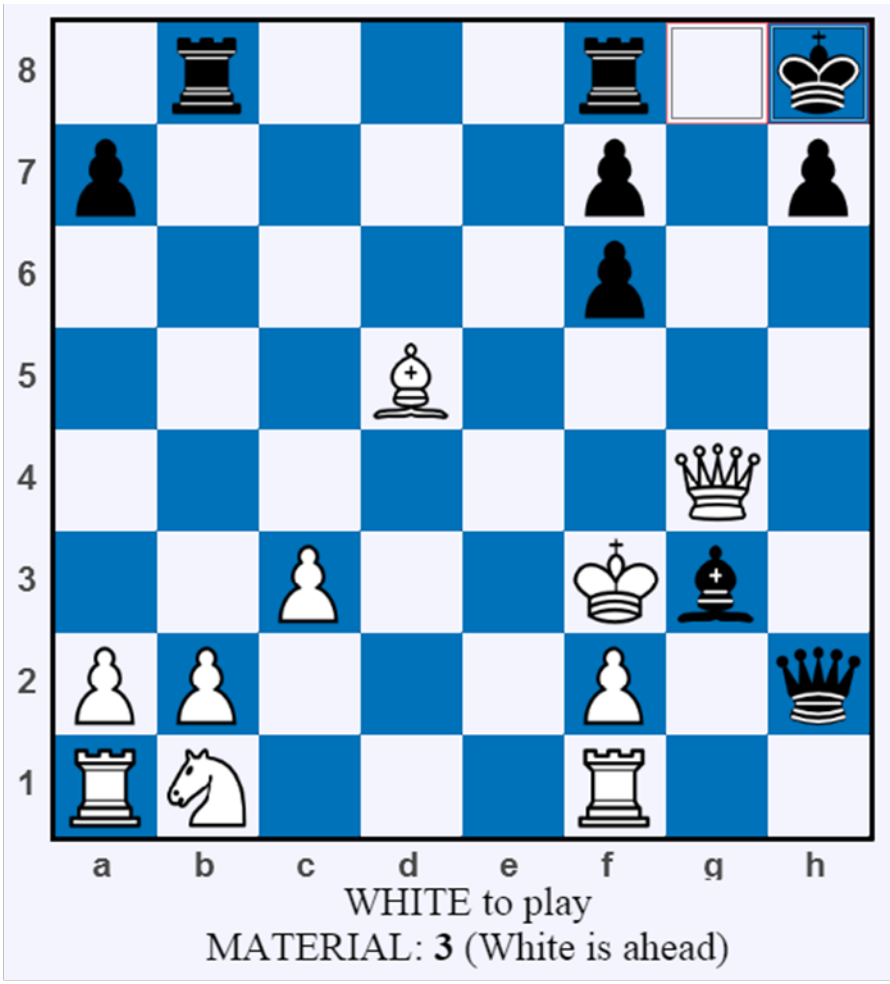
In order to test my theory, I need a multi-dimensional task where one dimension of performance is both reportable and verifiable but where other dimensions of performance cannot be verifiably reported. Furthermore, I need a task that has room for rich narrative reports. Chess works well for testing my theory. Chess is a complex, multi-dimensional game where the outcome of a game is a function of each player's material and the strategic positioning of the material. There is a simple, standard measure of the quantity of material, where a queen is worth 9 points, rooks are worth 5 points, bishops and knights are worth 3 points, and pawns are worth 1 point. Material count (MC) is the sum of White's remaining material minus Black's remaining material. A positive (negative) material count indicates that White (Black) has more chess material remaining on the board than Black (White). This measure is widely used by chess players and reported in most chess tournaments. However, there is not an easily-calculable measure of strategic positioning. While chess players can see and talk about positions and strategic advantages, they can't measure their value. In contrast, Crafty Chess software, available at www.craftychess.com, computes an engine score (ES) that incorporates both material count and positioning based on complex computer simulations and databases and thus is a superior measure of performance that can plausibly be withheld in an experiment. **Appendix 2.A** provides evidence that material count has no incremental contribution to measuring performance and predicting outcomes of chess games when controlling for engine score.

Conceptually, my dependent variable is the extent to which an individual overvalues the reported measure of performance relative to overall performance. Material count is my reported measure of performance. Engine score is my proxy for overall performance. Using

engine score and material count, I identify pairs of games that have the same engine score (i.e., the same overall strength of White relative to Black) but different material counts (i.e., different reported strength of White relative to Black). I create 15 pairs of games and call the game in a pair with the more-positive material count (more favorable toward White) the *High Material Board* and I call the game with the more-negative material count (more favorable toward Black) the *Low Material Board*.[4] For each of the 30 boards, participants see a screen like the one shown in **Figure 2.1**. The screen shows the chess board along with a note stating whose turn it is to play and what the current material count is. Participants place a bet of up to $0.50 on either White or Black and then explain their bet. Participants must write at least 10 characters before advancing to the next board. I sign bets for White as positive and bets for Black as negative. If participants systematically place more-positive bets for the High Material Boards than for the Low Material Boards, their behavior is biased toward the reportable measure of performance, material count. Appendix A describes how I selected boards to use in the task and contains a graphic illustrating the pairs of boards.

[4] Engine score is on a continuous scale, so matches are not exact. The average signed difference (absolute difference) between the high and low boards is -0.015 (0.154) with the High Material Boards having a lower average engine score. The average signed difference is not statistically different from 0 and is economically insignificant as it represents a difference of less than 1/50 of a pawn.

FIGURE 2.1
Chess Task



What bet would you like to place?

$0.50 on Black $0.50 on White

$0.00

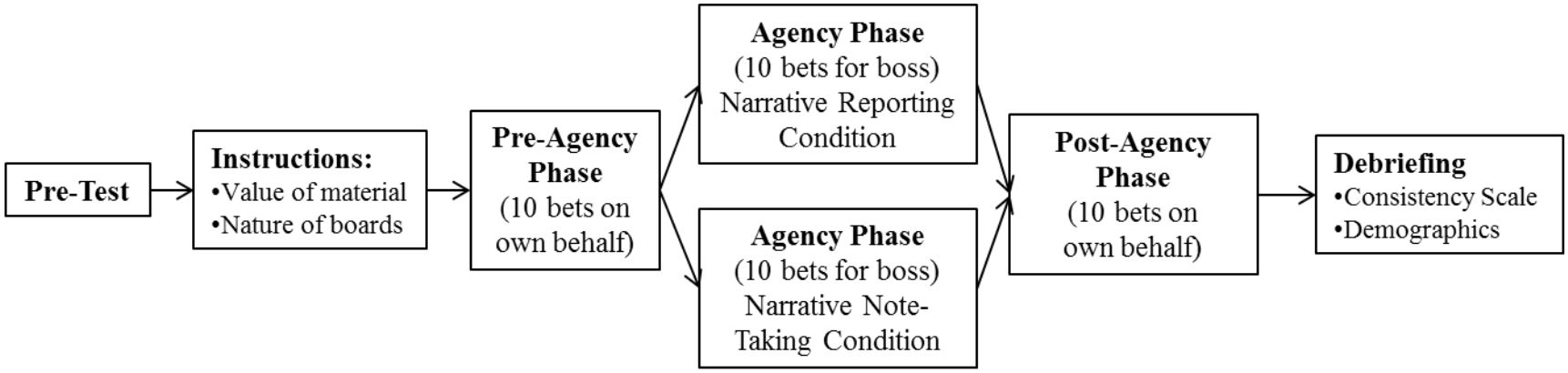
Write a few sentences indicating who you think is going to win, how sure you are, and why.

The task proceeds as follows. I solicit participants from Amazon Mechanical Turk who self-select as being familiar with chess and who pass a difficult chess pre-test. Participants are then told how the material count is calculated and are given instructions on how to place bets on the boards. The experiment then consists of three phases. In each phase, participants see 10 boards. I randomize which set of boards participants see in each phase.[5] I also randomize the order of boards in each phase for each participant. In the pre-agency phase, participants place 10 bets on their own behalf. In the agency phase, they place 10 bets on behalf of a boss.[6] The participants are told that the boss will evaluate four participants' bets at the same time. He will see each participant's material count and bet and then split a $1 bonus between the four participants however he sees fit. I use a fixed bonus pool to eliminate the moral hazard problem with respect to the principal's allocation decisions (Baiman and Rajan 1995). In the post-agency phase, participants again place bets on their own behalf. At the conclusion of the experiment, participants answer debriefing questions. Principals allocate bonuses a few days later, and participants are paid less than one week after participating. **Figure 2.2** contains an experimental timeline. The next subsection describes my independent variables.

[5] Order is not significantly correlated with any of my independent or dependent variables and is not discussed further.

[6] I did not design the study to shed light on principals' bonus allocations. Amazon Mechanical Turk participants expect to be paid very quickly. In order to meet their expectations, I had two graduate students act as principals for all participants. One graduate student was the principal for all participants in the narrative reporting condition and the other graduate student was the principal for all participants in the narrative note-taking condition. These choices allowed me to conduct the experiment in a timely fashion without any deception, but prevent me from providing any meaningful analysis of how principals allocated bonuses, because the manipulation is perfectly confounded with individual differences between the two graduate students.

FIGURE 2.2
Experimental Timeline



The task was open on MTurk from Saturday afternoon (ET) through early Tuesday morning (ET). The bosses (graduate students) evaluated workers' actions the following week, and MTurk workers were paid on Saturday, less than 7 full days after the first participants completed the online study.

3.3 INDEPENDENT VARIABLES

I manipulate reporting condition between the pre-agency phase and the agency phase. All participants must write an explanation for each of their bets in every phase. I manipulate whether or not the principal sees the 10 explanations written during the agency phase when allocating the subjective bonus. Participants in the *narrative note-taking* condition are told that their boss will see their bets and material count but will not see their explanations of the games, the outcomes of the games, nor the chess boards. Thus, the boss will base his allocation decisions on how well the participants' bets line up with material count. In contrast, participants in the *narrative reporting* condition are told that their boss will see their explanations of the games in addition to their bets and material count. Thus, these participants can explain why their bets may differ from material count. **Appendix 2.B** provides the instructions participants saw at the beginning of the agency phase in both conditions and shows an example game with some participant explanations.

I also measure participants' preference for consistency. At the conclusion of the experiment, participants take a modified version of Cialdini, Trost, and Newsom's (1995) Brief Form Preference for Consistency Scale, which is reproduced in **Figure 2.3**.[7] The Cialdini, et al. (1995) scale has been shown to be a stable measure of individuals' preference for consistency that is useful in predicting how much an individual dislikes experiencing cognitive inconsistency.[8]

[7] The original brief form preference for consistency scale had nine possible responses for each question. I reduced it to seven possible responses for each question so that the scale would fit comfortably on a computer screen.

[8] The experiment concluded with a single question asking subjects how acceptable it would be for a salesperson to engage in operational distortion by offering a discount to a customer in order to hit a sales target. This variable was moderately correlated with preference for consistency, but did not add any explanatory power to the model. Because it was intended only for exploratory purposes, I do not discuss it further.

FIGURE 2.3
Preference for Consistency Questionnaire (Adapted from Cialdini, et al. 1995)

	Mean	SD	Factor Loading
It is important to me that those who know me can predict what I will do	0.07	1.50	0.71
I want to be described by others as a stable, predictable person	0.59	1.54	0.75
The appearance of consistency is an important part of the image I present to the world	0.85	1.42	0.84
An important requirement for any friend of mine is personal consistency	0.83	1.50	0.86
I typically prefer to do things the same way	0.47	1.70	0.67
I want my close friends to be predictable	0.83	1.44	0.79
It is important to me that others view me as a stable person	1.37	1.26	0.73
I make an effort to appear consistent to others	1.13	1.30	0.77
It doesn't bother me much if my actions are inconsistent (reverse scored)	0.69	1.46	0.61
Sum of the 9 questions	6.83	9.78	N/A

Above is the Cialdini, et al. (1995) brief form Preference for Consistency Scale. As administered by Cialdini, et al. (1995) the scale had 9 options. I reduced the scale to have 7 options (removing the options “Somewhat Disagree” and “Somewhat Agree” in order to have the scale fit comfortably on a single webpage. Responses were coded on a scale of Strongly Disagree, Disagree, Slightly Disagree, Neither Agree nor Disagree, Slightly Agree, Agree, and Strongly Agree. I converted these responses to a -3 to +3 scale. Factor analysis confirms that the 9 questions load onto a single factor in the predicted directions.

IV. RESULTS

4.1 SAMPLE

Because participants need good familiarity with chess in order to provide interpretable responses, I solicit experienced chess players from Amazon Mechanical Turk (AMT). The task was advertised as "Evaluate Chess Games. Earn up to $24.00" and began with a five-question pre-test. Participants were told that they would only be allowed to complete the longer, more profitable task if they passed the pre-test. Eight-hundred and eleven AMT workers took the pre-test. Eighty-two passed the initial screen and were allowed to complete the full task. As some AMT participants may have passed the pre-test by chance, I drop 6 participants whose pre-agency behavior appears to be random (i.e. was not positively associated with engine score which suggests that they were unqualified or they chose not to exert effort on the task).[9,10] I also drop one participant whose pre-agency behavior indicates that he is an outlier.[11]

My final sample consists of 75 participants. These participants were well-qualified for the task. Over 70% of participants indicated that they play chess or solve chess puzzles at least a few times a month, and 30% indicated that they play chess or solve chess puzzles at

[9] In designing the pre-test I needed to balance the efficiency and the effectiveness of the pre-test. A longer pre-test would be more effective at screening out unqualified workers, but would increase the cost of the prescreening. I chose a 5-question pre-test where participants had to pass 4/5 questions to pass. Three of the questions had a binary response while the other two had participants select the best move. If a participant made completely random, but legal guesses (i.e. a legal chess move), he would have a 1.12% chance of passing the pre-test. If a participant made completely random guesses without caring for the legality of a move, he would have a 0.40% chance of passing the pretest. Thus, under the null hypothesis that all participants guessed, I would have between 3 and 9 of my participants pass the pre-test completely by chance. Thus, my exclusion of 7 people is in line with my ex-ante expectations of the false positive pass rate for the pre-test.

[10] If I keep in these 6 random-behavior participants, my test of H1 and H3 remain significant at $p<0.05$, my test of H2 is significant at $p=0.08$, but my tests of H4 and H5 are insignificant ($p=0.18$ and $p=0.12$ respectively).

[11] This participants' *pre-agency bias* (defined in section 4.2) was -200, which was 3.17 standard deviations below the mean of the remaining participants, and 1.19 standard deviations lower than the next-lowest participant's *pre-agency bias*. If I include this participant in my analyses, my test of H1, H3, and H5 remain significant at $p<0.05$, my test of H2 is significant at $p=0.07$, but my test of H4 is insignificant ($p=0.11$).

least a few times a week. They spent a mean (median) of 111 (89) minutes on the entire task, and a mean (median) of 2.7 (1.9) minutes per board, suggesting they took the task quite seriously. The reason for the skew of mean time on task appears to be breaks. Many participants acknowledged taking breaks to use the restroom, eat meals, etc. Thus, the median time per game (more precisely, the median participant's median time per game) is a more accurate measure of time-on-task than the mean time per game.[12]

Participants earned a mean (median) of $9.33 ($9.17), which translates to an effective hourly wage of approximately $6.48 ($6.01). This hourly wage is well above the median reservation wage reported in Horton and Chilton (2010) of $1.38 per hour and is comparable to or higher than hourly wages reported in other Amazon Mechanical Turk studies (e.g., Rennekamp 2012; Grenier, Pomeroy, and Stern 2014; Paolacci, Chandler, and Ipeirotis 2010).

4.2 DEPENDENT VARIABLE

Recall that in each phase participants place bets on 10 chess games that were paired based on engine score. When participants systematically make more-positive bets on the High Material Boards than the Low Material Boards, their bets are biased[13] in the direction of material count. As such, I proxy for bias in each 10-board phase by summing each participant's bets on the five High Material Boards and subtracting the participant's bets on the five Low Material Boards. The resulting number, *Material Bias*, represents the degree to

[12] Not surprisingly, participants spend a little bit longer in the agency phase when they are allowed to provide a narrative report (Mean=25.8 minutes after winsorizing at the 95th percentile) than when they are not allowed to provide a narrative report (Mean=20.5 minutes after winsorizing, p_{diff}=0.08). They also write more in the agency phase when they are allowed to provide a narrative report (Mean=1594 characters) than when they are not allowed to provide a narrative report (Mean=1002 characters, p_{diff}<0.01). These differences do not persist into the post-agency phase (both p>0.15), suggesting that the effect of narrative reporting on surrogation cannot be explained by effort.

[13] Note that I use the term bias to refer to a bias from the residual claimant's perspective. Thus, it's a bias relative to the participant's own perspective for the pre-agency and post-agency phases and a bias relative to the principal in the agency phase.

which a participant overemphasizes material count in a particular phase, relative to what the computer says is optimal (the optimal allocation for any level of risk aversion or probability matching is to place the same bet for both games in a pair).

TABLE 2.1
The Effect of Narrative Reporting on Material Bias

Panel A – Descriptive Statistics of Material Bias by Condition and Phase

Mean [St Dev] of *Material Bias*			
	Pre-Agency Phase	**Agency Phase**	**Post-Agency Phase**
Narrative Reporting	25.36 [74.62]	39.21 [104.9]	23.90 [75.95]
Narrative Note-taking	33.06 [70.57]	88.44 [125.6]	52.14 [107.4]
Difference	-7.70 [72.70]	-49.24 [115.3]	-28.24 [92.35]

Panel B – Hypothesis Tests[a]

	Estimated Difference	**t-Test**	**Wilcoxon Rank-Sum Z-test**
H1: Narrative Reporting < Narrative Note-taking in Agency Phase	-49.24	-1.85 p=0.03	1.90 p=0.03
H2: Narrative Reporting < Narrative Note-taking in Post-Agency Phase	-28.24	-1.32 p=0.09	1.72 p=0.04

[a] Reported p-values are 1-tailed equivalents reflecting directional hypotheses. See Appendix 2.C for variable definitions

Pre-agency bias is a baseline measure, and represents participants' bias in in the pre-agency phase. *Agency Bias* is my proxy for operational distortion and represents participants' bias in the agency phase. *Post-agency bias* is my proxy for surrogation and represents participants' bias in the post-agency phase.

Table 2.1 Panel A presents descriptive statistics of how biased participants' bets are in each phase and condition (3 x 2). In all six cells the average bias is greater than 0. While not hypothesized, this result is unsurprising, as material count is presented in a very salient fashion and is therefore likely to be relied upon when placing bets.

4.3 HYPOTHESIS TESTS

H1 predicts that participants will distort operations less when they are allowed to provide a narrative report than when they are not allowed to provide a narrative report. I test this hypothesis using a non-parametric Wilcoxon rank-sum test. As reported in Panel B of Table 2.1, I find that participants who were allowed to provide a narrative report have a significantly lower agency bias than participants who were not allowed to provide a narrative report ($Z=1.90$, $p<0.05$),[14] supporting H2. A t-test yields similar inferences when imposing distributional assumptions on the variables ($t=-1.85$, $p<0.05$).

An alternative method of testing H1 involves looking for a treatment effect while controlling for participants' pre-agency biases. I run a regression (untabulated) in which I find similar results as discussed above. Participants who were allowed to provide narrative reports are significantly less-biased in the agency phase than participants who were not allowed to

[14] All hypothesis tests are one-tailed, reflecting directional predictions.

provide narrative reports when controlling for the individual's bias in the pre-agency phase ($t=-1.78$, $p<0.05$).[15]

H2 predicts that narrative reporting reduces surrogation. To test this, I use a Wilcoxon rank-sum test to compare the post-agency bias of participants who were or were not able to provide narrative reports when they were agents. I find that participants who were able to provide narrative reports as agents make post-agency bets that are significantly less biased than participants who were not able to provide narrative reports as agents ($Z=1.72$, $p<0.05$). A t-test yields similar, but less-significant inferences ($t=-1.32$, $p<0.10$). **Figure 2.4** presents the tests of H1 and H2 graphically.

H3 predicts that the more someone distorts operations in the agency phase, the more they surrogate in the post-agency phase. Untabulated results show that agency bias is a significant predictor of post-agency bias ($\beta=0.20$, $t=2.28$,$p=0.01$). Thus, I find support that there is a positive relationship between operational distortion and surrogation. As expected, given my theory that surrogation is driven by preferences for consistency, support for H3 is considerably stronger than support for H2, because surrogation is predicted only among the subset of participants who engage in substantial operational distortion.

[15] The pre-agency phase was included in the experiment to provide a baseline for each participant's performance. However, as the within-subject correlation between pre-agency, agency, and post-agency biases is less than 0.5, a mixed model results in lower statist Section 5.2 presents alternative model specifications to control for pre-agency bias.

FIGURE 2.4
The Effect of Narrative Reporting on Operational Distortion and Surrogation

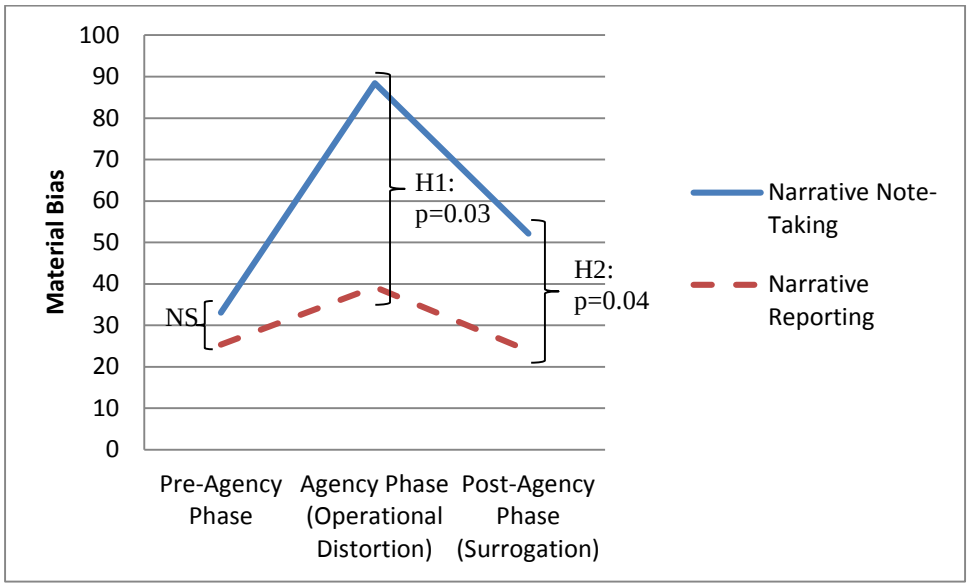


Figure 2.4 graphically shows how the narrative reporting manipulation affects the weight that participants place on material count when making decisions on behalf of others (i.e., operational distortion in the agency phase) and when making decisions on their own behalf after acting as agents (i.e., surrogation in the post-agency phase). The reported p-values for the agency and post-agency phases are one-tailed to reflect directional hypotheses. See Figure 2.2 for an experimental timeline and Appendix 2.C for variable definitions.

H4 predicts that the effect of narrative reporting on surrogation is larger for participants with a high preference for consistency. To measure preference for consistency, I take the average of the nine preference-for-consistency questions shown in Figure 2.3 (the last question is reverse-scored) and then create a 3-level variable to reduce the influence of outliers. The variable *preference for consistency (PFC)* is equal to 1 for participants whose response is at least 1 standard deviation above the mean, -1 for participants who are at least 1 standard deviation below the mean, and 0 otherwise. Prior research has shown that the preference for consistency questionnaire measures a stable individual characteristic (e.g. Cialdini, Trost, and Newsom 1995; Nail, et al. 2001). Nonetheless, as this variable was

measured at the end of the experiment, I test whether preference for consistency is affected by the reporting manipulation. Neither a t-test nor a Wilcoxon rank-sum test identifies a significant difference between conditions (both $p>0.50$), providing evidence that the preference for consistency scale measures a stable individual trait and that random assignment to condition was successful.

Table 2.2, Panel A shows the average post-agency bias for participants in each condition at the three levels of preference for consistency. To test H4, I run an ANCOVA in which I interact narrative reporting and preference for consistency to predict post-agency bias. **Panel B** reports the results of this ANCOVA. As predicted, the interaction is significant. Narrative reporting has a larger effect on surrogation for participants who have a high preference for consistency ($F=3.46$, $p<0.05$, one-tail equivalent). **Figure 2.5** presents the interaction in graphical form along with p-values for the simple effects. **Panel C** of **Table 2.2** presents the simple effect of preference for consistency in each condition as well as estimates and statistics for the simple effects at high (preference for consistency =1) and low (preference for consistency = -1) levels of preference for consistency.

TABLE 2.2
The Interactive Effect of Narrative Reporting and Preference for Consistency on Surrogation

Panel A – Descriptive Statistics of Post-Agency Bias by Condition and Preference for Consistency (PFC)

Mean [St Dev] of *Post-Agency Bias*			
	PFC=-1	**PFC=0**	**PFC=1**
Narrative Reporting	20.57 [73.23]	31.85 [80.97]	-14.40 [43.23]
Narrative Note-taking	-24.67 [174.80]	62.00 [83.21]	85.57 [93.87]
Difference	45.24 [129.70]	-30.15 [82.00]	-99.97 [77.68]

Panel B – Analysis of Covariance

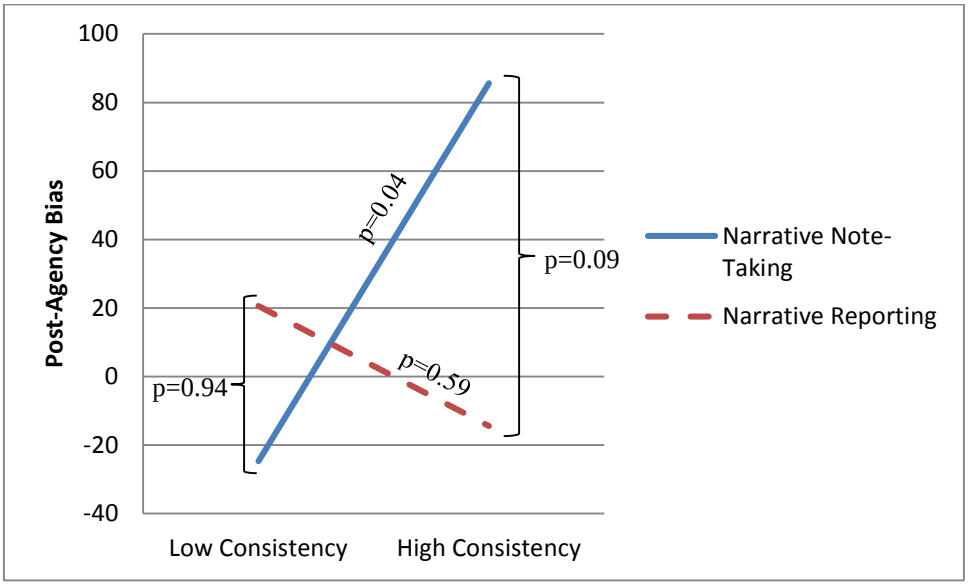
Factor	**df**	**MS**	**F**	***p*-value[a]**	
Narrative Reporting	1	14061.08	1.71	0.19	
PFC	1	9647.03	1.17	0.28	
Narrative Reporting X *PFC*	1	28445.72	3.46	**0.03**	H4
Error	71	583025.37			

Panel C – Tests of Simple Effects

Simple Effect	**Estimate**	**t**	p_t	$p_{Wilcoxon}$
PFC in the *Narrative Note-taking* Condition	53.57	2.13	0.04	N/A
PFC in the *Narrative Reporting* Condition	-14.14	-0.54	0.59	N/A
Narrative Reporting when *PFC*=1	-99.97	-2.20	0.05	0.09
Narrative Reporting when *PFC* = -1	45.24	0.63	0.54	0.94

[a] p-values in a bold face are one-tailed equivalents as they are tests of directional hypotheses
See Appendix 2.C for variable definitions

FIGURE 2.5
The Interactive Effect of Narrative Reporting and Preference for Consistency on Surrogation



All p-values are 2-tailed to represent the lack of a-priori predictions of simple effects
See Figure 2.2 for an experimental timeline and Appendix 2.C for variable definitions

The simple effect of preference for consistency is significant for participants who cannot provide narrative reports ($t=2.13$, $p<0.05$, two-tailed) but insignificant for participants who can provide narrative reports ($t=-0.54$, $p>0.50$). Similarly there is a statistically-significant difference between the behavior of participants who can and cannot provide reports who have a high preference for consistency ($t=-2.20$, $p_t=0.05$, two-tailed; $Z=-1.87$, $p_{Wilcoxon}<0.10$, two-tailed), but there is no difference in behavior for participants who have a low preference for consistency (both $p>0.50$). These simple effects support the notion that narrative reporting is most beneficial for participants who have a high preference for consistency. However, neither the interaction nor the simple effects are statistically robust to alternative specifications of the preference for consistency variable (e.g. continuous variable,

quartiles). Tests with alternative specifications of preference for consistency have effects in the same direction, but are not significant at the $p<0.05$ level.[16] Thus, I find only weak support for H4.

H5 predicts that participants who distort operations in the agency phase will surrogate more in the post-agency phase if they have a high preference for consistency. **Table 2.3** presents the results of an ANCOVA that support this hypothesis ($F=3.75$, $p<0.05$, one-tailed equivalent). This result is robust to alternative specifications of preference for consistency (e.g. continuous, quartiles). As expected, support for H5 is considerably stronger than support for H4, because preference for consistency should only matter for the subset of participants who engage in substantial operational distortion.

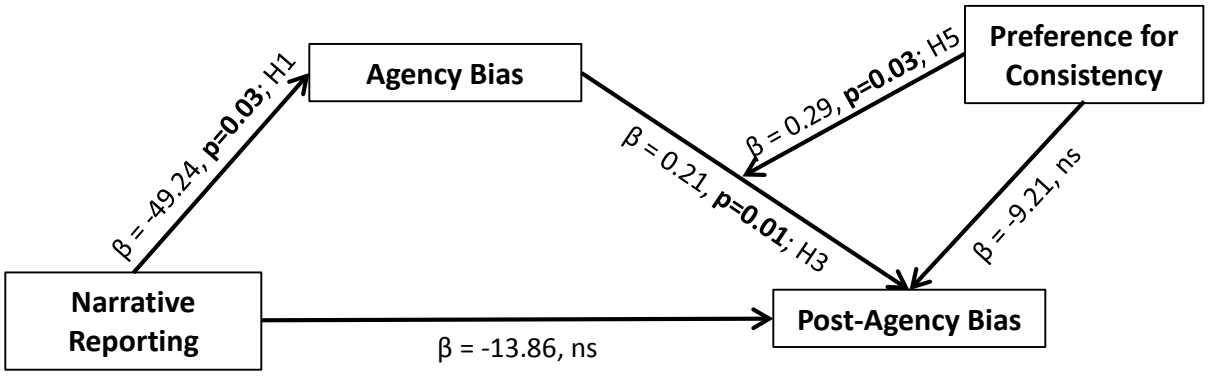
TABLE 2.3
The Effect of Operational Distortion on Surrogation
at High or Low Preference for Consistency

Factor	**df**	**MS**	**F**	***p*-value[a]**	
Narrative Reporting	1	3391.31	0.43	0.51	
Agency Bias	1	40430.88	5.14	0.03	
PFC	1	1267.57	0.16	0.69	
Agency Bias X *PFC*	1	29527.20	3.75	**0.03**	H5
Error	70	550567.06			

[a] p-values in a bold face are one-tailed equivalents as they are tests of directional hypotheses
See Appendix C for variable definitions

[16] For example, a test with preference for consistency as a continuous variable has an insignificant effect in the same direction as the reported results ($p=0.35$). Dividing the variable into quartiles yields a marginally-significant interaction ($p=0.075$), an insignificant simple effect of PFC for participants in either condition (both $p<0.17$), and a significant simple effect of narrative reporting for participants with a high preference for consistency ($p=0.02$).

FIGURE 2.6
Structural Equation Model



The betas are parameter estimates from a 2-equation structural equation model. The first equation uses narrative reporting to predict agency bias. The second equation uses narrative reporting, agency bias, preference for consistency, and the interaction between agency bias and reference for consistency to predict post-agency bias. p-values in boldface font are one-tailed reflecting directional hypotheses.

See Figure 2.2 for an experimental timeline and Appendix 2.C for variable definitions

Finally, I test for an overall moderated-mediation model. **Figure 2.6** diagrams the structural equation model. A moderated-mediation model tests if the overall effect of narrative reporting on surrogation through operational distortion is larger for participants with a stronger preference for consistency. Figure 6 presents the parameter estimates and p-values from the structural equations. To test the significance of the moderated-mediation model, I use a bootstrap analysis (Hayes 2013). A bootstrap analysis involves resampling the data multiple times with replacement and performing the specified set of regressions for each sample. In each repetition, the analysis tests whether preference for consistency moderates the indirect effect of narrative reporting on post-agency bias. A bootstrap analysis is superior to the traditional Sobel test because it does not make distributional assumptions about the data and it is less-easily influenced by outliers (Preacher and Hayes 2008). Furthermore, bootstrap

analysis allows me to test the more-complex moderated mediation model rather than just a simple mediation model.[17]

I conduct 100,000 repetitions of the bootstrap analysis. The bias-corrected 90% confidence interval for an indirect effect is [-49.41, -0.05], providing evidence that the effect of narrative reporting on surrogation is larger for participants with a stronger preference for consistency (one-tailed equivalent $p<0.05$). This analysis is robust to alternative specifications of preference for consistency (e.g. continuous, quartiles).

V. SUPPLEMENTAL ANALYSES

5.1 NARRATIVE EXPLANATIONS

The tests of surrogation in section 4 use participants' betting decisions to infer their beliefs about the importance of material count relative to other factors that affect the strength of a chess position. In this section, I analyze the content of narrative explanations in the post-agency phase to assess participants' beliefs more directly.

I use Perl regular expressions to count the number of explanations in which a participant refers to aspects of chess other than material count. Specifically, I code each explanation as a 1 (0) if the explanation contains (does not contain) at least one of the following strings: advance, aggress, bait, castle, center, check, close, control, corner, cramp, develop, defen, double, edge, expos, fork, gambit, guard, lone, middle, mobil, moment, move, offen, open, pin, position, promot, protect, struct, tactic, take, threat, turn.[18] These words describe aspects of the board that are not incorporated into the objective measure of material

[17] A Sobel test finds marginal significance that operational distortion mediates the relationship between narrative reporting and surrogation ($p=0.075$).

[18] I use word roots to allow for variations. For example, someone could say that White is more developed or that White's piece development is more advanced. Similarly, White could be on offense or on the offensive. Examples of the full words: developed, mobile, defense, offense, center, castled, threaten, check, positioning, control, momentum.

count, but are important for predicting the winner (e.g. how aggressive one side is, whose turn it is, who has control of the center of the board, etc.).

I construct the variables *Pre-Agency Explanations*, *Agency Explanations*, and *Post-Agency Explanations*, as the sum of the number of explanations in the pre-agency, agency, and post-agency phases respectively, where a participant uses at least one of the above words to refer to something other than material count. The variables can take on a value of 0 (the participant never refers to any of these words in any of the 10 explanations) to 10 (the participant refers to at least one of these words in all 10 of his/her explanations in a particular phase). **Table 2.4 Panel A** presents by condition and phase the mean number of sentences in which a participant talks about something other than material count. **Panel B** presents t-tests and non-parametric Wilcoxon rank-sum tests. Participants talk about non-material aspects of the game more when they are in the narrative reporting condition (Agency Explanations = 8.08, Post-Agency Explanations = 7.67) than when they are in the narrative note-taking condition (Agency Explanations = 6.33, Post-Agency Explanations = 5.89). The differences are significant at the p=0.01 level using both parametric and non-parametric tests. There is no difference in the extent to which participants talk about material count (all p>0.18, untabulated).[19] These results suggest that participants who are allowed to provide narrative reports as agents take a more holistic view of the games, focusing on both material and non-material aspects, as compared to participants who are not allowed to provide narrative reports of the games.

[19] I count an explanation as referring to material count if it contains at least one of the following strings: material, point, has more, more pawn, more knight, more bishop, more rook, fewer pawn, fewer knight, fewer bishop, fewer rook, less knight, less bishop, less rook, less pawn.

TABLE 2.4
The Effect of Narrative Reporting on Participants'
Discussion of Non-Material Aspects in Their Explanations

Panel A – Descriptive Statistics by Condition and Phase

Mean [St Dev] of *Explanation*			
	Pre-Agency Phase	**Agency Phase**	**Post-Agency Phase**
Narrative Reporting	7.90 [2.19]	8.08 [1.72]	7.67 [1.95]
Narrative Note-taking	7.17 [3.06]	6.33 [3.23]	5.89 [3.09]
Difference	0.73 [2.64]	1.74 [2.56]	1.78 [2.56]

Panel B – Statistical Tests[a]

	Estimated Difference	**t-Test**	**Wilcoxon Rank-Sum Z-test**
Narrative Reporting > Narrative Note-taking in Agency Phase	1.74	2.95 p=0.002	2.38 p=0.01
Narrative Reporting > Narrative Note-taking in Post-Agency Phase	1.78	3.01 p=0.002	2.52 p=0.007

[a] Reported p-values are 1-tailed reflecting directional predictions

See Appendix C for variable definitions

5.2 ROBUSTNESS TESTS

In my primary analyses reported in section 4, I find, as expected, that there is no statistical difference between participants' pre-agency biases. For simplicity and parsimony section 4 presents all results without controlling for participants' pre-agency bias. In this subsection I present results using two alternative model specifications. The first relies on a mixed design by differencing out pre-agency bias when calculating dependent variables. The second controls for pre-agency bias as a covariate in ANCOVA models.

One possible way to deal with the pre-agency bias is to run a mixed model and use each participant as his/her own baseline control. Many researchers assume that using a participant as his/her own control will always increase statistical power. However, that is only the case when the within-subject correlation is greater than 0.5 (McKenzie 2012). When $\rho<0.5$, a mixed analysis results in lower statistical power than a between-subjects design that does not difference out the baseline measurement because the variance added from the additional measure is larger than the within-subject variance that is removed.[20]

In my data, the Pearson correlation between the pre-agency bias and agency bias is 0.24 and the Pearson correlation between the pre-agency bias and post-agency bias is 0.41 (untabulated). While these correlations are statistically significant ($p<0.05$ and $p<0.001$ respectively), they are less than 0.5 in magnitude and so reduce rather than increase statistical power. Nonetheless, I conduct analyses differencing out the pre-agency bias acknowledging that for my data these tests will be less powerful than tests that rely on a between-subjects analysis. Using these variable transformations I find marginally-significant support for H1

[20] See McKenzie (2012) for a proof or Uri Simonsohn's blog for a numerical and graphical illustration and a detailed proof. http://datacolada.org/2015/06/22/39-power-naps-when-do-within-subject-comparisons-help-vs-hurt-yes-hurt-power/

($p_{Wilcoxon}$=0.07), marginally-significant support for H2 ($p_{Wilcoxon}$=0.09), significant support for H3 (t=2.46, p<0.01), marginally-significant support for H4 (interaction t=-1.35, p=0.09), a significant simple effect of PFC on surrogation when participants do not provide narrative reports (β=51.1, t=2.05, p=0.04), an insignificant simple effect of PFC on surrogation when participants do provide narrative reports (β=2.44, p=0.93), and marginal support for H5 (interaction t=1.13, p=0.13).

McKenzie (2012) recommends that when within-subject correlations are less than 0.5 researchers control for baseline measures using covariates in ANCOVA models instead of using a mixed design. The inclusion of a control variable in the ANCOVA never reduces statistical power, but it may limit what statistical tests are possible. For my data, I rely on Wilcoxon rank-sum tests for H1 and H2. These tests are not allow covariates and so are not conducive to controlling for participants' pre-agency biases. However, if I include participants' pre-agency bias in all regression tests, I find significant support for H1 (t=-1.78, p=0.04), marginally-significant support for H2 (t=-1.23, p=0.11), marginally-significant support for H3 (t=1.46, p=0.07), significant support for H4 (t=-1.73, p=0.04), and marginally-significant support for H5 (t=1.57, p=0.06). Thus, my results become weaker when I include pre-agency bias as a covariate, but they generally remain consistent with my hypotheses.

5.3 ROBUSTNESS TEST ON NARRATIVE EXPLANATIONS

The within-subjects correlation on the content of explanations is 0.66 between the pre-agency and agency explanations and 0.69 between the pre-agency and post-agency explanations, suggesting that a mixed design will improve statistical power for the analyses shown in section 5.1. I conduct untabulated analyses testing if participants write less about non-material aspects of the games in the agency and post-agency phases relative to the pre-

agency phase if they are not allowed to provide narrative reports in the agency phase than if they are allowed to provide narrative reports in the agency phase. As predicted, the within-subjects difference between pre-agency explanations and agency explanations is more negative for participants in the note-taking condition (difference = -0.83) than for participants in the narrative reporting condition (difference = 0.18). The difference in differences is significant at the $p<0.05$ level (d-in-d effect=1.01, t=2.04, one-tail p=0.02). Similarly the within-subjects difference between the pre-agency explanations and post-agency explanations is more negative for participants in the note-taking condition (difference = -1.28) than for participants in the narrative reporting condition (difference = -0.23). The difference in differences is significant at the $p<0.05$ level (d-in-d effect=1.05, t=2.20, one-tail p=0.02).

VI. CONCLUSION

Prior research has found that measuring one aspect of performance can lead agents to overweight it in their operational decisions (e.g. allocation of effort or other limited resources), leaving managers with the tradeoff between measuring/motivating performance and distorting performance. I present an intervention that reduces this tradeoff. My results indicate that agents who know that they will be allowed to provide narrative reports of their performance are less likely to distort operations in an attempt to improve reported performance. I also find evidence that after agents distort operations, they change their beliefs to rationalize their behavior. Choi, et al. (2012, 2013) argue that incentive compensation directs agents' attention which causes surrogation. In contrast, I present evidence that incentive compensation causes agents to distort operations and that agents who distort operations surrogate. This effect is larger for agents who have a stronger preference for consistency.

While prior academic and practitioner literature has expressed reasonable concerns that subjective, unverifiable narrative reports simply provide an opportunity for self-serving cheap talk (e.g. Keusch, et al. 2012; Clatworthy and Jones 2003; Agarwal, et al. 2009), I document that narrative reporting also has two important benefits. First, agents who are allowed to provide narrative reports are less likely to distort operations than agents who cannot provide narrative reports. Second, agents who are allowed to provide narrative reports are less likely to use the performance measure as a surrogate for true performance. My results have implications for both managerial and financial reporting. For example, they suggest that performance reporting within the firm can be improved by letting agents know that they will be able to provide narrative reports as part of their performance reviews. My results also suggest managers who anticipate providing external narrative reports (e.g. conference calls, MD&A, management commentary) will be less likely to engage in real earnings management or have distorted views about the firm's ideal model of success.

My study suggests several directions for future research. First, my moderating variable, preference for consistency, was a measured variable rather than a manipulated variable. Cialdini, et al. (1995) provide evidence that preference for consistency is a stable individual trait, and the variable does not differ by condition. Nonetheless, future research could add internal validity by manipulating preference for consistency. Future research could also investigate more fully the cognitive process underlying the relationship between narrative reporting, operational distortion, and surrogation. Third, future research could replicate this effect in a multi-period setting. In a multi-period setting, agents may be concerned about reputation, which would increase the degree to which they provide informative narrative reports and would likely increase the effect of narrative reporting on operational distortion

and surrogation. Finally, future research could investigate how principals react to narrative reports to see how the reports affect both principal welfare and agent compensation.

APPENDIX 2.A
Board Selection Procedures

I downloaded a sample of 828 games played by subscribers to www.chessclub.com subject to the conditions that (1) at least one player had an ELO rank between 1400 and 2700, (2) the game ended with one side winning (i.e., not a stalemate, draw, system shut-down, time flag, etc.), and (3) it was an open or semi-open game (i.e., a designation of games that suggests that White is taking an offensive strategy—this category of opening tends to result in more interesting, strategic moves such as gambits).

For each of these games, I analyzed the set of moves using Crafty Chess software to calculate the material count (MC) and engine score (ES) after every move (see Appendix C for definitions of material count and engine score). My dataset was the 53,763 moves from the set of 828 games.

I then analyzed the distributional properties of material count. My goal was to find pairs of chess boards that had the same engine score but different material count and were comparable in their level of unusualness. For example, when material count = -1, 15 percent of the observations have an engine score less than -4.13. When material count = -4, 15 percent of observations have an engine score greater than -4.13. Thus, I selected a pair of observations, one where material count = -1 and engine score was approximately -4.13 (within 0.25 on either side), and one where material count = -4 and engine score was approximately -4.13. I randomly selected chess boards that fit these parameters.

This selection process gave me 15 pairs of chess boards. In each pair, the game with the more positive material count was designated the High Material Board while the game with the lower material count was designated the Low Material Board. The difference between engine score of the High Material Boards (average engine score=0.55) and the engine score of the Low Material Boards (average engine score=0.57) was neither statistically nor economically significant (*p*-value from paired t-test > 0.75; effect size is less than 1/50 of a pawn).

From these 15 pairs of boards I created 3 sets of 5 pairs each. Each set of boards had at least two pairs where engine score favored Black and two pairs where engine score favored White.

Finally, I had Crafty Chess software play out the remainder of each of the 30 games identified. If the game ended in a draw, I replaced the game with a new game until I had 30 games that did not end in a draw.

My method of board construction assumes that engine score and material count are correlated and that engine score contains all of the information in material count plus additional information not incorporated in material count. To test this assumption, I run a logistic regression on the 30 boards, using engine score and material count to predict the winner. As expected, material count is a significant predictor of who will win the game when it is included in the model and engine score is not included in the model ($p<0.01$, pseudo R-square = 0.318). However, when engine score is added to the model, material count is an insignificant predictor of who will win the game ($p=0.50$) while engine score is a significant predictor (model pseudo R-square = 0.656). Furthermore, the model performs no better than when engine score alone is included in the model (pseudo R-Square=0.651). Thus, I find

strong support for my assertion that engine score dominates material count in predicting the winner of the game.

To confirm my analysis, I had Crafty Chess play out all boards where (1) engine score was between -6 and -1 or between 1 and 6 and (2) material count was not equal to 0.[21] I repeated my logistic regression on the 19,496 boards that did not end in a draw. Inferences are identical to those reported above. In a model with both engine score and material count, engine score loads as significant while material count does not load.

[21] I performed the analysis on a subset of board in order to reduce processing time. It is computationally intensive to have computers simulate games as each move involves analyzing thousands of potential future moves. My filter requirements were designed to capture games that were not obvious nor likely to end in a draw.

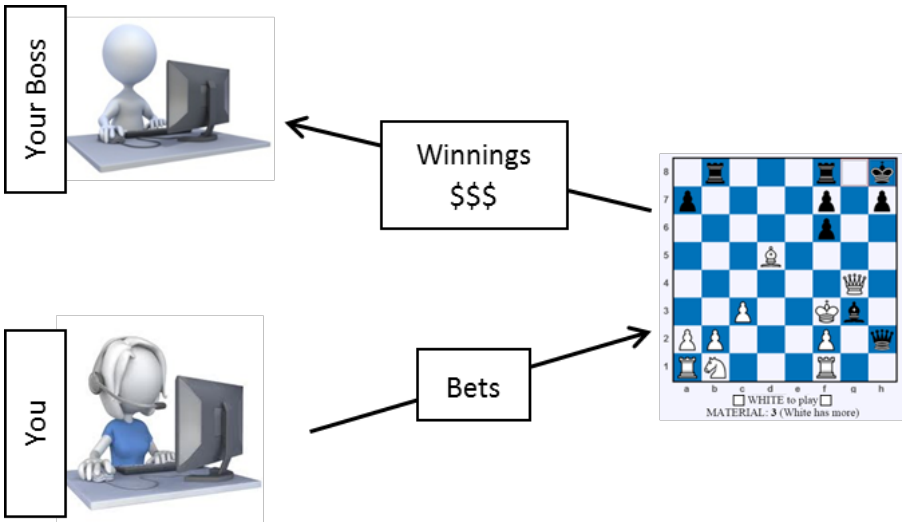
One set of 5 pairs of boards:

ES	Low MC Board	High MC Board
-4.2	MC= -4 	MC= -1 
-2.3	-3 	0 
-3.3	-3 	-1 
2.4	0 	2 
4.5	0 	4 

APPENDIX 2.B

Excerpts of Experimental Instructions

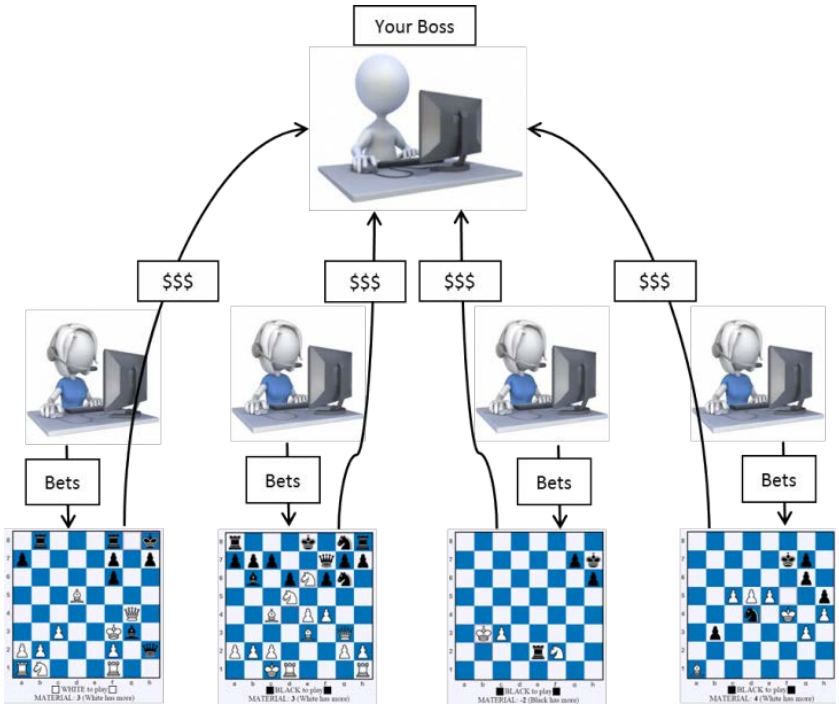
You will now place bets on behalf of another person (your boss). If you bet correctly, **your boss will win** the amount of money you bet. If you bet incorrectly, **your boss will lose** the amount of money you bet.



[NEW PAGE]

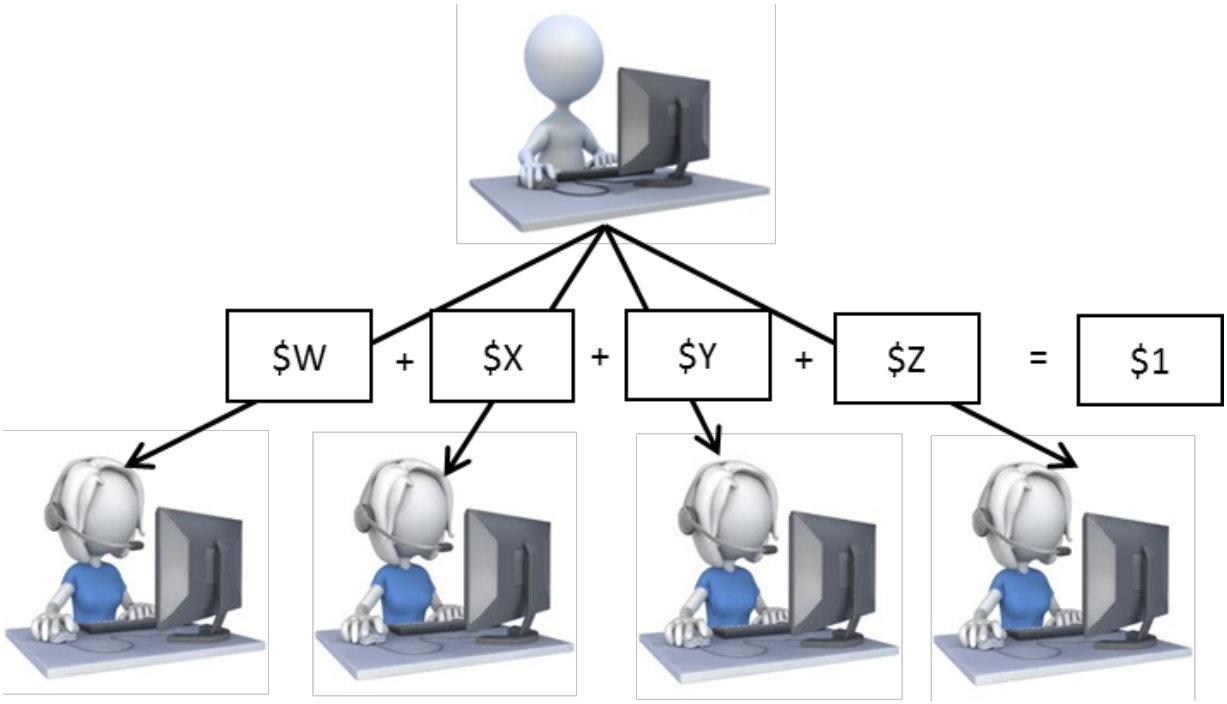
You and three other MTurk workers are employed by the same boss. You are each placing bets on **different games.**

Your boss and the other workers are all real people. There is no deception in this study.



[NEW PAGE]
After each chess game, your boss will evaluate your performance. He'll look at the bet you placed on your game and the bets that the other workers placed on their games, and decide how to split up a $1.00 bonus among the four of you. Your boss must pay out the entire $1.00, but can allocate it among the workers however he/she sees fit. After he/she makes his/her allocation decision, we will give you the money as an MTurk bonus.

Again, you are evaluating a different chess game than the other workers. Your boss knows that.



How your boss allocates the bonus

Your boss **will** see:

- The bet that you placed on your game
- The bets that other employees placed on their games
- The difference in material between White and Black. Whenever a bet is in the opposite direction as material, the bet appears **red.**
- [NARRATIVE REPORTING CONDITION: Your explanation of who you think is going to win and why]

Your boss **will not** see:

- The chess boards
- [NARRATIVE NOTE-TAKING CONDITION: Your explanation of who you think is going to win and why]
- Who won the game (i.e., whether White or Black wins)

[GRAPHIC IN NARRATIVE REPORTING CONDITION]

$W $X $Y $Z

Worker #1 Game #2049		
Bet	Mat	Explanation
-50	-1	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nam nibh. Nunc varius facilisis eros. Sed erat. In in velit quis arcu ornare laoreet. Curabitur adipiscing luctus massa. Integer ut purus ac augue commodo.

Worker #2 Game #4183		
Bet	Mat	Explanation
+45	-2	Sed malesuada, enim ut congue pharetra, massa elit convallis pede, ornare scelerisque libero neque ut neque. In at libero. Curabitur molestie. Sed vel neque. Proin et dolor ac ipsum elementum malesuada.

Worker #3 Game #256		
Bet	Mat	Explanation
+30	2	Nunc at velit quis lectus nonummy eleifend. Curabitur eros. Aenean ligula dolor, gravida auctor, auctor et, suscipit in, erat. Sed libero. Maecenas at est. Nam ut metus et orci fermentum nonum.

Worker #4 Game #1710		
Bet	Mat	Explanation
0	2	Nulla mauris odio, vehicula in, condimentum sit amet, tempus id, metus. Donec at nisi sit amet felis blandit posuere. Aliquam erat volutpat. Cras lobortis orci in quam porttitor cursus..

[GRAPHIC IN NARRATIVE NOTE-TAKING CONDITION]

$W $X $Y $Z

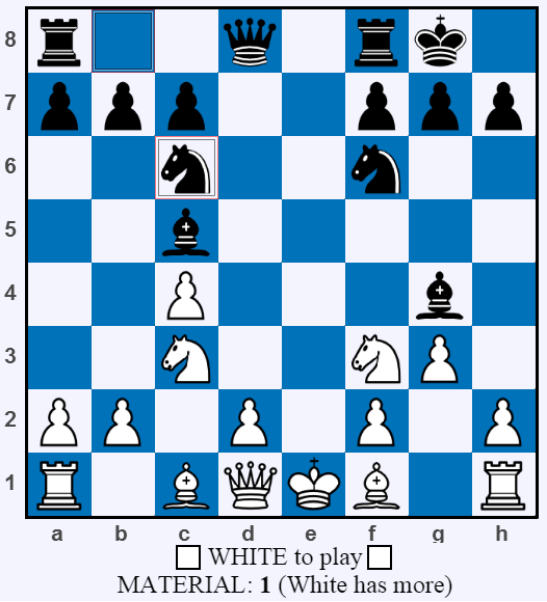
Worker #1 Game #2049	
Bet	Mat
-50	-1

Worker #2 Game #4183	
Bet	Mat
+45	-2

Worker #3 Game #256	
Bet	Mat
+30	2

Worker #4 Game #1710	
Bet	Mat
0	1

EXAMPLE GAME:



Material count = 1; engine score = –1.23

Expert's description of the game (not shown to participants): White has a pawn on the d file and Black does not, giving White a material advantage of one point (material count = 1). However, the engine score of -1.23 indicates that Black has a positional advantage equivalent to 2.23 pawns (engine score of -1.23 minus material count of 1), despite the fact that White has the opportunity to move next (WHITE to play). Black's positional advantage arises because White's King is exposed on the e file, White's Knight is pinned to the Queen (if it moves, Black will take White's Queen with his Bishop), Black has already castled his King to safety, moving his Rook toward the center, and Black has developed both Knights and Bishops, while both of White's Bishops have still yet to move.

Sample explanations by participants for the board shown above:

Condition and Phase	**Bet**	**Explanation**
Narrative Reporting (Agency Phase)	-31	Black seems to be winning. Even though White has one more pawn but Black's pieces are better developed and controlling the center of board with their bishops Bg4 & Bc5. Black King on g8 is well guarded by 3 pawns, rook n queen, While White King on e1 is lot more exposed. In fact Queen can give white king check with Q to e8
Narrative Note-taking (Agency Phase)	-15	White has a pawn advantage but no real development to speak of. Black is missing their central pawns but have a castled king and some development. The game is still early on but looks to largely be a toss up with the benefit of the doubt going to black as they have some initial tempo gainers against the white king.
Narrative Reporting (Post-Agency Phase)	-25	black's solid position should be able to carry the day, at least get the pawn back for a draw. Hope the computers are really good though...
Narrative Note-taking (Post-Agency Phase)	50	White will probably win. They have a minor material advantage although they are behind on development.

APPENDIX 2.C
Definitions

Task Definitions	
Variable/Term	**Definition**
Board	A board for an in-progress chess game. At the bottom of the board participants see whose turn it is to move and the material count. The games were begun by players with an ELO rating of at least 1400. At the board position shown, computers took over and finished the game. By design, none of the games ended in a draw.
Material Count (MC)	The amount of chess material that White has minus the amount of chess material that Black has. Material values follow the most common chess scoring, with Q=9, R=5, B/N=3, and P=1.
Engine Score (ES)	A sophisticated chess program simulates the next 10 possible moves to determine the true strength of position of the two sides. A positive engine score indicates that White has a stronger position (including both material advantages and non-material advantages) and therefore is more likely to win. Engine score is on the same scale as material.
Bet	On each board, participants can bet up to 50 cents on White (+50 bet) or 50 cents on Black (-50 bet) in 1 cent increments.
Phase	All participants make operational decisions for (i.e., place bets on) 30 boards. Throughout the paper (but not to participants), I refer to the first 10 boards as the pre-agency phase, the next 10 boards as the agency phase, and the final 10 boards as the post-agency phase. In the pre-agency phase and the post-agency phase, participants place bets on their own behalf. They personally receive the winnings/losses from the bets that they place. In the agency phase, participants act as agents, placing bets on behalf of a principal (referred to as their boss in the experiment). Their boss receives the winnings/losses from the bets. All participants see the same 30 boards, but the order of the boards is partially randomized. The 30 boards were created as 3 sets of 10 boards (set 1, set 2, and set 3). Each phase consists of one set, but which set a participant sees in each phase is randomized (e.g., a participant could see set 3 for the pre-agency phase, set 1 for the agency phase, and set 2 for the post-agency phase). Furthermore, within a phase the order of the boards is randomized. I hold constant the sets of boards to ensure that there is sufficient diversity in engine scores and material count within a phase to allow for useful inferences about participants' behavior.
High Material Board	In each set there are 5 pairs of boards. Every pair consists of two boards that have the same engine score but different material counts. Thus, the computer program says that the strength of White's position (including material) relative to Black's position (including material) is the same for the two boards. The High Material Board is the one with the higher (i.e., more favorable to White) material count.
Low Material Board	The Low Material Board is the one with a lower (i.e., less favorable to White) material count.

Statistical Definitions	
Material Bias	How much a participants' bets are biased by material count, after controlling for engine score. Equal to the sum of the bets (color-signed) on the High Material Boards minus the sum of the bets on the Low Material Boards, summed for each participant over all boards in a given phase.
Narrative Reporting Condition	Participants in this condition write an explanation for their bet and are told that their boss *will* see their explanation when assigning bonuses for agency-phase bets.
Narrative Note-taking Condition	Participants in this condition write an explanation for their bet, but are told that their boss *will not* see their explanation when assigning bonuses for agency-phase bets.
Preference for Consistency (PFC)	Measure of a participants' preference for cognitive consistency, adapted from Cialdini, et al. (1995). For use in testing, consistency = -1 if participants are at least one standard deviation below the mean, 1 if they are at least one standard deviation above the mean, and 0 otherwise. The questions are reproduced in Figure 2.3.
Explanation	For each board, participants "Write a few sentences indicating who you think is going to win, how sure you are, and why." I code each response 1 if the participant uses at least one of the following terms: develop, mobil, defen, offen, center, castle, threat, check, position, control, moment. Explanation represents the number of responses for which a participant uses at least one of these terms in a given phase (e.g. in the pre-agency phase), and is a count variable bounded at 0 and 10.

REFERENCES TO CHAPTER 2

Agarwal, R., A. Anand, P. Mohanty, and R.S. Shekhawat. 2009. *Comments on IASB's exposure draft on management commentary*. Bhubaneswar, India.

Baiman, S., and M.V. Rajan. 1995. The informational advantages of discretionary bonus schemes. *Accounting Review* 70(4): 557-79.

Bentley, J.W., R.J. Bloomfield, S. Davidai, and M.J. Ferguson. 2015. Drinking your own Kool-Aid: The role of beliefs, belief-revision, and meetings in persuasion. *Working Paper*.

Bhojraj, S., and R. Libby. 2005. Capital market pressure, disclosure frequency-induced earnings/cash flow conflict, and managerial myopia. *The Accounting Review* 80 (1): 1-20.

Bloomfield, R.J. 2012. A pragmatic approach to more efficient corporate disclosure. *Accounting Horizons* 26 (2): 357-70.

———. 2015. What counts and what gets counted. *Available at SSRN 2427106*.

Bol, J.C. 2008. Subjectivity in compensation contracting. *Journal of Accounting Literature* 27: 1-24.

Bonner, S.E., and G.B. Sprinkle. 2002. The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society* 27 (4): 303-45.

Brown, C. 1990. Firms' choice of method of pay. *Industrial and Labor Relations Review*: 165S-82S.

Campbell, D.T. 1979. Assessing the impact of planned social change. *Evaluation and Program Planning* 2 (1): 67-90.

Choi, J.W., G.W. Hecht, and W.B. Tayler. 2012. Lost in translation: The effects of incentive compensation on strategy surrogation. *The Accounting Review* 87 (4): 1135-63.

———. 2013. Strategy selection, surrogation, and strategic performance measurement systems. *Journal of Accounting Research* 51 (1): 105-33.

Cialdini, R.B., M.R. Trost, and J.T. Newsom. 1995. Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology* 69 (2): 318-28.

Clatworthy, M., and M.J. Jones. 2003. Financial reporting of good news and bad news: Evidence from accounting narratives. *Accounting and Business Research* 33 (3): 171-85.

Downs, J.S., M.B. Holbrook, S. Sheng, and L. F. Cranor. 2010. Are your participants gaming the system?: Screening mechanical turk workers. Paper presented at *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Evans III, J.H., R.L. Hannan, R. Krishnan, and D.V. Moser. 2001. Honesty in managerial reporting. *The Accounting Review* 76 (4): 537-59.

Festinger, L. 1957. *A cognitive theory of dissonance*. Evanston, IL: Row Petersen.

Forsythe, R., R. Lundholm, and T. Rietz. 1999. Cheap talk, fraud, and adverse selection in financial markets: Some experimental evidence. *Review of Financial Studies* 12 (3): 481-518.

Graham, J.R., C.R. Harvey, and S.Rajgopal. 2005. The economic implications of corporate financial reporting. *Journal of Accounting and Economics* 40, (1): 3-73.

Grenier, J.H., B. Pomeroy, and M. Stern. 2014. The effects of accounting standard precision, auditor task expertise, and judgment frameworks on audit firm litigation exposure. *Contemporary Accounting Research* 32 (1): 336-357.

Hannan, R.L., G.P. McPhee, A.H. Newman, and I.D. Tafkov. 2013. The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review* 88, (2): 553-75.

———. 2014. The effect of relative performance information temporal aggregation and detail level on effort allocation in a multi-task environment. *Available at SSRN 2481246*.

Hayes, A.F. 2013. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Holmstrom, B. 1979. Moral hazard and observability. *The Bell Journal of Economics* 10 (1): 74-91.

Holmstrom, B., and P. Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*: 7 (Sp): 24-52.

Horton, J.J., and L.B. Chilton. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*: 209-218.

International Accounting Standards Board. 2010. IFRS practice statement "Management commentary: A framework for presentation."

Ittner, C.D., and D.F. Larcker. 2003. Coming up short on nonfinancial performance measurement. *Harvard Business Review* 81 (11): 88-95.

Ittner, C.D., D.F. Larcker, and M.W. Meyer. 2003. Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. *The Accounting Review* 78 (3): 725-58.

Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3 (4): 305-60.

Kaplan, R. S., and D. P. Norton. 2000. Having trouble with your strategy? Then map it. *Harvard Business Review* 78 (5): 167-76.

Keusch, T., L.H.H. Bollen, and H.F.D. Hassink. 2012. Self-serving bias in annual report narratives: An empirical analysis of the impact of economic crises. *European Accounting Review* 21 (3): 623-48.

Kraft, A., R. Vashishtha, and M. Venkatachalam. 2014. Real effects of frequent financial reporting. *Working Paper*. City University London and Duke University.

Libby, R. 1981. *Accounting and human information processing: Theory and applications*. Englewood Cliffs, NJ: Prentice-Hall.

Libby, R., R.J. Bloomfield, and M.W. Nelson. 2002. Experimental research in financial accounting. *Accounting, Organizations and Society* 27 (8): 775-810.

McKenzie, D. 2012. Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics 99* (2): 210-221.

Nail, P.R., J.S. Correll, C.E. Drake, S.B. Glenn, G.M. Scott, and C. Stuckey. 2001. A validation study of the preference for consistency scale. *Personality and Individual Differences* 31 (7): 1193-202.

Paolacci, G., J. Chandler, and P.G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5 (5): 411-9.

Preacher, K.J., and A.F. Hayes. 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods* 40 (3): 879-91.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37 (1): 7-63.

Rajan, M.V., and S.Reichelstein. 2009. Objective versus subjective indicators of managerial performance. *The Accounting Review* 84 (1): 209-37.

Rennekamp, K. 2012. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research* 50 (5): 1319-54.

Tafkov, I.D. 2012. Private and public relative performance information under different compensation contracts. *The Accounting Review* 88 (1): 327-50.

Tayler, W.B. 2010. The balanced scorecard as a strategy-evaluation tool: The effects of implementation involvement and a causal-chain focus. *The Accounting Review* 85 (3): 1095-117.

CHAPTER 3

THE EFFECT OF MORAL LICENSING AND REPORTING STRUCTURE ON OPERATIONAL DISTORTION

Abstract:

Prior research has found that agents sometimes make operational decisions in a way that increases their own compensation at the expense of the firm, a phenomenon referred to as operational distortion (Bloomfield 2015). In a 2x2 between-subjects experiment using Amazon Mechanical Turk Workers I show that this tendency is more likely after people have done something out of the ordinary to help the firm. I also find that the relationship between extraordinary pro-firm actions and operational distortion is greater when the two actions are reported to a single supervisor than when the two actions are reported to different supervisors. I conclude this essay by proposing two follow-up experiments in which I plan to manipulate the way that an accounting system brackets performance: (1) at project ends or at regular time intervals regardless of project start/stop dates and (2) via report aggregation. The first of these new experiments provides an interesting differential prediction between the moral credits and moral credentials models of moral licensing as the former, but not the latter predicts that performance reporting on a project-oriented basis will reduce moral licensing relative to performance reporting at regular time intervals. Furthermore, both experiments, if successful, will inform practitioners of relatively simple methods to reduce moral licensing.

I. INTRODUCTION

When faced with an imperfect performance reporting system, agents sometimes take advantage of the imperfections to increase their own compensation at the firm's expense. One method agents sometimes employ is to make operational decisions that increase the reported performance rather than the underlying performance that the report is intended to capture. For example, a salesperson who is responsible for generating profitable sales, but is paid a large bonus based on sales volume, may provide lower margins than if he were paid a bonus based on the total profitability of his sales. Following Bloomfield (2015) I use the term *operational distortion* to refer to agent decisions that are intended to increase reported performance (sales volume) more than true performance (profitable sales). Prior research has found that executives state that their firms are likely to engage in operational distortion to meet earnings benchmarks (Graham, Harvey, and Rajgopal 2005) and that executives and lay people believe that operational distortion is a fairly ethical and acceptable activity (Bruns and Merchant 1990; Bentley, Bloomfield, Bloomfield, and Lambert 2015).

In this paper, I use the theory of moral licensing to predict that agents will be more likely to engage in operational distortion if they previously took exceptional actions to benefit the firm. I also predict that the relationship between pro-firm actions and operational distortion will be larger if a reporting system brackets the pro-firm action and the operational distortion opportunity together than if the reporting system separates the two actions. I argue that the reason why pro-firm actions lead to operational distortion occurs through two psychological processes: moral credits and moral credentials. The moral credits model states that individuals who do pro-firm actions feel that they deserve to distort operations. The moral credentials model states that individuals who do pro-firm actions feel that they and others will

favorably construe the operational distortion, interpreting the action as more-beneficial for the firm than if the individuals had not done pro-firm actions. These two processes results in a "moral license" to engage in operational distortion.

To test my theory, I have 168 Amazon Mechanical Turk workers play the role of a salesperson in a fictitious company. They are told that they work for two divisions of the company and report to two different bosses. They are then told about a need for out-of-the-ordinary work in one of the two divisions. I have participants write either what they would do to help the division or what they believe another employee (the HR manager) would do to help the division. I then tell them to imagine that they (or the HR manager) did those actions and the division benefitted as a result. Finally, I present them with an opportunity to increase their pay through operational distortion in either the same division as the extraordinary act or in the other division. Consistent with my predictions, I find that when the participant was responsible for the initial extraordinary act they are more likely to distort operations than if the HR manager was responsible for the extraordinary act. I also find that the relationship between responsibility for the extraordinary act and operational distortion is larger if the initial act and the operational distortion are in the same firm division than if they are in different firm divisions.

My results contribute to the accounting literature on operational distortion, the psychology literature on moral licensing, and the management literature on reporting structure. First, my paper contributes to the literature on operational distortion (a generalization of real earnings management) by showing that employees are more likely to engage in operational distortion when they have previously gone above and beyond their normal duties to help the firm, particularly when the operational distortion and prior helpful

behavior are bracketed together by the accounting system. Next, my results contribute to the psychology literature on moral licensing. Prior research has had participants make anti-racist, anti-sexist, or pro-environment decisions which then license their future racist, sexist, or wasteful behaviors. In contrast, I demonstrate a setting where an action that isn't inherently moral or immoral licenses operational distortion, suggesting that "moral" licensing can occur even when the licensing task doesn't involve a moral issue.

I also contribute to the psychology literature by showing that licensing is stronger when the licensing and licensed behaviors are more-closely linked in people's minds. I use the word domain to refer to the way in which actions are bracketed. Domains can vary across multiple dimensions such as type of action, temporal bracketing, who is affected, or to whom the actions are reported. Prior research in psychology has struggled to test if licensing is different within vs. between domains because it has focused on nature of action as the relevant method of bracketing performance. It is difficult to compare moral acts of different types. Racism, selfishness, sexism, etc. are fundamentally different concepts that cannot be objectively ranked, making it difficult to test if licensing varies based on a "type of action" domain. I overcome this challenge by holding constant the type of action domain (i.e. the operational distortion decision remains constant across settings) and instead varying domain by altering to whom the actions are reported. I find that reporting domain is an important moderator of moral licensing. Other dimensions of bracketing actions may also be important moderators. Finally, I contribute to the management literature on reporting structure by demonstrating a previously unknown benefit of matrix reporting and narrow bracketing of tasks. Specifically, in a matrix reporting structure an employee reports one type of action to

one supervisor and another type of action to another supervisor. This organizational design can reduce the likelihood that pro-firm actions lead to operational distortion.

The remainder of the paper proceeds as follows. Section 2 presents background research and develops my hypotheses. Section 3 describes my experimental methods. Section 4 presents my results. Section 5 presents a proposed second experiment, and Section 6 concludes.

II. BACKROUND AND HYPOTHESES

2.1 OPERATIONAL DISTORTION

Most of the time principals are unable to perfectly monitor agents' actions and resort to a second-best solution to the principal-agent problem (Jensen and Meckling 1976). Specifically, principals design incentive contracts using an imperfect measure of performance. Theoretical models (e.g. Holmstrom 1979; Holmstrom and Milgrom 1991) argue that agents will respond to imperfect incentive contracts by taking real actions intended to increase the measure of performance rather than the true performance that the measure imperfectly captures, a phenomenon that Bloomfield (2015) refers to as *operational distortion*. Prior research has found that many, but not all agents engage in operational distortion. For example, Graham, Harvey, and Rajgopal (2005) find that slightly more than half of surveyed financial executives said that they would be willing to delay starting a new project, even if the delay entailed a sacrifice in value, if doing so would cause them to hit a short-term earnings benchmark. Similarly, Bentley, et al. (2015) find that approximately half of New York State residents would advise a friend to offer a sales discount that hurts the firm if it would allow the friend to get a sales bonus. These results confirm that many agents are willing to distort operations in order to achieve personal gain but that others are reluctant to distort operations

for personal gain. Prendergast (1999) argues that firms can reduce operational distortion by using subjective rather than objective performance evaluations. While the literature has generally supported Prendergast's argument, operational distortion continues to exist even when firms follow Prendergast's recommendations (see Bol 2008 for a review of the literature), suggesting that further research is warranted to understand when and why agents engage in operational distortion.

2.2 MORAL LICENSING

One factor that may affect agents' propensity to engage in operational distortion is the history of their prior actions. Employee actions are not taken in isolation. Rather, they are a part of a series of actions that have helped or hurt the firm. The psychology literature on moral licensing (also referred to as self licensing) is relevant to this discussion. The term moral licensing refers to someone being more likely to engage in immoral behavior after engaging in moral behavior, all else equal. The moral licensing literature has proposed the moral credits and moral credential models as two related possible explanations for why initial moral actions lead to subsequent immoral actions (see e.g. Miller and Effron 2010 for a review of these two models). The purpose of this paper is not to distinguish between these models but rather to leverage both models to motivate my predictions. Below I briefly discuss the process theorized in each model and how that process supports my predictions.

The moral credits model states that individuals have a type of moral bank account. They get utility from having a sufficiently positive balance and from having others think that they have a sufficiently positive balance. Moral actions credit the account, increasing the balance, and providing a balance from which an individual can make subsequent withdrawals (debits). Thus, early moral actions lead to subsequent immoral actions because there are funds

available for withdrawal at the second decision point. If individuals bracket their decisions narrowly by thinking only about the decision and immediately-related decisions (e.g. Tversky and Kahneman 1981), then the moral credits model can be extended to allow individuals to have multiple moral bank accounts, one for each party with whom they interact or for each type of decision they make. Thus, an employee who takes unselfish actions that benefit Supervisor X credits the Supervisor X moral account and can later make withdrawals against the Supervisor X account by engaging in selfish actions when dealing with Supervisor X. However, under narrow bracketing the unselfish actions towards Supervisor X do not credit the Supervisor Y moral account and therefore will not make an employee more likely to engage in selfish actions when dealing with Supervisor Y. Similarly, Supervisor X cares more about the balance that the employee holds with him than about the balance that the employee holds with Supervisor Y.

In contrast, the moral credentials model states that moral licensing occurs not due to a positive moral balance, but rather because early moral actions allow actors and observers to construe subsequent ambiguous actions as more moral. For example, someone who votes for Obama has shown that he/she is not a racist, which provides credentials to show that hiring a white person over a black person is less likely due to race (Effron et al. 2009).[1] The moral credentials model predicts that early actions can only affect the construal of subsequent actions if the actions are perceived to be related. For example, anti-racism behavior provides

[1] The moral credits model argues that licensing occurs because the pro-black action (voting for Obama) cancels out the discrimination action (hiring a white person over a black person). The second action is still immoral, but the actor has earned the license to do the immoral action because of his previous good actions. In contrast, the moral credentials model argues that voting for Obama provides credentials such that hiring a white person over a black person no longer appears to be an immoral action – the action must have been taken because of the candidates' qualifications and not because of race, because the actor has already provided evidence they are not racist.

credentials for subsequent ambiguously racist actions, but not for subsequent ambiguously sexist actions. Similarly, an employee who engages in unselfish behavior towards Supervisor X establishes credentials for acting in the interests of Supervisor X, but not for acting in the interests of Supervisor Y. Thus, when the employee considers an ambiguously selfish action with regards to Supervisor X, he/she can take the action and (1) more favorably interpret the action him/herself and (2) be more confident that Supervisor X will interpret the ambiguity in a favorable manner (i.e. not accuse the employee of selfishness) whereas the employee does not have similar credentials and confidence with regards to Supervisor Y.

Key to this paper is that both the moral credits and the moral credentials models offer the same directional prediction that unselfish employee behavior will lead to subsequent operational distortion. Furthermore, both models also predict that the relation between initial unselfish behavior and subsequent operational distortion will be stronger when both actions are reported to a single supervisor than when they are reported to different supervisors. Thus, I make the following predictions:[2]

H1: Participants who previously took a pro-firm action will engage in more operational distortion than participants who did not take a pro-firm action.

H2: Participants who previously took a pro-firm action will engage in even more operational distortion if the operational distortion decision is reported to the same supervisor who observed the pro-firm action than if it is reported to a different supervisor, as compared to participants who did not take a pro-firm action.

[2] I test only the effect of positive performance on future operational distortion. It is possible that negative initial performance may reduce the likelihood of future operational distortion. However, I think the latter effect would be smaller and less likely. Most research on biased processing has found that individuals interpret information in a way that helps them achieve a desired outcome. In the case of operational distortion, the desired outcome is to be able to improve compensation through operational distortion. Thus, participants will search for reasons allowing them to distort operations rather than reasons why they shouldn't distort operations. I leave it to future research to test these conjectures.

An alternative theory to moral licensing relies on motivated reasoning, preference for consistency, and escalation of commitment. An agent who takes initial pro-firm actions may try to rationalize this unselfish action by convincing him/herself that the principal is deserving of help. The agent would then be less selfish in a subsequent action as they try to be consistent with their prior actions. This theory works against my hypotheses. As discussed but untested in Miller and Effron (2010), ambiguity in the second act is likely to decrease the extent of motivated reasoning because there is less need to be consistent between the initial and subsequent actions, but increase the extent of moral licensing because there is more room to favorably interpret the motivation behind the second action. In my setting, I examine a relatively ambiguous type of measure management: operational distortion (i.e. making operational decisions, such as deep sales discounts, which improve reported performance but harm true performance). Future research could test if these effects differ for a less ambiguous type of measure management: reporting distortion (i.e. making reporting decisions, such as underreporting expenses, which improve reported performance but do not affect true performance).[3]

III. METHODS

3.1 OVERVIEW

To test my predictions, I conduct a 2x2x2 between-subjects experiment manipulating whether or not the participant feels responsible for the initial pro-firm behavior, whether the initial pro-firm behavior occurs in the shoes division or the tools division of the firm, and whether the opportunity to distort operations occurs in the shoes division or the tools division

[3] See Bentley, et al. (2015) for a discussion of distinctions between operational distortion and reporting distortion.

of the firm. I collapse the final two manipulations to form one of my variables of interest: the operational distortion opportunity is in the same division or a different division as the initial pro-firm action.

3.2 RECRUITING PROCEDURES

The experiment was conducted via Amazon Mechanical Turk (AMT). AMT workers were allowed to participate in the survey if they had a human intelligence task (HIT) approval rate of at least 98%.[4] Participants were paid $0.50 for a task that was advertised as less than 10 minutes. The mean (median) time spent on the task was 8.9 (7.0) minutes for an effective hourly wage of $3.39 ($4.31) per hour, which is well above the median reservation wage reported in Horton and Chilton (2010) of $1.38 per hour and is comparable to the hourly wages reported in other AMT studies (e.g., Rennekamp 2012; Grenier, Pomeroy, and Stern 2014; Paolacci, Chandler, and Ipeirotis 2010).

3.3 EXPERIMENTAL DETAILS

Participants accepted the HIT and then took the experiment in Qualtrics. Participants were told to imagine that they were a salesperson for a company that makes tools and shoes. They were told that they reported to each division manager separately and that the two division managers rarely interacted and never talked about sales personnel. Participants then saw the first task which occurred in either the tools division or the shoes division. Participants were explicitly told which division the second task occurred in and a picture of tools or a shoe appeared on the page to make salient the division. They were asked to imagine what they or the HR manager would do in the following situation:

[4] Peer, Vosgerau, and Acquisti (2014) suggest that HIT approval rate is the best possible method for screening out inattentive workers.

> Imagine that the [tools | shoes] division is trying to get a new supplier. The [tools | shoes] manager has found a supplier that has better products at a lower price than the old supplier. However, the normal supplier normally only ships very large orders. Your firm places small orders, so you probably couldn't make a deal without any personal connections. However {you | the HR manager} went to school with the owner of the supply company. The [tools | shoes] manager asked {you |the HR manager} to use {your | his} connection to help convince the supplier to sell you supplies, even though this is outside of the scope of {your | the HR manager's} normal duties.
>
> Please take a few minutes to write what activities {you | the HR manager} will do to help convince the supplier.
>
> Is this in {your |the HR manager's} job description? What extra work is involved?
>
> How do these extra actions help the [tools | shoes] division?

Participants were then told that they (or the HR manager) had done a great job and the division had secured a long-term relationship that would save the division money and improve quality. Thus, my manipulation involved a "forced compliance" paradigm (e.g. Festinger and Carlsmith 1959) to avoid self-selection by participants and I was able to manipulate whether or not the participant felt responsible for engaging in an out-of-the ordinary act that benefitted the firm.

Participants then moved on to the second task. Independent of the division in which the first task occurred, the second task occurred in either the tools division or the shoes division. Participants were explicitly told which division the second task occurred in and a picture of tools or a shoe appeared on the page to make salient the division. In this task, participants were told the following:

> You have worked hard, but the poor economy has left you just short of your sales goal this year. It looks like you won't get the [tools | shoes] sales bonus.

> A customer placed an order large enough for you to hit your goal. Unfortunately, they requested that it be shipped the first of next year, so it won't count for this year's sales goal.
>
> You've asked the customer if they would be willing to move this one order up. They said that they will if you give them a large enough discount, which they understand would be a one-time discount applying only to this order.
>
> It's common to offer discounts of 5-10% for large orders and 10-15% for charities. The firm makes 17% profit on the retail price of the order.
>
> The customer has proposed that you give a 15% discount on their [tools | shoes] order if they accept early delivery. How likely are you to accept the customer's proposal?

Participants then responded on a 7-point Likert scale from "Very Unlikely" to "Very Likely."

After stating how likely they were to accept the customer's proposal, participants answered questions designed to measure their degree of moral credits and moral credentials. They then answered attention check questions and demographic variables.

IV. RESULTS

4.1 SAMPLE SELECTION

Two hundred AMT workers participated in my survey. Chandler, Mueller, and Paolacci (2014) express a concern about non-naïveté among AMT participants. To address their concern, I exclude 18 participants (9%) who in debriefing state that they could explain the concept of moral/self licensing.[5] I also exclude 14 participants (7%) who failed one or

[5] If I include these participants the test of H1 is marginally significant (β=0.35, t=1.50, one-tailed p=0.07) and the test of H2 is also marginally significant (β=0.60, t=1.27, one-tailed p=0.10). The simple effect of responsibility when the two activities are in the same division is significant (one-tailed p=0.03) but no other simple effects are significant.

more attention checks following recommendations by Downs, et al. (2010). [6,7] This level of attention-check failures is consistent with or lower than prior AMT research.[8] Goodman, Cryder, and Cheema (2013) express concern about the English language ability of AMT workers and suggest that many AMT workers lack the English ability to understand instructions and manipulations. Six participants self-report as not being native English speakers.[9] I exclude them from my main analyses. Results including these participants and language dummy variables are shown in section 4.4. My final sample consists of 162 participants. **Table 3.1** presents the number of participants in each cell (Panel A), the total number of participants excluded in each cell (Panel B), the number of participants who are non-native English speakers in each cell (Panel C, also included in the number in Panel B), and the average likelihood of offering a discount in each cell (Panel D).

[6] If I include these participants the test of H1 is marginally significant (β=0.32, t=1.38, one-tailed p=0.09) and the test of H2 is also marginally significant (β=0.60, t=1.30, one-tailed p=0.10). The simple effect of responsibility when the two activities are in the same division is significant (one-tailed p=0.04) but no other simple effects are significant.

[7] I originally had 7 attention check questions, 5 early in the experiment and 2 late in the experiment and intended to exclude any participant who failed one or more question. Four questions followed the same format: "If you sell some [tools | shoes] it will increase, decrease, or have no effect on your chance of getting a [tools | shoes] bonus." I had intended for the correct answer to be that selling tools would increase the chance of getting a tools bonus and have no effect on the chance of getting a shoes bonus, and vis-à-vis. However, one of the respondents pointed out that the correct answer depends on the relationship between the sales operations. If some customers purchase both products at the same time, then a tools sale may cause a shoe sale which would increase the chance of both bonuses. In contrast, if a worker's time is limited and customers don't purchase both types of products, then a tools sale means less time is spent selling shoes which decreases the chance of a shoes bonus. Due to these unanticipated, but technically correct answers, I do not count these questions as attention check questions for purposes of exclusions. However, I do exclude any participant who misses one or more of the three remaining attention check questions following recommendations by Oppenheiemer, et al. (2009) and others (e.g. Downs, et al. 2010, Goodman, et al. 2013).

[8] For example, Kaufmann, et al. (2011) find that over 30% of participants fail attention-check questions. Downs, et al. (2010) find that 39% of participants fail one or more difficult attention-check questions. Peer, et al. (2014) find that 16.7% of participants with high approval ratings and high productivity fail one or more attention check questions. Thus, my attention-check questions may have been easier than average. Failing to exclude inattentive participants should bias against me finding results.

[9] Native languages of these participants were Chinese (N=1), Hindi (N=1), Sourastra (N=1), Laotian (N=1), and Tamil (N=2).

TABLE 3.1 – Descriptive Statistics

Panel A: Final Sample Size

		Responsibility: Was the Participant or the HR Manager Responsible for the "Extra Mile" Performance in the First Task?					
		HR Manager			Participant		
		Discount in the Shoes Division	Discount in the Tools Division	Total	Discount in the Shoes Division	Discount in the Tools Division	Total
SameDivision: Was the Second Task in the Same Division as the First Task?	No	20	20	40	22	22	44
	Yes	23	17	40	18	20	38
	Total	43	37	80	40	42	82

Panel B: Sample Exclusions by Condition

		Responsibility: Was the Participant or the HR Manager Responsible for the "Extra Mile" Performance in the First Task?					
		HR Manager			Participant		
		Discount in the Shoes Division	Discount in the Tools Division	Total	Discount in the Shoes Division	Discount in the Tools Division	Total
SameDivision: Was the Second Task in the Same Division as the First Task?	No	6	2	8	2	5	7
	Yes	7	7	14	4	5	9
	Total	13	9	22	6	10	16

Panel C: Non-Native English Speakers by Condition

		Responsibility: Was the Participant or the HR Manager Responsible for the "Extra Mile" Performance in the First Task?					
		HR Manager			Participant		
		Discount in the Shoes Division	Discount in the Tools Division	Total	Discount in the Shoes Division	Discount in the Tools Division	Total
SameDivision: Was the Second Task in the Same Division as the First Task?	No	1	0	1	0	2	2
	Yes	1	0	1	1	1	2
	Total	2	0	2	1	3	4

Panel D: Mean (Standard Deviation) Likelihood of Offering a Product Discount

		Responsibility: Was the Participant or the HR Manager Responsible for the "Extra Mile" Performance in the First Task?					
		HR Manager			Participant		
		Discount in the Shoes Division	Discount in the Tools Division	Total	Discount in the Shoes Division	Discount in the Tools Division	Total
SameDivision: Was the Second Task in the Same Division as the First Task?	No	4.45 (1.64)	4.15 (1.60)	4.30 (1.60)	4.50 (1.41)	4.09 (1.60)	4.30 (1.51)
	Yes	3.83 (1.44)	3.76 (1.60)	3.80 (1.49)	5.06 (1.35)	4.10 (1.33)	4.55 (1.41)
	Total	4.12 (1.55)	3.97 (1.59)	4.05 (1.56)	4.75 (1.39)	4.10 (1.46)	4.41 (1.46)

4.2 HYPOTHESIS TESTS

I test both hypotheses using an ANCOVA. My independent variables are (1) whether the participant or the HR manager was responsible for securing the supplier relationship (*Responsibility*), (2) whether the discount task was in the same division or a different division from the supplier task (*SameDivision*), and (3) an interaction between *Responsibility* and *SameDivision*. I include a dummy variable (*Tools*) for the division of the discount task (tools vs. shoes) as it is marginally significant in the model for an unknown reason (F= 3.28, two-tailed p=0.07).[10] Tools does not have a significant interaction with any other variables (all p>0.25) so interactions are excluded from the model. All variables are mean-centered such that the variable=0.5 when at the high level (participant was responsible; tasks were in the same division; discount opportunity in the tools division) and -0.5 when at the low level (HR manager was responsible; tasks were in different divisions; discount opportunity in the shoes division). **Figure 3.1** presents my results graphically, and **Table 3.2** presents the results of the ANCOVA.

[10] Participants are less willing to offer a tools discount than a shoes discount. My post-hoc reason for this difference is that participants may think that a shoes discount is more common than a tools discount due to personal experience. For example, they may be familiar with significant sales at retail stores (e.g. JCPenney and Kohls regularly offer storewide discounts of 30% or more) but not at hardware stores (e.g. Lowes and Home Depot rarely offer large storewide discounts).

FIGURE 3.1: The Effect of Responsibility and Reporting Domain on Participants' Likelihood of Offering a Sales Discount

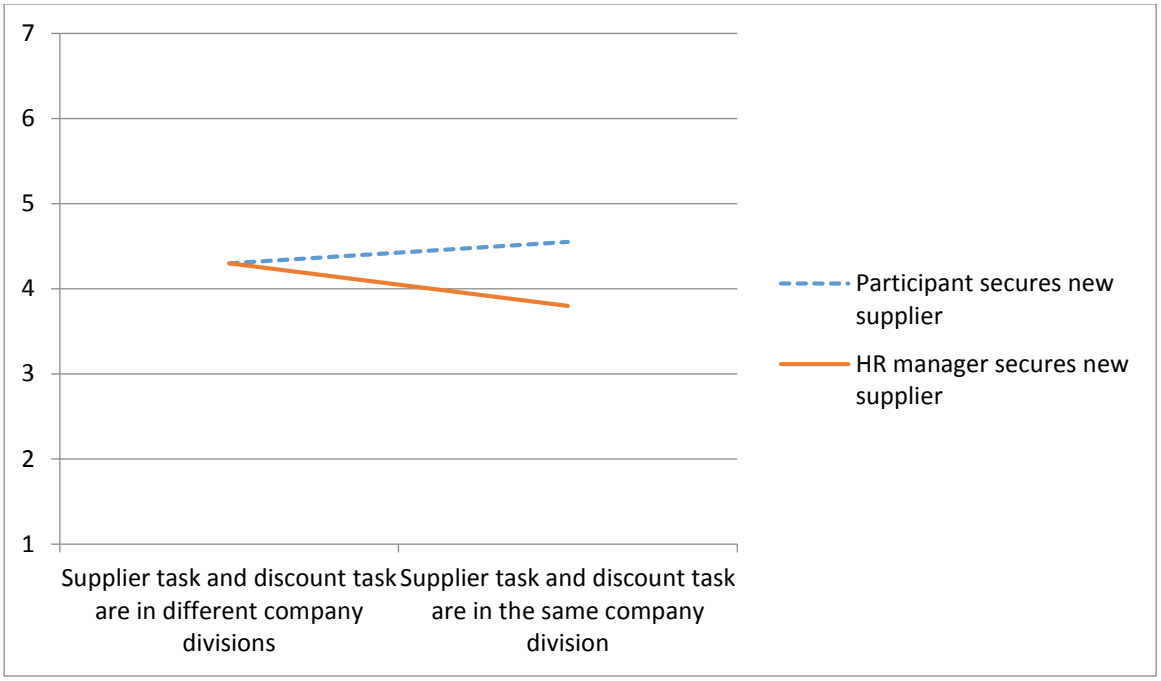


Figure 3.1 shows the average likelihood that a participant will give a discount without controlling for which division the discount is in.

TABLE 3.2: The Effect of Responsibility and Reporting Domain on Participants' Likelihood of Offering a Discount

	Likelihood of Offering Discount				
Source	**DF**	**Regression Parameter**	**Mean Square**	**F Statistic**	**Two-tailed *p*-value**
Intercept	1	4.23			
Responsibility	1	0.40	6.31	2.83	0.09
SameDivision	1	-0.13	0.70	0.31	0.58
Responsibility*SameDivision	1	0.80	6.45	2.89	0.09
Tools	1	-0.43	7.31	3.28	0.07
Error	157		SS= 350.04		
Hypothesis Test					
H1:*Responsibility*>0		β= 0.40	t= 1.68	p= 0.047	
H2:*Responsibility*SameDivision*>0		β= 0.80	t= 1.70	p= 0.045	
Simple Effects					
Responsibility when the two tasks are in the same domain		β= 0.75	t= 2.21	**p = 0.01**	
Responsibility when the two tasks are in different domains		β= 0.00	t= 0.00	p = 0.99	
SameDivision when the participant is responsible for the initial task		β= 0.26	t= 0.77	p = 0.44	
SameDivision when the HR manager is responsible for the initial task		β= -0.50	t= -1.49	p = 0.14	

Hypothesis tests and the bolded simple effect have one-tailed p-values reflecting directional predictions. Simple effects are shown without controlling for the control variables. Inferences are similar when looking at the marginal effects, adjusting for the division dummy variable – the effect of responsibility when the two tasks are in the same domain is significant at the p=0.01 level (one-tailed) while none of the other effects are significant at the p<0.20 level (two-tailed).

In support of H1, I find a significant main effect of Responsibility. Participants are more likely to distort operations (i.e., give a client a discount for accepting early delivery) when they were responsible for the initial pro-firm action than when the HR manager was responsible for the initial pro-firm action (β=0.40, one-tail $p<0.05$). H2 is also supported. The effect of responsibility for the initial pro-firm action is stronger when the initial action and the discount happen in the same division of the firm than when they happen in different divisions of the firm (interaction β=0.80, one-tail $p<0.05$). Simple effects tests reveal a significant effect of responsibility when the initial pro-firm action and the discount opportunity are in the same division of the firm (β=0.75, one-tail $p=0.01$), but no effect when the two actions are in different divisions of the firm (β=0.00, ns).

4.3 MEDIATION ANALYSIS

My theory predicts that responsibility for initial pro-firm actions increases a participant's likelihood of offering a discount, particularly if the discount is in the same division as the initial action, because the initial pro-firm action gives the participant moral credentials and/or moral credits. The moral credentials argument suggests that when a participant helps the firm in the initial task he has gained a pro-firm credential which makes it so that he and the supervisor will construe a future sales discount as being in the best interest of the firm. To test for this psychological process, I included two debriefing questions in my experiment, reproduced as the first two questions in **Figure 3.2**. The first question seeks to assess the degree to which the pro-firm behavior provides credentials in the participant's own mind (i.e. credentials as seen by the self) while the second question seeks to assess the degree to which the participant believes that the pro-firm behavior establishes credentials for the supervisor (i.e. credentials as seen by the supervisor). The moral credits argument suggests

that a participant who has helped a division has accrued a credit balance with the division against which he can make a withdrawal in the form of a selfish action. To test for this psychological process I include two additional debriefing questions, reproduced as the third and fourth questions in Figure 3.2. The first of these seeks to assess the degree to which the pro-firm behavior provides credits in the participant's own mind (i.e. credits as seen by the self) while the second of these seeks to assess the degree to which the participant believes that the pro-firm behavior establishes credits for the supervisor (i.e. credits as seen by the supervisor). Finally, I ask a fifth debriefing question to test for the alternative hypothesis that initial pro-firm behavior will result in an escalation of commitment.

FIGURE 3.2: Mediating Questions

Question (Psychological process to be assessed)	**Mean**	**St. Dev**
Offering a discount is in the best interest of the firm. (credentials – self)	3.57	1.74
If the [tools | shoes] manager were to find out that I offered a discount, he would assume that my actions were in the best interest of the firm. (credentials – supervisor)	3.82	1.61
Because of the hard work I've done for the firm in the past, I deserve a [tools | shoes] bonus. (credits – self)	5.03	1.45
If I slacked off a little bit, the [tools | shoes] manager wouldn't be mad because of the hard work I've done in the past. (credits – supervisor)	3.22	1.34
The [tools | shoes] manager deserves my hard work. (escalation of commitment)	5.36	1.23

The words in parentheses were not included in the instrument but rather are presented here to show the process that the question was intended to capture. The [tools | shoes] brackets were replaced with the name of the division for which the participant had the opportunity to provide a discount.

TABLE 3.3: The Effect of Responsibility and Reporting Domain on Credentials and Credits

Panel A: Mediating Questions						
	Responsibility		**SameDivision**		**Interaction**	
	β	**p-val**	**β**	**p-val**	**β**	**p-val**
Q1: Credentials – Self	0.36	0.19	-0.23	0.40	0.60	0.27
Q2: Credentials – Other	0.49	0.05	-0.09	0.73	0.06	0.91
Q3: Credits – Self	0.61	0.01	-0.25	0.27	0.73	0.11
Q4: Credits – Other	0.31	0.14	-0.22	0.31	0.35	0.40
Q5: Escalation of Commitment	0.05	0.78	0.12	0.53	0.02	0.95
Panel B: Aggregate Variables						
	Responsibility		**SameDivision**		**Interaction**	
	β	**p-val**	**β**	**p-val**	**β**	**p-val**
(Q1+Q3)/2 – Average of Self	0.48	0.02	-0.24	0.23	0.67	0.09
(Q2+Q4)/2 – Average of Other	0.40	0.03	-0.15	0.42	0.20	0.58
(Q1+Q2)/2 – Average of Credentials	0.43	0.07	-0.16	0.49	0.33	0.48
(Q3+Q4)/2 – Average of Credits	0.46	0.01	-0.23	0.19	0.54	0.13
(Q1+Q2+Q3+Q4)/4 – Overall Licensing	0.44	0.01	-0.20	0.25	0.44	0.20

All p-values are two-tailed with no adjustment for multiple comparisons.

I use regressions to test if responsibility for the pro-firm action is a significant predictor of these five mediating variables and if the effect is larger when the pro-firm action is in the same domain as the discount opportunity. **Table 3.3** presents the results of these regressions. In each regression the independent variables are the same as those in Table 3.2: Responsibility, SameDivision, Responsibility*SameDivision, and suppressed controls for language and whether the discount opportunity was in the shoes or tools domain. In Panel A the dependent variable for each regression is the question itself, which was on a scale of 1 to 7. I find that participants feel more credentials and credits when they were responsible for the initial pro-firm action than when the HR manager was responsible. Two-tail p-values range from 0.01 to 0.19. There is also weak evidence that the effect of responsibility on credentials and credits is larger when the initial pro-firm action and the discount opportunity are in the

same division. The interaction p-value is marginally significant and positive for the Credits-Self question (two-tailed p=0.11). The interaction p-value is insignificant for the other three variables (all p>0.27). Neither Responsibility, SameDivision, nor their interaction have a significant effect on the Escalation of Commitment question.

Panel B presents results when the variables are combined according to question type. Responsibility for the pro-firm action increases the degree to which participants feel licensed for themselves (p=0.02) and the degree to which they believe that others will license them (p=0.03). Responsibility has an effect on both credentials (p=0.07) and credits (p=0.01). Responsibility for the pro-firm action appears to have a marginally larger effect on participants' self-licensing and moral credits when the pro-firm action and discount opportunity are in the same domain than when they are in different domains (self p-value=0.09; credits p-value=0.13). One limitation of these analyses is that all four questions were asked on the same page using the same format. If participants suffered from experimental fatigue or had carry-through effects they may have answered the questions in a correlated fashion. Untabulated analysis finds a significant positive correlation between the first four variables. Correlations range from 0.24 (the correlation between the 2nd and 3rd questions) to 0.57 (the correlation between the 1st and 2nd questions), and all correlations are significant at the p<0.01 level.

The significant correlations make it difficult to statistically distinguish between the various psychological processes that the questions are intended to capture. As such, I test if the average of the four variables mediates the relationship between Responsibility and the Responsibility*SameDivision interaction on participants' likelihood of offering a discount. **Table 3.4** conducts the same regression as presented in Table 3.2, but includes *License* as a covariate.[11] License is the average of the four mediating questions. I find that License is a significant predictor of participants' likelihood of offering a discount (β=0.76, p<0.0001). Furthermore, neither Responsibility nor the Responsibility*SameDivision interaction is significant after controlling for License (p=0.79 and p=0.24, respectively).

TABLE 3.4: The Role of Moral Licensing on Likelihood of Offering a Discount

	Likelihood of Offering Discount				
Source	**DF**	**Regression Parameter**	**Mean Square**	**F Statistic**	**Two-tailed *p*-value**
Intercept	1	1.23			
Responsibility	1	0.05	0.11	0.07	0.79
SameDivision	1	0.02	0.01	0.01	0.93
Responsibility*SameDivision	1	0.47	2.16	1.38	0.24
License	1	0.77	106.99	68.67	<0.0001
Tools	1	-0.29	3.34	2.15	0.14
Error	156		SS= 243.05		

[11] An extra sum of squares (Chow) F-test finds that a model that includes all four mediators separately does not perform better than a model that includes just the average of the four questions (F=1.24; Deg. of Freedom=3,153; p=0.30).

To formally test for mediation, I use bootstrap analysis following Hayes (2013). Bootstrap analysis involves resampling the data multiple times with replacement and performing the specified set of regressions for each sample. In each repetition, the analysis tests whether License mediates the relationship between the manipulated variables and participants' likelihood of offering a sales discount. A bootstrap analysis is superior to the traditional Sobel test because it does not make distributional assumptions about the data and it is less-easily influenced by outliers (Preacher and Hayes 2008). Furthermore, bootstrap analysis allows me to test the more-complex mediated moderation model rather than just a simple mediation model.

I first conduct 100,000 bootstrap repetitions testing if License mediates the main effect of Responsibility, with Tools included as covariates but without the SameDivision variable or Responsibility*SameDivision interaction. The model confirms that License mediates the relationship between Responsibility and Discount ($p<0.01$). When participants feel responsible for the initial pro-firm action they feel a greater moral license and that moral license makes them more willing to distort operations. I next conduct 100,000 bootstrap repetitions testing if License mediates the interaction between Responsibility and SameDivision. In 90% of repetitions the effect of Responsibility on Discount through License is larger when the initial pro-firm task and the discount task are in the same division of the firm than when they are in different divisions of the firm, suggesting that mediated moderation is significant at the $p<0.10$ (one-tail) level. Thus, there is marginally significant evidence that responsibility for a pro-firm action increases participants' moral license for the discount more if the discount is in the same division as the initial pro-firm action than if the

discount is in a different division. The moral license then drives the difference in likelihood of offering a discount.

4.4 ROBUSTNESS TESTS

In this section I briefly discuss my decision to exclude participants who were not native English speakers and the effect of said decision on my results. At the essence of my task is the notion that offering a discount isn't in the company's best-interest. The experimental materials provide examples of common discount levels. However, individuals may have pre-conceived ideas about what is an appropriate discount. The appropriateness and levels of haggled discounts varies by cultures. Many travel books discuss approaches and discount levels in different cultures.[12] Native language, then, may be a good proxy for participants' understanding of socially-acceptable levels of bargaining. Of my N=168 sample, six participants had a non-English native language: two spoke Tamil, and one each spoke Hindi, Chinese, Laotian, and Saurashtra.[13] In Panel A of **Table 3.5** I reproduce my main results but include the six participants who don't natively speak English along with dummy variables for the five native languages. The effects and significance of these observations are of little individual meaning as they are driven by one or two observations each. In Panel B I show the discount-likelihood for each of these six participants. Untabulated results find that the average discount likelihood by native English speakers is 4.23 while the average discount likelihood by non-English speakers, all of whose native language is predominately spoken in Asia, is 5.33. The difference between the average likelihood of these two groups is marginally

[12] See e.g. http://wikitravel.org/en/How_to_haggle, or http://www.business2community.com/travel-leisure/the-cultural-differences-in-haggling-0352522 for examples.

[13] The participant said that his/her native language was sourastra, which I assume is an alternative spelling or misspelling of Saurashtra, a language spoken but rarely written in southern India.

significant (t=1.75, two-tail p=0.08; Wilcoxon Z=1.77, two-tail p=0.08). Thus, there is some evidence that native language is correlated with how likely a participant is to offer a discount.

TABLE 3.5: Language and Likelihood of Offering a Discount

Panel A: Regression Results

	Likelihood of Offering Discount				
Source	**DF**	**Regression Parameter**	**Mean Square**	**F Statistic**	**Two-tailed *p*-value**
Intercept	1	4.02			
Responsibility	1	0.40	6.32	2.85	0.09
SameDivision	1	-0.13	0.67	0.30	0.58
Responsibility*SameDivision	1	0.81	6.66	3.01	0.08
Tools	1	-0.43	7.53	3.40	0.07
Sourastra (sic)	1	0.92	0.83	0.37	0.54
Chinese	1	3.02	8.84	3.99	0.05
Hindi	1	-1.51	2.23	1.00	0.32
Tamil	1	1.07	2.24	1.01	0.32
Laotian	1	1.65	2.63	1.19	0.28
Error	158		SS= 350.09		
Hypothesis Test					
H1:*Responsibility*>0		β= 0.40	t= 1.69	p= 0.047	
H2:*Responsibility*SameDivision*>0		β= 0.81	t= 1.73	p= 0.042	

Hypothesis tests have one-tailed p-values reflecting directional predictions.

Panel B: Individual Responses

Language	Discount
Chinese	7
Hindi	3
Tamil	6
Tamil	5
Laotian	6
Sourastra	5

V. PROPOSED SECOND EXPERIMENT

One limitation of my first experiment is that it brackets performance in only a single way: to whom actions are reported. I find that moral licensing is more likely to occur when the licensing and licensed actions are reported to the same person in the same firm division. The goal of my proposed second experiment is to extend this research by using an alternative bracketing mechanism. Below I briefly discuss two possible bracketing mechanisms.

5.2 TEMPORAL BRACKETING

One possible way a reporting system can bracket performance is to do so in a temporal fashion. Performance evaluations can be conducted on more or less frequent bases (e.g. montly vs. quarterly vs. annual performance evaluations). I propose to conduct an experiment manipulating how frequently performance evaluations occur. I predict that good performance in an initial task leads people to feel more licensed before the next performance evaluation than after the next performance evaluation. Thus, reporting systems that have more-frequent performance evaluations may see less moral licensing.

This extension has two major advantages. First, it may be easier for some firms to alter how frequently performance evaluations occur than to whom different activities are reported (i.e. the manipulation in my first experiment). Thus, this new experiment may provide insights that can be more easily incorporated into practice. Second, this new experiment may provide a theoretical contribution by helping tease apart the moral credits and moral credentials models. Under the moral credits model, performance evaluations could serve to zero out people's moral bank accounts. Thus, good performance only provides a positive net balance until a withdrawal is made, either by choice (i.e. the licensed behavior) or by force (i.e. the performance evaluation). In contrast, the moral credentials model suggests that credentials persist across performance evaluations. The positive credentials that someone establishes in one time period are not nullified but rather may be formalized or reinforced by a performance evaluation. Thus, the moral credits and moral credentials models make different predictions for frequent vs. infrequent performance evaluations which may help us distinguish between the two models empirically.

5.3 TASK BRACKETING

Another possible extension is to manipulate how tasks are bracketed in an accounting report. Practitioners use tools like the Balanced Scorecard (Kaplan and Norton 1992) or dashboards (Bloomfield 2015) to aggregate and display performance metrics in a coherent fashion. I predict that participants will feel greater license to distort operations if the performance reporting system aggregates the licensing and licensed behavior together than if it reports them separately. This experiment also represents a practical method for practitioners to reduce moral licensing as they may have more control over how information is aggregated in performance reporting than they do over who conducts performance evaluations.

VI. CONCLUSIONS

Prior research has found that most financial executives (Graham, et al. 2005) and US residents (Bentley, et al. 2015a) are willing to accept actions that sacrifice firm value in order to improve their own welfare, actions that Bloomfield (2015) refers to as operational distortion. However, little is known about what circumstances encourage or discourage operational distortion. In this paper, I present theory and evidence to suggest that people are more likely to engage in operational distortion after they have "gone the extra mile" in helping the firm. In my experiment, participants are told to imagine that either they or the HR manager went above and beyond the normal line of duty in order to help secure a contract with a new supplier, where the contract saves the firm money and improves the quality of the firm's product. I find that when participants feel responsible for securing the new supplier they feel that they have proved their worth to the firm (i.e., moral credentials) and that they deserve a bonus (i.e., moral credits). Together, the credentials and credits license participants to offer a sales discount that improves their own welfare but hurts firm value. I also manipulate whether the sales discount opportunity is in the same vs. different firm division and reported to the same vs. different supervisor as the initial action (securing a new supplier). I find that responsibility for the initial pro-firm action has a larger effect on participants' likelihood of offering a discount if the two tasks occur in the same firm division with the same supervisor than if they occur in different firm divisions with different supervisors. I also propose two follow-up experiments with alternative manipulations of performance bracketing. If the experiments yield the results I predict, the combination of the three experiments will provide evidence that firms may be able to reduce operational distortion by creating accounting systems that narrowly bracket performance.

This study is subject to several limitations. First, all decisions were hypothetical and did not involve real incentives. The hypothetical nature of the task may reduce my observed effect because participants only imagined what they would do to help the firm rather than actually performing said actions. The lack of real incentives could result in a demand effect to the extent that individuals guessed my hypotheses and tried to do what was expected of them. On the other hand, the lack of real incentives may reduce individual's incentives to say that they would distort operations and thereby weaken my effects. In future research I would like to replicate these results with real-world incentives. Second, I used AMT participants who may have been familiar with the psychological phenomenon being studied. Chandler, et al. (2014) document that AMT workers are often very familiar with common experimental paradigms and may be able to guess the hypotheses. I used an abbreviated funneled debriefing technique to see if any participants guessed my hypotheses. None correctly guessed that I was manipulating reporting domain. However, a large number thought that the two tasks were not independent (e.g. guessing that they would be more likely to offer a discount if the firm was doing well than if it was doing poorly) or stated that they could explain the moral/self licensing theory. Future research could replicate these findings with more naïve participants. Another limitation of this study is that I rely on a single manipulation of reporting domain to test the theory that the domain of performance moderates moral licensing. I show that moral licensing is greater if the licensing and licensed tasks are both reported to the same supervisor. Future research could try alternative methods of bracketing performance. For example, researchers could manipulate the type of action (e.g. reporting expenses, reporting time, offering a discount) to see if licensing is greater within a single domain than across domains.

APPENDIX TO CHAPTER 3
Experimental Instrument

RECRUITING MATERIALS

About this study: In this HIT, you will pretend that you are a salesperson for a company that makes tools and shoes. You will make a series of decisions in this job. The HIT pays $0.50 and takes 5-10 minutes.

Risks and Benefits: I do not anticipate any risks greater than everyday use of the internet. There are no direct benefits other than being paid and having fun.

If you have questions: The researcher conducting this study is Jeremiah Bentley. You may contact him at jwb282@cornell.edu. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) for Human Participants at 607-255-5138 or access their website at http://www.irb.cornell.edu. You may also report your concerns or complaints anonymously through Ethicspoint online at www.hotline.cornell.edu or by calling toll free at 1-866-293-3077. Ethicspoint is an independent organization that serves as a liaison between the University and the person bringing the complaint so that anonymity can be ensured.

Please click the next button (right arrows) to indicate that you have read and agree to this consent document. If you do not agree, please close the survey and return the HIT.

EXPERIMENT

For this research study, imagine that you work as a salesperson for a company that makes tools and shoes. You work for both the tools division and the shoes division as you sell both types of products to retail stores.

You have two managers: a tools manager and a shoes manager. The two managers are located in different buildings and rarely talk to each other. They never talk about you.

If you sell a lot of tools, the tools manager will give you a bonus. If you sell a lot of shoes, the shoes manager will give you a bonus. You could get zero, one, or two bonuses.

To confirm your understanding of your job, answer the following questions:

- If you sell some tools it will
    - Increase your chance of getting a tools bonus
    - Decrease your chance of getting a tools bonus
    - Have no effect on your chance of getting a tools bonus
- If you sell some tools it will
    - Increase your chance of getting a shoes bonus
    - Decrease your chance of getting a shoes bonus
    - Have no effect on your chance of getting a shoes bonus

- [NEW PAGE] If you sell some shoes it will
    - Increase your chance of getting a tools bonus
    - Decrease your chance of getting a tools bonus
    - Have no effect on your chance of getting a tools bonus
- If you sell some shoes it will
    - Increase your chance of getting a shoes division bonus
    - Decrease your chance of getting a shoes division bonus
    - Have no effect on your chance of getting a shoes division bonus
- What is the relationship between the shoes manager and the tools manager?
    - One reports to the other
    - The share an office
    - They rarely see each other and never talk about sales reps

This task occurs in the [shoes | tools] division. [graphic of tools or shoes]

Imagine that the [*tools division | shoes division*] is trying to get a new supplier. The [*tools | shoes]* manager has found a supplier that has better products at a lower cost than the old supplier. However, the supplier normally only ships very large orders. Your firm places small orders, so you probably couldn't make a deal through the normal sales route. However, [you | the HR manager] went to school with the owner of the supply company. The [*tools | shoes*] manager asked [you | the HR manager] to use [your | his] school connection to help convince the supplier to sell you supplies, even though this is outside of the scope of [your |the HR manager's] normal duties.

Please take a few minutes to write what activities [you | the HR manager] will do to help convince the supplier. _____

Is this in [your | the HR manager's] job description? What extra work is involved?
_____

How do these extra actions help the [*tools division | shoes division*]?
_____

[NEW PAGE]

[You | The HR manager] did a great job! The [*tools division | shoes division*] was able to land a long-term relationship with the supplier, resulting in considerable savings and a significant improvement in product quality. The [*tools division | shoes division*] manager said that [you |the HR manager] really proved [your | his] commitment to the [*tools division | shoes division*].

This task occurs in the [tools division | shoes division]. [graphic of tools or shoes]

You have worked hard, but the poor economy has left you just short of your sales goal this year. It looks like you won't get the [tools | shoes] sales bonus.

A customer placed an order large enough for you to hit your goal. Unfortunately, they requested that it be shipped the first of next year, so it won't count for this year's sales goal.

You've asked the customer if they would be willing to move this one order up. They said that they will if they are given a large enough discount, which they understand would be a one-time discount applying only to this order.

It's common to offer discounts of 5-10% for large orders and 10-15% for charities. The firm makes 17% profit on the retail price of the order.

The customer has proposed that you give a 15% discount on their [tools | shoes] order if they accept early delivery. How likely are you to accept the customer's proposal?

Very Unlikely	Unlikely	Somewhat Unlikely	Undecided	Somewhat Likely	Likely	Very Likely

Please indicate the degree to which you agree/disagree with the following statements (all 7-point scales):

- Offering a discount is in in the best interest of the firm.
- If the [tools | shoes] manager were to find out that I offered a discount he would assume that my actions were in the best interest of the firm.
- Because of the hard work I've done for the firm in the past, I deserve a [shoes | tools] bonus.
- If I slacked off a little bit, the [tools | shoes] manager wouldn't be mad because of the hard work I've done in the past.
- The [tools | shoes] manager deserves my hard work.

[NEW PAGE]

The first task (helping secure a supplier) took place in the

- Tools division
- Shoes division
- Both divisions

The second task (potential sales discount) took place in the

- Tools division
- Shoes division
- Both divisions

How old are you?

- 18-25
- 26-35
- 36-45
- 46-55
- 56-65
- Older than 66

What is your gender?

- Male
- Female

How many MTurk HITs do you do a week on average?

- 0-2
- 2-10
- 10-50
- More than 50

Native language

- English
- Spanish
- Chinese
- Korean
- Hindi
- Other (free text response)

Would you have been more or less likely to give a discount if the second task was in the in the [shoes | tools] division instead of the [tools | shoes] division?

- Much less likely
- Less likely
- Neither more nor less likely
- More likely
- Much more likely

[NEW PAGE]

What do you think was the purpose of this study? (free text response)

Have you ever heard of moral licensing (also known as self licensing)?

- Yes, I could explain it
- Sounds familiar, but I couldn't explain it
- No

Do you have any other comments or suggestions regarding this study? (free text response)

REFERENCES TO CHAPTER 3

Bentley, J.W., M. Bloomfield, R.J. Bloomfield, and T. Lambert. 2015a. The morals of unethical reporting. *Working Paper*

Bentley, J.W., R.J. Bloomfield, S. Davidai, and M.J. Ferguson. 2015b. Drinking your own Kool-Aid: The role of beliefs, belief-revision, and meetings in persuasion. *Working Paper*.

Bloomfield, R.J. 2015. What counts and what gets counted. *Available at SSRN 2427106*.

Bol, J.C. 2008. Subjectivity in compensation contracting. *Journal of Accounting Literature* 27: 1-24.

Bruns, W.J., and K.A. Merchant. 1990. The dangerous morality of managing earnings. *Management Accounting* 72 (2): 22-25.

Chandler, J., P. Mueller, and G. Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods* 46 (1): 112-130.

Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. 2010. Are your participants gaming the system?: Screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems:* 2399-2402.

Effron, D. A., J.S. Cameron, and B. Monin. 2009. Endorsing Obama licenses favoring whites. *Journal of experimental social psychology 45* (3): 590-593.

Festinger, L., and J.M. Carlsmith. 1959. Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology* 58 (2): 203.

Goodman, J. K., Cryder, C. E. and Cheema, A. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decis. Making*, 26: 213–224.

Graham, J.R., R.H. Campbell, and S. Rajgopal. 2005. The economic implications of corporate financial reporting. *Journal of Accounting and Economics* 40 (1): 3-73.

Grenier, J.H., B. Pomeroy, and M. Stern. 2014. The effects of accounting standard precision, auditor task expertise, and judgment frameworks on audit firm litigation exposure. *Contemporary Accounting Research* 32 (1): 336-357.

Hayes, Andrew F. 2013. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Holmstrom, B. 1979. Moral hazard and observability. *The Bell Journal of Economics* 10 (1): 74-91.

Holmstrom, B., and P. Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*: 7 (Sp): 24-52.

Horton, J.J., and L.B. Chilton. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*: 209-218.

Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3 (4): 305-60.

Kaplan, R. S., and D. P. Norton. 1992. The balanced scorecard - Measures that drive performance. *Harvard Business Review* 70 (1):71-79.

Kaufmann, N., T. Schulze, and D. Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. *AMCIS*. Vol. 11.

Miller, D.T. and D.A. Effron. 2010. Chapter three-psychological license: When it is needed and how it functions. *Advances in experimental social psychology 43*: 115-155.

Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology 45*(4): 867-872.

Paolacci, G., J. Chandler, and P.G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5 (5): 411-419.

Peer, E., J. Vosgerau, and A. Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46 (4): 1023-1031.

Preacher, K.J., and A.F. Hayes. 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods* 40 (3): 879-91.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37 (1): 7-63.

Rennekamp, K. 2012. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research* 50 (5): 1319-54.

Tversky, A. and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science 211* (4481): 453-458.

CHAPTER 4

TREATISE ON CHALLENGES WITH AMAZON MECHANICAL TURK RESEARCH IN ACCOUNTING

Amazon Mechanical Turk (henceforth AMT) is an increasingly popular method of conducting academic research. It is less expensive than traditional laboratory-based experiments because (1) AMT workers have a lower reservation wage than the average lab participant (see Horton and Chilton 2010 for a discussion of wages) and (2) AMT workers don't need to travel to a lab to participate so there is a lower fixed cost on their end. AMT research is also more flexible because the pool of participants is large enough that data collection can happen extremely quickly without complicated signups or recruiting procedures. Furthermore, the subject pool is more diverse than student populations (Kaufmann, Schulze, and Veit 2011), helping with external validity. However, AMT research is not without its share of challenges and problems. In this chapter, I discuss four challenges that arise when conducting research using AMT: (1) lack of participant effort, (2) non-naïveté among participants, (3) lack of experimental control, and (4) researcher degrees of freedom. The fourth challenge is one that researchers, reviewers, and the academic community at large need to be prepared to deal with in interpreting research conducted with AMT. I also suggest techniques to help researchers and reviewers deal with each challenge.

Challenge #1: Lack of Attention or Effort

The first and most prominent challenge with AMT research is the fact that workers may not be fully attending to the task or giving their best effort. For example, Kittur, Chi, and Suh (2008) find that some participants answer questions faster than is possible given the

reading requirements and that others copy and paste answers for free-response questions,[1] Callison-Burch (2009) finds that participants perform only slightly better than chance at translation tasks, and Kaufmann, et al. (2011) find that over 30% of participants fail simple attention-check questions. These extreme levels of inattention generate so much noise in observations that many experiments fail not due to a lack of a real effect but rather due to a lack of power to overcome the noise.

Downs, Holbrook, Sheng, and Cranor (2010) discuss two possible techniques to test for inattentive participants. The first technique involves asking explicit attention-check questions such as "to show you are paying attention, please select the third option below." Oppenheimer et al. (2009) calls these types of questions Instructional Manipulation Checks (IMC). Downs, et al. (2010) state that while these types of questions are the most common technique, but they present two major problems. First, they only catch the most egregious behavior and violate Gricean norms by forcing participants to pay careful attention only to predictable information. Thus, they aren't actually checking whether participants are paying attention to the important information in the task, but rather if participants know how to respond to IMCs. Second, these types of questions set up a tone of distrust for the remainder of the task: "I, as a researcher, don't trust you, as a worker, and you probably shouldn't trust me either." This distrust may carry over into future research, poisoning the AMT pool of participants.

The second technique presented by Downs, et al. (2010) consists of more natural questions that have a factual answer but appear to be a mere formality of the study. For

[1] Worker effort was so low in Kittur, et al.'s first experiment that the researchers ended up flagging 58.6% of all observations as potentially invalid due to overly-fast responses, blank/nonsense responses to free response questions, or copy/paste responses to free response questions.

example, Downs, et al. (201) have participants read an email and then answer two factual questions from the email. These questions are presented as though they were ACT/SAT questions with no reference to the inattentiveness or low work ethic of prior workers. They appear to readers as though they are simple review questions aimed to make sure that the article is understandable. Downs, et al. (2010) find that only 61% of participants correctly answer both questions, but recognize that the accuracy of responses depends on the difficulty of question being asked. These types of questions are superior to IMCs because they (1) ask questions that are relevant to the research and (2) don't prompt distrust or insult the participants.

Peer, Vosgerau, and Acquisti (2014) propose an alternative method of screening out inattentive participants: reputation. AMT requesters can approve/reject work performed by AMT workers. The AMT software allows requesters to specify criteria for workers to participate in a particular job (called a Human Intelligence Task, or HIT). One criteria is approval rating. For example, a requester can say that workers can only participate if their approval rating is at least 95%. Peer, Vosgerau, and Acquisti (2014) conducted two experiments. In the first experiment, they had participants with a high (>95% approval rating) or low (<95% approval rating) complete a task that had three IMC-style attention check questions. These questions were copied or adapted from prior research, so participants may have seen them before. Peer, et al. found that 97.4% of participants with a high approval rating passed all three attention check questions while only 66.1% with a low approval rating passed all three questions. In their second experiment they used a novel attention check question and found much poorer accuracy by participants. They also found that participants who had high productivity (i.e. had completed at least 500 HITs) performed much better than

participants who had low productivity (i.e. completed no more than 100 HITs). 83.3% of high-productivity workers passed all 3 novel questions while only 70.9% of low-productivity workers passed all 3 novel questions.

Peer, et al. (2014) recommend that researchers only allow AMT workers who have a high accuracy (e.g. 95% HIT approval) and a high productivity (e.g. at least 500 HITs completed). This approach seems to be an effective way to reduce, but not eliminate inattentive workers. One problem with this screening method is that it may be a bit tautological. Participants with high accuracy are those who know how to respond to the tricks of IMC-style attention check questions. They continue to be successful at responding to these types of questions from future researchers. Another potential problem is that high productivity workers may be less naïve with respect to research (see challenge #2 below).

I recommend that researchers combine the Peer, et al. (2014) prescreening method with attention check questions (following the guidelines in Downs, et al. 2010) to have the best chance of high accuracy. Ideally the majority of the screening occurs through the Peer, et al. (2014) method with the Downs, et al. (2010) attention-check questions serving as primarily a backup. My reading of the literature suggests that relying only on attention-check questions will result in 30-60% of data being excluded from final analyses due to worker inattention. This number could reasonably drop to less than 30% (e.g. 16.7-29.1% in Peer, et al. 2014) when preceded by accuracy and productivity qualifications.

Challenge #2: Non-Naïveté Among Participants

Another challenge is also possible in research involving AMT workers: hypothesis guessing or attempting to guess the unseen experimental conditions. Because of the inexpensive nature of studies, many researchers conduct multiple studies using the same AMT

participant pool, assuming that the large size of the pool will prevent repeat participation in related experiments. That assumption is simply not accurate. Chandler, et al. (2014) compile data from 132 academic studies and find that the average worker had participated in 2.44 of the studies, and 33% of the workers seek out academic studies. More concerning, however, was the fact that the most prolific 10% of workers completed 41% of all HITs. These workers are highly likely to have seen related research in the past. Chandler, et al. (2014) state that reusing participants is likely to be more of a concern for commonly-used paradigms than for less-used paradigms. For example, they find that 56% (52%) of participants said that they had previously seen the prisoner's dilemma (ultimatum game) while only 7% of participants said that they had previously seen the p-beauty contest.

Chandler, et al. recommend that researchers avoid commonly used paradigms for MTurk research and that they share a list of worker IDs for less commonly used paradigms. Unfortunately, to my knowledge no such list currently exists, and IRBs at some schools may be reluctant to allow researchers to distribute unique identifiers (i.e. worker IDs) of the participants. Chandler, et al. also suggest that researchers create their own lists which they use to exclude future participants (see also Paolacci, Chandler and Ipeirotis 2010). However, that doesn't address the concern that workers may have participated in related experiments by other researchers.

The fact that many participants are familiar with research techniques and theories is particularly concerning because AMT workers score higher on the social desireability scale (a measure of how anxious someone is to please others) than students (Behrend, et al. 2011) and therefore may exhibit higher experimenter demand (Berinsky, Huber, and Lenz 2012). As a brief anecdote, before conducting my first AMT study, I participated as a worker in a dozen

or so studies posted by other researchers. With few exceptions I was able to guess the hypothesis, which made me unable to respond to the stimuli in a natural way. I found myself debating whether I should (1) answer in the biased fashion the way that I believed an uninformed participant would respond, (2) answer in a way that proved I wasn't biased, or (3) answer at the midpoint of the scale to try and not affect the research. Answering truthfully was out of the question as I no longer knew what "truth" was for myself. Luckily, the majority of these studies followed a technique common in social psychology research: funneled debriefing.

Bargh and Chartrand (2000; see also Aronson, et al. 1990), referring to general psychology research not just AMT research, recommend funneled debriefing at the end of experiments to give participants multiple opportunities to state what they believe is the purpose of the study. They state that

> "In general, if a participant evidences any genuine awareness of a relation between the prime and experimental task, his or her data should not be included in the analyses. By 'genuine awareness' we mean any answer in the debriefing which is 'in the ballpark' as to what could have affected responses. In our research, we take a conservative stance and err on the side of over-exclusion if there is any doubt."

Funneled debriefings may be a bit lengthy for AMT tasks. For example, the funneled debriefing in Chartrand and Bargh (1996) consists of 7 free response questions, effectively doubling the length of some of my studies. To simplify the process, I recommend a simple two-step debriefing process. In the first step, participants are asked to guess the purpose of the study using free response questions. In the second step (subsequent page), participants are asked if they have ever heard of XXXXXXX (the primary theory motivating the research such as "the prisoner's dilemma" or "moral licensing"). Participants respond to the second question with either "Yes, I could explain it," "Sounds familiar but I couldn't explain it," or

"No." Participants are excluded if their free response to the first question is "in the ballpark" or if they say that they could explain the theory.

AMT also has a large forum base wherein workers regularly discuss jobs. My experience is that most workers avoid discussing hypotheses or specifics of studies. Furthermore, several of the forums explicitly prohibit posting study details. Nonetheless, researchers should actively monitor popular forums (e.g. mturkforum.com, turkopticon, mturkgrind, reddit, twitter) for the duration of the study to ensure that their hypotheses/manipulations were not revealed during the study. As a brief example, one participant posted how he thought I must be looking for the next Bobby Fischer. Thus, he revealed an incorrect guess about the purpose of my study. A correct guess could have compromised the integrity of the research. Most AMT workers know they should not actively discuss the specifics of studies, but some discussion could occur inadvertently. The appendix to this chapter provides examples of comments about my chess experiment (Chapter two of this dissertation).

Challenge #3: Lack of Experimental Control and Heterogenous Population

In a lab environment, the researcher can control a large number of variables that could influence the results. In contrast, on AMT many things are simply uncontrollable. For example, Jim Cannon, Todd Thornock and I attempted to use AMT to test how different reporting environments affected participants' effort levels. To do so, we used a mouse-clicking task so that there would be very little, if any, ability for participants to learn (participants were paid based on how many clicks they made in a 20-minute period). We realized, however, that we couldn't adequately control for an important aspect of performance: type of mouse used. Touchpads, traditional mice, gaming mice, and

autoclicking software produce extremely different rates of clicking. This variation would not exist in a traditional lab environment where all subjects have the exact same mouse on the exact same computer. This example is an extreme form of an individual difference that can create a large source of noise. Noise reduces experimental power, sometimes to the point that it becomes impossible to find a real effect.

Kaufmann, et al. (2011) document a number of demographic variables that vary tremendously in AMT studies but do not vary in most lab studies: age, country,[2] education level, employment status, household income, time on MTurk (e.g. <1 week to several years), and hours/HITs per week. While none of these variables by themselves are cause for excluding participants, they are added "Box 5" variables (Libby, Bloomfield, and Nelson 2002) that may affect individuals' responses in AMT studies more than in lab studies as the latter has considerably less variation. Therefore, researchers should ask for these variables as applicable and include them as control variables in analyses to reduce noise.

Goodman, Cryder, and Cheema (2013) express concern about the English language ability of some MTurkers. They follow recommendations by Oppenheimer et al. (2009) and include Instructional Manipulation Checks in their studies. They found a large, significant difference in the rate of IMC questions answered correctly between native English speakers and non-native English speakers (hereafter ESL). In study 1 they use a relatively simple IMC and find that 81% of native English speakers answered the IMC correctly while only 61% of ESL participants answered correctly. In study 2 their results were even more dramatic: 71% of native English speakers correctly answered the difficult IMC while only 29% of ESL

[2] Country is now mostly a moot concern as AMT has made it increasingly difficult for people outside of the US to work for AMT due to tax implications. See AMT's terms and conditions for details.

participants answered the IMC correctly. The difference between the two groups was significant at the $p<0.001$ level. They suggest that IMC filters are more efficient than ESL filters, but acknowledge that IMCs need to be at the appropriate difficulty level to ensure attentiveness. English language ability may be especially problematic in psychology-style studies that require participants to detect subtle nuances in the instructions.

English language ability can also serve as a proxy for a number of cultural demographic variables that influence behavior. A large stream of literature had documented cross-cultural personality differences (see e.g. Taras, Rowney, and Steel 2009 for a recent review). Culture plays a significant role in paradigms such as the endowment effect (Maddux, Yang, Falk, et al. 2010) and prisoner's dilemma (Wong and Hong 2005) and basic psychological theories such as cognitive dissonance and the fundamental attribution error (see e.g. Kasima 2001, Lehman, Chiu and Schaller 2004). The magnitude of cultural effects suggests that a few ESL participants in an otherwise native-English participant pool could dramatically influence results, particularly if the ESL participants are not evenly distributed across conditions. Researchers may want to limit their sample to workers who are native English speakers or include controls variables for participants' native language.

Challenge #4: Research Degrees of Freedom

The final challenge with using AMT for research is not a concern about the data or AMT workers, but rather with how researchers respond to the previous three challenges. The previous three challenges present problems that cannot be fully dealt with from an ex-ante perspective, forcing researchers to make decisions about how to treat data from in an ex-post setting. Researchers could opportunistically used these ex-post discretionary choices to increase their chances of finding significant results. These discretionary choices are

sometimes referred to as “researcher degrees of freedom,” something that is very concerning to the social sciences as a whole (see e.g. Simmons, Nelson, and Simonsohn 2011). Chandler, et al. (2014) in discussing this issue, point out that the vast majority of exclusions can and should be done before allowing workers to take a study rather than after in order to prevent researcher degrees of freedom. The Peer, et al. (2014) method dramatically reduces the number of participants who need to be excluded after data has been collected by setting screening criteria before allowing workers to participate.

For example, in my first dissertation experiment (Chapter 2), I needed experienced chess players. I could have allowed hundreds of AMT workers to complete my study and then excluded unqualified participants ex-post. Instead, I used a simple, but imperfect, five-question chess pre-screening test. Workers who did not pass the test were not allowed to participate in the actual experiment – I never collected data from them and there was no discretion in excluding them from my results. Thus, I “dropped” 729 unqualified participants without ever collecting their data. However, this screening process was imperfect. With a 2-minute task using multiple-choice questions I was attempting to test if someone was qualified to participate in a 30-minute task that required a high level of skill. Some participants appear to have passed the pre-screen without being qualified (see the appendix to this chapter for self-reports of unqualified passing). A longer pre-screening task would have been more accurate, but it would have also increased the cost of prescreening and the likelihood of cheaters. In the end, I identified seven participants who appear to have passed the pre-test by chance and needed to be excluded from my analyses. By including the pre-test, I was able to do 99% (729/736) of my exclusions before collecting data, meaning I eliminated 99% of my possible researcher degrees of freedom.

Whenever possible, researchers should screen out participants before rather than after collecting data. Below are three screens that should be conducted before AMT workers are allowed to accept a HIT:

1. High accuracy/productivity (Peer, et al. 2014)
2. Repeat participation (Chandler, et al. (2014)
3. Custom pre-screening such as accounting knowledge (or chess ability in my dissertation)

Some researchers may appropriately be concerned that pre-tests will "tip their hand" to participants. In my chess study, I was comfortable with participants knowing that the study required chess ability, but I didn't want participants to know that I cared about reporting relationships. Therefore, I made my pre-test about chess and not about reporting. However, some things cannot be pre-tested. For example, funneled debriefing needs to occur after an experiment rather than before an experiment to avoid causing the very suspicion it is intended to detect. If, in the experiment reported in Chapter three of this dissertation I had asked if participants knew about moral licensing, some participants may have tried to look up moral licensing and had unnatural behavior as a result of the prescreen. Thus, researchers should conduct prescreening whenever possible, but only if said prescreening will not induce experimenter demand or suspicion.

Even when researchers conduct adequate prescreening, there may still be cause for excluding observations after data collection has occurred. Participants who fail attention-check questions and/or guess the researcher's hypotheses (as revealed by funneled debriefing) should be excluded. In these cases, researchers should footnote what the results are before these exclusions. Furthermore, researchers should be diligent in applying these exclusions regardless of whether or not the pre-exclusion data yields a significant p-value.

On a similar note, demographic variables should be included in a model if they significantly affect the results or interact with independent variables. However, researchers should state how the inclusion/exclusion of these variables affect results.

As a final note, Simmons, Nelson, and Simonsohn (2011) point out that research degrees of freedom decrease with larger sample sizes. They provide the rule of thumb that there should be at least 20 participants per condition. However, Simmons, et al. (2011) structured their recommendations solely on reducing the rate of false positive results. Furthermore, Simmons, et al. (2011) were focused reducing the false positive rate for a single study rather than considering the false positive rate of research as a whole. Ideally, researchers should conduct power analyses based on pilot data or expected results when determining sample sizes. Realistically, however, it is difficult to do a power analysis when the effect size is unknown.

One solution, then, is to conduct power analyses based on economically-significant effect sizes from prior research (see e.g. Fraley and Vazire 2014). For example, Richard, et al. (2003) conduct a meta-analysis of social psychology research and find an average effect size (r) of 0.21. For a 2-condition experiment to have an 80% chance of finding an effect of this magnitude, 88 participants per cell are needed. For most AMT studies, the cost per participant is quite low,[3] so having 100 participants per cell is a reasonable benchmark that allows sufficient statistical power while simultaneously reducing researcher degrees of freedom.

[3] The median reservation wage of AMT workers is $1.38 per hour. For sake of illustration, a 10-minute study that pays the equivalent of $6.00 per hour (a relatively high AMT wage), would cost $1.00 per participant or $400 for an experiment with a 2x2 between-subjects design with 100 participants per condition. Longer experiments (e.g. Chapter 2 of this dissertation where the average time spent was 111 minutes) or experiments that require repeat participation (e.g. Asay 2015) or expert participants who demand higher wages may not be able to achieve the 100/cell benchmark, but should present compelling evidence to reviewers why the sample size had to be reduced.

Conclusion

Amazon Mechanical Turk is a growing playground for accounting research. The low cost, easy access, and diverse population make it an enticing laboratory. In this chapter, I presented four concerns with conducting AMT research: lack of attention/effort, hypothesis guessing, lack of experimental control/heterogeneous population, and researcher degrees of freedom. I then presented recommendations from the literature and my own thoughts on how to deal with these problems. Applying these recommendations will allow accountants and other behavioral researchers to continue to conduct research using Amazon Mechanical Turk while simultaneously increasing experimental power and decreasing the risk of false positive results.

APPENDIX TO CHAPTER 4
EXAMPLES OF AMT WORKER DISCUSSIONS

Below are comments I observed on various AMT forums regarding the experiment presented in the second chapter of this dissertation. Minor bullet points represent responses to the original comment. With the exception of one vulgar word which I replaced with ****, all comments are taken verbatim including any errors.

- They must be looking for the next Bobby Fischer for this HIT. I'm not a Chess Master, but I do know the game well enough to know which side was winning in each case and I wasn't approved either.
    - I know absolutely nothing about chess and I qualified for the HIT which is an automatic 2.00 bonus :p
        - Congratulations! Interesting that the requester is going to allow a person that knows nothing about chess to judge chess matches. Their loss, you're gain.
    - I agree, just based on the "points" each is worth it made it simple imo to see which sides were winning but whatever
        - I don't know why it rubbed me the wrong way as I've been turned down for higher paying HIT's before. Lol But I've played chess enough in my life to know what's going on. Oh well, on to the next HIT.
- **** I play a lot of chess, those last questions are a little tricky. Point wise tied yet slightly different. I am confused
    - Edit: I just realized that's probably a spoiler
- "You passed the pretest. You may continue with this HIT. Passing the pretest has earned you $4.00" Whoa, I'll report back when I finish.
- Has anyone gotten paid for this yet? I got the $0.25 for the pre-test and did qualify for the long one, but I would be seriously surprised if I missed every single guess
    - I got paid over the weekend around $6. Not much more than the pretest, but then again I suck at Chess.
- failed but my chess is so so at best
- Didn't pass, but the pretest is super fast
- Pretest was pretty common chess things, not too hard. If you pass it, some of the questions get hard because of how close the games are but you can bet 0.
    - See, did I miss any instructions about what to do if you think it could be a draw? There's been two or three positions that should end in a draw

        - I didn't see any either, but 0 is an option on the slider, so I bet it a couple times.
- Passed the pretest. The actual survey took about 30 minutes but it wasn't too bad. Got at least $4.25 but the rest of the money depends on betting, so we'll see how I do. Either way, the hit was interesting.
    - You're betting with the bonus. If you lost every bet and bet the max, you won't be getting that $4.
        - That is true, but many of the games are pretty safe bets. If two good computers are playing an obvious advantage should result in a win.
- That was fun! I'm interested to see how I end up.
- Took me 1 hour 7 minutes. Quite long but hope it was worth it.
- Did not expect to pass that pretest, but I'm hoping this works out in a week.
- Very enjoyable suvey! Earned over $8.00. Could have made more but I was conservative with my bets since I'm rusty at chess. Most of my bets were less than half what was allowed. Base and bonus paid in 5 days.
- Amazing chess hit. Took a little bit to get it approved & bonused, but given the task and the reward it was totally understandable.
- Awaiting approval and payment. It's nice that they pay you for the screener. Will update once approved. Approved and paid.
- Well, I had fun trying. I haven't played chess in over 20 years, and attempted to Google a couple of the moves, but was unsuccessful in the end. This is the ONLY time in 9+ years I wished my ex-husband was around to help me earn money with his superior chess skils. Overall, I had fun trying for a quarter. If I could try it again I would surely do some reading to refresh my skills before attempting. The pay was great for a screener, and fun to boot. Payment was received within 24-hours from completion.
- I failed the prescreen, emailed him, he replied quickly and politely. Looks like I need to brush up on chess. :P
    - I failed too. I haven't played chess in over 20 years, and tried Googling two of the moves quickly, but in the end I failed. It was fun trying anyhow.

REFERENCES TO CHAPTER 4

Aronson E, Ellsworth P, Carlsmith J, Gonzales M. 1990. Methods of research in social psychology. New York, NY: McGraw-Hill.

Asay, H.S. 2015. Horizon-induced optimism as a gateway to earnings management. *Working Paper*. The University of Iowa.

Bargh, J.A., and T.L. Chartrand. 2000. The mind in the middle. In *Handbook of research methods in social and personality psychology*: 253-285.

Behrend, T. S., Sharek, D. J., Meade, A. W., and E. N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior research methods*,*43*(3): 800-813.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* 20 (3): 351-368.

Callison-Burch, C. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk, In Proc. EMNLP 2009, ACL and AFNLP (2009), 286–295.

Chandler, J., P. Mueller, and G. Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods* 46 (1): 112-130.

Chartrand, T. L., and J.A. Bargh. 1996. Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, *71*(3): 464-478.

Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. 2010. Are your participants gaming the system? Screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems:* 2399-2402.

Fraley R.C. and S. Vazire. 2014 The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE* 9(10): e109019. doi:10.1371/journal.pone.0109019

Goodman, J. K., Cryder, C. E. and Cheema, A. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decis. Making*, 26: 213–224.

Kashima, Y. 2001. Culture and social cognition: Toward a social psychology of cultural dynamics. In D. Matsumoto (Ed.), Handbook of culture and psychology (pp. 325–360). New York: Oxford University Press.

Kaufmann, N., T. Schulze, and D. Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.*AMCIS*. Vol. 11.

Kittur, A., E.H. Chi, and B. Suh. 2008. Crowdsourcing user studies with Mechanical Turk. *Proceedings of the 26th Annual ACM Conference on Human Factors in Computing Systems (CHI '08)*; 2008 April 5-10; Florence, Italy. NY: ACM; 2008: 453-456.

Lehman, D., C. Chiu, and M. Schaller. 2004. Culture and psychology. Annual Review of Psychology, 55: 689–714.

Maddux, W.W., H. Yang, C. Falk, H. Adam, W. Adaiar, Y. Endo, Z. Carmon, and S.J. Heine. 2010. For whom is parting with possessions more painful? Cultural differences in the endowment effect. *Psychological Science* 21(12): 1910-1917.

Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872.

Paolacci, G., J. Chandler, and P.G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5 (5): 411-419.

Pedersen, E.J., R. Kurzban, and M. E. McCullough. 2013. Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society of London B: Biological Sciences* 280.1758: 20122723.

Peer, E., J. Vosgerau, and A. Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46 (4): 1023-1031.

Richard, F. D., C.F. Bond Jr, and J.J. Stokes-Zoota. 2003. One hundred years of social psychology quantitatively described. *Review of General Psychology 7*(4): 331-363.

Taras, V., J. Rowney, and P. Steel. 2009. Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management* 15(4): 357-373.

Wong, R. Y., and Y. Hong. 2005. Dynamic influences of culture on cooperation in the prisoner's dilemma. *Psychological Science* 16(6): 429-434.