Carnegie Mellon
Tepper
SCHOOL OF BUSINESS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
INDUSTRIAL ADMINISTRATION
(ORGANIZATIONAL BEHAVIOR AND THEORY)

Titled

"Forecasting Unethical Behavior Using The Hidden Information Distribution and Evaluation (HIDE) Model"

Presented by

Yeonjeong Kim

Accepted by

Taya R. Cohen 4/26/18

Chair: Prof. Taya R. Cohen **Date**

Approved by The Dean

Robert M. Dammon 4/26/18

Dean Robert M. Dammon **Date**

Forecasting Unethical Behavior Using
The Hidden Information Distribution and Evaluation (HIDE) Model

Yeonjeong Kim

Organizational Behavior & Theory

Tepper School of Business

Carnegie Mellon University

Dissertation Committee:

Professor Taya R. Cohen (Chair)

Professor Laurie R. Weingart

Professor Denise M. Rousseau

Professor Rebecca L. Schaumberg

ACKNOWLEDGEMENT

As a graduate student, I encountered many challenging moments while working on my dissertation. However, I was extremely lucky to have many wonderful individuals help me overcome those difficulties, support this research, and guide me along my path to an academic career. I am especially grateful to my family, my committee members, and my collaborators.

First and foremost, I want to express my appreciation for my family who has supported me from the very beginning, long before I even considered graduate school. I am deeply thankful to my mother, Guibun Jeong, an exceptionally strong person who overcame a lot of difficulties. Despite not being able to receive much education herself while growing up, she instilled the value of education in my brother and I. She made great sacrifices raising and educating us by herself, in a society that did not support single mothers, either economically or culturally. I also want to thank my older brother, Young Min Kim, for the many nights he stayed up helping me debug my messy simulation code when I first started learning to program.

I am deeply indebted and thankful to my partner, Chris Olivola, for his affection, generosity, commitment, and support. He has been my greatest source of happiness and encouragement during my final years of graduate school. Our discussions helped shape the direction of my dissertation, both theoretically and methodologically. His family members (Maye, Ken, and Maëlle Olivola, as well as Maryanne Neil) also deserve recognition for their kindness, generosity, and warmth. They always made me laugh and feel welcomed, as a part of their family, which provided wonderful additional emotional support.

Academically, I was extremely fortunate to have had both Taya Cohen and Laurie Weingart as advisors throughout graduate school. I am grateful to Taya for taking a risk by accepting me—someone without prior behavioral science research experience—into the Ph.D.

program. From the very start, she has always been there when I needed her help, not only for our research projects, but also for personal matters. More than anyone else, Taya gave me a lot of important and stimulating research opportunities. I especially love her energy, rapid feedback, and direct communication style, which made our collaboration fun and exciting.

I am grateful to Laurie for many reasons. She provided me with a great role model that I admire both academically and personally. Despite her copious accomplishments, stellar reputation, and many responsibilities, she remained available, approachable, empathetic, and dedicated to her students. She made sure to always be present when I presented our research at conferences. Whenever I encountered a difficult situation, she made time for me and helped me navigate through those challenges. Finally, Laurie showed me the value of practicing what she learned from her research on team communication, by being a wonderful listener and a great communicator.

I would also like to express my gratitude to Rebecca (Becky) Schaumberg and Denise Rousseau. I greatly appreciate Becky for her many insightful comments that helped improve my dissertation. One of Becky's many great qualities is her unparalleled positive and enthusiastic attitude. She is one of the most energetic and kindest people I have ever met, and therefore someone that I greatly enjoy interacting with. Despite her busy schedule, she twice traveled to Pittsburgh for my dissertation.

Denise is someone that I have admired for a long time—even before applying to the Ph.D. program I had heard great things about her from her former students. So it has been a great honor to have her on my committee, to be encouraged by her, and especially to know that she appreciates this dissertation work.

In addition to my committee members, I would like to thank Kristin Behfar, who has been an amazing mentor, collaborator, and friend during the course of my doctoral work. She is one of the most respectful, fair, warm, and empathetic people I have ever met. Despite her busy schedule, she was always available when I needed a thought partner or caring advice. I also very much appreciate her generously recognizing my contributions to our research projects.

Finally, I would like to thank the numerous sources of financial support that enabled me to conduct eight empirical studies and recruit more than 5,000 participants for this dissertation. This work was generously supported by the Heritage Dissertation Award (from the Foundation for Personality and Social Psychology), the Paul S. Goodman Endowed Doctoral Dissertation Award (from Carnegie Mellon University), a CBDR small research grant (from Carnegie Mellon University), and the GuSH graduate student grant (from Carnegie Mellon University). This research was also supported by grants, awarded to Taya Cohen, from the Notre Dame Deloitte Center for Ethical Leadership and from the Carnegie Bosch Institute. Finally, I want to thank Abigail Panter, who served as a collaborator and also helped fund data collection for this project.

ABSTRACT

The ability to correctly judge moral character—an individual's disposition to think, feel, and behave ethically—is critical considering the negative consequences of misjudgment (e.g., being betrayed or swindled). However, it is currently unknown whether people can reliably detect strangers' moral character, nor is it known how to best elicit relevant information from strangers to determine their moral character. This research is designed to remedy this dearth in our understanding of moral character judgments, particularly in settings where we need to make prompt evaluations of strangers based on limited information that we obtained from them. The biggest challenge in assessing another person's moral character is that it is extremely socially desirable, and therefore highly susceptible to distorted self-perceptions and impression management. To address this problem, I propose and test a new person-perception theory: the *hidden information distribution and evaluation* (HIDE) model.

In chapter 1, I develop the HIDE model, which posits that there are aspects of information that individuals do not correctly know about themselves (which I call the ***hidden-self***), as well as aspects of information individuals misrepresent to others (which I call the ***hiding-self***). This model articulates when and why judges (i.e., evaluators) not personally acquainted with targets of evaluation (e.g., job applicants) can reliably detect these targets' moral character and predict their future unethical behavior. In particular, I propose that the impromptu thinking and language usage that arises when a person answers specially designed interview questions reveal information about his/her hidden-self and hiding-self, enabling a group of judges to make valid judgments about his/her moral character. Additionally, the HIDE model predicts that judges' evaluations using this written interview method will be more valid than evaluations provided by targets' acquaintances. This is because social relationships can lead

people to form biased impressions of targets they are acquainted with, so that they are unable to see the targets' hidden selves as clearly as judges who do not know the targets.

In chapter 2, I test the HIDE model's prediction that groups of judges can reliably predict targets' unethical behavior by evaluating their moral character using the written interview method. In studies 1 and 2, large groups of judges were crowd-sourced online. I show that their average moral character evaluations successfully predicts targets' frequency of unethical behaviors in the laboratory (study 1) and the workplace (study 2). Study 3 extends these findings by determining the minimum number of judges (six) required to make moral character evaluations that predict unethical behavior.

In chapter 3, I test the HIDE model's prediction that judges' evaluations based on the written interview method can capture unique information about targets' ***hidden-self***. Three empirical studies (studies 4, 5, and 6) show that these evaluations indeed capture unique variance in targets' moral character that is missed by both self-reports and ratings provided by targets' acquaintances. Consequently, these evaluations are more predictive of targets' unethical behavior than the ratings provided by either the targets themselves or their acquaintances.

In chapter 4, I investigate the HIDE model's prediction that judges' evaluations using the written interview method can capture unique information about targets' ***hiding-self***. This occurs because responses to the interview questions reveal implicit aspects of moral character that targets cannot control or fake, even when they want to. In study 7, I manipulated whether targets had an incentive to answer the interview questions in a positively biased manner. I show that judges' evaluations of targets (based on the interview questions) are actually more predictive of their unethical behavior when targets were motivated to respond in a positively biased manner.

Finally, in chapter 5, I carried out text analyses to explore how human judges utilize linguistic cues in written responses to form impressions of moral character, and how these cues predict targets' unethical behavior. The goal of this chapter is to identify linguistic cues that human judges fail to correctly detect or utilize, and thus to identify ***shared biases in human perceptions of ethicality***. Building on these exploratory text analyses, I discuss the future directions of this research program, especially the potential value of ***combining human judgments and machine algorithms*** to boost the accuracy of unethical behavior forecasts.

Key words: unethical behavior; moral character; interviews; text-analysis; person perception.

CHAPTER I

Unethical Behavior Forecasting Using the Hidden Information Distribution and Evaluation (HIDE) Model

The ability to predict whether an individual is likely to behave unethically is critical considering the negative consequences of misjudgment. This is particularly true in organizational settings. Although employees' unethical work behaviors can be triggered by negative situations such as workplace mistreatment or other stressful environments (Kim, Cohen, & Panter, 2016; Kish-Gephart, Harrison, & Treviño, 2010), numerous studies reveal that certain types of employee behavior, including unethical behavior, are stable over time controlling for the effect of organizational circumstances (Ashton & Lee, 2007, 2008; Berry, Carpenter, & Barratt, 2012; Berry, Ones, & Sackett, 2007; Cohen, Kim, Jordan, & Panter, 2016; Cohen, Panter, Turan, Morse, & Kim, 2013, 2014). Therefore, one way to mitigate unethical work behavior is to identify individuals who are likely to engage in unethical behaviors during the hiring process. However, it is currently unknown whether, or how, we can detect peoples' tendencies to behave unethically if we do not know the person well. The goal of this dissertation is to answer the question of whether, and how, people can reliably detect strangers' (e.g., job candidates) unethical tendencies in order to predict a broad set of unethical behaviors, especially in contexts where judges (e.g., interviewers) need to make prompt evaluations based on limited information that they might obtarin from strangers.

Individual differences in likelihood to engage in unethical behavior are studied in the literature on moral character (Cohen & Morse, 2014; Cohen et al., 2015; Fleeson, Furr, Jayawickreme, Meindl, & Helzer, 2014; Kim & Cohen, 2015; Lee & Ashton, 2012; Peterson & Seligman, 2004). Recent psychological research has approached the study of morality and ethics from a personality perspective. In this work, the terms *morality* and *ethics* refer to standards of right and wrong conduct that provide guidance on what we should and should not do (Janoff-Bulman & Carnes, 2013). The term *personality* refers "an individual's characteristic patterns of thought, emotion, and behavior, together with the psychological mechanisms— hidden or not— behind those patterns" (Fast & Funder, 2010, p. 669). Therefore, moral character can be thought of as characteristic patterns of thought, emotion, and behavior that are associated with morality and ethics (Cohen & Morse, 2014; Cohen et al., 2015; Fleeson, Furr, Jayawickreme, Meindl, & Helzer, 2014; Kim & Cohen, 2015; Lee & Ashton, 2012; Peterson & Seligman, 2004).

Consistent with the existing literature on moral character, I define *moral character* as *an umbrella term referring to individual differences in thinking, feeling, and behaving in an ethical or unethical manner across diverse situations.* According to this definition, valid evaluation of strangers' moral character should predict their likelihood of engaging in unethical behavior across diverse situations. I adopt the widely accepted definition of *unethical behavior* as "either illegal or morally unacceptable to the larger community" (Jones, 1991, p.367). However, I further refine the scope of unethical behavior by excluding behavior that might violate some moral principles (e.g., honesty, integrity), but are conducted with benevolent intentions (e.g., white lies to help others) (Levine & Schweitzer, 2015). This restriction arises from recent debates in contemporary moral philosophy (Korsgaard, 1986; Strudler, 2016; Nichols, 2015), which discusses how not all seemingly morally unacceptable behaviors (e.g., bribes) are immoral when they can be justified (e.g., Schindler's bribes that saved the lives of Jewish workers; Nichols, 2015). Therefore, *unethical behavior* in this dissertation refers to *acts that are conducted without benevolent intentions for others and only can hurt other people or the larger community*. Examples of such behaviors include cheating, stealing, or lying that can hurt others with a goal of selfish gains. Using this smaller, but cleaner, set of unethical behaviors as criteria, this dissertation aims to identify the conditions under which evaluations of strangers' moral character is reliable and valid.

Understanding individual differences in moral character allows us to predict and possibly prevent unethical behaviors that harm people, organizations, and society. Indeed, measures that capture information relevant to moral character reliably predict observable unethical behaviors in anonymous research settings. For example, self-reports of honesty-humility—one of the "Big Six" factors from the HEXACO model of personality structure, which encompasses sincerity, fairness, modesty, and greed-avoidance—predicts not only self-reported delinquency and unethical decisions but also observable dishonesty, such as in behavioral economics games (Hilbig & Zettler, 2015), and coworker-reported workplace deviance (Cohen, Panter, Turan, Morse, & Kim, 2013). Judge reports of honesty-humility also predict self-reported delinquency, self-reported unethical decisions, and coworker-reported workplace delinquency (Cohen et al., 2013). Likewise, self-reported guilt proneness—an individual difference indicative of whether a person would feel guilty about committing transgressions even if no one were to find out—also predicts self-reported and observable unethical behaviors (Cohen, Kim, Jordan, & Panter, 2016; Cohen, Wolf, Panter, & Insko, 2011), and both self- and judge reports of guilt proneness predict self-reports and coworker-reports of workplace deviance (Cohen et al., 2013). Even more striking is the observation that guilt proneness measured from self-reports in children aged 10 to 12 correlates negatively with illegal behavior during young adulthood and with involvement in the criminal justice system through ages 18 to 21, providing powerful evidence of the importance of this moral character trait for predicting consequential harmful behaviors (Stuewig et al., 2015).

Although previous research has shown that self-reported moral character traits and assessments made by well-acquainted others predict unethical behaviors in anonymous research settings, we currently do not know whether we can predict strangers' moral character, nor do we know how to elicit relevant information from strangers in order to make reliable and accurate predictions about their unethical behavior. The biggest challenge in assessing strangers' moral character is that moral character is an extremely evaluative, if not the most evaluative, trait (i.e., high in social desirability). Moral character plays a central role in shaping how we view ourselves (Fernandez-Duque & Schwartz, 2016) as well as how others view us (Goodwin, Piazza, & Rozin, 2014; Goodwin, 2015). Therefore, both how people view themselves and how they convey their moral character toward others are likely to be distorted because of ego-protection motivations (Asendorpf & Ostendorf, 1998; Vazire, 2010). This means that to accurately judge strangers' moral character, we need a way to reveal aspects of targets' moral character beyond what those targets report themselves.

As a first step toward answering the question of whether and how we can make valid judgments of strangers' moral character, this dissertation introduces a new person-perception framework, the *hidden information distribution and evaluation* (HIDE) model. This model distinguishes two qualitatively different aspects of information that can determine the validity of perceptions and/or evaluations about targets of judgment: information that individuals do not correctly know about themselves (which I call the *hidden-self*) and information individuals know about themselves but misrepresent to others (which I call the *hiding-self*). The HIDE model articulates when and why groups of judges who do not know targets can make more valid judgments of targets than either the targets themselves or their acquaintances. In particular, this model predicts that the impromptu thinking and language usage that arises when people answer specially designed interview questions reveal information about their hidden-self and/or hiding-self that the targets cannot control, thereby enabling judges who are unacquainted with the targets to reliably detect targets' moral character, even in the contexts where people are highly motivated to convey good impressions.

The HIDE model and its implications for moral character judgments have the potential to make several groundbreaking theoretical and applied contributions to organizational psychology and related fields. For example, in many interview settings, judges (e.g., potential employers) are limited to evaluating targets' (e.g., job candidates') moral character from small samples of linguistic cues (e.g., candidates' responses to interview questions). Yet, we currently do not know whether character judgments based on verbal or written linguistic cues are diagnostic of unethicality, and if they are, we do not know how to elicit particularly relevant linguistic cues from targets. These are critical issues for organizations considering that interview methods are a centerpiece of employee selection procedures (Huffcutt, Iddekinge, & Roth, 2011) and that moral character judgments can be an important means to identify individuals who might harm organizations and the people within them. More broadly, this research paves the way toward theoretical development in our understanding of what moral character is, how it is revealed in written responses to interview questions, and how to assess it.

The Hidden Information Distribution and Evaluation (HIDE) Model

At its highest level, the HIDE model, presented in Figure 1, compares the evaluation process of two rating sources: self and judge. In the HIDE model, "judges" refer to those people who provide other-reports (as opposed to self-reports) of the targets being evaluated. These judges could be the targets' acquaintances, or they could be strangers. The HIDE model differentiates *perception* from *reporting decision*, and in doing so, helps us understand obstacles to accurate person evaluation. On the one hand, invalid ratings can stem from incorrect perceptions, such as when an evaluator does not correctly know a target. This type of invalidity does not involve an evaluator's intention; it happens unconsciously. On the other hand, invalid ratings can stem from conscious, biased reporting decisions, such as when an evaluator perceives a target correctly, but nonetheless decides to misrepresent or omit certain information about the person in their reports. The HIDE model incorporates these two different mechanisms to build a deeper understanding of when and why a particular rating source has more validity (i.e., is more accurate) than others.

For each rating source, the HIDE model assumes that these two distinctive processes—conscious and unconscious—jointly determine how the available information of interest (i.e., evaluation domain) about the target is distributed into three categories: 1) valid information (*correctly-identified information*); 2) invalid information (*incorrectly-identified information*), which is comprised of errors and reporting biases; and 3) no information (*hidden information*).

The HIDE Model for Self-Reports

The top half of Figure 1 depicts the HIDE model for self-reports. The center, solid-lined oval represents the information offered through a self-report. The dashed ovals to the left and right represent two distinct processes—conscious and unconscious— that intervene and affect the validity of the report. The various segments created by the overlapping ovals represent the full range of information that is available about the individual, whether it is reported (shown as shaded) or not (shown as unshaded).

The correctly-identified-self component (blue area located in the center of the model) describes information that targets know about themselves and that they report correctly. Researchers often assume that self-reports measure the correctly-identified-self component, but it is important to recognize that self-reports often include invalid information as well. Invalid information is captured by the incorrectly-identified-self components of the model, in the two red areas created by the overlap of a dashed oval and the center solid oval: self-deception and impression management. Self-deception refers to errors in how targets understand themselves. Impression management refers to targets having a correct understanding of themselves but misrepresenting that information to others, usually (but not always) in a positive manner. The incorrectly-identified-self component, therefore, captures both controllable (impression management) and uncontrollable (self-deception) aspects of invalid information. The combination of the correctly-identified-self and incorrectly-identified-self components of the model capture the total information available in a self-report.

Information not included in self-reports is captured by the self-ignorance and self-screening components of the model, depicted as the unshaded portions of the dashed ovals. Self-ignorance (on the left) is information that the target is unaware of and therefore cannot report. Self-screening (on the right) describes information that the target is aware of but decides not to report.

In summary, invalidity in self-reports is driven by two separate mechanisms, an unconscious process (represented by the dashed oval on the left) and a conscious process (represented by the dashed oval on the right). The unconscious process results in the self-ignorance and self-deception components of the HIDE model, which together comprise the ***hidden-self***. The conscious process results in the self-screening and impression management components of the HIDE model, which together comprise the ***hiding-self***. The hiding-self reflects targets' conscious decision making, which further contributes to invalidity in self-reports.

The HIDE Model for Judge Reports

Other-reports from judges can capture information that is hidden from or incorrectly identified by the self, thus providing insights that self-reports miss. The bottom half of Figure 1 depicts the HIDE model for judge reports, which parallels the self-report section of the model using three overlapping diamonds. As in the self-report section, the judge report model shows that information about the target is distributed into three categories: 1) valid information that the judge perceives about the target's characteristics of interest (i.e., correctly-identified-target component, shown in green in the center of the model), 2) invalid information that the judge perceives because of errors or biases (i.e., incorrectly-identified-target component, shown as the areas where the center diamond is overlapped by the dashed diamonds on the right and left), and 3) information that judges do not know and therefore cannot report (i.e., judge-ignorance, shown as the unshaded area within the dashed diamond on the left) or that they know but choose not to report (i.e., judge-screening, shown as the unshaded area within the dashed diamond on the right).

The judge-error section of the incorrectly-identified-target component captures information about the target that judges are not able to correctly recognize, whereas the judge-bias section captures the information that judges are able to correctly recognize but are motivated to misreport in an effort to make the target look better or worse than they actually believe them to be. This might happen after a job interview, for example, when a judge is motivated to make his or her favored candidate look particularly good, or make a disfavored candidate look particularly bad.

The combination of the correctly-identified-target and incorrectly-identified-target components (the total shaded areas of the model) capture the totality of the information available in the judge report. The combination of the ***hidden-target*** (judge-ignorance and judge-error) and ***hiding-target*** (judge-screening and judge-bias) components together capture the information that judge reports do not report completely or correctly.

In summary, similar to self-reports, invalidity in judge reports is drive by two separate mechanisms, an unconscious (represented by the dashed diamond on the left) and a conscious process (represented by the dashed diamond on the right). The unconscious process results in the judge-ignorance and judge-error components of the HIDE model, which together comprise the ***hidden-target***, the aspects of targets judges cannot perceive targets correctly. The conscious process results in the judge-screening and judge bias components of the HIDE model, which together comprise the ***hiding-target***. The hiding-target reflects judges' conscious misreporting about targets, which contributes to invalidity in judge reports.

Well-Acquainted Judges versus Unacquainted Judges

Judge reports can be provided by people who know the target well (i.e., well-acquainted others) or by strangers who have no relationship with the target but nonetheless have access to information about targets' characteristics relevant to the evaluation dimension. We often assume that well-acquainted others will be better judges of personality than strangers, and studies generally support this claim (Funder 1995, Kenny, Albright, Malloy & Kashy, 1994). However, the HIDE model suggests that, in some circumstances, evaluations made by strangers can be more informative than those provided by well-acquainted others, even though acquaintances have the opportunity to observe targets in various situations over time. In particular, strangers are likely to be more accurate than friends or other well-acquainted others when the relationship hinders the ability to form correct impressions or to report correctly about targets. Using the language of the HIDE model, judgments by strangers are likely to be more accurate than judgments from well-acquainted others in circumstances in which the latter's evaluations can be heavily influenced by the hidden- and hiding-target components. On one hand, the relationship that judges have with targets can distort judges' ability to correctly perceiving targets' characteristics, thereby decreasing the validity of evaluations of socially desirable traits. On the other hand, judges' conscious, purposeful misrepresentation is also more likely to happen when a relationship exists between the judge and the target. Even when well-acquainted judges are able to correctly recognize targets' characteristics, they might decide to misreport in an effort to make the target look better (or worse) than they actually believe them to be in an effort to help (or harm) targets.

Balance theory (Heider, 1958; Insko, 1981) explains why people tend to perceive or believe good things about their friends and bad things about their enemies. According to balance theory, individuals' perceptions of others depend on the social relationships these individuals share (Insko, 1981). When a judge has a positive relationship with a target (e.g., friendship), the judge tends to ascribe high value to the target on positive traits, but lower value on negative traits. This pattern achieves balance (i.e., consistency) between the positive "unit relationship" of having a friendship with the target and the positive evaluations people have of good traits. In other words, the following three cognitions are balanced: This person is my friend (+); I value this trait (+); my friend has high standing on this trait (+). Imbalance in this triad of cognitions leads to cognitive dissonance and motivation to reduce the inconsistency. Judges acting on this consistency motive might unconsciously misconstrue targets to maintain balance in their cognition about their friends and avoid the discomfort of inconsistency. Also, balance theory can explain why judges might consciously misrepresent targets to others in a more positive (or negative) manner than they believe them to be: This person is my friend (+); I want to view them positively (+); I will ensure my friend is positively evaluated by others (+). Well-acquainted others' evaluations, therefore, can be susceptible to both conscious or unconscious bias in evaluations. By definition, unacquainted judges do not have relationships with targets, and thus their evaluations should be less susceptible to the hiding-target zone of the HIDE model. Of the various person perception theories and frameworks in the psychology literature (e.g., Vazire, 2010; Funder, 1995, 1999, 2012), the HIDE model is the only one to specifically make this prediction that strangers can, in some circumstances, make more accurate evaluations of targets than can well-acquainted others.

Validity and Reliability Using Unacquainted Judges

In situations where unacquainted judges evaluate multiple targets and we can assume that their evaluations reflect their perception (i.e., no misrepresentation occurs), the HIDE model of judge reports is reduced to the combination of hidden-target and correctly-identified target zones (See Figure 2). The validity of judges' evaluations is determined, then, by how much information about targets is captured in the correctly-identified-target zone compared to the hidden-target zone.

Funder's (1995; 1999; 2012) Realistic Accuracy Model (RAM) model provides a theoretical base that informs the HIDE model, specifically regarding how to reduce the hidden-target zone so that judges' evaluations reflect correctly-identified information about targets. The RAM describes four necessary steps for a valid interpersonal judgment. First, relevance: the target must provide cues relevant to the trait being judged. Second, availability: the trait-relevant information must be available to the judge. For example, judges need to have an opportunity to observe targets' behaviors that are associated with the focal trait. Third, detection: the judge must be able to detect available and relevant information about the trait, meaning that the judge must have sufficient ability and motivation to see and understand the information and not ignore it. Finally, utilization: the judge must use the trait-relevant, available, and detected information correctly and not misinterpret it. The HIDE model extends the RAM in two key ways. First, the HIDE model articulates that implicit cues should be elicited to uncover information that is located in the hidden-self or hiding-self zone. Also, the HIDE model provides insight on how to satisfy the necessary conditions of the detection and utilization phases in the RAM: using multiple judges to combat unreliability.

Eliciting Relevant Cues

To ensure that judge reports reflect correctly-identified knowledge about targets, we need to provide judges with a sufficient amount of relevant and valid cues about targets' characteristics (Funder, 1995; 1999). The HIDE model of self-reports provides a useful guideline how to generate relevant and valid cues about targets. If we could assume that targets' hidden-self and hiding-self zones are minimal, then generating cues from targets would be relatively simple. For example, judges could directly ask targets about their characteristics. However, self-perception and direct representation are not always trustworthy, and hidden and hiding-self zones can be substantial. Cue generation should focus, then, on implicit aspects of targets' characteristics that targets do not correctly know about themselves or that targets cannot fake. Methods for generating implicit cues are discussed in a later section of this chapter.

Using a Group of Judges for Reliable Detection and Utilization of Cues

Even when provided with a sufficient amount of valid and relevant cues about targets, an individual judge might not be able to fully process that information well enough to make a reliable or valid evaluation. In the literature on judgment and decision making, it is well-established that an individual's judgment is often unreliable, as human judges inconsistently detect and utilize cues across targets of judgment (e.g., Arvey & Campion, 1982; Highhouse, 2008). The inconsistent detection and utilization of cues, then, increase the risk of having a large hidden-target zone, and thereby harm the validity of judge reports. One simple solution for this type of unreliability is to use aggregated evaluations from multiple judges to offset individual judges' idiosyncratic detection or utilization errors. Consistent with this reasoning, research on the "wisdom of crowds" shows that collectives composed of independent judges often make more accurate judgments and decisions than do solo individuals (Davis-Stober et al., 2014; Larrick and Soll, 2006; Mannes, 2009). Therefore, while each individual might not be able to judge targets reliably, collectives might be able to make reliable and valid judgments.

Relationship of the HIDE Model to Existing Person Perception Models

The difference between targets' self-perception versus well-acquainted judges' perception of targets is closely related to Hogan and Shelter's concept of inner and outer personality (Hogan, 1996; Hogan & Shelton, 1998). Inner personality is measured by self-reports and captures one's internal motivation and identity. Outer personality, in contrast, is measured by other-reports and captures how the target is viewed by his or her acquaintances based on the target's observable behaviors in social interactions. The HIDE model extends the understanding of self- and other-perceptions by providing a mechanism for understanding when and why one's inner and outer personality converge or diverge from one another. The more that the judge reports capture knowledge in the hidden-self and incorrectly-identified-self components, the more likely the self- and judge reports are to diverge. In this case, self- and judge reports can provide complementary and non-overlapping information. In contrast, when judge reports are closely aligned with the correctly-identified-self component (self-knowledge) and self-reports are closely aligned with the correctly-identified-target component (judge knowledge), self- and judges-reports are more likely to converge.

The HIDE model also extends prior research comparing the accuracy of self-perceptions compared with others' perceptions. To understand the relative validity of self- and judge reports, several interpersonal judgment models have been proposed. For example, the Johari Window (Luft & Ingham, 1955) partitions personality knowledge into four categories: aspects that the target and others both know (arena), aspects that only the target knows (facade), aspects that only others know (blind spot), and aspects that neither knows (unknown). Building on the Johari Window, Vazire (2010) developed the Self-Other Knowledge Asymmetry (SOKA) model, suggesting that targets are more accurate than others in judging traits that are low in observability (e.g., neuroticism) but that others are more accurate than targets when the traits are evaluative (i.e., highly socially desirable; e.g., intellect-related traits). In general, empirical studies in this area support the idea that evaluativeness and observability are important determinants of the accuracy of self- and other-perceptions across traits (e.g., Asendorpf & Ostendorf, 1998; Connelly & Ones, 2010; Connelly & Hűlsheger, 2012; Gosling, John, Kenneth, & Robins, 1998; Human & Biesanz, 2011; Kolar, Funder, & Colvin, 1996; Oh, Wang, & Mount, 2010; Vazire, 2010; Vazire & Mehl, 2008).

An assumption inherent in the previous models of person perception is that an invalid rating from an evaluator signifies inaccurate knowledge about the target. However, it is possible that the evaluator perceives the target correctly, but nonetheless decides to misrepresent or omit certain information about the person, resulting in invalid reports. Therefore, invalid ratings can stem from two different mechanisms: incorrect perception and biased reporting decision. The HIDE model differentiates the *perception* from the *reporting decision*, and thus incorporates these two mechanisms to build a deeper understanding of when and why a particular rating source has more validity than others.

Finally, the HIDE model extends Funder's (1995; 1999; 2012) Realistic Accuracy Model (RAM) about how a valid interpersonal judgment can occur. The RAM describes four necessary conditions for a valid interpersonal judgment: *relevant* cues about targets' characteristics should be *available* to judges, and judges should *detect* and *utilize* those cues. Each condition influences the extent to which the target's trait is connected to the judge's correct evaluation of that trait. Therefore, the validity of interpersonal judgments is likely to be high when the information provided is strong in quantity and quality ("good information"), the focal trait is visible and easily judged ("good trait"), the target is judgeable ("good target"), and the judge is well-calibrated ("good judge") (Funder, 2012). The "good trait" component of the RAM is connected to the SOKA model (Funder, 2012): in the RAM, trait evaluativeness is detrimental to accurate person perception because self-deception and impression management tactics distort the availability and relevance of cues. Trait observability, on the other hand, improves the accuracy of person perception because more visible traits are more available to judges and easier to detect, hence judges' evaluations are more likely to be accurate (Funder, 1995). The HIDE model extends the RAM in two key ways. First, the HIDE model articulates that implicit cues should be elicited to uncover information that is located in the hidden-self or hiding-self zone. Also, the HIDE model provides insight on how to satisfy the necessary conditions of the detection and utilization phases in the RAM: using multiple judges to combat unreliability.

In sum, the HIDE model extends prior models that theorize about the relative validity of self-perceptions and others' perceptions (i.e., inner vs outer self, the Johari and SOKA models, and the RAM) by distinguishing between perception and reporting decisions, both in self- and judge reports. In the HIDE model, the relative validity of self- and judge reports are jointly determined by the extent to which evaluators perceive the target correctly and the extent to which they decide to represent those perceptions correctly. While judges' evaluations of targets can be insightful when they reveal information that is hidden or incorrectly identified by the targets themselves, judge reports can also suffer from hidden and incorrectly identified information. Accordingly, the relative validity of self- and judge reports depends on the distribution of information that is detectable to and reported by targets and judges. If one party's knowledge about a target is aligned with the other party's incorrectly-identified or hidden information, then the former's evaluation is informative above and beyond the latter's (i.e., incremental validity).

Applying the HIDE Model to Moral Character Judgment

Social desirability is a critical factor when considering the relative validity of self-reports versus judge reports. Prior research has shown that people often hold biased perceptions of themselves on desirable dimensions such as attractiveness or intelligence (Vazire 2010, Vazire and Mehl 2008). Using the language of the HIDE model, the more desirable the trait of interest, the more self-reports will reflect the incorrectly-identified-self components (i.e., self-deception and impression management). Moral character is an extremely desirable trait, if not the most socially desirable trait; people have a strong desire to see themselves as moral, which leads to self-deception. Moreover, people want to be seen by others as moral, leading to impression management when representing themselves to others. Together, self-deception and impression management increase the likelihood that information captured by self-reports will reflect the incorrectly-identified-self components of the HIDE model. Consequently, judge reports could complement or replace self-reports to the extent that they tap into valid information from the incorrectly-identified-self component. Moreover, judge reports could capture information in the hidden-self (i.e., self-ignorance) and hiding-self (i.e., self-screening) components that are not accessible to or not revealed by the individuals providing self-reports.

Judge reports of moral character can be provided by people who know the target well (i.e., well-acquainted others) or by strangers who have no relationship with the target but nonetheless have access to information about their moral character. The HIDE model predicts that strangers can produce more accurate evaluations than well-acquainted others when the relationship hinders a judge's ability to correctly construe or report targets' moral character. Moreover, even when well-acquainted others can form accurate evaluations of targets, they still can decide to misrepresent them to others to help or harm targets. By definition, strangers do not have relationships with targets, and thus their evaluations of targets' moral character should be less likely to be pushed into the judge-bias and judge-error zones of the HIDE model.

Instead, reducing the hidden-target zone of the model is an important condition that could allow strangers to form accurate moral evaluations of targets. For the judge reports to complement or replace self-reports, the correctly-identified-target component should include knowledge contained in the hidden-self and/or incorrectly-identified-self components of the HIDE model. It follows, then, that it is necessary to develop a tool that judges can use to accurately extract information about moral character traits that the targets themselves are unaware of or less able to control.

Moral Character Judgment Using the Written Interview Method

An interesting and practical method that judges might use to evaluate strangers' moral character is to ask open-ended questions designed to reveal the "hidden" or "hiding" aspects of unethical tendencies—implicit aspects of moral character that job applicants are less able to control. In pilot studies of this dissertation with Taya Cohen and Abigail Panter, we wrote several interview questions to covertly elicit peoples’ unethical tendencies through their spontaneous written responses. In this dissertation, I use a subset of three of those questions to test the HIDE model predictions and understand how best to evaluate strangers’ moral character.

I focus on written responses to interview questions for several reasons. Most importantly, previous studies have shown that performing an expressive task (i.e., writing) requires an individual to engage in impromptu thinking, and the dispositions reflected in such expressions are difficult to counterfeit (Hojbotă, 2015). Second, evaluations based on written responses (compared to other media, such as face-to-face conversations) helps to reduce certain factors that might bias judges, such as a candidate’s attractiveness (Cann, Siegfried & Pearce, 1981).

The interview questions used in this dissertation are presented in Appendix 1. The questions were modeled after behavioral interview questions commonly employed in research and practice (Blackman, 2002; Hoevemeyer, 2005). Each interview question can reveal implicit aspects of moral character that can be predictive of targets’ unethical behavior across diverse contexts. What targets choose to write about (e.g., past events that are salient to them), whether they consider others’ needs in difficult situations, and how they feel when their behaviors might influence others (e.g., feeling guilty when their behaviors negatively influence others) are likely to provide judges with explicit and implicit information that enables them to make accurate moral character judgments. For example, the “mistake questions” asks job applicants to recall a mistake they made at work and to report how they felt and behaved at the time. Prior research has shown that unethical individuals experience less guilt following wrongdoing (e.g., Cohen et al., 2016). Although targets may not overtly identify as unethical, their responses to this question could reveal that they elaborate much less on past experiences of guilt following a mistake, and this response pattern would make it possible to identify them.

HIDE Model Predictions for the Written Interview Method

The HIDE model enables us to predict when and why judges who are strangers to targets can provide more valid evaluations than targets themselves or people who know targets well. Applying the HIDE model to moral character judgment, this dissertation proposes the written interview method to detect strangers' moral character and predict their unethical behavior. In this section, I summarize the HIDE model predictions about moral character evaluations using the written interview method.

First, the HIDE model predicts that a group of judges should be able to forecast targets' unethical behaviors by evaluating targets' moral character. The assumption of the written interview method is that judges should be able to detect and utilize linguistic cues related to targets' moral character from their written responses. However, individual judges might not be able to do so, as human judges are often inconsistent. Using the HIDE model terms, when we use only one judge's ratings to evaluate targets, we risk a large incorrectly-identified-target zone. One simple way to reduce this error is to use average-ratings from multiple judges. Accordingly, based on the HIDE model, I predict *the wisdom of crowds* in forecasting unethical behavior, such that although one judge cannot predict a target's unethical behavior, groups of judges' average evaluations can.

Second, the written interview method is designed to enable a group of judges to see implicit aspects of moral character that targets themselves are *unaware of*, and therefore judges' evaluations should predict targets' unethical behavior above and beyond targets' self-perceptions of moral character. When researchers evaluate targets using self-reports in anonymous settings, the resulting measures often capture targets' *explicit* views about themselves (i.e., self-perception), especially when targets are not given an incentive to misrepresent their characteristics. Considering that moral character is a highly desirable trait, targets' explicit views are likely to be distorted by targets' *hidden-self*. People want to see themselves as highly moral, and therefore, their self-perception of moral character is likely to be distorted. In contrast, targets' *implicit* unethical tendencies revealed in written responses enable judges to uncover information located in the hidden-self zone. Therefore, judges' evaluations of targets' moral character should capture some information unavailable in self-reports. Accordingly, based on the HIDE model, I predict that the average of judges' evaluations of targets should predict targets' unethical behavior above and beyond targets' perceptions of their own moral character.

Third, the HIDE model predicts that judges who do not personally know the targets can provide more valid evaluations of targets' moral character from written interview responses than can targets' acquaintances. This is because social relationships between targets and judges might hinder judges' ability to correctly perceive targets or honestly report their perceptions about extremely socially desirable traits. Therefore, based on the HIDE model, I predict that unacquainted judges' evaluations of targets using the written interview method can predict targets' unethical behavior to a greater degree than evaluations provided by targets' well-acquainted others.

Finally, the written interview responses are not only useful for revealing targets' hidden-self but also their *hiding-self*. This is because answers to the behavioral interview questions used in this dissertation can reveal implicit aspects of moral character that targets cannot control even when they want to. In other words, individuals with lower levels of moral character might try to convey a positively biased image of themselves to judges, but nonetheless, judges can see through this attempt because aspects of moral character revealed in written interview responses are implicit and not controllable. For example, the employer question (see Appendix 1) asks targets about how their employer would describe them. Targets who are low in the character trait of humility might over-emphasize their positive qualities, and in doing so, expose to judges their self-focus and low levels of humility. Therefore, the more they engage in impression management, ironically, the more their responses to the interview questions reveal their true moral character. Accordingly based on the HIDE model, I predict that judges' evaluations of targets can be more predictive of unethical behavior for targets who misrepresented themselves to judges to a greater degree. I test this prediction in chapter 4.

Overview of Empirical Studies

The remaining chapters of this dissertation report on a total of eight empirical studies. Chapters 2, 3, and 4 examine four HIDE model predictions for the written interview method using seven empirical studies. Chapter 5 is an exploratory chapter that demonstrates the possibility of shared biases between human judges and introduces the future direction of my research program.

Chapter 2 consists of three empirical studies that investigate the wisdom of crowds in forecasting unethical behavior. In studies 1 and 2, large groups of judges were crowd-sourced online to test whether the average rating of judges' moral character evaluations predicts targets' frequency of unethical behavior. Study 3 extends the findings of studies 1 and 2 by directly examining the minimum number of judges required to make reliable judgments about strangers using the written interview method.

Chapter 3 seeks to achieve two main goals via three empirical studies. First, I investigate whether judges' evaluations using the written interview method can capture information located in the hidden-self zone—aspects of targets' moral character that targets do not correctly know about themselves and thus fail to accurately capture in self-reports. Second, I investigate whether unacquainted judges can capture information located in well-acquainted judges' hidden-target zone—distorted perceptions about targets that arise from a preexisting relationship between targets and judges. I show that unacquainted judges' evaluations indeed capture unique variance about targets' moral character that self-reports or ratings by well-acquainted judges cannot, and thus that unacquainted judges' evaluations can be more predictive of targets' unethical behavior that the other two evaluation methods.

Chapter 4 describes an empirical study to investigate whether judges' evaluations using the written interview method can uncover information located in targets' hiding-self zone—information that targets purposefully misrepresent about themselves. In Study 7, I manipulate the incentive structure of the interview questions and measure the extent to which targets answer in a positively biased manner. I examine whether judges' evaluations are more predictive of targets' unethical behavior when targets try to answer questions in a more positively biased manner.

Finally, chapter 5 of this dissertation does not conduct hypothesis testing, but rather takes an exploratory approach to investigate how to further increase the predictive validity of unethical behavior using the written interview method. While using multiple judges can be an important means of reducing the unreliability of human judgments and therefore increasing the validity of judge reports, certain aspects of invalidity in human judgments—those that are due to basic limitations in cognitive capacity or to widely shared cognitive biases—are unresolvable. In these situations, machine algorithms (e.g., automatized text analyses, machine learning) can complement human judgments. In Study 8, I conduct automated text analyses to compare the relationship between certain linguistic cues, targets' unethical behavior, and judges' moral character evaluations. Based on these exploratory text analyses, I discuss the future directions of this research program, especially concerning the possibility of combining human judgments and machine algorithm to further increase the accuracy of unethical behavior forecasts using the written interview method.

CHAPTER II

The Wisdom of Crowds in the Detection of Moral Character

Information about targets' moral charcter extracted via behavioral interview questions may be located in the correctly-identified-target or hidden-target zones of the HIDE model of judge reports. Because the judges do not know the targets in the written interview method, the hiding-target zone, which encompasses the judge-screening and judge-bias zones of the model, should be minimal because there is no relationship between the target and the judge. The validity of judges' moral character evaluations using the written interview method is determined, then, by how much information about targets is captured in the correctly-identified-target zone compared to the hidden-target zone.

The hidden-target zone of the HIDE model for judge reports encompasses both judge-ignorance and judge-error. Judge-ignorance can happen when behavioral interview questions do not produce any relevant cues about targets' moral character, and thus judges cannot make any viable inferences. Therefore, cue generation is the most fundamentally necessary condition for an accurate judgment. An expressive task like writing is difficult to counterfeit, and thus the written interview method was proposed as a way to reveal targets' hidden-self or hiding components of moral character (Hojbotă, 2015). My prediction is that targets' impromptu thinking and language usage captured in written responses to specific interview questions will reveal enough information about targets' moral character to allow judges to make valid character assessments. For example, the subjects that targets choose to talk about (e.g., past events that are salient to them), whether they consider others' needs in difficult situations, and how they feel when their behaviors might influence others (e.g., feeling guilty when their behaviors negatively influence others) are likely to provide judges with cues that enable them to make valid moral character judgments.

Even when judges have access to ample valid cues about targets' moral character, judge-error can happen when judges do not consistently detect and utilize cues across targets, resulting in incorrect ratings (Funder, 2012). Reliability is a prerequisite condition of validity: judge-error stemming from individual idiosyncrasies in detecting and utilizing cues can significantly lower the validity of judges' evaluations. Consistent with this reasoning, prior studies examining the methodology of interviewing suggest that judges often lack reliability, which results in low validity (e.g., Arvey & Campion, 1982; Highhouse, 2008). For example, Highhouse (2008) explained that the interrater reliability of interviews is low because interviewers often focus on irrelevant factors, apply different standards to different applicants, and inconsistently use evaluation criteria. Therefore, the HIDE model describes another necessary condition for an accurate judgment: using aggregated evaluations from multiple judges (rather than a single judge). This recommendation is consistent with research on the "wisdom of crowds," which explains why collectives tend to make better decisions than individuals (Davis-Stober, Budescu, Dana, & Broomell, 2014; Larrick, Mannes, & Soll, 2012; Larrick & Soll, 2006; Mannes, 2009). The "wisdom of crowds" is based on the premise that the aggregate of multiple independent judgments will be more valid than a single judgement because high and low errors offset each other. For example, a very positive or lenient judge who rates all candidates highly will be offset by a very negative or conservative judge who rates all candidates poorly. Building upon the wisdom of crowds literature, the HIDE model articulates that by using aggregated evaluations from multiple judges, individual judges' idiosyncratic errors in detecting and utilizing cues across targets can be canceled out, resulting in more reliable evaluations.

In sum, the HIDE model suggests that we can make valid evaluations of strangers' moral character if the behavioral interview questions produce a sufficient amount of valid cues and if we use multiple judges to reliably detect and utilize those cues. If these two conditions are satisfied, judges' evaluations can be valid, and thus predict targets' unethical behaviors in various contexts.

In this chapter, I describe two empirical studies that I conducted to examine how well groups of naïve judges predict targets' unethical behaviors observed in the laboratory (studies 1 and 3), and one study that examines how well such judgments predict reports of unethical behavior in organizational settings (Study 2). By showing that judges' aggregated evaluations of targets' moral character predict targets' unethical behavior, I provide initial evidence in support of the HIDE model as well as the written interview method as a method for predicting strangers' unethical behavior.

Large sets of judges were crowd-sourced online in studies 1 and 2 to test the assumption that using multiple judges can be an important means to reduce unreliability. However, adding judges increases costs, so organizations interested in using this methodology need information on how many judges are required to produce reliable predictions. Therefore, in Study 3, I investigate the number of judges required to make sufficiently reliable evaluations using the written interview method and show that small groups of naïve judges (namely, six) can predict targets' unethical behavior by evaluating their moral character using the written interview method.

Study 1

In Study 1, I investigated the predictive validity of judges' evaluations of targets' moral character based on reading targets' written responses to interview questions. The targets' unethical behaviors were observed in a laboratory experiment in which they had the opportunity to over-report their performance on a problem-solving task to earn additional money. I examine whether the aggregated evaluations of multiple judges predict how frequently targets engage in cheating.

Study 1 Method

First, two behavior-based interview questions were developed to extract moral character information from interviewees based on pilot studies I conducted with Taya Cohen and Abigail Panter. The questions were modeled after behavioral interview questions commonly employed in research and practice (Blackman, 2002; Hoevemeyer, 2005):

- Please tell us about a time when you made a mistake at work. How did you feel when this occurred? What did you do? What, if anything, did you learn from this experience? [*Mistake*]
- Please describe an experience in which you were faced with a difficult dilemma at your job—a situation where you found it hard to decide what to do. What factors did you consider? What did you do? What, if anything, did you learn from this experience? [*Dilemma*]

Each interview prompt was developed to reveal aspects of traits diagnostic of unethical tendencies. The mistake questions ask respondents to recall a mistake they made at work and to report how they felt and behaved at the time. Prior research has shown that individuals with low levels of moral character tend to experience little guilt following wrongdoing (e.g., Cohen et al., 2014). Although these individuals may not overtly admit to feeling relatively little guilt following a mistake, if their responses to this interview prompt lack elaboration on their feelings, this pattern could make it possible to identify them.

The dilemma interview questions give targets the opportunity to reveal the extent to which they are considerate of others and mindful of how their decisions and actions affect other people. We designed this question based on the assumption that high-moral-character targets would be more likely than low-moral-character targets to mention such considerations.

Each target responded to one of the two questions after reading the following instructions.

> *Imagine that you have been selected to interview for your dream job. The employers want to conduct an online interview before you meet them face to face. You will be asked questions about yourself and past experiences you may have had. Please use real examples from your life when responding. Please do not include last names or any other personally identifiable information in your response. Remember: you need to answer the following questions honestly, but in a way that makes you look like the best possible job candidate.*

Data Collection from Targets

The targets in this study were 195 U.S. adults who participated in an experiment in a mobile research laboratory parked in the city of Pittsburgh, Pennsylvania[1]. In addition to answering one of the two interview questions, participants completed a problem-solving task in which they had the opportunity to lie about their performance, and a computerized survey in which they answered questions capturing demographic information as well as two invetories that capture information about targets' self-perceptions of their moral character traits: the five-item guilt proneness scale (GP-5; Cohen, Kim, & Panter, 2014), and the HEXACO-60 personality inventory (Ashton & Lee, 2009). The guilt proneness scale and the honesty-humility and conscientiousenss scales in the HEXACO inventory each capture distinctive aspects of moral character (Cohen et al., 2014). Self-reports of honesty-humility, conscientiouenss, and guilt proneness were analyzed to examine whether judge reports of targets' moral character predict objectively measured unethical beheaivor above and beyond targets' self-perceptions of their moral character traits. The four other HEXACO scales included in the 60-item inventory are less relevant to moral character (Cohen et al., 2014), and were included in the survey to mask the fact

[1] Two additional participants completed the study but were excluded from the analyses because they answered 19 out of 20 items correctly on the problem-solving task, and therefore had little opportunity to cheat compared to other participants. The decision to exclude targets who answered 19 (out of 20) questions correctly was made prior to data analysis.

that the purpose of the study was to examine moral character traits. Given that this research is moral character evaluations and unethical behavior, I did not analyze the data from the other HEXACO scales.

The problem-solving task was based on methods used by Shu, Mazar, Gino, Ariely, and Bazerman (2012). Participants were given a worksheet containing 20 matrices with 12 three-digit numbers within each matrix. They had five minutes to find two numbers in each matrix that added to 10.00. Each correctly identified pair of numbers was worth $0.25 in earnings, for a maximum bonus payment of $5.00. Participants learned that they would work on the task for five minutes and then would be asked to calculate the number of problems they solved correctly and indicate this number, and how much money they should be paid, on a payment form after they had recycled the matrices worksheet. Unbeknownst to the participants, we were able to link each participant's problem-solving performance to his or her payment form by a three-digit identifier contained in each of the documents: one three-digit number in the bottom matrix on the problem-solving worksheet was identical to three digits in the payment form number. At the end of each day of data collection we gathered all the matrices worksheets from the recycling bin and compared each participant's reported performance on the payment form to his or her actual performance on the worksheet. Participants were considered to have cheated when the number of problems they reported solving was greater than the number they actually solved correctly on the worksheet.

After participants completed the problem-solving task and filled out the payment form, they completed a computerized survey that included a question asking them to describe themselves, one of two interview questions (either the dilemma or the mistake questions), the GP-5, the HEXACO inventory, and demographic questions. Following the computerized survey, participants handed their payment forms to the experimenter, were paid according to the number of problems they indicated solving on the payment form, and were provided with a debriefing form that explained that the true purpose of the study was to examine cheating.[2]

Data Collection from Judges

One hundred and fifty-two participants (55.9% were female) were recruited from a university-administered subject pool to complete a web-based study, in which they judged the study 1 targets' moral character. They were given class credit for their participation. Judges read the following instructions:

> *In making your judgment of moral character, please consider the following definition.*
>
> *Moral character is a term used to describe an individual's disposition to think, feel, and behave in an ethical manner. People with high levels of moral character consider the needs and interests of others, and how their own behavior affects other people. When they do something wrong they feel guilty and try to correct for what they did, even if no one knows about it. In general, those with high moral character are benevolent, trustworthy, and compassionate. In contrast, people with low levels of moral character are callous, manipulative, and more focused on themselves than on other people. When they do something wrong they are unlikely to feel bad about their behavior or attempt to correct for their mistakes. In general, those with low moral character are cruel, dishonest, and inconsiderate.*

Each judge rated moral character by responding to the question: Do you consider the author of this response to be a moral person? *[1 (Extremely weak moral character), 2 (Weak moral character), 3 (Neither weak nor strong), 4 (Strong moral character), 5 (Extremely strong moral character)]*. Each judge rated interview responses from 20 randomly selected targets. Each interview response was rated by an average of 15 judges; the ratings for each target were averaged before conducing analyses.

[2] We administered the computerized survey after the problem solving task so as not to influence the assessment of unethical behavior. We recognize the limitations of this design choice. Accordingly, in subsequent studies we reversed the assessment order such that the interview questions were asked prior to the assessment of targets' unethical behavior.

Study 1 Results

The criterion variable, cheating, is operationalized as the number of matrices the participants claimed they solved minus the number they actually solved correctly. Descriptive statistics and correlations among targets' cheating frequencies, self-reported moral character traits, and judges' average-moral-character-rating are presented in Tables S1A and S1B in Appendix 2.

Figures 3 and 4 depict the relationship between judges' average-moral-character-ratings and the extent to which targets cheated on the problem-solving task in the mistake and dilemma questions conditions. As expected, the more a target cheated in the problem-solving task, the lower the moral character rating that target received from judges.

I conducted negative binomial regressions for each interview question condition. Each analysis controlled for the number of correctly solved matrices, because participants who solved more matrices correctly had less opportunity to cheat. In total, three different sets of analyses were conducted. The results were similar regardless of whether the mistake and dilemma questions were analyzed together or separately. The results from the separate analysis for each question are presented in Table 1.

The first analysis included only the judge reports. The results indicated that judge-reported moral character negatively and significantly predicted the extent to which targets cheated in the problem-solving task, regardless of whether those ratings were made from targets' written interview responses to the mistake or dilemma questions. These results, therefore, support hypothesis 1: groups of judges can predict targets' unethical behavior by evaluating their moral character from written responses to behavioral interview questions about mistakes or dilemmas.

The second model analyzed only the self-reports as measured by the honesty-humility, conscientiousness, and guilt proneness scales. Self-reported conscientiousness negatively, and marginally significantly, predicted the extent to which targets cheated in the mistake questions condition only. Finally, in the third model, targets' frequency of cheating was regressed on both judge- and self-reports to test which rating source is more predictive. The results indicated that only the judge reports had incremental validity, which means that the judge reports were more informative than the self-reports in predicting cheating. The net effects of judges' moral character judgments were negative and significant for the mistake questions condition and negative and marginally significant for the dilemma questions condition.

Study 1 Discussion

In Study 1, judges' moral character evaluations significantly predicted the extent to which targets cheated on the problem-solving task, and this held when controlling for targets' self-reported moral character traits. Therefore, the findings of Study 1 are consistent with the prediction of the HIDE model, that groups of unacquainted judges are able to predict targets' unethical behavior by evaluating their moral character using the written interview method. Considering that judges' evaluations of moral character were based only on a brief paragraph of information provided by the target (word count average=76, See Table S1B in Appendix 2), this is a very powerful finding.

It is noteworthy that judge reports of targets' moral character were positively correlated with self-reports of honesty-humility (r=.15, p=.03), conscientiousness (r=.26, p<.001), and guilt proneness (r=.23, p=.001). Also, albeit less predictive of unethical behavior than judge reports, self-reports of honesty-humility, conscientiousness, and guilt proneness had negative relationships with cheating (See Table 1 and Table S1A in Appendix 2). Finally, the predictive validity of self-reports was reduced when the model simultaneously included judge-reported moral character. Together, these results indicate that the correctly-identified-target zone in the HIDE model of judge reports and the correctly-identified-self zone in the HIDE model of self-reports have some alignment.

Although the criterion I used to measure unethical behavior in Study 1—lying about one's performance on a laboratory task—has strong internal validity and was directly observable by the experimenter (as opposed to self-reported), it lacks clear external validity. The specific form of cheating we examined and the laboratory context in which it occurred do not correspond exactly to the kinds of cheating that occur in real-life settings. Therefore, investigating the predictive validity of this text-based interview method in more naturalistic settings would increase the generalizability of the laboratory findings.

Study 2

In Study 2, I investigated the predictive validity of the written interview method by focusing on full-time employees with counterproductive work behavior (CWB) as a criterion. Also known as workplace deviance, CWB is defined as employees' volitional behaviors that harm or intend to harm the people in an organization and the organization itself. CWB is perceived as unethical by employees in general (Cohen et al., 2014) and includes a wide range of unethical work behaviors, such as falsification of expense reports, stealing, and interpersonal abuse. An advantage of using CWB as a criterion variable is that it is not limited to a particular type of unethical behavior, such as cheating. Rather, it encompasses a wide range of harmful acts, including aggression (physical and verbal), sabotage, theft, and withdrawal.

Study 2 Method

Interview Questions

In addition to the mistake and dilemma questions, Study 3 used an additional question:

- How would your current or last employer describe you? *[Employer]*

I reasoned that targets' assessments of their employer's perceptions about them might be indicative of targets' humility, with high-moral-character targets being more modest and unassuming compared to low-moral-character targets.

Data Collection from Targets

The target participants in Study 2 were 495 employed U.S. adults recruited by an online survey firm (Qualtrics). These target participants were randomly assigned to answer one of the interview prompts[3]. Employees' CWB was measured using the 32-item inventory developed by Spector and his colleagues (2006). Related to the fact that CWB was measured with self-reports in this study, a meta-analysis found that self-reported CWB is positively correlated with coworker-reports of CWB, therefore providing validity evidence of CWB when it is measured by self-reports (Berry, Carpenter, & Barratt, 2012). Moreover, this meta-analysis suggested that although both self-reports and coworker-reports of CWB each have some reliability and validity, self-reports are likely to be more valid than coworker-reports because employees try to hide their CWB from coworkers, and thus coworkers do not have as much information about employees' CWB as employees' themselves do. Finally, participants were administered the HEXACO-60 revised personality inventory (Ashton & Lee, 2009) and the five-item guilt proneness scale (GP-5; Cohen et al., 2015) to measure self-perceptions of moral character traits. As in the prior study, the full HEXACO 60-item inventory was administered to mask the fact that the moral character was the focus of the study. Only the honesty-humility and conscientiousness scales from the

[3] These participants are a subset of participants in a larger project that investigated a larger number of interview questions. The results for other questions are available from the author.

HEXACO are relevant to investigating moral character and unethical behavior, so the other four HEXACO scales were not analyzed.

Data Collection from Judges

U.S. residents were recruited via Amazon's Mechanical Turk website (www.mturk.com) to provide ratings for targets who answered the mistake, the dilemma, or the employer interview questions. Each participant rated interview responses from 20 randomly selected targets from one of the three interview question conditions. Eligible participants were those with an at least 90% approval rating on previous tasks. We excluded two participants prior to data analysis who did not meet our a priori selection criteria, which were to complete the study and pass the attention checks embedded in the survey. Our final sample size eligible for analysis was 409 participants, 52% of whom were female. Each interview response was rated by an average of 17 judges. The rating instructions and definitions of moral character traits were identical to those used in Study 1.

Study 2 Results

The predictive validity of the judges' rating average was tested using negative-binomial regression analyses. For each interview prompt condition, three sets of analyses were conducted. The first set of analyses examined the predictive validity of judge-reported moral character while the second examined the predictive validity of self-reported moral character traits. Finally, in the third set of analyses, both self- and judge reports were entered simultaneously. The results are presented in Table 2. The results indicate that, across all interview question conditions, judges' average rating of moral character negatively and significantly predicted the frequency with which targets reported engaging in CWB. Therefore, hypothesis 1 was supported with CWB as a criterion.

Self-reported honesty-humility, conscientiousness, and guilt proneness also negatively and significantly predicted CWB. Finally, when judge- and self-reported moral character traits were entered simultaneously, only self-reports provided incremental validity. Nonetheless, judges' moral character ratings, while not significant at the standard $\alpha < .05$ level, showed the expected negative patterns for all interview question conditions, and they were marginally significant for the Employer question condition.

Study 2 Discussion

I found that judges' average rating of targets' moral character predicts targets' workplace deviance, a criterion of unethical behavior in the workplace. Study 2 replicated the findings of Study 1, supporting the notion that judge reports based on targets' written interview responses reflect the correctly-identified-target knowledge rather than the hidden-target information. Therefore, studies 1 and 2 verified that two conditions of the HIDE model (i.e., generating valid cues, reliably detecting and utilizing available cues by using multiple judges) are satisfied.

In Study 2, self-reported moral character traits and the criterion, CWB, were both measured from the same reporting source (i.e., targets themselves). Therefore, the predictive validity evidence of self-reported moral character traits is somewhat vulnerable to common method-bias (i.e., shared variance) compared to a situation in which CWB would be measured from other reporting sources. However, the use of judge reports and CWB does not have this common method bias, and so the negative relationships between judge's moral character rating and targets' self-reports of CWB provides validity evidence of judges' moral character rating. In study 6 of chapter 3 of this dissertation, I measure CWB using coworker-reports to replicate the findings of study 2 and address the limitations of the research design used in this study.

Similar to the findings of Study 1, I found that self-reports and judge reports have shared variance, and that both reports predict targets' CWB, at least when analyzed independently. This means that the correctly-identified-self zone in the HIDE model of self-reports and the correctly-identified-target zone of the HIDE model of judge reports are somewhat aligned. This conclusion is based on correlations between judge reports, self-reports, and CWB, and the pattern observed when CWB is regressed on self-reports, on judge reports, or on both (See Tables A3 and A4 in Appendix 2). First, both self- and judge reports were predictive of CWB in the regression models. Second, judge-reported moral character was positively correlated with honesty-humility (r=.19, p<.001) and guilt proneness (r=.24, p<.001). Third, the strength of judge reports was reduced when self-reports were added to the model (Table 2), meaning that self- and judge reports explained common variance of CWB in the negative binomial regressions.

In Study 2, judges' moral character evaluations based on the employer question explained unique variance that self-reports did not explain. When controlling for self-reported honesty-humility, conscientiousness, and guilt proneness, the coefficient of judges' average-moral-character rating was marginally significant in the employer question (Table 2). This is quite interesting given that targets' written interview responses for this question were the shortest of the three (See Table S2B in Appendix 2). These results suggest that what is important to the predictive validity of judge reports is not how many, but what kind, of cues are available in written interview responses. Still, given that the criterion of Study 2 is not objectively measured, the relative predictive validity across different interview questions needs to be investigated further and is discussed in later sections of this dissertation.

Study 3

Study 3 aimed to achieve two main goals. First, I examined whether small groups of unacquainted judges can predict targets' unethical behavior. Second, I examined how many judges are required to reliably evaluate targets' moral character using the written interview method. In studies 1 and 2, a large number of participants were recruited to play the role of judges. Using multiple judges increases the reliability of aggregated evaluations; however, it also entails substantial costs (in terms of time, money, etc.). Indeed, most organizations employ relatively small groups of interviewers to evaluate job candidates. Therefore, it is important that we determine the minimum number of judges required to form reliable character judgments using this written-interview-response method. To do so, I used generalizability theory (Cronbach et al., 1963) to calculate the changes in inter-rater reliability as the number of judges varies. Generalizability theory analyses require that the same set of targets be evaluated by the same set of judges. In studies 1 and 2, calculating inter-rater reliability was not possible because judges were randomly assigned to different sets of targets. In Study 3, however, six judges read and evaluated the entire set of targets, thus allowing me to conduct generalizability theory analyses.

Study 3 Method

Six undergraduate research assistants were recruited to read the entire set of interview responses from Study 1, then rate each target's specific moral character traits, overall moral character, and other characteristics[4]. Each judge indicated their rating of overall moral character by responding to the question: *Do you consider the author of this response to be a moral person?*

[4] After indicating their judgment of moral character, each judge answered three additional questions related to moral character: *Do you think this person considers the needs and interests of others, and how his/her own actions affect other people?*; *Do you think this person values morality and wants to see himself or herself as a moral person?*; and *This person participated in a laboratory experiment in which they could cheat by over-reporting their performance in a problem-solving task to earn money. Do you think this person cheated in the experiment? [No, this person was honest (did not cheat at all); Yes, this person cheated a little; or Yes, this person cheated a lot].* These judgments were measured for other, exploratory purposes, so they are not reported in this dissertation. However, the results are available from the author.

[1 (Extremely weak moral character), 2 (Weak moral character), 3 (Neither weak nor strong), 4 (Strong moral character), 5 (Extremely strong moral character)]. In contrast to the prior two studies, judges were not provided with a specific definition or criteria for evaluating overall moral character. However, each judge also made a number of other ratings of the targets, which may have influenced their judgment of moral character.

Prior to judging each target's overall moral character, the judges were given definitions of guilt proneness, conscientiousness, honesty-humility, and agreeableness, and were asked to rate each target on these traits relative to a typical job applicant (ranging from extremely low to extremely high). The following definitions were provided to the judges to facilitate their ratings of the targets' moral character traits and agreeableness.

> ***Guilt Proneness****: Guilt proneness is a personality trait indicative of a disposition toward experiencing negative feelings about personal wrongdoing, even when the wrongdoing is private. In judging guilt proneness, think about whether the person would feel bad about making a mistake or committing a transgression even if no one knew about what they did. A person high on Guilt Proneness feels bad about their behavior when they do something wrong; a person low on Guilt Proneness does not feel guilty about wrongdoing.*
>
> ***Conscientiousness****: Conscientiousness is a personality trait indicative of a disposition toward organization, diligence, perfectionism, and prudence. In judging Conscientiousness, think about whether the person is hard-working, careful, and thorough when working or completing tasks. A person high on Conscientiousness is dependable and self-disciplined; a person low on Conscientiousness is disorganized and careless.*
>
> ***Honesty-Humility****: Honesty-Humility is a personality trait indicative of a disposition toward fairness, sincerity, modesty, and greed-avoidance. In judging Honesty-Humility, think about whether the person is truthful and humble in their interactions with others. A person high on Honesty-Humility is honest and fair; a person low on Honesty-Humility is boastful and greedy.*
>
> ***Agreeableness****: Agreeableness is a personality trait indicative of a person's forgivingness, gentleness, flexibility, and patience. In judging Agreeableness, think about whether the person is tolerant and peaceful in their interactions with others. A person high on Agreeableness is sympathetic and warm; a person low on Agreeableness is critical and quarrelsome.*

Study 3 Results

To determine the number of judges required to reliably evaluate targets' moral character traits, I used generalizability theory (Cronbach, Nageswari, & Gleser, 1963), which enables us to estimate how the reliability of judgments varies with the number of judges. Generalizability theory consists of two studies, which are referred to as the Generalizability (G) study and the Decision (D) study. The main purpose of the G study is to estimate the relative influence of various factors on evaluations. In this study, judges' evaluations are influenced by targets themselves (i.e., main effect of targets; their characteristics revealed in the written interview responses), judges' overall strictness or leniency (i.e., main effect of judges), and interactions between judges and targets (i.e., interaction between targets and judges; judges rank the same set of targets differently). The latter two components are considered to be rating errors, and the first component captures the degree of consensus among judges in evaluating targets.

The degree of consensus among judges in evaluation of targets (i.e., the main effect of targets) was 41% in the mistake questions and 30% in the dilemma questions, suggesting that judges agree more about relative ordering of targets' moral character in the mistake questions. To put these levels of consensus into perspective, we can look at previous studies examining levels of consensus in interpersonal perceptions using generalizability theory. In their meta-analysis, Kenny, Albright, Malloy, and Kashy (1994) also used a generalizability theory approach to examine levels of consensus in interpersonal perception. They found that levels of consensus for the conscientiousness evaluation were 3% when judges evaluated targets based on brief face to face interactions, and 26% among judges who had known targets for an extended time. In comparison to these prior meta-analytic findings, the results of the current study suggest strong consensus among the judges.

In the D study stage, reliability is calculated for hypothetical situations such as varying the number of raters or items. I calculated inter-rater reliability by varying the number of judges; the results are presented in Table 3. Both interview questions showed decent levels of consensus and accordingly had good levels of inter-rater reliability in general. The six judges who participated in Study 3 had greater than .70 reliability. The analyses indicate that increasing the number of judges becomes decreasingly beneficial as the number of judges increases.

To formally test the HIDE model's prediction that a group of judges can predict targets' u nethical behavior by evaluationg their moral cahracter, negative binomial regression analyses were conducted for each interview question. For each question condition, two different sets of analyses were conducted (see Table 4). In the first set of analyses, targets' cheating frequency was regressed on the average-moral-character-rating from the six judges. These analyses revealed that six judges' average-moral-character-ratings significantly and negatively predicted targets' unethical behavior for both interview question conditions, thereby replicating the findings of Studies 1 and 2 in support of the HIDE prediction.

In the second set of analyses, targets' cheating frequency was regressed on moral character judgments from each judge *individually* to investigate the possibility that each individual judge was able to detect target's moral character as well as the group. The second set of analyses provided partial support for individual-level accuracy in judging strangers' moral character based on written interview responses. Each judge's moral character judgments significantly and negatively predicted targets' cheating frequency in the mistake questions condition, although their predictive power was weaker than when average-ratings was used. The predictive validity of individual-level moral character judgments was weaker and less robust for the dilemma questions condition, with only one judge having a statistically significant prediction.

Although somewhat inconsistent between the dilemma and mistake interview question conditions, the individual level accuracy observed in this study is thought-provoking, considering the limited information provided to judges (a brief paragraph consisting of an average of 76.21 words).

Study 3 Discussion

Replicating the findings of studies 1 and 2, the results of Study 3 support the HIDE model prediction that a group of unacquainted judges can predict targets' unethical behavior by evaluating moral character using the written interview method. The most striking and interesting finding of Study 3 is that even a small number of judges (i.e., six judges) can reliably estimate targets' moral character from reading the written interview responses.

I compared an individual judge's predictions with the aggregate-level predictions and showed that the predictive validity (i.e., effect size) was much higher for the latter. This finding demonstrates that the "wisdom of crowds" phenomenon (i.e., that the quality of human judgment increases as the number of judges increases) also applies to moral character judgments: collectives of individuals detected strangers' moral character more accurately than individuals did alone (Larrick, Mannes, & Soll, 2012). Although it is beneficial to have a large number of judges to reduce unreliabilty, this method entails substantial costs of time and money that may be prohibitive for organizations seeking to use this method. Therefore, in Study 3, I used generalizability theory to formally investigate the extent to which reliability increases with increasing the number of judges. I found that increasing the number of judges up to six enhances the reliability greatly; however, increasing the number of judges becomes decreasingly beneficial as the number of judges increases beyond 6. This phenomenon has important practical implications for organizational contexts. In interview settings, for example, one interviewer might not be able to detect moral character accurately by him/herself, but a small set of independent interviewers (e.g., six judges) might be able to do so.

The judges in both Study 1 and Study 3 evaluated the moral character of the same set of targets (i.e., the Study 1 targets). However, Study 1 judges were given a formal definition of moral character while Study 3 judges were not. Despite this difference, there was strong correlation between these two sets of judge ratings ($r = .80$, $p < .001$), suggesting that people's intuition of the definition of moral character is likely similar to the definition we provided to judges in Study 1. Also, regarding the correlation between self- and judge reports, similar patterns were found: they are positively and significantly correlated, suggesting that judges' ratings share common variance with targets' self-perceptions of their moral character traits (See Tables SB3 in Appendix 2).

General Discussion

Moral character judgments are among the most important interpersonal judgments that people make. If we can reliably and accurately detect strangers' moral character from written interview responses, it will have important practical applications in selection and promotion contexts within organizations, as well as important theoretical implications for understanding how we come to know individuals, and specifically whether others are likely to behave ethically.

The HIDE model provides us with guidance on how to reliably detect targets' moral character and forecast unethical behavior. The main goal of this chapter was to test the HDE model's prediction that groups of judges can reliably predict targets' unethical behavior by evaluating targets' moral character using the written interview method. Together, the three empirical studies described in this chapter provide compelling initial evidence in support of the HIDE model and the validity of the written interview method. In Study 1, I found that judges' average ratings of targets' moral character significantly predicted the extent to which targets cheated on a problem-solving task. In Study 2, I replicated this finding with a different criterion, CWB, which includes a wide range of harmful work behaviors, such as falsification of expense reports, stealing, absenteeism, and interpersonal abuse. In line with Study 1, I found that judges' average ratings of targets' moral character significantly predicted targets' self-reported engagement in CWB. Study 3 provided compelling evidence of the validity of the written interview method by showing that even a small number of judges (i.e., six judges) can reliably estimate targets' moral character and predict their unethical behavior.

Beyond providing initial evidence supporting the HIDE model and moral character judgment using the written interview method, the findings in this chapter contribute to the literature of wisdom of crowds in two important ways. First, while the wisdom of crowds has been demonstrated for various prediction and estimation tasks such as prediction markets (e.g., Hastie & Kameda, 2005; Susnstein, 2006; Surowiecki, 2004), to the best of my knowledge, this dissertation is the first study that demonstrates the wisdom of crowds in evaluating people's character and forecasting people's unethical behavior. Second, this study extends our understanding of the wisdom of crowds by directly demonstrating how the reliability of human judgements improves as "crowd" size varies. I found that increasing the number of judges up to six dramatically enhanced reliability, but groups of judges greater than six show decreasing benefits from adding an additional judge. This finding has clear practical implications for designing crowd-source methodologies, for which knowledge of both minimum group size necessary for reliable estimates and strategies for managing evaluation costs by not over-recruiting past the point of a group that can produce satisfactory reliability.

Across three studies, I found that self-reported moral character traits (i.e., honesty-humility, conscientiousness, guilt proneness) were positively correlated with judges' evaluations of targets' moral character. For example, judges' moral character ratings were positively correlated with targets' self-reported honesty-humility (r=.15), conscientiousness (r=.26), and guilt proneness (r=.23) in Study 1. Moreover, self-reported moral character traits were predictive of targets' unethical behavior, consistent with prior studies (e.g., Ashton & Lee, 2007, 2008; Berry, Carpenter, & Barratt, 2012; Berry, Ones, & Sackett, 2007; Cohen, Kim, Jordan, & Panter, 2016; Cohen, Panter, Turan, Morse, & Kim, 2013, 2014). In models that include both self-reports and judge reports, the unique variance of unethical behavior explained by each reduced. These findings mean that the correctly-identified-self zone measured by self-reports are somewhat overlapping with the correctly-identified-target zone captured by judge reports. Therefore, although the written interview method was proposed to enable judges to see implicit aspects of targets' moral character, the judges' evaluations seemed to pick up on not only the implicit aspects, but also on the explicit aspects that targets' can report on as well.

Finally, in terms of the relative validity of self-reports versus judges' evaluations of moral character, the HIDE model suggests that the latter is a relatively more valid methodology when relevant cues can be elicited. This is because moral character is socially desirable and thus targets' self-perceptions and self-reporting can be distorted. Evaluations from unacquainted judges can be informative when behavioral interview questions produce cues that shed light on targets' hidden-selves. When the criterion (i.e., unethical behavior) was measured objectively in a laboratory setting, I found that judges' moral character evaluations was more successful in predicting unethical behavior, with predictive power above and beyond self-reports. Therefore, the findings in this chapter support the notion that judges' evaluations can be more valid than self-reports of moral character. However, it is important to note that there are limitations to this conclusion because the self-reported moral character traits and judges' moral character ratings used different evaluation dimensions. In particular, judge reports were evaluating overall moral character, while self-reports were made for three separate traits (i.e., honesty-humility, conscientiousness, guilt proneness). In Study 5, discussed in the following chapter, I compare the predictive validity of judge reports and self-reports in a more robust way by having both rating sources evaluate targets on honesty-humility, conscientiousness, and guilt proneness.

CHAPTER III

Revealing the Hidden-self through Written Interview Responses

The studies described in chapter 2 utilized anonymous research settings: targets answered questionnaires without a particular motivation to misrepresent themselves. In anonymous research settings, the hiding-self zone—targets' purposeful misrepresentation of themselves in self-reports—should be relatively inconsequential, meaning that self-reports should be fairly accurate in reflecting targets' self-perceptions. In such situations, the degree of invalidity in the self-reports depends on the relative amount of information that is located in the hidden-self zone rather than the correctly-identified target zone (See the HIDE model of self-reports in Figure 2). Accordingly, for judges' evaluations to complement or potentially replace self-reports, the correctly-identified target zone should capture information that is located in the hidden-self zone of the self-reports model. While chapter 2 confirmed one of the HIDE model's prediction—that groups of judges can make valid evaluations of targets using the written interview method—it is not yet clear whether judges' evaluations of targets can indeed capture unique information about targets that cannot be captured in self-reports. The incremental variance accounted for by the judge reports and self-reports in the regression models in chapter 2 suggest that self-reports and judge reports do pick up on somewhat, but not wholly different sources of information. Nonetheless, the data from chapter 2 are not conclusive in this regard given limitations in the research designs.

Therefore, in this chapter, I focus on another prediction of the HIDE model: that judges' evaluations based on targets' written interview responses can capture information that is located in the hidden-self zone—aspects of moral character that targets cannot recognize or incorrectly recognize about themselves. The HIDE model suggests that because moral character is highly socially desirable, the hidden-self zone may be quite large, containing a substantial amount of information about the target that would be missed if only self-reports were used. If this prediction is true, judges' evaluations should show stronger predictive validity for unethical behavior compared to self-reports.

The HIDE model additionally predicts that social relationships can hinder people from forming valid impressions of their acquaintances, meaning that judges who are acquaintances of targets might not be able to correctly identify targets' hidden-self as much as judges who do not know the targets. Therefore, in this chapter, I also investigate whether unacquainted judges' evaluations based on the written interview method can be more accurate than well-acquainted judges' evaluations by testing whether the former predicts targets' unethical behavior to a greater degree than the latter.

In sum, this chapter focuses on examining whether unacquainted judges' evaluations of targets' moral character based on targets' written interview responses capture unique and valid information, as compared to the information captured in targets' self-reports and in well-acquainted judge reports. To make this comparison, the three rating sources need to provide ratings of targets using the same evaluation dimensions. Therefore, whereas judges in chapter 2 evaluated targets on their moral character as a whole, in this chapter, all raters evaluate targets on honesty-humility, conscientiousness, and guilt proneness.

There are several important reasons why I focus on these three specific moral character traits. First, these three constructs represent non-overlapping elements of moral character, and taken together, they can well-subsume the space of individual differences in moral character (Cohen et al., 2014; Kim & Cohen, 2015). According to personality frameworks such as the Big Five or the HEXACO personality framework, people's individual differences can be summarized by a set of five or six distinguishable factors (Ashton & Lee, 2007; Lee & Ashton, 2012). Among these broad personality factors, honesty-humility and conscientiousness are key indicators of people's moral character that do not overlap each other. Honesty-humility is an individuals' tendency to be sincere, fair, modest, and to avoid greed (Ashton & Lee, 2007). Conscientiousness is an individuals' tendency to follow socially prescribed norms for impulse control (John & Srivastava, 1999). In addition to honesty-humility and conscientiousness, guilt proneness is another key indicator of moral character (Cohen et al., 2013, 2014). Guilt proneness is an individual's tendency to feel bad about personal wrongdoing, even when the wrongdoing is private (Cohen, Wolf, Panter, & Insko, 2011). Guilt proneness does not map exclusively onto the Big Five or HEXACO factors, although it does contain a mix of honesty-humility, conscientiousness, and to a lesser extent agreeableness (Cohen, Wolf, Panter, & Insko, 2011). Guilt proneness is a more specific moral character trait than are the more general personality dimensions from five-factor and six-factor (i.e., HEXACO) frameworks.

Second, well-established tools already exist for self-reporting and acquaintance rating for these traits (e.g., HEXACO-60 inventory, GP-5 scale), enabling comparison of self-reports and ratings provided by acquaintances. These scales were developed from rigorous psychometric procedures (Ashton & Lee, 2009; Cohen et al., 2011). Also, the predictive validity of these traits for unethical behavior is well established: both self-reported ratings and ratings provided by well-acquainted others have been shown to predict targets' unethical behaviors across various situations (e.g., Cohen et al., 2013; Hilbig & Zettler, 2015; Stuewig et al., 2015). Finally, the behavioral interview questions we used in our pilot study were developed in an exploratory manner with the expectation that the responses they elicit could potentially extract information relevant to at least one of these moral character traits or possibly a combination of them.

This chapter utilizes three sets of interview prompts: the mistake, dilemma, and employer questions described in chapter 2. The mistake questions ask targets how they felt and behaved after making a mistake. Answers to this prompt might reveal targets' guilt proneness: people who are low on guilt proneness might not elaborate as much on negative feelings compared to others higher in guilt proneness. Answers to the mistake questions might also reveal targets' conscientiousness, because highly conscientious individuals are more likely to expend effort correcting for their mistakes and thus may elaborate more on what they did and what they learned from their past mistakes compared to less conscientious individuals.

The dilemma questions ask targets what they considered and what they did facing a difficult dilemma at work. This prompt was initially designed to extract targets' honesty-humility. In answering the dilemma questions, people who are modest (i.e., not narcissistic) and generous (i.e., high in greed-avoidance) might talk about how their decisions impacted others, or could have impacted others, rather than focusing narrowly on themselves. Because guilt proneness is related to having a strong sense of responsibility to others (Levine, Bitterly, Cohen, & Schweitzer, 2018; Schaumberg & Flynn, 2012), it is possible that the dilemma questions could reveal information related to guilt proneness, and some of the results from chapter 2 allude to this possibility.

The employer question asks targets to describe how their current or last employer would describe them. This question was initially included in this program of research to be a more neutral, standard interview question that could be used as a comparison, but through Study 2 in chapter 2 it was revealed that it could be used to extract relevant moral character information. This might be the case if targets who are more honest and humble give a more balanced account of how their employers would view them, compared to less honest or modest individuals who might describe themselves in only the most positive terms.

The findings I report in Chapter 2 indicate that judges were able to reliably evaluate targets' moral character as a whole. However, it is an open question as to how judges make global assessments of character, and whether they can evaluate targets' honesty-humility, conscientiousness, and guilt proneness specifically. Therefore, as a first step toward answering these questions, and to develop a better understanding whether unacquainted judges' evaluations of honesty-humility, conscientiousness, and guilt proneness can capture more information than self-reports or ratings provided by acquaintances, we sought to identify which interview questions (if any) should be used to evaluate which of these three moral character traits.

This chapter consists of three empirical studies. In Study 4, I explore the extent to which each of the three interview prompts (i.e., the mistake, dilemma, and employer questions) reveals information about targets' honesty-humility, conscientiousness, and guilt proneness. The results of Study 4 attempt to match interview questions to evaluation traits, indicating which questions should be used as the basis for judges to specifically evaluate targets' honesty-humility, conscientiousness, and guilt proneness. Studies 5 and 6 use the matched questions from Study 4 to compare the predictive validity of unacquainted judges versus self-reports of moral character (Study 5) and unacquainted judges versus well-acquainted judges' evaluations (Study 6).

Study 4

The main purpose of Study 4 was to understand the relative ability of each interview question to elicit cues that help judges reliably evaluate targets on each of three moral character traits: honesty-humility, conscientiousness, and guilt proneness. I take an exploratory approach in Study 4 because not much is known about what is revealed by the specific interview questions I used in this research. The results of Study 4 inform which interview questions should be used for judges to evaluate honesty-humility, conscientiousness, and guilt proneness in Studies 5 and 6.

In addition to providing a basis for testing the HIDE model predictions in Studies 5 and 6, Study 4 has its own unique contributions to our understanding of the HIDE model and the written interview method. In the HIDE model, selecting an appropriate scope for the evaluation dimension is an important means of increasing the reliability and validity of judge reports. If each interview question extracts information relevant to specific aspects of moral character, the HIDE model posits that judges asked to focus on smaller sets of cues (i.e., cues of targets' conscientiousness specifically rather than moral character as a whole) will produce more valid evaluations than judges' evaluations based on a larger scope (e.g., cues of targets' conscientiousness, honesty-humility, guilt proneness, and etc.). This is because an unnecessarily wide evaluation dimension can increase the size of the hidden-target zone in the judge reports model by increasing its two component zones: judge-ignorance—some aspects of targets cannot be known to judges since there are no cues for detecting those aspects—and judge-error –even when available cues cover the breadth of the evaluation scope, simultaneously detecting and utilizing too many cues can increase inconsistency in judges' evaluation of targets. To reduce the judge-ignorance and judge-error zones, it is important to identify which traits (if any) can reliably be evaluated from each interview question.

To explore which traits judges can reliably evaluate from responses to each interview question, I conduct two types of analyses. First, I use generalizability theory (Cronbach, Nageswari, & Gleser, 1963) to measure the levels of consensus in judges' evaluations of targets' honesty-humility, conscientiousness, and guilt proneness from targets' written responses to each interview question. If judges' evaluations of targets' traits do not show a high level of consensus, this may indicate a lack of sufficient cues available about those traits in the written interview responses. The generalizability theory results, therefore, could be particularly informative for determining which traits cannot be reliably evaluated using each interview question. Accordingly, using the generalizability analyses, I aim to rule out moral character traits that cannot be evaluated by judges for each interview question.

Second, in order to check whether judges' evaluations of targets' traits are valid, I compare how much shared variance exists between self-reports and unacquainted judges' evaluations of honesty-humility, conscientiousness, and guilt proneness. Consensus among judges (i.e., interrater reliability) is a necessary condition of validity, but not a sufficient condition of validity. Therefore, I seek convergent validity evidence by examining the correlation of self-reports and judges' evaluations on the same trait. Some information captured by self-reports is valid (i.e., the correctly-identified self zone of the HIDE model), and some (but not all) of this same valid information is likely also conveyed in targets' written interview responses. Therefore, if an interview question generates ample cues about targets' traits, judges' evaluations are expected to capture a degree of overlapping information with targets' self-perceptions. That is, although the written interview method enables judges to capture information that targets' self-reports cannot correctly measure, written interview responses also contain information located in targets' correctly-identified self zone, and judges should be able to capture that as well. A small to moderate sized positive correlation between self-reports and judges' evaluations of a particular trait would support the assumption that the interview questions generate valid cues about targets' traits.

Study 4 Method

Data Collection from Targets

The targets who provided the responses to the interview questions in Study 4 were 296 U.S. adults recruited from Amazon Mechanical Turk. Each of the participants (i.e., targets) answered three randomly chosen interview questions out of a set of five possible questions. To provide consistency with the other studies in this dissertation, in the current study, I focus on targets' responses to three questions (the mistake, dilemma, employer questions).[5] Some targets provided incomplete or overly short answers to the subset of interview questions they were given; interview responses with fewer than 20 words were excluded from the current study because such short responses might reflect targets' lack of motivation in answering that question.

Following the open-ended interview questions, the targets answered several personality questionnaires, including the Ten Item Personality Measure (TIP; Gosling, Rentfrow, & Swann, 2003) and GP-5 (Cohen et al., 2014). Similar to how conscientiousness and other personality traits are measured in the TIPI, honesty-humility was measured with two pairs of traits: "honest, fair," and "boastful, greedy." Participants indicated the extent to which they believed each pair of characteristics applied to themselves by choosing a category ranging from 1 (Disagree Strongly) to 7 (Agree Strongly).

Data Collection from Judges

Five undergraduate research assistants read the entire set of interview responses in a randomized order and rated each target's honesty-humility, conscientiousness, agreeableness, and guilt-proneness. Agreeableness is not a key indicator of moral character in the HEXACO

[5] The results for the other two questions are available from the author upon request. The other two questions were: [Mistreatment] Please describe an experience you have had in which you or a coworker was mistreated at your job. How did you feel when this occurred? What did you do? What, if anything, did you learn from this experience?, and [Negative emotion] Please tell us about a time at work when you felt a strong negative emotion, such as anger, guilt, or sadness. What made you feel this way? What did you do? What, if anything, did you learn from this experience?

model (Ashton & Lee, 2008; Cohen et al., 2014; Hilbig & Zettler, 2009) but was included as a comparison evaluation dimension.

The judges were asked: Compared to a typical job applicant, do you consider the author of this essay to be low or high on [*Honesty-Humility, Conscientiousness, Agreeableness, and guilt-proneness*]? They could indicate: 1 (Extremely Low), 2 (Low), 3 (Neither Low nor High), 4 (High), 5 (Extremely High). Judges read the following instructions for each trait:

> ***Guilt Proneness****: Guilt proneness is a personality trait indicative of a disposition toward experiencing negative feelings about personal wrongdoing, even when the wrongdoing is private. In judging guilt proneness, think about whether the person would feel bad about making a mistake or committing a transgression even if no one knew about what they did. A person high on Guilt Proneness feels bad about their behavior when they do something wrong; a person low on Guilt Proneness does not feel guilty about wrongdoing.*
>
> ***Conscientiousness****: Conscientiousness is a personality trait indicative of a disposition toward organization, diligence, perfectionism, and prudence. In judging Conscientiousness, think about whether the person is hard-working, careful, and thorough when working or completing tasks. A person high on Conscientiousness is dependable and self-disciplined; a person low on Conscientiousness is disorganized and careless.*
>
> ***Honesty-Humility****: Honesty-Humility is a personality trait indicative of a disposition toward fairness, sincerity, modesty, and greed-avoidance. In judging Honesty-Humility, think about whether the person is truthful and humble in their interactions with others. A person high on Honesty-Humility is honest and fair; a person low on Honesty-Humility is boastful and greedy.*
>
> ***Agreeableness****: Agreeableness is a personality trait indicative of a person's forgivingness, gentleness, flexibility, and patience. In judging Agreeableness, think about whether the person is tolerant and peaceful in their interactions with others. A person high on Agreeableness is sympathetic and warm; a person low on Agreeableness is critical and quarrelsome.*

Initially, based on the findings of Study 3 regarding judge group size, six judges were recruited; however, one of them dropped out of the study. Therefore, one limitation of the judge data collection in Study 4 is that type 1 error can be increased because the average-ratings of judges' evaluations are expected to be lower than having more judges. However, this does not increase the type 1 error.

Study 4 Results

The descriptive statistics and correlations among targets' self-reported and judge-reported honesty-humility, conscientiousness, guilt proneness, and agreeableness measures are presented in Tables S4A and S4B in Appendix 2.

First, I ran generalizability theory analyses; results are presented in Table 5. The generalizability theory analysis consists of two studies: the G study and D study. The G study quantifies the levels of consensus among judges regarding targets' relative ordering on each evaluation trait. The G study results for Study 4 indicated that judges' evaluations of honesty-humility had the lowest levels of consensus across all interview prompts. In the mistake questions prompt, the levels of consensus of honesty-humility evaluations was 16%, while that of other three traits were over 20%. Similarly, in the dilemma questions prompt, honesty-humility evaluations had the lowest levels of consensus. Finally, in the employer question, the honesty-humility evaluation had particularly low levels of consensus (i.e., 10%), which was the lowest levels of consensus across all interview question prompts and evaluation dimensions.

The D study estimates the reliability of judges' scores as the number of judges varies. As shown in Table 5, similar to the G study results, the honesty-humility evaluation showed the lowest reliability across all conditions. In particular, for the employer question, even when the number of judges is significantly large, the reliability of judge reports was still quite low for the honesty-humility dimension. These results suggest that honesty-humility judgments were not as reliable (and thus not as valid) as conscientiousness and guilt proneness judgments when using the three interview questions examined in this research

Second, I examined which interview questions revealed shared variance between judge reports and self-reports of the same traits (i.e., convergent validity). These results are presented in Table 6. Considering that the generalizability theory analysis resulted in low levels of consensus in honesty-humility evaluations, I was particularly interested to see which interview question might be useful for judges to evaluate targets' conscientiousness and guilt proneness. I found that the mistake questions resulted in positive correlations for conscientiousness, and the dilemma and employer questions resulted in positive correlations for guilt proneness.

In this study, agreeableness was included as a comparison trait. Given that the interview questions used in this dissertation were designed to capture targets' tendency to think, feel, and behave ethically, whereas agreeableness in the HEXACO framework is not strongly related to these characteristics (Hilbig & Zettler, 2009), it is not expected that judges' agreeableness evaluations should be valid in this setting. Although judges' ratings of agreeableness revealed consensus levels somewhat comparable to those for conscientiousness and guilt proneness, self-reported and judge-reported agreeableness did not reveal any significant correlations. These findings provide some evidence of divergent valdity of moral chracter judgments using the written interview method, such that constructs that should not have a relationship, indeed, do not show any relationship.

Study 4 Discussion

Based on the generalizability theory results and correlation patterns, it seems that among the three dimensions explored, written interview responses to the mistake questions are most informative about targets' conscientiousness. Initially, we reasoned that the mistake questions would extract targets' guilt proneness, given that people who are high in guilt proneness might report that they felt bad after making a mistake. The results of Study 4, however, suggested that targets' guilt proneness was not revealed effectively by responses to the mistake questions, but was instead better captured by judge reports from the dilemma or employer questions. The reason for this might be that guilt proneness, as a personality trait, is more closely related to people's internal sense of personal responsibility, as opposed to their anticipated guilty feelings about particular actions (Levine, Bitterly, Cohen, & Schweitzer, 2018). As such, it is possible that the dilemma and employer questions are better able to elicit information about this, than are the mistake questions.

The generalizability theory analysis indicated that judges' ratings of honesty-humility had the lowest levels of consensus between judges among all the interview question prompts. Furthermore, honesty-humility evaluations did not reveal statistically significant and positive correlations with self-reported honesty-humility either for the dilemma or the employer questions, which we initially reasoned might to be able to extract honesty-humility information.

There are several possible reasons why the honesty-humility judgments showed low levels of consensus. It might be the case that there are fewer available cues for honesty-humility in written responses compared to cues for conscientiousness and guilt proneness, or that judges were unable to detect or utilize available cues of honesty-humility in a consistent way across targets. Related to the latter, it is possible that the scope of the honesty-humility evaluation trait may still be too broad for judges to evaluate targets. In the HEXACO framework, self-reported ratings for four more specific elements (fairness, sincerity, greed-avoidance, and modesty) form the global honesty-humility factor. In other words, when measured by self-reports, fairness, sincerity, greed-avoidance, and modesty share a strong variance, and are interpreted together as the honesty-humility factor. However, it is possible that judges' evaluations of these component elements are not similar enough to form one factor; therefore, combining these four elements into one overarching factor may create an evaluation dimension that is too broad and varied to be useful in producing valid evaluations. These possibilities are further investigated in later chapters of this dissertation. Because of the inconsistency in the honesty-humility evaluations using the current interview questions, in studies 5 and 6 in this chapter, I focus only on judges' evaluations of conscientiousness and guilt proneness.

Study 5

The purpose of Study 5 was to test the HIDE model's prediction that judges' evaluations using the written interview method can capture information about targets' moral character located in the hidden-self zone of the HIDE model, which self-reports cannot adequately measure. Written interview responses reveal implicit aspects of targets' moral character. While targets' explicit self-views of their morality can be biased, targets' responses to written interview questions might reveal information located in the hidden-self zone. For example, a target might incorrectly perceive himself or herself as a highly conscientious person, but their written responses may covertly reveal their hidden-self, perhaps showing that they elaborate much less about how they tried to fix a mistake compared to others, indicating lower conscientiousness. Therefore, the HIDE model predicts that judges' evaluations based on the written interview method can reveal aspects of targets' hidden-self, and thus should be able to explain variance of unethical behavior that self-reports cannot.

In Study 5, two interview questions were used: the mistake and dilemma questions. The results of Study 4 indicated that while responses to the mistake questions aided judges in making valid evaluations of targets' conscientiousness, the dilemma questions aided judges in making valid evaluations of targets' guilt proneness. Therefore, in Study 5, I investigate whether judges' conscientiousness evaluations from the mistake questions capture unique information that self-reported conscientiousness ratings do not. Additionally, I investigated whether judges' guilt proneness evaluations using the dilemma questions capture unique information that self-reported guilt proneness ratings do not. Based on the findings from Study 4, judges' evaluations of honesty-humility are expected to be less valid than conscientiousness and guilt proneness evaluations.

Study 5 Method

I recruited 500 participants from an online participant pool (Amazon's Mechanical Turk) to read and evaluate the interview responses from targets in Study 1. This study used a 2 by 3 between-conditions design. Judges were randomly assigned to one of two interview question conditions (i.e., mistake, dilemma) and one of three evaluation dimension conditions (i.e., honesty-humility, conscientiousness, guilt proneness).

The rating instructions and definitions for honesty-humility, conscientiousness, and guilt proneness were the same ones used in Study 4. Each judge rated interview responses from 20 randomly selected targets. Each interview response was rated by an average of 16 judges.

Targets' frequency of cheating, as measured in Study 1, was used as a criterion of unethical behavior to determine the relative predictive power of judges' ratings versus self-reports of honesty-humility, conscientiousness, and guilt proneness. Also, this study compared the predictive validity of judges' ratings of these dimensions to judges' ratings of moral character as a whole, measured as the average of judges' moral character ratings for each target from Studies 1 and 3.

Study 5 Results

The descriptive statistics for judges' average ratings of honesty-humility, conscientiousness, and guilt proneness are presented in Table S5A in Appendix 2.

To formally test the HIDE model's prediction that judges' evaluations of targets using the written interview method can capture information in the hidden-self component that self-reports cannot correctly measure, I conducted negative binomial regression analyses for each interview question. If the HIDE model's prediction is true, judges' evaluations should be able to predict targets' unethical behavior above and beyond targets' self-reports.

Three different analyses were conducted for each interview question condition; results are presented in Table 7. In the first model, targets' frequency of cheating was regressed on judge-reported and self-reported honesty-humility. In the second model, targets' frequency of cheating was regressed on judge-reported and self-reported conscientiousness. In the third model, targets' frequency of cheating was regressed on judge-reported and self-reported guilt proneness. The results indicated that across all evaluation dimensions, judges' evaluations predicted targets' unethical behavior above and beyond self-reports. Together, these results support the HIDE model's prediction that judges' evaluations of targets using the written interview method can capture information that self-reports cannot. In particular, when judges' evaluations and self-reports were entered together in the model, only the judges' evaluations were predictive of targets' unethical behavior. Similarly, when judges' evaluations of guilt proneness and self-reports were entered together in the prediction model of unethical behavior, only the judges' evaluations were predictive of targets' unethical behavior, meaning that judges' evaluations were more valid.

Study 5 Discussion

The HIDE model predicts that judges' moral character evaluations from targets' written interview responses should be able to capture information that targets themselves cannot recognize or incorrectly recognize about themselves. However, the HIDE model suggests two caveats: this can only happen when judges have access to a sufficient amount of valid cues, and multiple judges are necessary to reliably detect and utilize those cues. Results from Study 4 indicated that the mistake questions generate a sufficient amount of valid cues about targets' conscientiousness and the dilemma questions generate a sufficient amount of valid cues about targets' guilt proneness. Therefore, based on targets' responses to the mistake questions, judges' evaluations can potentially capture information about targets' conscientiousness that self-reports cannot. Likewise, judges' evaluations based on the dilemma questions can potentially capture information about target's guilt proneness that self-reports cannot. Results from Study 5 confirmed this possibility: judges' conscientiousness evaluations from responses to the mistake questions and judges' guilt proneness evaluations from the dilemma questions captured unique variance that self-reports failed to capture and predicted targets' unethical behavior to a greater degree than self-reports.

One interesting finding from Studies 4 and 5 is that although the mistake questions were initially expected to capture guilt proneness, judges' ratings of targets' guilt proneness from the mistake questions were the least informative among all rating dimensions. The lack of validity of judges' guilt proneness evaluations based on responses to the mistake questions can be explained in several ways. First, it is possible that targets did not talk about guilt or related cognitive or affective responses, and therefore, no cues were available in the written responses to judge guilt proneness. Second, it is possible that targets talked about experiencing guilt after making mistakes but that judges did not associate targets' state-level of guilt after making a mistake with their guilt proneness. These issues are further examined and discussed in chapter 5.

Study 5 found that targets' guilt proneness was better revealed through the dilemma questions rather than the mistake questions, which is consistent with the findings in Study 4.

Initially, the dilemma questions were developed to give targets an opportunity to talk about how their decisions might affect others. Therefore, I expected that the dilemma questions would reveal aspects of targets' honesty–humility. For example, a person who is not narcissistic (i.e., high in modesty) and generous toward others (i.e., high in greed-avoidance) might be expected to talk about how his or her decisions influence others rather than focusing on himself or herself. However, results from both Studies 4 and 5 suggest that the dilemma questions are not particularly useful for revealing targets' honesty-humility, but are useful for revealing targets' guilt proneness. It is possible that either targets did not talk about their consideration of others in response to the dilemma questions (e.g., if they chose to report on dilemmas that did not involve other people) or that regardless of whether the targets talked about consideration of others, this query was not indicative of honesty-humility. These possibilities are explored in Chapter 5.

It is noteworthy that among the three self-reported moral character traits examined in this study, conscientiousness had the strongest predictive power for cheating. It is possible that the criterion variable in this laboratory task had a particularly strong relationship with conscientiousness. I investigate the relative predictive validity of judges' evaluations of conscientiousness versus guilt proneness using a broader criterion of unethical behavior in Studies 6, 7, and 8.

Study 6

We often assume that well-acquainted others will be better judges of an individual's personality than strangers, and studies generally support this claim (Funder 1995, Kenny, Albright, Malloy & Kashy, 1994). However, the HIDE model suggests that, in some circumstances, evaluations made by strangers can be more informative than those provided by well-acquainted others, even though acquaintances have the opportunity to observe targets in various situations over time. In particular, the HIDE model predicts that unacquainted judges are likely to be more accurate than well-acquainted others when the relationship hinders the latter's ability to form correct impressions or to report objectively about targets. This is most likely to happen when the evaluation dimension is socially desirable. In chapter 1, I used balance theory (Heider, 1958; Insko, 1981) to explain why people tend to perceive or believe good things about their friends and bad things about their enemies, and therefore, judge-target relationships can hinder the validity of well-acquainted judges' evaluations.

While unacquainted judges' evaluations of targets are not susceptible to errors arising from relationship, in order to make valid evaluations unacquainted judges still require a sufficient amount of relevant and valid cues about targets' characteristics and the ability to reliably detect and utilize those cues in forming impressions. The results of Studies 4 and 5 revealed that the mistake questions and the dilemma questions produced cues that enabled judges to make reliable and valid evaluations of targets' conscientiousness and guilt proneness, respectively. Therefore, to test the relative validity of acquainted versus unacquainted judges in Study 6, I compared unacquainted versus well-acquainted judges' conscientiousness evaluations based on the mistake questions and guilt proneness evaluations based on the dilemma questions. I examined the relative extent to which ratings from those two types of judges predict targets' unethical behavior. If unacquainted judges' evaluations of targets are more accurate than well-acquainted others' evaluations, it is expected that the former would predict targets' unethical behavior to a greater degree than the latter.

In this study, I also included the employer question. Study 4 revealed that judges' honesty-humility judgments using this question were invalid. However, lacking testing with objectively measured unethical behavior as a criterion in Study 5, the employer question was included in Study 6 only for exploratory purposes to further understand whether the employer question can be used to evaluate targets' conscientiousness and guilt proneness. Therefore, for the employer question, I do not test hypotheses; rather, I explore the predictive validity of judges' evaluations of conscientiousness and guilt proneness.

In sum, in Study 6, I investigate the HIDE model's prediction that moral character evaluations from unacquainted judges' will outperform those from targets' acquaintances in predicting targets' unethical behaviors. The criterion for unethical behavior in Study 6 is peer-reported counterproductive work behavior (CWB), which involves a wide range of unethical work behaviors (e.g., falsification of expense reports, stealing, interpersonal abuse).

Study 6 Method

Data Collection from Targets

The target participants in Study 6 were 174 full-time U.S. employees recruited from an online participant pool maintained by the university research center. Respondents answered one of three interview questions (mistake, dilemma, employer) and completed the HEXACO-60 personality inventory (Ashton & Lee, 2009) and GP-5 (Cohen, Kim, & Panter, 2014) via a computerized survey.[6]

Data Collection from Judges

Targets' coworkers served as acquaintance judges: target participants who completed the study were invited via email to participate in a study in which they asked their coworkers to take

[6] The other two questions were: [Mistreatment] Please describe an experience you have had in which you or a coworker was mistreated at your job. How did you feel when this occurred? What did you do? What, if anything, did you learn from this experience?, and [Negative emotion] Please tell us about a time at work when you felt a strong negative emotion, such as anger, guilt, or sadness. What made you feel this way? What did you do? What, if anything, did you learn from this experience?

surveys about them. Coworkers provided peer-reports on targets' CWB and rated targets on the HEXACO-60 personality inventory. In total, 87 coworkers participated in the study.

The role of unacquainted judges was filled by six undergraduate research assistants recruited to read and evaluate targets' written responses. Judges were presented with targets' responses to the mistake and dilemma questionss in a randomized order, followed by responses to the employer question last. Only four judges completed ratings for the employer question. In this study, therefore, I focused on testing theoretical predictions for judges' ratings of the responses to the mistake and the dilemma questions. The rating instructions and definitions of honesty-humility, conscientiousness, and guilt proneness were the same ones used in Studies 4 and 5.

Study 6 Results

The descriptive statistics for the self-reports, peer-reports, and unacquainted judge reports of honesty-humility, conscientiousness, and guilt proneness are presented in Tables S6A, S6B, and S6C in Appendix 2. The self–judge correlations and judge–peer correlations for these traits were consistent with the findings in Studies 4 and 5 (See Table S6D Appendix 2).

I formally tested the relative predictive validity of peer-reports versus judges' evaluations of conscientiousness in the mistake questions prompt and guilt proneness in the dilemma questions prompt by conducting negative binomial regression analyses (see Table 9). Consistent with predictions from the HIDE model, judges' evaluations of conscientiousness had stronger predictive power than peer-reports of targets' CWB. Further, the predictive power of judge-reported guilt proneness was stronger than the ratings provided by targets' peers for the dilemma questions prompt. However, none of judges' evaluations or peer-reports were statistically significant at α=.05.

I conducted exploratory negative binomial analyses for each evaluation dimension and each interview prompt. The results are presented in Table 10. In the mistake questions prompt, only the conscientiousness evaluation showed a negative relationship with targets' CWB. In the dilemma questions prompt, both the conscientiousness and guilt proneness dimensions showed negative relationships, but the latter was stronger. Finally, in the Employer question prompt, only the conscientiousness evaluation showed a negative relationship albeit they were not significant at α=.05

Study 6 Discussion

Consistent with the prediction from the HIDE model, Study 5 found that ratings based on written interview responses provided by judges unacquainted with targets were more predictive of targets' CWB than reports from targets' coworkers. Although the coefficients from the unacquainted and coworker judges' evaluations were not statistically significant for predicting CWB, the former were descriptively stronger than the latter. This study's limitations included its use of a relatively small sample size and a criterion that was not objectively measured, but rather provided by coworkers. However, coworker-reported CWB would be more likely to inflate the predictive validity of coworker-reports of moral character because of the shared method variance. Therefore, this does not increase type 1 error in testing whether judges' evaluations of targets are more predictive of CWB, but increases type 2 error (i.e., less power). In future studies, the results should be replicated with a greater number of dyads using more objective criteria of unethical behavior.

The HIDE model's prediction that unacquainted judges can outperform well-acquainted others in forming and reporting correct impressions of targets has important implications theoretically and practically. We often assume that well-acquainted others will be better judges of personality than strangers, and studies generally support this claim (Funder 1995, Kenny, Albright, Malloy & Kashy, 1994). Challenging this notion, the HIDE model incorporates the idea that the relationships between well-acquainted judges and targets can hinder judges' ability to form correct impressions or to report correctly about targets. The HIDE model is unique in making the prediction that unacquainted judges can, in some circumstances, make more accurate evaluations of targets using the written interview method than well-acquainted others.

This prediction holds implications for various organizational practices, including peer-referral programs or evaluation of job candidates based on references from well-acquainted others, such as former employers or coworkers. In Study 6, well-acquainted others evaluated targets via an anonymous survey, therefore, their ratings were not likely to be influenced by the hiding-target component, or judges' intentional misrepresentation of targets. However, in real-life situations where well-acquainted others' identities are known and their references could substantially help or harm targets, the HIDE model predicts that the validity of well-acquainted others' evaluations will be diminished by the role of the hiding-target component.

While Study 4 revealed that the employer question might be useful for evaluating targets' guilt proneness, because the employer question was not included in Study 5, much less is known about which contents of targets' moral character it reveals. In Study 6, only four out of six judges provided ratings for targets' written interview responses to the employer question, and the order of judgments on the employer question were not randomized. Therefore, I did not conduct formal hypothesis testing for judges' evaluations of responses to the employer question. Nonetheless, I explored the predictive validity of judges' reports for the employer question and found that only the conscientiousness evaluation was predictive. However, it is noteworthy to mention that in Study 4, judges' evaluations of conscientiousness were not positively correlated with self-reported conscientiousness. Therefore, it is currently unclear whether the employer question is useful for revealing targets' conscientiousness or guilt proneness. Exploration of whether the employer question can be used to evaluate targets' conscientiousness or guilt proneness is continued in study 7 in the next chapter.

General Discussion

The HIDE model describes when and how judges who are unacquainted with targets can provide more valid evaluations of targets' moral character than targets themselves or well-acquainted judges. Two important findings in chapter 2 confirmed the HIDE model's suggestions that behavioral interview questions generate cues about targets' moral character, and that relevant cues can be reliably detected and translated into predictions of unethical behavior by small groups of judges.

The findings in this chapter build upon chapter 2 in several important ways. First, this chapter compared information about targets' moral character extracted by different behavioral interview questions to further understand which questions produced cues that enabled judges to form valid evaluations of specific moral character components. Across three studies, it was found that information about targets' conscientiousness is revealed through their written responses to the mistake questions and that guilt proneness is revealed through their answers to the dilemma questions. Therefore, while judges in chapter 2 made evaluations of targets' moral character as a whole, the results of this chapter indicated that each interview question extracted different, specific elements of targets' moral character. To evaluate specific aspects of moral character, then, different interview questions should be used so that the available information is assessed to ensure the validity of evaluations (Funder, 1995; 2012).

Second, findings from Study 5 revealed that judges' evaluations of targets' conscientiousness using the mistake questions and evaluations of targets' guilt proneness using the dilemma questions captured information that self-reports failed to correctly capture. While numerous studies have found that moral character information captured by self-reports predict targets' unethical behaviors in various situations (e.g., Ashton & Lee, 2007, 2008; Berry, Carpenter, & Barratt, 2012; Berry, Ones, & Sackett, 2007; Cohen et al., 2013, 2014), the results in this chapter indicated that unacquainted judges' evaluations can be more predictive of targets' unethical behaviors than self-reports. These findings suggest that there is a substantial amount of hidden-self information—characteristics that individuals do not correctly know about themselves—in moral character. Currently, the field of personality psychology has largely been built on targets' self-reports (Connelly & Hűlsheger, 2012; Oh, Wang, & Mount, 2011) and not much is known about aspects that self-reports cannot correctly capture. The HIDE model and the findings of this chapter have promising implications for further research focusing on unacquainted judges' evaluations of targets using the written interview method to advance our understanding of people's self-perception blind spots.

Third, results of Study 6 provided initial evidence that unacquainted judges can be more accurate than well-acquainted judges in evaluating targets. I found that judges' evaluations of conscientiousness and guilt proneness based on targets' written responses to the mistake and the dilemma questions respectively were more predictive of targets' unethical behavior than ratings provided by targets' coworkers. Although prior studies revealed that well-acquainted others' evaluations of targets' moral character are valid (Cohen et al., 2013; Helzer et al., 2014), findings of this chapter indicate that there are areas where unacquainted judges outperform well-acquainted others in making predictive assessments. The HIDE model explains that well-acquainted others might not be able to form correct impressions of targets when the evaluation dimension is highly socially desirable. Related to this claim, Lee and colleagues' (2009) compared the similarities among self-reports, peer-reports, and peers' own self-reports of the HEXACO personality dimensions. They found that people tend to overestimate similarity in personality characteristics between themselves and their close others when the evaluation dimension is central to defining their identity. In particular, among six broad factors from the HEXACO model, the assumed similarity effect (i.e., overestimation of similarity of themselves and close others) were strongest for the honesty-humility dimension. Building upon this finding, the current research compared the validity of unacquainted and well-acquainted judges using the objectively measured behavioral criterion, and showed that the former's judgments was more valid than the latter's.

It is noteworthy that the employer question resulted in some mixed findings so far. The results of Study 4 in chapter 3 suggested that the employer question might be useful in informing judges' evaluations of targets' guilt proneness but not conscientiousness. However, the results in Study 5 in this chapter suggested that the employer question might be useful for conscientiousness. It is not yet clear why we observe these mixed findings for the employer question; there are several possibilities. One possibility is that these differences reflect the difference in self-report personality scales used in Study 4 (i.e., TIPI; Gosling, Rentfrow, & Swann, 2003) and Study 6 (i.e., HEXCAO-60; Ashton & Lee, 2009). Another possibility is that the employer question can extract both the conscientiousness and guilt proneness information of targets. Therefore, in the following chapter, I seek to unravel the mixed findings for the employer question by comparing evaluations of conscientiousness and guilt proneness using objectively measured unethical behavior.

Finally, Studies 4, 5, and 6 consistently demonstrated that judges' honesty-humility evaluations are not as reliable or valid as judges' evaluations of conscientiousness and guilt proneness across all interview question conditions. In particular, judges' honesty-humility evaluations had the lowest consensus in the employer question. This is particularly intriguing as initially the employer question was expected to reveal this specific dimension of character. There are several possible reasons for this. For example, it might be the case that while several elements of honesty-humility (i.e., fairness, sincerity, greed-avoidance, and modesty) form one general factor (i.e. honesty-humility factor) using self-reports, they do not form one factor when evaluated from judges' perspectives. This possibility will be explored in chapter 4 to further understand the written interview method and differences of self- and other-perceptions.

CHAPTER IV

Revealing the Hiding-Self through Written Interview Responses

The main purpose of this chapter is to investigate the HIDE model's prediction that judges' evaluations using the written interview method can uncover information located in targets' hiding-self zone of the model—information that targets purposefully represent about themselves in a biased manner. In addition to demonstrating that judges' evaluations can capture information located in the hiding-self zone, in this chapter I investigate two additional research questions based on the findings in chapter 3. First, I formally test whether judges' specific evaluations of targets' conscientiousness (using the mistake questions) and guilt proneness (using the dilemma questions) are more valid than judges' evaluations of targets' moral character as a whole. Second, I investigate possible reasons why judges' honesty-humility evaluations show low levels of consensus and lack validity.

Uncovering Information Located at the Hiding-Self Zone

In chapter 1, I proposed that targets' self-ratings of moral character are likely influenced by their hidden-self and hiding-self. People not only want others to see themselves as moral (and thus try to hide information that might lead others to think otherwise: i.e., the hiding-self), but also want to see themselves as moral (and thus there may be aspects of their personalities that are hidden even from themselves: i.e., the hidden-self). I further propose that judges' ratings based on targets' written responses to specific behavioral interview questions can uncover information located in the hidden- and hiding-self zones of the HIDE model because written responses to those questions can reveal implicit aspects of targets' moral character, which they cannot control. In chapters 2 and 3, I demonstrate that judges' evaluations uncover information located in the hidden-self zone by showing that judges' evaluations explain unique variance in unethical behavior that targets' self-perceptions do not explain. One important remaining question is whether the written interview method can also uncover information located in targets' hiding-self zone— people's purposeful misrepresentation of themselves. The main purpose of this chapter, therefore, is to investigate the HIDE model's prediction that judges can make valid evaluations even when targets represent themselves in a positively biased manner.

The hidden-self zone has two components: impression management and self-screening. These two components respectively describe information incorrectly captured by self-reports, and information that self-reports fail to capture at all. An example of self-screening is targets skipping certain questions in a survey because they do not want to provide particular types of information. An example of impression-management is targets answering certain questions about their characteristics extremely positively (or negatively) than they actually believe them to be.

While self-reports might be less valid when targets are more motivated to represent themselves in a positively biased manner, judges' evaluations using the written interview method in this scenario can actually increase in validity compared to instances where targets are less motivated to represent themselves positively. While counterintuitive, there is sound reasoning behind this effect. The written interview method is expected to expose targets' implicit aspects of moral character. Then, the more targets attempt to misrepresent themselves, the more judges' evaluations should capture implicit aspects of targets' moral character—especially information located in impression management and self-screening zones of the HIDE model. In the written interview method, impression management takes the form of targets answering questions in an overly positive (or negative) way. It is possible that targets with low moral character might decide to represent themselves in an overly positive manner, but judges might see through targets' impression management based on their written interview responses. For example, when targets are asked to describe how their employer would describe them, low moral character targets might choose to say their employer would describe them as extremely skilled in every possible dimension, but judges could rate these targets more negatively due to the demonstrated lack of humility. In the written interview method, self-screening takes the form of targets not talking about something related to questions that judges expect them to elaborate on. For example, in the mistake questions, targets are asked to describe how they felt after making a mistake. If they choose not to talk much about the types of negative feelings that followed the mistake (to hide the fact that they did not feel bad), judges may infer that they did not feel bad, and may give them lower ratings compared to targets who elaborate on their negative feelings. In sum, I predict based on the HIDE model that, ironically, the more targets' hiding-self plays a role in answering questions, the more targets' moral character can be revealed through their written responses to the interview questions used in this research. Using this method, judges' evaluations of targets can still be predictive of targets' unethical behavior for targets who misrepresent themselves, even if self-reports in such settings are not.

The hiding-self component should be especially influential in situations in which targets know that they are being evaluated by other people and that the evaluations will affect their future. In this chapter, I manipulate the incentives targets have to make a favorable impression and measure the extent to which targets try to answer in a positively biased manner. I examine whether judges' evaluations of targets using the written interview method are able to predict targets' unethical behavior to a greater degree for people who are incentivized to make favorable impressions on others (i.e., "fake good") than for people who are not incentivized to make favorable impressions.

In addition, in this chapter, I also aim to replicate the findings of chapter 3 that showed that judges' evaluations of targets' conscientiousness (in the mistake questions) and guilt proneness (in the dilemma questions) can be more valid than self-reports of those same traits. Whereas chapter 3 demonstrated that judges' evaluations can be more valid than self-reports in situations where targets' hiding-self should be inconsequential, in this chapter I aim to test whether judges' evaluations can be more accurate than self-reports especially for targets who tried to answer questions in an extremely positive manner.

Effect of Evaluation Dimensions on Validity

If each interview question extracts specific contents of moral character (e.g., conscientiousness in the mistake questions), the HIDE model posits that judges' evaluations focusing on smaller sets of cues (i.e., cues about targets' conscientiousness) can be more valid than judges' evaluations focusing on larger sets of cues (e.g., cues about targets' moral character as a whole). This is because an unnecessarily wide evaluation dimension can increase the size of the judge-ignorance and judge-error components, which together comprise the hidden-target zone in the HIDE model for judge reports. When the evaluation scope is too broad, encompassing more cues than are available and relevant, it can weaken the validity in two different ways. First, a lack of cues available to cover the breadth of the evaluation scope can result in a large judge-ignorance zone in the HIDE model. Second, even when there are cues available across a wide evaluation scope, the task of simultaneously detecting and utilizing too many cues can increase inconsistency in judges' evaluation of targets, resulting in a large judge-error zone in the HIDE model.

Findings in chapter 3 indicated that among the three moral character traits that we examined (honesty-humility, conscientiousness, guilt proneness), the mistake questions extract targets' conscientiousness information the best whereas the dilemma questions prompt extracts targets' guilt proneness information the best. The evaluation dimension scope was tested in an exploratory manner in chapter 3; in chapter 4, I conduct confirmatory hypothesis testing to verify the initial findings from the previous chapter. I do this by testing whether judges' specific evaluations of targets' conscientiousness (using the mistake questions) and guilt proneness (using the dilemma questions) are more predictive of targets' unethical behavior than judges' evaluations of targets' moral character as a whole.

Honesty-Humility and Greed-Avoidance

An unexpected finding in chapter 3 was that judges' evaluations of targets' honesty-humility revealed low levels of validity for all interview questions. In this chapter, I investigate a possible reason for this by focusing on four elements of the honesty-humility factor: fairness, sincerity, modesty, and greed-avoidance (Ashton & Lee, 2009). I propose that judges' evaluations of these four elements are not similar enough to form an overarching factor (i.e., honesty-humility), and thus asking judges to evaluate honesty-humility as a whole yields invalid ratings.

In the self-reported personality literature, it is well-established that the four facets of honesty-humility (i.e., fairness, sincerity, modesty, greed-avoidance) have enough shared variance to form one general honesty-humility factor (Ashton & Lee, 2009). However, it is possible that when these four facets are measured with judge reports from strangers, they lack the same shared variance, and thus evaluating these elements simultaneously results in unreliable judgments. In particular, I expect the largest heterogeneity in ratings between greed-avoidance and the other three honesty-humility facets. To allow for more intuitive evaluations, I use the term "greed" rather than "greed-avoidance," given the latter is a double negative. The HEXACO model defines greed as a tendency to desire lavish wealth, luxury goods, and signs of high social status (Ashton & Lee, 2009). Although it is true that being extremely greedy can be a negative indicator of moral character, moderate levels of greed might be perceived positively if it is interpreted as a signal of targets' achievement focus. In other words, greed can be interpreted as one's desire to get ahead, or ambitiousness, which can be a positive trait to certain degree. Therefore, I propose that targets' greed might be somewhat positively evaluated by judges, and thus do not reveal expected shared variance with honesty-humility.

To test the possibility that judges' ratings about greed are associated with targets' achievement focus, I conduct two types of analyses. First, I examine whether judges' greed evaluations are more strongly and positively correlated with conscientiousness than the other three elements of honesty-humility (fairness, sincerity, and modesty). This is done because conscientiousness is also well connected to one's achievement focus. Conscientiousness describes one's tendency to be organized, diligent, thorough and control impulses, which can all be positively associated with one's achievement focus. Therefore, I investigate whether judges' ratings of targets' greed compared with conscientiousness ratings, and ratings of other honesty-humility facets. To do so, in Study 7, I randomly assigned judges to make ratings of one of various evaluation dimensions, including greed, fairness, sincerity, modesty, honesty-humility, and conscientiousness.

Second, I conduct text analyses of written interview responses to examine the density of linguistic cues of achievement focus, using the Linguistic Inquiry and Word Count (LIWC) approach (Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC categorizes each word (e.g., I, my, me) into higher-order categories (e.g., first-person pronouns). Previous scholars have developed higher-order categories related to specific psychological processes (see Pennebaker et al., 2015), including achievement. This category measures usage of words that are related to achievement striving (e.g., win, success, better; Pennebaker, Booth, Boyd, and Francis, 2015). In the written interview method, judges are asked to make evaluations based on linguistic cues in targets' written responses. Therefore, if judges' ratings of greed are associated with their perceptions of targets' achievement focus, greed ratings should be positively correlated with the presence of linguistic cues related to achievement. I therefore test whether targets' word usage in the achievement category in LIWC is positively and more strongly correlated with judges' greed ratings compared with judges' ratings on the other three facets of honesty-humility (i.e., fairness, sincerity, and modesty).

By conducting these two different types of analyses, I examine the possibility that greed-avoidance does not map onto the overarching general factor of honesty-humility when measured by judge reports from strangers based on the written interview method, resulting in the observed invalid judge evaluations of honesty-humility observed in the previous chapter.

Study 7

The HIDE model predicts that the more targets try representing themselves in a positively biased manner, the more their implicit aspects of moral character are revealed through their written responses to behavioral interview questions. If this prediction is true, judges' evaluations should be more predictive of unethical behavior for targets who try harder to answer questions in a positive manner. In order to empirically demonstrate a link between validity of judges' evaluations and positively biased reporting, targets' levels of biased reporting should be varied. Therefore, in Study 7, I conducted an online experiment where targets were randomly assigned to two conditions with different levels of incentives to make (or fake) good impressions. Targets answered one of three interview questions (the mistake, dilemma, or employer question) and survey questionnaires. Then targets participated in two tasks in which they had opportunity to earn additional money by cheating and lying to other people.

Study 7 Methods

Data Collection from Targets

The target participants in this study were 606 U.S. full-time employees recruited from an online participant pool. These participants were assigned to one of two conditions, which were differentiated by the presence or absence of a reward for earning a positive evaluation. Targets in both conditions answered one of either the mistake, the dilemma, or the employer questions. Before writing their responses, targets in both conditions read the following instructions.

> *Imagine that you have been selected to interview for your dream job. The employers want to conduct an online interview before you meet them face-to-face. You will be asked questions about yourself and past experiences you may have had. Please use real examples from your life when responding. Please do not include last names or any other personally identifiable information in your response.*

Targets in the reward condition read the following instructions in addition to the above.

> *When responding to the interview questions and the survey that follows, we would like you to answer as if you are actually applying for a job and attempting to present yourself in the best possible way. The goal is to answer the interview questions in a way that you think would make you appear to be a good person with admirable qualities.*
>
> *Your interview responses and your answers to the personality questions will be evaluated by judges in the future (anonymously). The judges will determine the best job candidates among the participants in this study, based on these responses. Participants who score in the top 5% of the judges' evaluations will be sent a $25 Amazon gift card in a few weeks.*

After answering the interview questions, participants in both conditions completed two questionnaires: the HEXACO-60 personality inventory (Ashton & Lee, 2009) and the GP-5 (Cohen et al., 2014). Before answering the personality questionnaires, the participants in the reward condition were reminded that they should answer the personality questionnaires as if they were actually applying for a job and attempting to present themselves in the best possible way.

To measure the extent to which the target participants tried to answer the interview questions and personality questionnaires in a postively biased manner, two questions were administered: In responding to the written interview question, to what extent did you try to answer in a way that would make you appear to be a good person with admirable qualities? *[1 (Not at all), 2 (Slightly), 3 (Moderately), 4 (Quite a bit), 5 (Extremely)]*; In responding to the personality surveys, to what extent did you try to answer in a way that would make you appear to be a good person with admirable qualities? *[1 (Not at all), 2 (Slightly), 3 (Moderately), 4 (Quite a bit), 5 (Extremely)]*.

Finally, participants completed two online tasks (the number task and the problem-solving task) in a randomized order. They were given instructions that once they complete two types of online tasks, one of these tasks would be randomly chosen and then they would be paid a bonus payment based on their decisions in the chosen task.

The number task was based on methods used by Gneezy (2005). In this task, participants were led to believe that they were assigned to one of two possible roles (sender or receiver) and were paired with another participant who played the other role. In reality, all participants were assigned to the sender role. As the sender, participants needed to decide whether to send a deceptive message to the receiver to increase their chances of earning a bonus payment of $0.25. After participants were given instructions, they completed a comprehension-check test. If they failed the comprehension-check test, they were given the instructions again. If they failed the comprehension check a second time, they were informed that they could not participate in the number task. After passing the comprehension check, they were instructed that they would be paired with five different participants in five number tasks. The complete instructions for the number task are presented in Appendix 2.

The problem-solving task used in Study 1 was modified for online administration. Participants were shown a matrix for 7 seconds and asked to find two numbers that add up to 10. Each correctly identified pair of numbers was worth $0.25 in earnings, for a maximum bonus payment of $1.25. Participants indicated whether they solved the matrix after 7 seconds passed. They were given five matrices. In reality, all matrices were unsolvable, and thus participants who reported that they solved any matrices were considered to have cheated. The complete instructions for the problem-solving task are presented in Appendix 3.

Data Collection from Judges

In total, 2,390 participants recruited online served the role of judges. These judges were randomly assigned to read written responses to one of the three interview question conditions and to evaluate one of the evaluation dimensions. They read the definition of each evaluation dimension and then rated targets on 5-point rating scales ranging from extremely low to extremely high[7]. Each judge rated the interview responses of 20 randomly selected targets; judges' evaluations for each target are averaged before conducting statistical analyses. Each interview response for each evaluation dimension is rated by about 10 judges. Judges were assigned to one of the following evaluation dimension instructions.

> Please evaluate this respondent's tendency to think, feel, and behave in an ethical manner as compared to a typical job applicant. [*Moral Character*]
>
> Please evaluate this respondent's tendency to be fair, sincere, modest, and avoid greed as compared to a typical job applicant. [*Honesty-Humility*]
>
> Please evaluate this respondent's tendency to be organized, diligent, thorough, and inhibit impulses as compared to a typical job applicant. [*Conscientiousness*]
>
> Please evaluate this respondent's tendency to feel bad about his/her mistakes and wrongdoings even if no one knows about them as compared to a typical job applicant. [*guilt proneness*]
>
> Please evaluate this respondent's tendency to be genuine and truthful in his or her interpersonal relations as compared to a typical job applicant. [*Sincerity*]

[7] Judges were randomly assigned to 10 dimensions, including eight dimensions of traits and two questions that ask judges to predict the frequency of targets' cheating and lying. Results for the latter questions were not included in this study, but information about those two questions are available from the author.

Please evaluate this respondent's tendency to be fair and avoid fraud/corruption as compared to a typical job applicant. [*Fairnes*s]

Please evaluate this respondent's tendency to desire lavish wealth, luxury goods, and signs of high social status as compared to a typical job applicant. [*Greed*]

Please evaluate this respondent's tendency to be modest, humble, and unassuming as compared to a typical job applicant. [*Modesty*]

Study 7 Results

The descriptive statistics for Study 7 judges' average ratings of the eight dimensions are presented in Table S7C in Appendix 2. The two manipulation-check questions were highly correlated (r = .70, p < .001), and thus averaged to represent the extent to which targets tried representing themselves to judges in a positively biased manner. The mean of this average score was 2.59 (SD = 1.17) in the control condition and 3.29 (SD = 1.25) in the reward condition. The mean difference between the control condition and reward condition was significant (t=-.7.08, p<.001). While the levels of positively biased reporting in the two conditions differed significantly at the group-level, within each condition, individuals' positively biased reporting varied significantly (See Figure 9). More importantly, there was significant overlap of individuals' levels of positively-biased reporting between the two conditions (See Figure 9). Therefore, I used individuals' level of positively biased reporting score as a continuous predictor in the prediction model rather than using reward condition as a dummy variable.

The frequency-of-lying variable in the target data had missing values because a number of targets (1.5%) failed to pass the comprehension checks in the number task. Because of the existence of these missing values, rather than using the summed count score of lying and cheating, unethical behavior was operationalized as the average frequency of targets' lying and cheating. It was found that targets' frequency of cheating and lying did not exactly follow the shape of negative binomial or Poisson distribution. This was because a number of targets were concentrated at zero and five. This means that both right and left censoring occurred such that people who were unwilling to cheat or lie even when there was more opportunity were all categorized to zero and that people who were willing to cheat or lie more than five times were all categorized to five given the limited frequency that the 0 to 5 scale of the measure offers. Consequently, even when averaging cheating and lying, the data forms a w-shaped distribution, where the limits of categories (i.e., 0 and 5) had larger numbers of people concentrated than a usual normal distribution. Therefore, to deal with this simultaneous left- and right-censoring, in testing how well judges' evaluations predict targets' unethical behavior, I conducted two-sided censored regression analyses.

Do Judges' Evaluations Uncover Information Located in the Hiding-Self Zone?

The first sets of analyses answer the main research question of this chapter regarding whether judges' evaluations can correctly capture information located in the hiding-self zone of the HIDE model. If the written interview method elicits cues about implicit aspects of targets' moral character that targets cannot control or fake, then the more targets engages in positively biased reporting, the more their moral character should be revealed in their written responses. Then, judges' evaluations of targets based on those written responses should be more predictive of unethical behavior for targets who engage in more impression management and/or self-screening. The statistical representation of this pattern is the significant and negative interaction term of judges' evaluations of targets and targets' levels of positively biased reporting in the regression model of unethical behavior.

To examine whether judges' evaluations are more predictive for targets who answered interview questions in a more positively biased manner, four different models were analyzed for each interview prompt (See Table 10). Based on findings of chapter 3, conscientiousness evaluations were used for the mistake questions and guilt proneness evaluations were used for the dilemma questions. As reported in chapter 3, the employer question condition yielded mixed findings, and it remains unclear whether the employer question is useful for revealing targets' conscientiousness or guilt proneness. Therefore, I do not conduct confirmatory hypothesis testing for the employer question, and instead examined both the conscientiousness and guilt proneness evaluations. In the first model, only judges' ratings were entered. In the second model, targets' levels of positively biased reporting were added as well. Models 1 and 2 were conducted to provide baselines in interpreting the results of the third model, specifically regarding whether the addition of an interaction term – targets' levels of positively biased reporting and judges' evaluations – changes the main effect of judges' ratings or targets' positively biased reporting. A negative and significant coefficient of the interaction term would indicate that judges' ratings were more predictive of unethical behavior for targets with greater intentions to positively represent themselves. In HIDE model terms, this would indicate that judges were able to uncover information located in targets' hiding-self zone.

The results of models 1 and 2 indicate that judges' conscientiousness evaluations in the mistake questions had a negative and statistically significant prediction. The coefficient of targets' levels of positively biased reporting was positive and significant, meaning that the more targets engaged in impression management and/or self-screening, the greater the frequency at which they cheated and lied. Most importantly, in model 3, the interaction of judge's ratings and targets' positively biased reporting was negative, albeit not significant. The strength of the main effect of judges' ratings did not decrease substantially in models where targets' levels of positively biased reporting or the interaction term were added. Therefore, judges' evaluations seemed to capture variance that is not explained by targets' levels of positively biased reporting. Similar patterns were found for the dilemma questions, albeit the judges' evaluations and the interaction terms revealed weaker predictive power. For the dilemma questions, although the coefficients of judges' guilt proneness evaluations and the interaction terms were negative, they were not statistically significant. Finally, for the employer question, judges' conscientiousness ratings were negative but not statistically significant. Importantly, however, the interaction term was negative and marginally significant (b = -.38, p = .06). For the guilt proneness evaluation, the main effect of judges' ratings was not predictive of targets' unethical behavior, but nonetheless, the interaction term was negative and marginally significant (b = -.40, p = .07). This means that judges' evaluations of targets were more predictive of targets' unethical behavior for targets with greater intentions to positively represent themselves.

Are Judges' Evaluations More Valid than Self-Reports When Targets Represent Themselves in an Extremely Positively Manner?

In chapter 3, I demonstrated that judges' evaluations of targets' conscientiousness using the mistake questions were more predictive of targets' unethical behavior than self-reported conscientiousness. Likewise, results of chapter 3 indicated that judges' evaluations of targets' guilt proneness using the dilemma questions were more predictive of targets' unethical behavior than self-reported guilt proneness. An important but unanswered question is whether this pattern would be replicated in the situations where targets answer questions in an extremely positively biased manner, which resemble real-life situations such as job interview settings. Therefore, in Study 7, I replicate the findings of chapter 3 for targets who engaged in substantial amounts of positively biased reporting when answering interview questions and personality scales (top 25% on the positively biased reporting measure).

The results are presented in Table 11. For each condition, three sets of two-sided censored regression were conducted. In the first model, only the judges' ratings were entered; in the second, only the self-reports; the third model combines both judges' evaluations and self-reports. The results indicated that while self-reports of conscientiousness and guilt proneness were negative predictors, they were not significant at α=.05. However, judges' conscientiousness evaluations in the mistake questions and guilt proneness evaluations in the dilemma question were negative and statistically significant predictors of unethical behavior. Moreover, for the mistake questions, when both ratings were entered, judges' ratings exhibited stronger predictive power of unethical behavior, albeit not significant at α=.05. For the dilemma questions, when both ratings were entered together in the regression model, judges' evaluations significantly predicted targets' unethical behavior. These findings are in consistent with findings of chapter 3 that judges' evaluations of targets' conscientiousness (in the mistake questions) and guilt proneness (in the dilemma questions) can be more valid than self-reports of those same traits.

Finally, for exploratory purposes, I compared the judges' evaluations and self-reports for both the conscientiousness and guilt proneness dimensions in the employer question condition. The coefficients of judges' evaluations were negative but not significant for the conscientiousness dimension. For the guilt proneness dimension, judges' evaluations were marginally predictive of unethical behavior in both models 1 and 3.

Are Judges' Evaluations of Specific Traits More Valid than Evaluations of Moral Character?

To formally test whether judges' ratings of specific traits (conscientiousness in the mistake questions condition and guilt proneness in the dilemma questions condition) were more valid than ratings of moral character as a whole, I conducted three sets of two-sided censored regressions. In the first model, only the judges' evaluations of either conscientiousness or guilt proneness were entered. In the second, only the overall moral character evaluations were entered. In the third model, judges' evaluations of both dimensions were entered to see whether the conscientiousness or the guilt proneness evaluations explained unique variance of targets' unethical behavior. The results are presented in Table 12. Two-sided censored regression analyses revealed that only the conscientiousness evaluations were predictive of targets' unethical behavior in the mistake questions condition. Similarly, for the dilemma questions condition, guilt proneness evaluations revealed stronger coefficients, albeit both the moral character and guilt proneness evaluations were not significant at α=.05. In the exploratory analyses for the employer question condition, only the conscientiousness evaluations were negatively related to unethical behavior, but the coefficients were not significant.

Judges' Evaluations of Honesty-Humility and Greed

I examined possible reasons why judges' honesty-humility evaluations were not as valid as their conscientiousness or guilt proneness evaluations across all interview question conditions. First, I focused on the correlation patterns among four elements of honesty-humility (i.e., fairness, sincerity, modesty, greed), honesty-humility, and conscientiousness. The correlations among judges' evaluations of these various dimensions are presented in Table 13. Consistent with my prediction, I found that judges' greed evaluations were the most positively and strongly correlated with their conscientiousness evaluations. Greed evaluations were also positively and significantly correlated with moral character evaluations. Importantly, greed evaluations were not negatively correlated with judges' evaluations of honesty-humility, fairness, and sincerity. Greed evaluations were negatively and significantly correlated with modesty evaluations, but not strongly ($r = -.11$). Together, these patterns support the notion that judges' evaluations of four distinct elements of honesty-humility cannot be combined into a one general factor using the written interview method, and explain the invalid results of the honesty-humility evaluation in studies in chapter 3.

Second, I conducted text analyses to examine how linguistic cues about achievement focus might be associated with judges' ratings of the various evaluation dimensions. I proposed that the density of achievement focus cues should positively influence judges' conscientiousness and greed evaluations. Consistent with this prediction, I found that achievement cues were most positively and significantly correlated with judges' greed ratings (r = .23, p<.001), followed by conscientiousness evaluations (r = .13, p<.01) (See Table 13). Judges' evaluations of the other three elements of the honesty-humility factor were not correlated with achievement cues. These findings further support the notion that judges did not necessarily evaluate targets' greed negatively, because it can signal achievement focus; therefore, greed cannot be combined into honesty-humility factor.

Study 7 Discussion

The results of Study 7 indicated that the predictive power of judges' evaluations for unethical behavior tends to increase when targets answer interview questions with more positive bias. The written interview method was proposed as a way to elicit implicit aspects of targets' moral character that they cannot control or fake. I found evidence that the more targets engaged in misrepresentation in their written responses to interview questions, the more their answers revealed their unethicality. Therefore, the findings of this chapter support the HIDE model's prediction that the written interview method enables judges to uncover information located in the hiding-self zone. Further, the finding that only judges' evaluations of targets' conscientiousness and guilt proneness were predictive of unethical behavior for targets whose positively bias reporting score was in the top 25%. For those targets, self-reports were not predictive of their unethical behavior because their ratings are largely influenced by the hiding-self zone from the HIDE model of self-reports.

In Study 7, although the patterns found across the different interview question and evaluation conditions were largely similar to each other and supported the HIDE model's prediction, several coefficients were marginally significant or insignificant. I suspect that Study 7 in general has weak statistical power because the distribution of the dependent variable was not optimal: the criterion of unethical behavior (the average score of cheating and lying) was censored to both directions. Initially, I assumed that five opportunities for cheating and lying should be enough to differentiate targets who are high or low on ethicality. However, I found censoring in both directions, indicating an insufficient range in the variable. The left-side censoring means that targets who were extremely unwilling to cheat or lie were not differentiated from people who were unwilling to cheat or lie only for the five opportunities the study presented. The right-side censoring means that people who were willing to cheat or lie many more times than five were also not well-distinguished from others who were willing to cheat or lie only for the five opportunities presented. Therefore, this current criterion was insufficient for distinguishing people who are either extremely low or high in ethicality. This limitation of criterion range lowered the power in testing the two-tailed hypothesis testing. In future work, it is recommended to conduct a replication study using a criterion with more categories (e.g., 20 opportunities for cheating, similar to Study 1).

Findings in chapter 3 indicated that judges' evaluations of conscientiousness using the mistake questions and the guilt proneness evaluations using the dilemma questions were more valid than judges' evaluations of moral character as a whole. These results were replicated in

Study 7, in which targets were given an incentive to misrepresent themselves positively. These findings have important implications for the design of organizational interview processes: rather than evaluating targets using a holistic evaluation dimension (e.g., moral character), interviewers should evaluate candidates on specific dimensions with demonstrated validity based on interview questions that have been shown to elicit cues about those dimensions.

Also related to interview question design, results in Study 7 indicated that judges' evaluations using the mistake questions revealed the most stable and strongest predictive power of unethical behavior. This finding is consistent across most of the studies reported in chapters 2, 3, and 4. Future work should explore why the mistake questions prompt was particularly good at eliciting targets' moral character.

In previous chapters, judges' honesty-humility ratings did not show good predictive validity compared to the other two dimensions. Results of Study 7 revealed that one reason is, in contrast to self-reported personality measures, judges' evaluations of fairness, sincerity, modesty, and greed-avoidance do not share enough similarity to form an overarching general factor for honesty-humility. Therefore, judges' evaluations of honesty-humility based on four elements were not valid. This finding is quite important because it calls into question a common research assumption that the factor structure of personality measures holds across all constructs and for all different rating sources.

Finally, in Study 7, I found that targets in the reward condition produced more words, or longer responses to interview questions than those in the control condition (See Table S7B in Appendix). The mean differences of word count between the two experiment conditions were statistically significant for the mistake and the dilemma questions and marginally significant for the employer question. One possible reason for this is that more words revealed more information, which enhanced the predictive validity of the written responses. However, additional analyses revealed that this is not the case. First, targets' positively biased reporting score were not positively correlated with word count (See Table S7B in Appendix 2). Second, controlling for a word count variable in analyses did not change the main results (See Table S7E in Appendix 2). Therefore, it seems that the quantity of words is not what enables judges to evaluate targets' character. This in turn begs the question: what kinds of linguistic cues enable judges to evaluate targets? This question will be explored in chapter 5.

General Discussion

It would be reasonable to assume that job candidates are motivated to represent themselves to others in a positively biased manner rather than being honest about their negative traits. Consistent with this prediction, prior researchers estimate that 30-50% of job applicants engage in positive misrepresentation when answering questionnaires and surveys used in a selection process (Griffith, Chmielowski, & Yoshita, 2007). If all applicants misrepresent themselves by a similar degree, then this behavior does not change people's relative orders and therefore should not interfere with a valid evaluation process. However, a critical problem is that not all individuals distort their answers to a similar level; applicants who are dishonest and deceptive engage in higher levels of faking than their more honest peers (McFarland & Ryan, 2000; O'Neill et al., 2013). Because of this discrepancy, organizations are susceptible to hiring individuals who are dishonest and deceptive, and who will go on to contribute to increased levels of counterproductive work behavior that harms organizations and the people within them (Ones, Viswesvaran, & Schmidt, 1993; O'Neill et al., 2013).

Faking in real-world selection settings is pervasive and problematic, and numerous researchers have investigated how to reduce the invalidity of self-reports on desirable traits (e.g., Dwight & Donovan, 2003; Joubert, Inceoglu, Bartram, Dowdeswell, & Lin, 2015; Kluger & Colella, 1993; Landers, Sackett, & Tuzinski, 2011; Mcfarland, 2003; O'Neill et al., 2013; Salgado and Táuriz, 2012). Methods currently being explored include forced-choice response formats (e.g., Joubert et al., 2003; Salgado and Táuriz, 2012) and warning against faking before administering surveys (e.g., Dwight & Donovan, 2003; O'Neill et al., 2013). In a forced-choice format, respondents select one answer from a set of several equally desirable alternatives. This is in contrast to single-stimulus rating scale formats (e.g., Likert scale) where respondents indicate their agreement with a statement that could describe them. Salgado and Táuriz (2012) conducted a meta-analysis and found that forced-choice measures slightly increase the predictive validity or have similar levels of predictive validity of counterproductive work behavior compared to a single-stimulus format. With regard to warning against faking, researchers found that including warnings or instructions at the start of an assessment somewhat reduced faking of self-reports as well (Kluger & Colella, 1993; O'Neill et al., 2013).

While a rich set of research has been dedicated to dealing with faking or positive misrepresentation problems in selection settings using self-reports, the HIDE model provides an alternative method to reduce this concern. Specifically, the HIDE model suggests replacing self-reports with unacquainted judges' evaluations. Furthermore, the HIDE model suggests that when it is expected that substantial amounts of information can be located in targets' hidden or hiding-self zones, interview questions should be designed to elicit cues about implicit aspects of targets' characteristics. In this dissertation, the written interview method was used to elicit targets' moral character information, and the findings reported in this chapter indicated that the written interview method is particularly useful for identifying people who are engaged in faking their responses. Using the written interview method, judges' evaluations of targets' unethical behavior were more predictive of targets' unethical behavior among targets who misrepresented themselves more positively. Importantly, I found that among targets who represent themselves in a positively biased manner to a substantial degree (top 25%), judges' evaluations were predictive of targets' unethical behavior, but targets' self-reports were not. Furthermore, the HIDE model suggests that even when faking can be successfully reduced for self-reports, unacquainted judges' evaluations still offer a better choice for evaluating socially desirable traits such as conscientiousness. This is because invalidity of self-reports arises not only not only from the hiding-self but also by the hidden-self—aspects that individuals do not correctly know about themselves. In chapters 2 and 3, even under the situations where individuals are likely to fake their responses, judges' evaluations were more predictive of targets' unethical behavior than self-reports. Combined, the findings in chapters 2, 3, and 4 provide a promising direction for further research as well as evolution of organizational practice, namely, to replace or complement self-reports with unacquainted judges' evaluations to uncover information located in the hidden or hiding-self zone.

CHAPTER V

Shared Biases in Human Perceptions of Ethicality

The HIDE model describes when and how unacquainted judges can make valid evaluations of targets' moral character, and thus predict future unethical behavior. In the HIDE model of judge reports, invalidity of unacquainted judges' evaluations arises from the hidden-target zone—aspects of targets' characteristics that judges cannot correctly perceive. For the written interview method, the hidden-target zone can happen when no cues (or too few cues) about targets' characteristics are available from the written interview responses, or when cues are available, but judges do not reliably detect and utilize them. As a one simple way to reduce the hidden-target zone for the written interview method, the HIDE model proposes to use multiple judges. This is because while an individual judge might be unreliable in detecting relevant cues and utilizing those cues in a consistent way, groups of judges can reliably detect and utilize available cues based on the literature of the wisdom of crowds (Davis-Stober, Budescu, Dana, & Broomell, 2014; Larrick, Mannes, & Soll, 2012; Larrick & Soll, 2006; Mannes, 2009).

While using multiple judges can be an important means to reduce the unreliability of human judgments, certain aspects of unreliability and invalidity in human judgments are unresolvable due to basic limitations in cognitive capacity or to widely shared cognitive biases (Hammond et al., 1987). In particular, if most judges have a similar, systematic bias in utilizing certain cues, then aggregating evaluations from several judges will still yield an unreliable and biased estimate of targets' characteristics. In these situations, I propose that machine algorithms (e.g., automated text analysis, machine learning) for understanding texts can complement human judgments.

I do not conduct hypothesis testing in this chapter, but rather take an exploratory approach. In particular, I use a lexicon-based text analysis to examine whether the density of certain linguistic cues in targets' written responses to interview questions is predictive of unethical behavior and whether the density of those cues is predictive of judges' average moral character ratings. If there is shared biased in human judgment, at least two possible patterns are expected. First, while some linguistic cues are predictive of targets' unethical behavior, judges' moral character evaluations are not associated with those cues. Second, it is possible that while judges' moral character evaluations are associated with certain linguistic cues, targets' unethical behaviors are not associated with those cues.

In this exploratory chapter, I identify a small set of higher-order categories that theoretically are relevant to unethical behavior and moral character judgments for each interview question. I then compare the relationship among the density of those categories used in targets' written responses, unethical behavior, and judges' ratings. For example, I examine whether specific patterns of word usage (e.g., frequent usage of first-person pronouns compared to third-person pronouns) are diagnostic of unethical behavior. I use the approach developed in the LIWC (Linguistic Inquiry and Word Count) program, which categorizes individual words into predefined, higher-order categories, and then provides information on the extent to which those categories are used in given texts. One important note is that the categories that I choose to explore in this chapter are not an exhaustive set of all the categories that can predict targets' unethical behaviors or judges' moral character evaluation using the LIWC approach. Instead, they were selected to demonstrate initial evidence of shared biased in human perceptions of ethicality. In the following sections, I summarize the higher-order categories chosen to explore in each interview question condition.

Mistake Questions. The mistake questions ask targets about how they felt and behaved after making a mistake. For written responses to the mistake questions, I focus on linguistic cues related to negative emotions for several reasons. First, the question directly asked targets about how they felt after a negative experience (making a mistake), and therefore, it is reasonable to assume that targets would express negative emotions in their written responses. Second, negative emotions are theoretically associated with aggression and other harmful behaviors, which broadly can be considered unethical. For example, according to the stressor-emotion model of counterproductive work behavior (Fox & Spector, 2006), negative emotions that arise from stressful situations (e.g., frustration, anger) can lead individuals to engage in aggressive or harmful behaviors, including counterproductive work behavior (Fox & Spector, 2006). Therefore, targets' implicit tendencies to experience negative emotions after making a mistake (a source of stress), may be predictive of their unethical behavior. Third, in prior chapters, I found that the judges' conscientiousness ratings revealed the strongest predictive validity of the dimensions tested. One important aspect of conscientiousness is whether or not individuals can regulate their impulses that arise from negative emotions. How targets managed their negative emotions after making a mistake, therefore, should be diagnostic of their levels of conscientiousness.

In LIWC, three types of higher-order categories are developed for negative emotions: anger, anxiousness, and sadness. For example, words such as *hate*, *kill*, and *annoyed* are grouped into the anger category. Some example words of the anxiousness category are *guilt*, *nervous*, and *worried*. Words like *crying*, *grief*, and *sad* are grouped into the sadness category.

Initially, I expected that the mistake questions could reveal targets' guilt proneness since more ethical targets might express more guilt after making mistake. However, the findings reported in chapters 2, 3, and 4 revealed that judges' ratings of guilt proneness were the least valid for the mistake questions. In light of this, I suspected that state-level guilt (i.e., guilt following a mistake) is not a good measure of guilt proneness, as the latter is defined as anticipated guilt after future wrongdoing. It follows that targets' expressions of guilty feelings (or shameful feelings) in their written responses could have been perceived by judges as an indicator of guilt proneness, but they do not predict unethical behavior because they are state-level guilt, which could be mixed with shame, humiliation, embarrassment, anxiety, and/or other negative emotions that are not necessarily indicator of one's morality. Guilty feelings are captured by the anxiousness category in the LIWC. If the above explanation is true, I expect that the anxiousness category will be positively associated with judges' moral character ratings because guilt proneness is one indicator of moral character, but not negatively associated with unethical behavior.

Regarding the anger and sadness categories, it is helpful to consider the circumplex model of emotion to understand their differences. According to the two-dimensional structure of affect, emotion can be explained by the valence (pleasant vs. unpleasant) and the activation (high vs. low energy) dimensions (e.g., Feldman, Barrett & Russell, 1998). Anger is characterized by high energy where sadness is characterized by low energy. Because anger is an active emotion, it is more likely to activate relevant behavior as well; the positive link between anger and aggression is well-established in the literature (e.g., Deffenbacher, Deffenbacher, Lynch, & Richards, 2003). Therefore, I expect a positive association between anger and unethical behavior in the mistake questions condition. Also, considering that the link between anger and aggression is quite intuitive, I expect judges' moral character ratings to be negatively associated with anger. I expect that sadness would have a positive link with unethical behavior given that it can be also a source of stress that can activate negative behaviors. However, given that sadness is characterized by low energy, and that low energy can be associated with low levels of engagement of any behavior (including commitment of unethical behavior), judges might not associate cues of sadness with targets' unethicality.

Dilemma Questions. The dilemma questions ask targets what factors they considered in making a difficult decision. This set of questions was initially designed to capture targets' honesty-humility, because targets who are fair and modest would consider how their own decisions might influence others while targets who are greedy and selfish would only focus on themselves. Therefore, in general, I expected that the dilemma questions would reveal targets' other-orientation versus self-focus. This chapter explores whether written responses to the dilemma questions did contain linguistic cues that can be reflective of targets' other-orientation versus self-focus and whether judges' moral character evaluations were based on those cues.

In exploring linguistic cues that are possibly diagnostic of unethical behavior or judges' moral character ratings in the dilemma questions condition, I focus on the higher order categories in LIWC that can be associated with other-orientation. First, I investigate whether personal pronoun usage, especially third-person pronoun usage (e.g., she, they) versus first-pronoun usage (e.g., I, mine), influences judges' moral character evaluations and predict targets' unethical behavior. In LIWC, the social process category (e.g., words related to friends, female/male references, family) can also capture how much targets talked about other people. Finally, the prosocial dictionary (Frimer et al., 2014) category in LIWC consists of words or word stems that are indicative of content about collective interests and interpersonal harmony, which can also be closely related to targets' other-orientation. Therefore, I investigate whether third- or first-person pronoun usages, social process words, and prosocial words predict targets' unethical behaviors and judges' moral character evaluations of targets.

Several scenarios are possible. First, it is possible that these linguistic cues (or a subset of these cues) of other-orientation predict unethical behavior negatively and judges' moral character ratings positively. If this is the pattern, it would indicate that human judges were able to perceive targets' ethicality correctly from the available linguistic cues. Second, it is possible that these cues (or a subset of these cues) were predictive of unethical behavior, but judges did not use these cues in making their judgement. This would then mean there might be some cognitive limitation or biases in human judgment. Third, it might be the case that these cues were not predictive of unethical behavior, but were positive predictors of judges' ratings. In this third situation, we can infer that there is a shared-biases in human perception of ethicality such that judges perceived those cues are related to ethicality, but they were not.

Employer Question. This question asks targets about their perceptions about how their current or previous employers would evaluate them. Initially, this question was expected to capture targets' honesty-humility based on the speculation that targets who are high on honesty-humility would be less likely to assume that their employer would only perceive them in a positive manner. In particular, it was expected that a more honest person might reveal more negative information about herself or himself. Also, it was expected that a person who values fairness might describe herself or himself by referencing others, such as saying that they are good team player, while person who is self-focused might only focus on their own competence. Therefore, I expected that the lower a target's honesty-humility, the more extremely positive their answer to this question would be.

In exploring the relationship among linguistic cues, judges' moral character ratings, and unethical behavior for the employer question, I focus on three predefined LIWC categories that can be broadly associated with various elements of honesty-humility: achievement, affiliation, and prosocial dictionary. The achievement category in LIWC summarizes word usage in reference to success, failure, and achievement striving (e.g., *win*, *success*, *better*). In chapter 4, I argued that these cues can be associated with judges' perception of targets' greed. The affiliation category is another one that may capture cues for honesty-humility. While both the achievement and affiliation categories in LIWC summarize word usage that reveals one's desires, goals, and motivations, the former focuses on one's success, while the latter focus on interacting with others or belonging in a group (e.g., *ally*, *social*, *friend*). The third category I focus on for the employer question is the prosocial dictionary category (Frimer & Schaefer, 2014; Frimer, Aquino, Gebauer, Zhu, & Oakes, 2015), which captures people's other-orientations. I consider this category because affiliation motivation is closely related to collective interests. Therefore, in Study 8 I explore whether these three higher-order word categories—affiliation, achievement, and prosocial dictionary—predict targets' unethical behavior and judges' moral character ratings in the Employer question condition.

I conduct automated text analyses to identify the linguistic cues that are predictive of judges' evaluations of targets' moral character and cues that are predictive of targets' frequency of unethical behavior. I then compare the two sets of cues, and identify some linguistic cues that human judges failed to correctly detect or utilize, indicating possible shared biases in human perceptions of ethicality. By illuminating this possiblity, I aim to introduce some potential future directions of this research program regarding combining human judgments and machine algorithms to increase the accuracy of unethical behavior forecasts.

Study 8

I conducted text analyses for the three target data sets collected in studies 1, 2, and 7. The other data sets were not analyzed because either they did not utilize a criterion of unethical behavior or because the sample size is relatively small. Samples A, B, and C (i.e., targets data sets in studies 1, 2, and 7) differ from each other on several aspects, including types of criteria and target recruiting method. Importantly, the same sets of linguistic categories were used for analyzing all data sets and all interview questions. Therefore, by cross-validating the predictive patterns of linguistic cues in these different data sets, I aimed to establish the robustness of the findings. The ultimate goal of Study 8 was to identify linguistic cues that human judges do not detect or utilize in making moral character judgment, which in turn elucidates the shared-biases in human perceptions of ethicality.

In conducing analyses for sample C (i.e., Study 7 targets), I aimed to replicate as well as to extend the findings in samples A and B by examining whether the associations among linguistic cues, unethical behavior, and judges' ratings changed depending on targets' levels of intention to provide favorable answers. For example, it is possible that certain linguistic cues are used more or less frequently when targets were more motivated to convey good impressions.

Study 8 Method

Samples A & B. Participants in the first two target data sets answered interview questions without a financial incentive to represent themselves in an overly favorable manner. They were instructed to imagine they were applying for a job, but there was no financial incentive associated with this. In Study 1, targets responded to either the mistake or dilemma interview prompt and engaged in a problem-solving exercise in which targets' cheating was measured in a laboratory experiment. In Study 2, employees in various parts of the U.S. were recruited by a survey research firm, and responded to one of the mistake, dilemma, or employer interview prompts, then self-reported their CWB. In examining the relationship between the linguistic categories and unethical behavior, I used negative binomial regressions because the criteria in Studies 1 and 2 were count measures. For sample 1, I controlled for the number of correctly solved matrices, since the more targets correctly solved, the less opportunity they had to cheat. To analyze the relationship between linguistic cues and moral character ratings, I used ordinary least square regressions. Two sets of judges (judges in studies 1 and 3) evaluated the study 1 targets' moral character. Average scores from those two sets of judges were calculated prior to analyses.

Sample C. Analysis of sample C extends the findings of samples A and B by examining whether the relationships among linguistic cues and unethical behavior varied depending on a target individual's level of positively biased reporting. In study 7, targets answered one of the mistake, dilemma, or employer questions in an online experiment. Then, they participated in two online activities measuring cheating and lying. In study 7, targets' motivations to convey positive impressions were manipulated and their levels of positively biased reporting were measured via survey. Given that individuals' levels of positively biased reporting varied significantly within the control and the reward condition (See Figure 9), I used individual score of positively biased reporting as a continuous predictor in the prediction model rather than using the dummy variable of reward condition. Because the criterion (i.e., average frequency of cheating and lying) was censored in both directions, I conducted double-censored regressions when testing the relationship between the linguistic cues and unethical behavior. To examine whether the relationships among linguistic cues and unethical behavior varied depending on a target individual's level of positively biased reporting, I tested whether the interaction term of a given linguistic cue and target individuals' levels of positively biased reporting is significant.

To reduce the possibility of multicollinearity and to provide more meaning to the intercept in the model, variables were grand mean centered in testing these interactions. Finally, to analyze the relationship between linguistic cues and moral character ratings, I used ordinary least square regressions. Two sets of judges (judges in studies 7 and 8) provided moral character ratings of targets. I used the average of those two scores in analyses.

Study 8 Results

The correlations between targets' unethical behavior and judges' moral character judgments and selected LIWC categories for each interview question condition (i.e., first and third person pronoun ratio[8], three negative emotions, affiliation and achievement, social process words, and prosocial dictionary) are presented in Table S8A in the Appendix 2.

Mistake Questions. The results for the mistake questions are presented in Table 14. Across three samples, analyses revealed that the anger category positively predicted unethical behavior and negatively predicted judges' moral character ratings. The patterns for the anxiousness category were somewhat mixed. For samples B and C, anxiousness showed a positive relationship with unethical behavior (although not significant), though this was not the case in sample A. For samples A and B, anxiousness was shown to have a positive relationship with judges' moral character ratings. For sadness, the three samples provided consistent results showing cues in the sadness category were a positive predictor of unethical behavior. The coefficient was marginally significant for sample A, and significant for sample B. Although not significant, the coefficient in sample C was in the same direction. As predicted in the introduction, sadness was not associated with judges' moral character ratings. Therefore, the results of sadness category demonstrated that human judges were unable to utilize some valid cues in written responses.

Dilemma Questions. The results for the dilemma questions are presented in Table 15. Results regarding the first- and third-person pronoun usage did not reveal a consistent

[8] The ratios of first-person and third-person pronouns over the total personal pronouns were computed.

relationship with unethical behavior predictions, but they revealed consistent relationships with judges' moral character ratings. Across the three studies, the more the target used first-person pronouns, the more negatively they were evaluated by judges. When they were entered together in one model, these were not significant with α=.05 in samples A and B because the two types of pronoun usage were negatively correlated. When entered separately (See Table S9A in Appendix 2), coefficients of the first-person pronoun category were negative and significant in samples A, B, and C. Similarly, when entered separately, the third-person pronoun category revealed a positive and statistically significant relationship with moral character in samples A, and B. The social process category and prosocial dictionary revealed mixed findings, and their coefficients were much weaker than those from pronoun usage.

Employer Question. The results for the employer question are presented in Table 16. Written responses to the employer question were only available in samples B and C. Across the two samples, it was revealed that affiliation was a negative predictor of unethical behavior; this effect was statistically significant in sample A. Also, use of cues in the affiliation category was a positive and statistically significant predictor of moral character judgment. The achievement category revealed mixed findings. The prosocial category revealed consistent findings in samples B and C: the more prosocial words targets used in their written responses, the less unethical behavior they engaged in. This prediction was significant in sample A. Also, the more prosocial words targets used in their written responses, the more positively they were evaluated by judges. This prediction was significant in both samples. Interestingly and importantly, the interaction term of the linguistic cues of achievement and targets' positively biased reporting was positive and significant. This means that judges' ratings were more predictive of unethical behavior for targets who answered questions in a more positively biased manner and that in these cases targets used linguistic cues of achievement focus to a greater degree.

Study 8 Discussion

In this exploratory chapter, I examined the possibility of shared biases in human perceptions of ethicality that would interfere with the functioning of the written interview method for making valid judgements about targets' moral character. The analyses revealed that there were indeed certain valid cues available in written responses that human judges failed to utilize in making moral character judgments. For example, in the mistake questions condition, linguistic cues about negative emotions—especially anger and sadness—were predictive of targets' unethical behavior. Although the judges' moral character evaluations based on the mistake questions were negatively associated with cues for anger, there was no association with cues for sadness. Therefore, these results provide initial evidence that human judges may fail to detect or utilize some valid cues in written responses. This in turn indicates a shared bias in human perceptions of ethicality, such that people do not associate unethicality with sadness as much as they do with anger.

In the beginning of this chapter, I argued that state-level guilt would not negatively predict unethical behavior; nonetheless, judges' moral character ratings can be positively associated with it. Although not fully supported, patterns found in study 8 were consistent with these predictions. Anxiousness, which captured expression of guilt, had a positive relationship with moral character ratings in samples A and B, but also have a positive relationship with unethical behavior.

In the dilemma questions, the text analyses revealed that targets' third-person pronoun usage was a positive predictor of judges' moral character ratings, yet was also a positive predictor of unethical behavior in two studies out of three. These findings suggest that there are shared biased in human judgments regarding how to utilize detected cues in judging others' ethicality. Targets' first-person pronoun usage was a negative predictor of judges' moral character ratings across all studies, but similar to the third-person pronoun usage, the relationship between the first-person pronoun usage and unethical behavior was inconsistent.

Finally, in the employer question, an interesting and important pattern was found for the moderating effect of positive reporting intentions: the more targets were motivated to make a good impression, the more achievement linguistic cues were predictive of targets' unethical behavior. This means that unethical individuals tended to emphasize their achievement much more, especially when motivated to make good impressions. This finding provides some promising directions for future research to investigate whether and people's self-representation changes depending on contexts.

Although the sets of linguistic cues explored in this chapter were rather small, the findings were rich and stable enough to provide some initial evidence of the shared biases in human perceptions of ethicality and open a rich set of potential directions for this research program.

Summary of Dissertation & Conclusion

This dissertation investigates how we can identify individuals who are likely to engage in unethical behavior when we do not personally know them. In particular, this dissertation investigates when and how we can make valid evaluations of others' moral character—people's tendency to think, feel, and behave ethically—and thus predict unethical behavior. The biggest challenge in assessing moral character traits is that they are extremely socially desirable, and therefore highly susceptible to distorted self-perceptions and impression management.

To address this problem, in chapter 1 of this dissertation, I developed a new person-perception and reporting framework, the *hidden information distribution and evaluation (HIDE) model*. The HIDE model posits that there are aspects of information that individuals do not correctly know about themselves (which I call the *hidden-self*), as well as aspects of information individuals misrepresent to others (which I call the *hiding-self*). The model articulates when and why judges who do not personally know targets can make more valid judgments of targets' moral character than targets themselves or their acquaintances. In particular, the model predicts that the impromptu thinking and language usage that arises when a person answers specifically designed interview questions reveal information about targets' hidden-self and hiding-self, thereby enabling a small group of judges to reliably evaluate targets' moral character. Based on the HIDE model, I proposed four specific predictions for the written interview method, and I tested these predictions as well as the validity of the written interview method by conducting eight empirical studies with 1,876 targets and 3,555 judges (See Table 17).

In chapter 1, I examine the HIDE model's prediction that a group of judges will outperform a single judge in predicting targets' unethical behavior based on their written interview responses. I investigated this "wisdom of crowds" in forecasting unethical behaviors using the written interview method by conducting three empirical studies. In studies 1 and 2, I crowd-sourced large sets of judges online and these judges evaluated targets' moral character from written interview responses. In study 1, the judges' average moral character ratings negatively and significantly predicted the extent to which targets cheated on a problem-solving task. In study 2, the judges' average moral character ratings negatively and significantly predicted the frequency that targets reported engaging in workplace deviance (e.g., falsification of expense reports, stealing, and interpersonal abuse). Study 3 extended the findings of studies 1 and 2 by determining the judge group size at which the crowd effect occurred when forecasting unethical behavior using the written interview method. I found that six judges were enough to reliably estimate the moral character of targets and predict their unethical behavior.

Having established the possibility of predictive validity of unethical behavior for unacquainted judges using the written interview method, in chapter 3, I test two additional HIDE model predictions. First, I investigate the prediction that judges' evaluations can uncover information located in the hidden-self zone of the HIDE model of self-reports. I investigate this by comparing the predictive validity of unacquainted judges' evaluations with targets' self-reports. Second, I investigate the prediction that unacquainted judges' evaluations can uncover information located in well-acquainted judges' hidden-target zone of the HIDE model of judge reports. I test this by comparing the predictive validity of unacquainted judges' evaluations with ratings provided targets' acquaintances, namely, their coworkers. In order to compare the predictive validity of unethical behavior among these three rating sources (unacquainted judges, targets themselves, well-acquainted judges), all rating sources evaluated targets on the same evaluation dimensions: honesty-humility, conscientiousness, and guilt proneness. While the findings in chapter 2 show that judges reliably evaluated targets' moral character as a whole and these evaluations predicted targets' unethical behavior, in Study 4, I seek to understand what discrete aspects of moral character judges are detecting from each interview question. Specifically, I investigate judges' evaluations of targets' honesty-humility, conscientiousness, and guilt proneness and find that the mistake questions exposed targets' conscientiousness the best and the dilemma questions revealed targets' guilt proneness the best. Judges' honesty-humility evaluations were the least reliable and valid across all interview questions. Based on these findings in subsequent studies, the mistake questions prompt was used for judges to evaluate targets' conscientiousness and the dilemma questions prompt was used for judges to evaluate targets' guilt proneness. Findings of studies 5 and 6 indicate additional support for these question-evaluation dimension pairings. In Study 5, I find that judges' conscientiousness evaluations of targets using the mistake questions are more predictive of targets' unethical behavior than self-reported conscientiousness, and judges' evaluations of targets' guilt proneness using the dilemma questions are more predictive of unethical behavior than targets' self-reported guilt proneness. In Study 6, I find that judges' conscientious evaluations using the mistake questions and guilt proneness evaluations using the dilemma questions are more predictive of targets' unethical behavior than conscientiousness ratings provided by targets' coworkers (well-acquainted judges).

The main purpose of chapter 4 is to test the HIDE model's prediction that judges' evaluations can uncover targets' hiding-self—purposeful, positively biased representations of themselves to others. In Study 7, I manipulate the incentive structure in the interview questions and measure the extent to which targets answered interview questions in a positively biased manner. I examine whether judges' evaluations of targets using the written interview method can predict targets' unethical behavior to a greater degree among people who try to answer questions in more positively biased manners. I find that judges' conscientiousness evaluations using the mistake questions and judges' guilt proneness evaluations using the dilemma question are indeed more predictive of targets' unethical behavior among targets who answer interview questions positively to a greater degree. This finding indicates that judges can detect targets' hiding-self using the written interview method, and in fact that the larger the hidden-self component of the HIDE model is (or, the more targets have an incentive to misrepresent themselves in a positive manner), the more valid unacquainted judges' evaluations of targets' moral character are.

Combined, the seven empirical studies in this dissertation provide compelling evidence for the HIDE model and the written interview method for evaluation. This research finds that groups of judges are able to forecast targets' unethical behavior by evaluating their moral character via written interview responses. Importantly, unacquainted judges' ratings were more predictive of targets' unethical behavior than self-perception or reports from targets' acquaintances. Therefore, judges' evaluations using the written interview method are able to uncover targets' hidden-self, information that targets themselves do not correctly know or are unaware of themselves. Also, it was found that judge's evaluations were more predictive of unethical behavior among targets who answers interview questions in more positively biased manners. Therefore, judges are able to uncover targets' hiding-self, information that targets consciously try to hide or misrepresent to judges.

In cases when there are shared biases in human perceptions of ethicality, a different method (e.g., automated text analyses) could complement human judgments. In chapter 5, I report on Study 8, in which I conduct exploratory text analyses of the written interview responses from chapters 2 and 4 to detect linguistic cues that human judges failed to utilize in making moral character judgments. The analyses reveal that targets' negative emotions that are revealed in written responses to the mistake questions—especially anger and sadness—are diagnostic of unethical behavior among the targets. Although the judges' moral character evaluations in the mistake questions are negatively associated with anger, they are not associated with sadness. I find an interesting moderating effect of targets' positively biased reporting to linguistic cues. The positive predictive power of anger decreases as targets' positively biased reporting increases, whereas the positive predictive power of guilt proneness increases as targets' positively biased reporting increases. In the dilemma questions, the text analyses reveal that targets' third-person pronoun usage is a strong, positive predictor of judges' moral character ratings. However, third-person pronoun usage is a positive predictor of unethical behavior in two studies out of three. Finally, in the employer question, linguistic cues of affiliation are negative predictors of unethical behavior, and linguistic cues of achievement are a positive predictor of unethical behavior. Importantly, the predictive power of unethical behavior regarding linguistic cues of achievement increases as targets' level of positively biased reporting increases. Finally, while judges' moral character ratings were weakly positively associated with affiliation, they were not associated with achievement. Together, the results of Study 8 provide some initial evidence that there are shared biases in human perceptions of ethicality, such that judges do not correctly utilize some valid cues of targets' unethical tendencies.

Implications

The HIDE model and the written interview method have significant theoretical and applied contributions to the social, behavioral, and organization sciences. Virtually all managers desire an ethical workforce, yet little evidence-based guidance exists for forecasting unethical behavior of strangers. Previous work has not studied whether we can forecast strangers' unethical behavior, and if we can, then how to elicit the information necessary to make valid judgments. Predicting individuals' unethical behavior is particularly challenging because people want to see themselves and be seen by others as moral.

The most significant contribution of this research is that it shows that predicting strangers' unethical behaviors is possible and offers initial guidance on how valid predictions can be made. This dissertation has demonstrated that naïve groups of judges are able to predict targets' unethical behavior by using the written interview method and questions specially designed to elicit implicit aspects of targets' unethical tendencies.

Researchers could use the character-interview questions developed in this research to facilitate understanding of how an individual's ethicality is revealed via lingusitic cues and how people form impressions of others' ethicality, while practitioners could apply the findings from this research to improve personnel selection, promotion, and admissions procedures in organizations.

Future Research

Although I tested the HIDE model predictions and the validity of the text-based interview method by conducting eight empirical studies with a large number of participants (1,876 targets and 4,105 judges), still there are some limitations that future research should seek to remedy. In particular, the findings in chapter 5 provide promising directions for further research that could remedy the limitations of this dissertation and further extend the theory of person perception. In the following sections I summarize the limitations of this dissertation and introduce some ongoing projects as well as the potential direction of this research program.

Machine-Learning Algorithms to Understand Human Biases

In chapter 5, I present initial evidence that certain linguistic cues in written responses are not optimally detected or utilized in judges' evaluations. However, I focused on several small sets of linguistic cues, rather than simultaneously considering all linguistic cues available. Although those small sets of linguistic cues successfully demonstrated the existence of human biases in perception of ethicality, we still lack a full understanding of the relative predictive validity of human judges versus machine-algorithms in predicting unethicality based on texts.

In ongoing work, I use machine-learning to investigate comprehensive sets of verbal cues more systematically to predict targets' unethical behavior. Using latent Dirichlet allocation (LDA) analyses, I aim to identify topics that are predictive of unethical behavior from written responses to each interview question. LDA, which is conceptually similar to factor analysis, is a form of machine-learning for text data that extracts underlying dimensions (i.e., latent semantic clusters). In LDA, each dimension consists of several different linguistic cues, or words, that appear together in texts. For example, the use of certain keywords in written interviews (e.g., *others*, *concern*, *worry*, *need*, *care*, *help*, *empathize*) could reflect semantic factors that would allow us to identify targets who are considerate of others; conversely, targets prone to engage in unethical behaviors would rarely use those keywords. I aim to understand which cues are more predictive of targets' unethical behavior compared to judges' ratings of moral character as well as other dimensions to better understand human biases in perceptions of unthicality. Eventually, I aim to combine human judgment and machine algorithms to increase the predictive validity of the HIDE model-based written interview method.

Alternative Dimensions to Evaluate Written Interview Responses

Another important finding from chapter 5 is that the linguistic cues that are predictive of targets' unethical behavior and judges' ratings are quite different across interview questions (See Table S8F in Appendix). This is somewhat intuitive given that each question is designed to measure different elements of honesty-humility, conscientiousness, and guilt proneness. However, given the differences in available cues, the evaluation dimensions could be further refined to each interview question. This suggestion is consistent with the HIDE model prediction that smaller and more specific evaluation dimensions could potentially reduce judgment errors.

The results of chapter 5 provide some possible alternative dimensions. For example, in the employer question, linguistic cues for affiliation are shown to be negative predictors of unethical behavior and positive predictors of judges' ratings. Also, achievement cues are shown to be positive predictors of unethical behavior in the employer question. It follows that alternative evaluation dimensions for the employer question could be targets' affiliation focus versus achievement focus.

In the mistake questions condition, targets' negative emotions were a positive predictor of unethical behavior. While conscientiousness evaluations based on the mistake questions are valid and reliable, more refined evaluation dimensions could further enhance the validity. For example, judges' evaluations of targets' ability to regulate their negative emotions could be a good alternative dimension to examine in the future.

The dilemma questions prompt was initially designed to reveal targets' tendency to consider others, which I initially suspected would be captured by honesty-humility judgments. Honesty-humility judgments did not show validity, I used guilt proneness as the evaluation dimension for the dilemma questions. However, it is possible that there are other relevant constructs related to other-orientation. For example, empathic concern or benevolence are potential alternative dimensions to examine in future research.

Positive versus Negative Valance of Evaluation Dimensions

In these studies, most evaluation dimensions used were presented using a positive valence. For example, in the Conscientiousness dimension condition in study 8, judges were asked evaluate targets' tendency to be organized, diligent, thorough, and inhibit impulses compared to a typical job applicant. One exception was greed, given that greed-avoidance could be confusing because the latter is characterized with double-negative meaning. In future work, it should be examined whether the valance of the construct can influence judgment qualities. It is possible that by asking judges to focus on construct with a negative valence (e.g., dishonest, lazy, impulsive), judges would focus on characteristics of unethical individuals rather than focusing on characteristics of ethical individuals (e.g., honest, diligent, cautious) and this might lead to better detection of unethical individuals.

Finally, future work could examine whether the written interview method can also be used to predict *ethical* behavior. Given the definition of moral character—people's tendency to think, feel, and behave ethically—both ethical and unethical behavior can be criteria in evaluating the validity of moral character judgments. I only focused on the latter because I reasoned that moral character should have a stronger and more stable link with unethical behavior than ethical behavior in everyday life, because ethical behavior is more difficult to define clearly. The concept of ethicality or unethicality includes motivational elements (Cohen et l., 2014; Cohen & Morse, 2014; Hogan, 1973; Schwartz et al., 2012). This means that whether a behavior is right, or ethical, cannot be evaluated without considering the fundamental reason why the actor engages in such conduct. When the motivation is purely self-benefitting, the act is not considered ethical regardless of whether the behavior seems helpful to others on a surface level. For example, helping others can be sourced back to one's self-benefiting motivations such as cultivating social networks or building positive reputations, in addition to stemming from social norms. In contrast, harming others is less likely to be interpreted as having other-benefiting motivations. Therefore, ethical behavior is more interpretative and ambiguous in its motivations than unethical behavior. Future work examining the validity of judges' evaluations using clearly defined ethical behavior would provide further evidence of the theory proposed in this dissertation.

References

Asendorpf, J. B., & Ostendorf, F. (1998). Is self-enhancement healthy? Conceptual, psychometric, and empirical analysis. *Journal of Personality and Social Psychology, 74*, 955-966.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11,* 150-166.

Ashton, M. C., & Lee, K. (2008). The prediction of Honesty-Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality, 42*, 1216-1228.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Persoonel Psychology, 35*, 281-322.

Berry, C. M., Carpenter, N. C., & Barratt, C. L. (2012). Do other-reports of counterproductive work behavior provide an incremental contribution over self-reports? A meta-analytic comparison. *Journal of Applied Psychology, 97*, 613-636.

Berry, C. M., Ones, D. S., Sackett, P. R. (2007) Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology,* 92, 410–424.

Blackman, M. C. (2002). Personality Judgment and the Utility of the Unstructured Employment Interview. *Basic and Applied Social Psychology, 24,* 241-250.

Cann, A., Siegfried, W. D., & Pearce., L. (1981). Forced attention to specific applicant qualifications: Impact on physical attractiveness and sex of applicant biases. *Personnel Psychology, 34*, 65-75.

Cohen, T. R., Kim, Y., Jordan, K. P., & Panter, A. T. (2016). Guilt-proneness is a marker of integrity and employment suitability. *Personality and Individual Differences, 92*, 109-112.

Cohen, T. R., & Morse, L. (2014). Moral character: What it is and what it does. *Research in Organizational Behavior, 34*, 43-61.

Cohen, T. R., Kim, Y., & Panter, A. T. (2015). The five-item guilt proneness scale (GP-5).

Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2013). Agreement and similarity in self-other perceptions of moral character. *Journal of Research in Personality, 47*, 816-830.

Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2014). Moral character in the workplace. *Journal of Personality and Social Psychology, 107,* 943-963.

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology, 100*, 947-966.

Connelly, B. S., & Hülsheger, U. R. (2012). A narrower scope or a clearer lens for personality? Examining sources of observers' advantages over self-reports for predicting performance. *Journal of personality, 80*, 603-631.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity.*Psychological Bulletin, 136,* 1092-1122.

Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16*, 137-163.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise?. *Decision, 1,* 79-101.

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: acquaintance and the big five. *Psychological bulletin, 116*, 245-258.

Kim, Y., & Cohen, T. R. (2015). Moral character and workplace deviance: Recent research and current trends. *Current Opinion in Psychology, 6,* 134-138.

Kim, Y., Cohen, T. R., & Panter, A. T. (2016). Cause or consequence? The reciprocal model of counterproductive work behavior and mistreatment. *Academy of Management Annual Meeting Best Paper Proceedings*.

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *Journal of Applied Psychology, 95*, 1-31.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of personality, 64*, 311-337.

Fast, L. A., & Funder, D. C. (2010). Personality in social psychology. *Handbook of social psychology*, 668-697.

Fleeson, W., Furr, R. M., Jayawickreme, E., Meindl, P., & Helzer, E. G. (2014). Character: The Prospects for a Personality-Based Perspective on Morality. *Social and Personality Psychology Compass, 8*, 178-191.

Fernandez-Duque, D., & Schwartz, B. (2015). Common Sense Beliefs about the Central Self, Moral Character, and the Brain. *Frontiers in Psychology, 6*.

Funder, D. C. (2012). Accurate personality judgment. Current Directions in *Psychological Science, 21*, 177-182.

Funder, D. C. (1995). On the accuracy of personality judgment: a realistic approach., *Psychological review*, 102, 652-670.

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: properties of persons, situations, and behaviors. Journal of personality and social psychology, 60, 773.

Gneezy, U. (2005). Deception: The role of consequences. American Economy Review. 95, 384–394.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of personality and social psychology, 106*, 148-168.

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*, 38-44.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of personality and social psychology, 74*, 1337.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B.(2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37, 504-528.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. John Wiley & Sons.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1*, 333-342.

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality, 57*, 72-88.

Hoevemeyer, V. A. (2005). *High-Impact Interview Questions: 701 Behavior-Based Questions to Find the Right Person for Every Job* (1 edition ed.). New York: AMACOM: American Management Association.

Hogan, R. (1996). A socioanalytic interpretation of the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality* (pp. 163–179). New York: Guilford Press.

Hogan, R., & Shelton, D. (1998). A socioanalytic perspective on job performance. *Human Performance, 11*, 129–144.

Hojbotă, A. M. (2015). Investigating implicit theories of moral character and moral evaluative judgments: Testing psychometric properties of a set of scaleson a Romanian population. *Psihologia Socială, 35*, 85-100

Huffcutt, A. I., Van Iddekinge, C. H., & Roth, P. L. (2011). Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance. *Human Resource Management Review, 21*, 353-367.

Human, L. J., & Biesanz, J. C. (2011). Through the looking glass clearly: accuracy and assumed similarity in well-adjusted individuals' first impressions. *Journal of personality and social psychology, 100*, 349-364.

Insko, C. A. (1981). Balance theory and phenomenology. *Cognitive responses in persuasion*, 309-338. New Jersey: Erlbaum.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102-138). New York: Guilford Press.

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (pp. 227–242). New York: Psychology Press.

Lee, K., & Ashton, M. C. (2012). *The H Factor of Personality: Why Some People are Manipulative, Self-Entitled, Materialistic, and Exploitive—And Why It Matters for* Everyone. Waterloo, Canada: Wilfrid Laurier University Press.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106.

Luft, J., & Ingham, H. (1955). *Proceedings of the Western Training Laboratory in Group Development*. The Johari windo, a graphic model of interpersonal awareness.

Mannes, A. E. (2009) Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science 55,*1267-1279.

Oh, I. S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: a meta-analysis. *Journal of Applied Psychology, 96*, 762-773.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the Prediction of Consequential Outcomes. *Annual Review of Psychology, 57,* 401-421.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. USA: Oxford University Press.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science, 2*, 313-345.

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., . . . Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology, 103,* 663-688.

Spector, P. E., Fox, S., Penney, L. M, Bruursema, K., Goh. A., Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? *Journal of Vocationala Behavior, 68*, 446–460.

Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences, 109*, 15197-15200

Stuewig, J., Tangney, J. P., Kendall, S., Folk, J. B., Meyer, C. R., & Dearing, R. L. (2015). Children's proneness to shame and guilt predict risky and illegal behaviors in young adulthood. *Child Psychiatry & Human Development,46*, 217-227.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of personality and social psychology, 98*, 281-300.

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: the accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of personality and social psychology, 95*, 1202.

Table 1. Study 1: Negative Binomial Regression of Targets' Number of Cheating on Online Judges' Average-Moral-Character-Judgments and Self-Reported Moral Character Traits

	Judge-report	Self-report	Judge- and Self-reports
	B (S.E.)	B (S.E.)	B (S.E.)
Mistake Questions			
Intercept	3.88(.82)***	3.58(1.30)**	4.38(1.25)***
Number correctly solved	-.10(.04)*	-.10(.04)*	-.09(.04)*
Average-Moral-Character-Rating	-.98(.25)***		-.95(.27)***
Honesty-Humility (Self-Report)		.04(.30)	.18(.28)
Conscientiousness (Self-Report)		-.64(.37)+	-.36(.34)
Guilt proneness (Self-Report)		-.15(.18)	.00(.18)
Dilemma Questions			
Intercept	3.87(1.11)***	3.24(1.15)**	4.23(1.30)**
Number correctly solved	-.14(.04)*	-.15(.04)***	-.15(.04)***
Average-Moral-Character-Rating	-.72(.32)***		-.61(.36)+
Honesty-Humility (Self-Report)		-.10(.33)	-.02(.33)
Conscientiousness (Self-Report)		-.41(.26)	-.25(.27)
Guilt proneness (Self-Report)		.03(.27)	.06(.27)

*** p<.001, ** p<.01, * p<.05, +p<.10

Table 2. Study 2: Negative Binomial Regression of Targets' frequency of CWB on Judges' Moral Character Rating and Self-Reported Moral Character Traits

	Judge-Report	Self-Report	Judge- and Self-Reports
	B(S.E.)	B(S.E.)	B(S.E.)
Mistake			
Intercept	5.15(.87)***	8.52(.95)***	8.82(1.06)***
Average-Moral-Character-Rating	-.98(.24)***		-.14(.22)
Honesty-Humility (Self-Report)		-.42(.20)*	-.42(.20)*
Conscientiousness (Self-Report)		-1.17(.20)***	-1.14(.20)***
Guilt Proneness (Self-Report)		-.19(.16)	-.18(.16)
Dilemma			
Intercept	4.22(.85)***	8.44(.96)***	8.76(1.08)***
Average-Moral-Character-Rating	-.67(.24)**		-.16(.24)
Honesty-Humility (Self-Report)		-.40(.22)+	-.39(.22)+
Conscientiousness (Self-Report)		-.90(.22)***	-.90(.22)***
Guilt Proneness (Self-Report)		-.45(.14)**	-.40(.16)*
Employer			
Intercept	3.82(.83)***	6.70(.85)***	7.46(.96)***
Average-Moral-Character-Rating	-.55(.24)*		-.44(.25)+
Honesty-Humility (Self-Report)		-.52(.20)**	-.46(.20)*
Conscientiousness (Self-Report)		-.45(.19)*	-.39(.19)*
Guilt Proneness (Self-Report)		-.34(.13)*	-.26(.14)+

*** p<.001, ** p<.01, * p<.05, +<.10

Table 3. Study 3: Levels of Consensus and Reliability of Moral Character Judgment

	Mistake Questions	Dilemma Questions
G study: Target Variance Percentage	41%	30%
D Study: Reliability Estimate Varying Judge Size		
2	0.60	0.50
3	0.70	0.60
4	0.75	0.67
5	0.79	0.72
6	0.82	0.75
7	0.84	0.78
8	0.86	0.80
9	0.87	0.82
10	0.88	0.83
11	0.89	0.85
12	0.90	0.86
13	0.91	0.87
14	0.91	0.88
15	0.92	0.88

Table 4. Study 3: Negative Binomial Regressions of Targets' Number of Cheating on Moral Character Judgements

Models	Mistake Questions		Dilemma Questions	
	Estimate (S.E.)	p-value	Estimate (S.E.)	p-value
Average-Moral-Character-Rating	-.93 (.24)	<.001	-.78 (.32)	.02
Judge 1	-.56 (.23)	.02	-.16 (.23)	.47
Judge 2	-.86 (.26)	<.01	-.46 (.35)	.18
Judge 3	-.60 (.15)	<.001	-.38 (.19)	.05
Judge 4	-.39 (.17)	.02	-.42 (.14)	<.01
Judge 5	-.68 (.18)	<.001	-.08 (.25)	.76
Judge 6	-.47 (.19)	.01	-.15 (.22)	.49

Table 5. Study 4: Levels of Consensus of Judges' Evaluations on Honesty-Humility (HH), Conscientiousness (C), Guilt Proneness (GP), and Agreeableness (A)

	Mistake				Dilemma				Employer			
	HH	C	GP	A	HH	C	GP	A	HH	C	GP	A
G study: Target Variance Percentage												
	16%	20%	27%	23%	31%	37%	37%	37%	10%	38%	30%	47%
D Study: Reliability Estimate Varying Judge Size												
2	.30	.37	.48	.38	.50	.55	.56	.55	.19	.64	.48	.65
3	.40	.47	.58	.48	.60	.65	.66	.65	.26	.73	.58	.74
4	.47	.54	.65	.55	.66	.71	.72	.71	.32	.78	.65	.79
5	.52	.59	.70	.60	.71	.76	.76	.76	.37	.82	.70	.83
6	.57	.64	.73	.64	.75	.79	.79	.79	.42	.84	.74	.85
7	.60	.67	.76	.68	.78	.81	.82	.81	.46	.86	.77	.87
8	.64	.70	.79	.71	.80	.83	.84	.83	.49	.88	.79	.88
9	.66	.72	.81	.73	.82	.85	.85	.85	.52	.89	.81	.89
10	.69	.74	.82	.75	.83	.86	.86	.86	.55	.90	.82	.90
11	.71	.76	.83	.77	.85	.87	.87	.87	.57	.91	.84	.91
12	.72	.78	.85	.78	.86	.88	.88	.88	.59	.91	.85	.92
13	.74	.79	.86	.80	.87	.89	.89	.89	.61	.92	.86	.92
14	.75	.80	.87	.81	.87	.90	.90	.90	.63	.93	.87	.93
15	.77	.81	.87	.82	.88	.90	.90	.90	.64	.93	.88	.93

Table 6. Study 4: Correlations of Self-Reports and Judge Reports

	Honesty-Humility	Conscientiousness	Guilt Proneness	Agreeableness
Mistake	.24*	.32**	.01	.16
Dilemmas	.16	.14	.30**	.05
Employer	-.12	-.16	.43**	.12

**:p<.01, *:p<.05

Table 7. Study 5: Negative Binomial Regressions of Targets' Number of Cheating on Self- and Judge-reported Honesty-Humility, Conscientiousness, and Guilt Proneness

	Mistake Questions			Dilemma Questions		
	B (S.E.)	B (S.E.)	B (S.E.)	B (S.E.)	B (S.E.)	B (S.E.)
Honesty-Humility Model						
Intercept	3.70(.74)***	1.90(.98)	3.33(.99)**	3.24(1.14)**	2.30(.94)*	3.50(1.27)**
Number correctly solved	-.10(.04)**	-.11(.04)**	-.10(.04)*	-.14(.04)***	-.14(.04)***	-.14(.04)***
Honesty-Humility (Judge)	-.92(.22)***		-.96(.23)***	-.53(.33)		-.48(.34)
Honesty-Humility (Self)		-.28(.26)	.14(.25)		-.24(.26)	-.12(.27)
Conscientiousness Model						
Intercept	3.88(.65)***	3.25(1.20)**	4.02(1.03)***	3.95(.77)***	3.09(.94)**	4.32(1.00)***
Number correctly solved	-.09(.04)*	-.09(.04)*	-.09(.04)*	-.14(.04)***	-.15(.04)***	-.15(.04)***
Conscientiousness (Judge)	-1.09(.21)***		-1.07(.23)**	-.81(.23)***		-.77(.24)**
Conscientiousness (Self)		-.68(.33)*	-.05(.30)		-.44(.25)+	-.14(.24)
Guilt Proneness Model						
Intercept	2.65(.72)***	1.72(.74)*	2.90(.89)**	3.76(.91)***	1.87(.87)*	3.40(1.04)**
Number correctly solved	-.11(.04)**	-.11(.04)*	-.12(.04)**	-.14(.04)***	-.13(.04)**	-.15(.04)***
Guilt proneness (Judge)	-.57(.21)**		-.54(.22)*	-.70(.26)**		-.77(.28)**
Guilt proneness (Self)		-.21(.17)	-.10(.17)		-.10(.22)	.16(.23)

Table 8. Study 6: Negative Binomial Regressions of Targets' Frequency of CWB on Judge Reports of Honesty-Humility, Conscientiousness, and Guilt Proneness

	Honesty-Humility	Conscientiousness	Guilt Proneness
Mistake	.31 (.51)	**-.23 (.47)**	-.02 (.37)
Dilemmas	.17 (.74)	-.41 (.76)	**-.76 (.69)**
Employer	1.28 (.58)*	**-1.11 (.62)+**	3.08 (1.56)*

Note. Mistake: N=26, Dilemma: N=29, Employer: N=32; +: <.10

Table 9. Study 6: Negative Binomial Regressions of Targets' Frequency of CWB on Judge Reports versus Self- or Coworker-Reports

	Model 1	Model 2	Model 3	Model 4	Model 5
Mistake					
Conscientiousness (Judge)	-.23 (.47)			-.22(.47)	-.35 (.51)
Conscientiousness (Self)		-.17(.53)		-.16(.52)	
Conscientiousness (Coworker)			.14 (.33)		.23 (.35)
Dilemma					
Guilt proneness (Judge)	-.76 (.69)			-.64(.74)	-.76 (.69)
Guilt proneness (Self)		-.65(.57)		-.54(.58)	
Guilt proneness (Coworker)			-.01 (.55)		-.01 (.55)
Employer					
Conscientiousness (Judge)	-1.11 (.62)+			-1.15 (.61)+	-1.11 (.61)
Conscientiousness (Self)		-.77 (.56)		-.73 (.49)	
Conscientiousness (Coworker)			-.35 (.48)		-.34 (.44)
Employer					
Guilt proneness (Judge)	3.07 (1.40)*			.58 (1.53)	
Guilt proneness (Self)		-.70 (.21) **		-.66 (.23)**	
Guilt proneness (Coworker)			-.63 (.42)		-.61 (.36)+

Note. Mistake: N=26, Dilemma: N=29, Employer: N=32; + : <.10 * : <.05, ** : <.01

Table 10. Study 7: Two-Sided Censored Regression of Targets' Average Frequency of Cheating and Lying on Judges' Average Rating, Positive Reporting Intention and Interaction of Judges' Rating and Positive Reporting Intention

	Model 1		Model 2		Model 3	
	B (SE)	p	B (SE)	p	B (SE)	p
Mistake						
Intercept	2.32 (.16)	.00	2.32 (.16)	.00	2.34 (.16)	.00
Conscientiousness (Judge)	-.36 (.17)	.03	-.35 (.16)	.03	-.31 (.17)	.06
Positively-Biased-Reporting (PBR)			.31 (.15)	.04	.19 (.17)	.12
PBR × Conscientiousness (Judge)					-.24 (.16)	.12
Dilemma						
Intercept	2.88 (.15)	.00	2.88 (.16)	.00	2.87 (.16)	.00
Guilt Proneness (Judge)	-.15 (.16)	.34	-.15 (.16)	.37	-.15 (.16)	.35
Positively-Biased-Reporting (PBR)			.17 (.15)	.26	.17 (.15)	.26
PBR × Guilt Proneness (Judge)					-.11 (.16)	.48
Employer						
Intercept	2.34 (.18)	.00	2.36 (.18)	.00	2.40 (.18)	.00
Conscientiousness (Judge)	-.11 (.18)	.55	-.13 (.18)	.48	-.17(.18)	.35
Positively-Biased-Reporting (PBR)			.24 (.17)	.16	.36 (.18)	.05
PBR × Conscientiousness (Judge)					-.38 (.20)	.06
Employer						
Intercept	2.34(.17)	.00	2.34(.17)	.00	2.35(.17)	.00
Guilt Proneness (Judge)	.15(.22)	.51	.22(.17)	.19	.08(.22)	.71
Positively-Biased-Reporting (PBR)			.13(.22)	.56	.16(.17)	.37
PBR × Guilt Proneness (Judge)					-.40(.22)	.07

Table 11. Study 7: Two-Sided Censored Regression of Targets' Average Levels of Cheating and Lying on Self- and Judge Reports for Targets whose positive report intention reports were in top 25%

	Model 1		Model 2		Model 3	
	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value
Mistake						
Intercept	3.18 (.31)	.00	2.83 (.34)	.00	2.89 (.36)	.00
Conscientiousness (Self)	-.41 (.26)	.12			-.15 (.31)	.62
Conscientiousness (Judge)			-.62 (.30)	.04	-.52 (.37)	.16
Dilemma						
Intercept	3.00 (.34)	.00	2.96 (.33)	.00	2.96 (.33)	.00
Guilt Proneness (Self)	-.33 (.35)	.36			-.05 (.37)	.45
Guilt Proneness (Judge)			-.81 (.36)	.03	-.79 (.39)	.04
Employer						
Intercept	2.25 (.44)	.00	2.58 (.40)	.00	2.37 (.46)	.00
Conscientiousness (Self)	.28 (.46)	.54			.45 (.49)	.35
Conscientiousness (Judge)			-.36 (.45)	.43	-.50 (.48)	.29
Employer						
Intercept	2.45 (.46)	.00	2.36 (.34)	.00	2.39 (.45)	.00
Guilt Proneness (Self)	-.04 (.45)	.47			-.05 (.44)	.45
Guilt Proneness (Judge)			-.72 (.51)	.08	-.72 (.51)	.08

Note. Mistake: Target N=42; Dilemma: Target N=50; Employer: N=47.

Table 12. Study 7: Two-Sided Censored Regression of Targets' Average Frequency of Cheating and Lying on Moral Character Dimension versus Smaller Dimensions

	Model 1		Model 2		Model 3	
	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value
Mistake						
Intercept	2.46(.16)	.00	2.32(.16)	.00	2.33(.16)	.00
Moral Character Rating	-.07(.15)	.66			.29(.20)	.15
Conscientiousness Rating			-.36(.17)	.03	-.59(.22)	.01
Dilemma						
Intercept	2.87(.16)	.00	2.88(.16)	.00	2.88(.16)	.00
Moral Character Rating	-.12(.16)	.44			-.07(.17)	.69
Guilt Proneness Rating			-.15(.16)	.34	-.12(.18)	.50
Employer						
Intercept	2.31(.17)	.00	2.35(.18)	.00	2.35(.18)	.00
Moral Character Rating	.07(.19)	.73			.24(.25)	.35
Conscientiousness Rating			-.11(.18)	.56	-.25(.24)	.29
Employer						
Intercept	2.31(.17)	.00	2.34(.17)	.00	2.34(.18)	.00
Moral Character Rating	.07(.19)	.73			.01(.22)	.96
Guilt Proneness Rating			.15(.22)	.51	.14(.25)	.57

Table 13. Study 9. Correlations among Judges' Evaluations on Targets' Moral character, Honesty-Humility, Conscientiousness, Guilt Proneness, and Four Elements of Honesty-Humility

	Achievement Linguistic Cues	MC	HH	C	GP	Fairness	Sincerity	Modesty
Moral Character (MC)	.09*							
Honesty-Humility (HH)	.04	.65***						
Conscientiousness (C)	.13**	.68***	.57***					
Guilt Proneness (GP)	-.13**	.48***	.49***	.35***				
Fairness	-.02	.69***	.61***	.65***	.49***			
Sincerity	-.03	.55***	.57***	.52***	.51***	.59***		
Modesty	-.03	.49***	.61***	.37***	.50***	.49***	.50***	
Greed	.23***	.25***	.09*	.38***	.00	.23***	.15***	-.11**

***: p<.001, **: p<.01, *: p<.05

Table 14. Study 8 (Mistake Questions): Text Cue Predictions of Targets' Unethical Behavior on Judges' Moral Character Judgment

	Model 1		Model 2		Model 3		Model 4	
	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value
Prediction of Unethical Behavior								
Sample A								
Number Correctly Solved	-.13 (.04)	.00	-.10 (.04)	.02	-.09 (.04)	.03	-.13 (.04)	.00
Anger	.28 (.11)	.01					.25 (.11)	.02
Anxiousness			-.02 (.12)	.45			.06 (.12)	.60
Sadness					.23 (.13)	.07	.13(.13)	.34
Sample B								
Anger	.02 (.17)	.93					.17 (.18)	.35
Anxiousness			.15 (.11)	.16			.23 (.11)	.03
Sadness					.28 (.14)	.04	.35 (.14)	.01
Sample C								
PBR	.28 (.15)	.06	.31 (.15)	.04	.32 (.15)	.03	.27 (.15)	.07
Anger	.39 (.17)	.02					.41 (.17)	.01
PBR× Anger	-.21 (.16)	.18					-.24 (.16)	.13
Anxiousness			.21 (.15)	.15			.25 (.15)	.09
PBR× Anxiousness			.12 (.15)	.41			-.10 (.15)	.49
Sadness					.14 (.15)	.35	.18 (.15)	.24
PBR× Sadness					.01 (.14)	.97	.01 (.14)	.94
Prediction of Moral Character Judgment								
Sample A								
Anger	-.15 (.04)	.00					-.15 (.05)	.00
Anxiousness			.08 (.05)	.10			.08 (.05)	<.10
Sadness					-.02 (.05)	.70	.04 (.05)	.47
Sample B								
Anger	-.09 (.05)	.06					-.09 (.05)	.07
Anxiousness			.04 (.03)	.26			.03 (.03)	.16
Sadness					-.02 (.04)	.70	-.02 (.04)	.34
Sample C								
Anger	-.09 (.04)	.03					-.10 (.04)	.02
Anxiousness			-.01 (.04)	.70			-.02 (.04)	.64
Sadness					-.03 (.05)	.56	-.04 (.05)	.41

Table 15. Study 8 (Dilemma Questions): Text Cue Predictions of Targets' Unethical Behavior and Moral Character Rating

	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value
Prediction of Unethical Behavior								
Sample A								
Number Correctly Solved	-.14 (.04)	.00	-.15 (.04)	.00	-.16 (.04)	.00	-.17 (.04)	.00
First Person Pronoun Ratio	.11(1.00)	.90					.35(1.19)	.77
Third Person Pronoun Ratio	.67 (1.20)	.58					.04(1.24)	.98
Social Words			.03 (.03)	.26			.02 (.04)	.76
Prosocial Dictionary					.10 (.06)	.08	.10 (.08)	.18
Sample B								
First Person Pronoun Ratio	1.09 (.68)	.11					-.07 (.77)	.93
Third Person Pronoun Ratio	1.96 (.76)	.01					1.79 (.81)	.03
Social Words			-.07 (.02)	.01			-.09 (.03)	.00
Prosocial Dictionary					-.12 (.08)	.15	-.06 (.07)	.43
Sample C								
PBR	.18 (.15)	.23	.17 (.15)	.28	.19 (.15)	.23	.19 (.15)	.21
First Person Pronoun Ratio	-.30 (.19)	.12					-.34 (.21)	.11
PBR× First.	-.44 (.24)	.06					-.47 (.26)	.07
Third Person Pronoun Ratio	-.65 (.19)	.00					-.63 (.21)	.00
PBR× Third.	-.63 (.24)	.01					-.62 (.24)	.01
Social Words			-.21 (.16)	.18			-.03 (.22)	.87
PBR× Social Words			-.06 (.15)	.69			.00 (.20)	.99
Prosocial Dictionary					-.16 (.16)	.29	-.10 (.16)	.52
PBR× Prosocial Dic.					-.17 (.15)	.25	-.12 (.16)	.45
Prediction of Moral Character Judgment								
Sample A								
First Person Pronoun Ratio	-.24 (.29)	.41					-.07 (.37)	.85
Third Person Pronoun Ratio	.51 (.34)	.14					.51 (.35)	.15
Social Words			.02 (.01)	.00			.01 (.01)	.48
Prosocial Dictionary					.02 (.02)	.33	.00 (.02)	.98
Sample B								
First Person Pronoun Ratio	-.24 (.19)	.21					-.14 (.20)	.48
Third Person Pronoun Ratio	.25 (.24)	.30					.18 (.20)	.47
Social Words			.02 (.01)	.00			.01 (.01)	.19
Prosocial Dictionary					.02 (.02)	.29	.01 (.02)	.82
Sample C								
First Person Pronoun Ratio	-.45 (.20)	.03					-.51 (.23)	.03
Third Person Pronoun Ratio	-.03 (.25)	.92					.04 (.27)	.88
Social Words			.01 (.01)	.20			-.01 (.01)	.44
Prosocial Dictionary					.03 (.02)	.25	.02 (.03)	.34

*** p < .001, ** p < .01, * p < .05, +: p ≤.10

Table 16. Study 8 (Employer Question): Text Cue Predictions of Targets' Unethical Behavior and Moral Character Judgment

	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value
Prediction of Unethical Behavior								
Sample B								
Affiliation	-.16 (.04)	.00					-.15 (.05)	.00
Achievement			-.07 (.03)	.01			-.04 (.03)	.14
Prosocial Dictionary					-.08 (.04)	.04	.00 (.04)	.98
Sample C								
IIM	.23 (.17)	.18	.27 (.17)	.11	.24 (.17)	.08	.30 (.17)	.08
Affiliation	-.05 (.17)	.76					.06 (.19)	.76
PBR× Affiliation	.00 (.16)	.98					-.04 (.17)	.79
Achievement			.13 (.17)	.46			.10 (.17)	.56
PBR× Achievement			.39 (.18)	.03			.44 (.19)	.02
Prosocial Dictionary					-.26 (.17)	.14	-.31 (.19)	.11
PBR× Prosocial. Dic.					.04 (.19)	.85	.16 (.20)	.42
Prediction of Moral Character Judgment								
Sample B								
Affiliation	.02 (.01)	.05					.00 (.07)	.95
Achievement			.01 (.01)	.39			.01 (.01)	.44
Prosocial Dictionary					.04 (.01)	.00	.04 (.01)	.00
Sample C								
Affiliation	.05 (.01)	.00					.04 (.01)	.00
Achievement			-.01 (.01)	.39			-.01 (.01)	.42
Prosocial Dictionary					.04 (.01)	.00	.02 (.01)	.06

Table 17. Summary of Studies (modified)

	Target Sample	Judge Sample	Interview Questions	Criterion
Study 1	195 U.S. adults recruited from the streets using a mobile research laboratory, Datatruck	152 undergraduates recruited from a university-administered subject pool	Mistake, Dilemma	Frequency of cheating measured in a laboratory experiment
Study 2	495 U.S. full-time employees recruited by an online survey firm (Qualtrics panel)	409 U.S. adults recruited via Mturk	Mistake, Dilemma, Employer	Frequency of workplace deviance reported by targets
Study 3	Study 1 targets	Six undergraduate research assistants	Mistake, Dilemma	Frequency of cheating measured in a laboratory experiment
Study 4	406 U.S. adults recruited from Mturk	Five undergraduate research assistants	Mistake, Dilemma, Employer	Inter-judge agreements self-other agreement
Study 5	174 full-time employees recruited from the online participant pool maintained by a university research center	Six undergraduate research assistants	Mistake, Dilemma, Employer	Frequency of workplace deviance reported by targets' coworkers
Study 6	Study 1 targets	500 U.S. adults recruited via Mturk	Mistake, Dilemma	Frequency of cheating measured in a laboratory experiment
Study 7	606 U.S. adults recruited from Mturk	2,390 U.S. adults recruited via Mturk	Mistake, Dilemma, Employer	Frequency of cheating and lying measured in an online experiment
Study 8	Target data sets from studies 1, 2, and 7	Judge data sets from studies 1, 2, 3, 6 and 7	Mistake, Dilemma, Employer	

Note. Total non-over-lapping target N = 1,876, Total judge N = 3,555

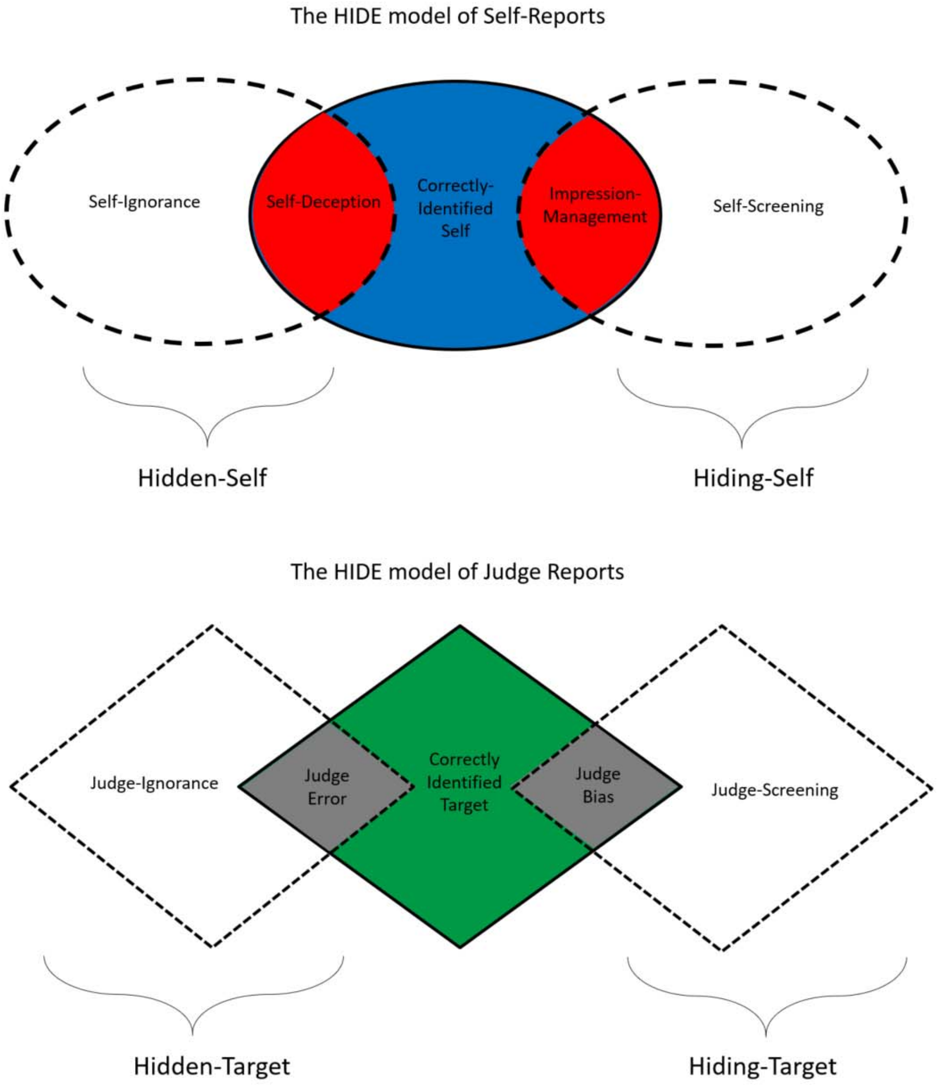
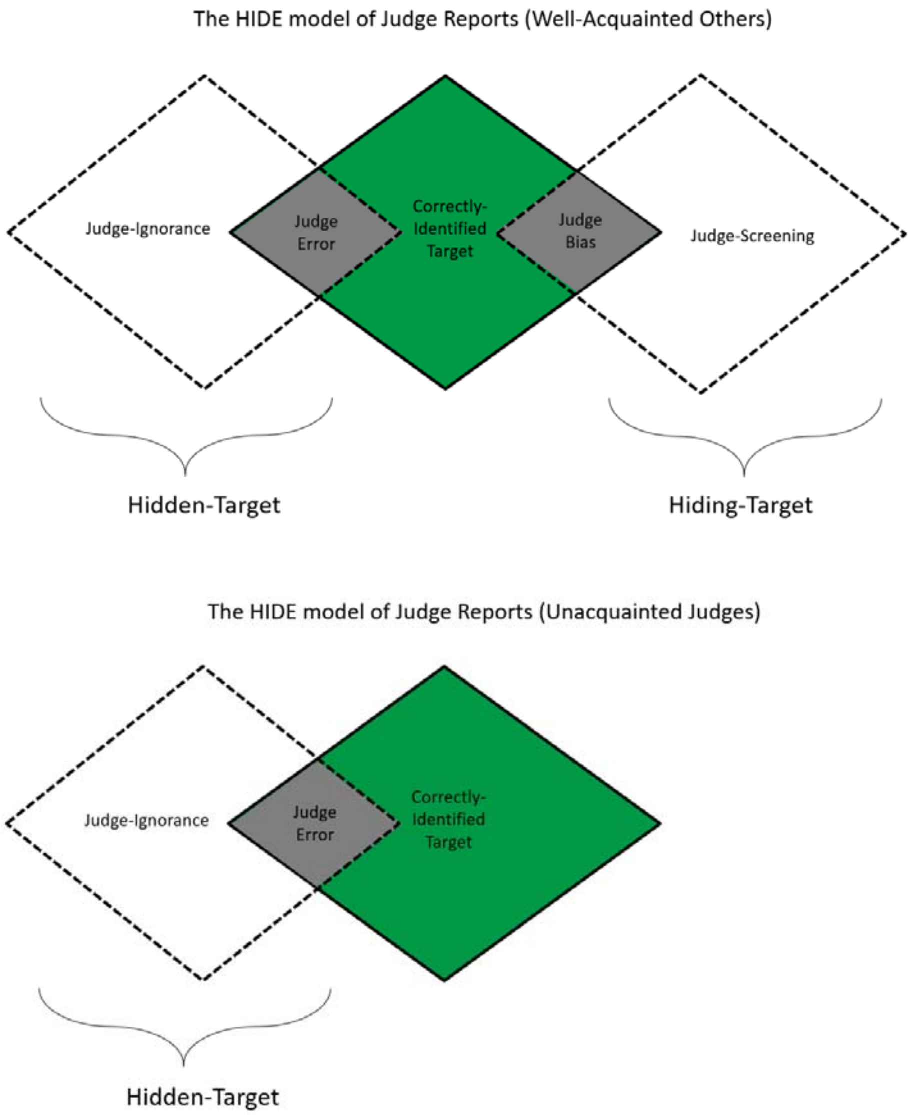


Figure 1. The Hidden Information Distribution and Evaluation (HIDE) Model of Person Perception and Reporting Decision



The HIDE model of Judge Reports (Well-Acquainted Others)

The HIDE model of Judge Reports (Unacquainted Judges)

Figure 2. The HIDE Model of Judge Reports for Well-Acquainted Others and Unacquainted Judges

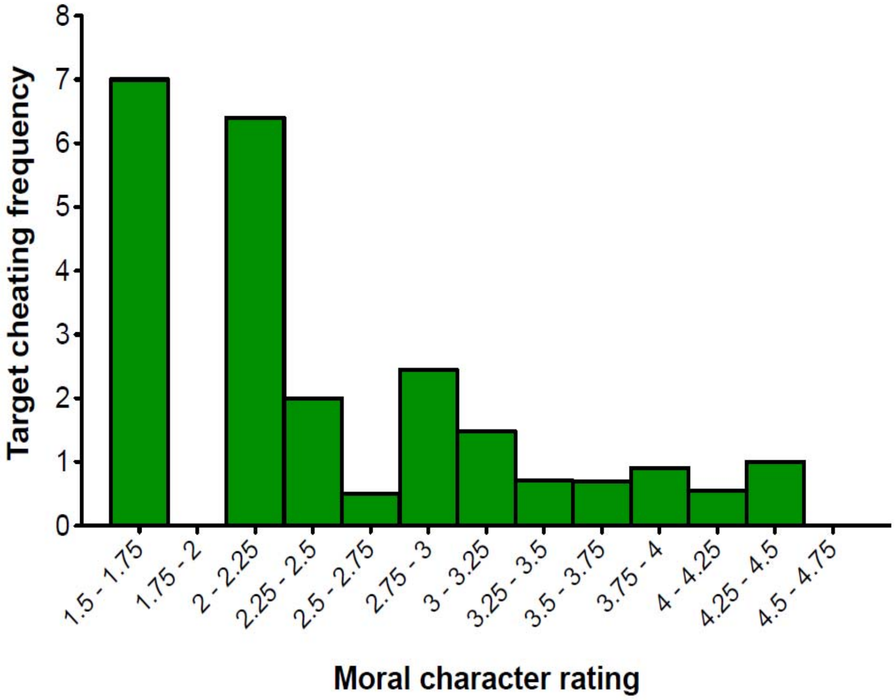


Figure 3. Study 1: Cheating in the Problem-Solving Task as a function of Judges' Ratings of Moral Character for Mistake Question

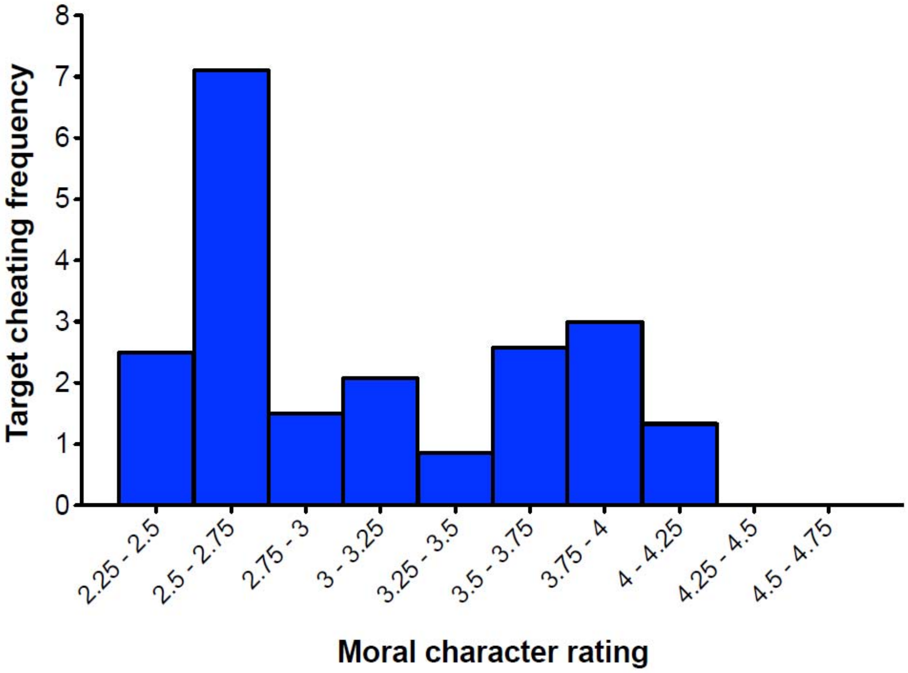


Figure 4. Study 1: Cheating in the Problem-Solving Task as a function of Judges' Ratings of Moral Character for Dilemma Questions

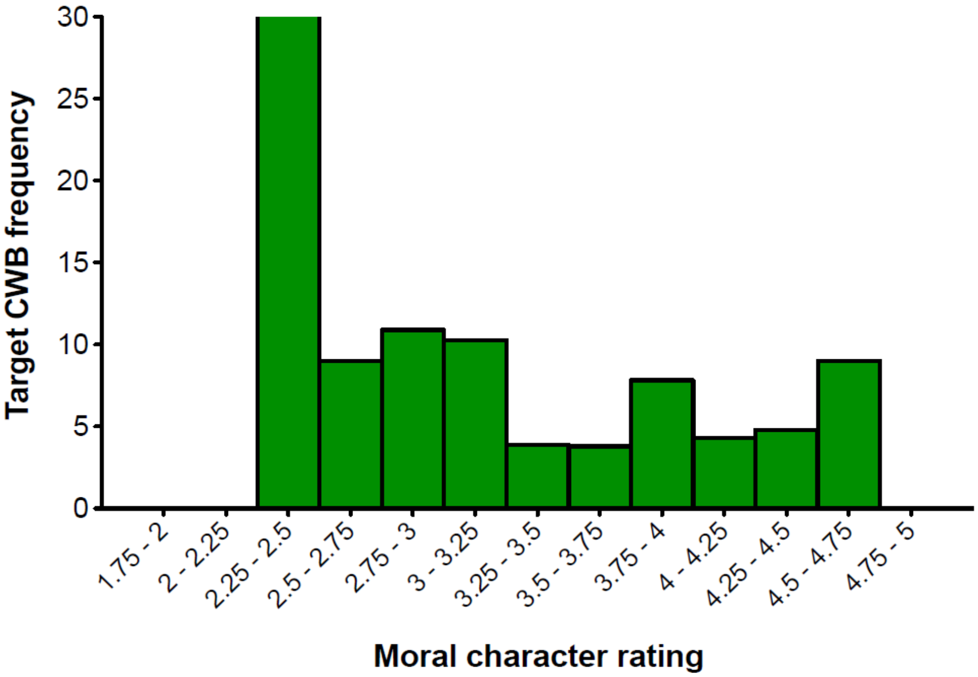


Figure 5. Study 2: Frequency of CWB as a function of Judges' Ratings of Moral Character for Mistake Questions

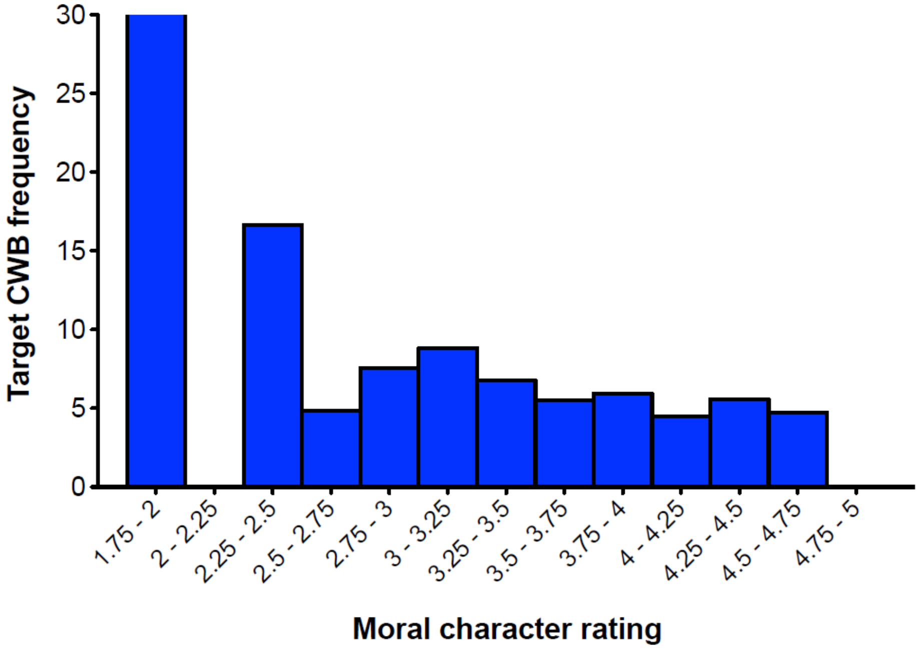


Figure 6. Study 2: Frequency of CWB as a function of Judges' Ratings of Moral Character for Dilemma Questions

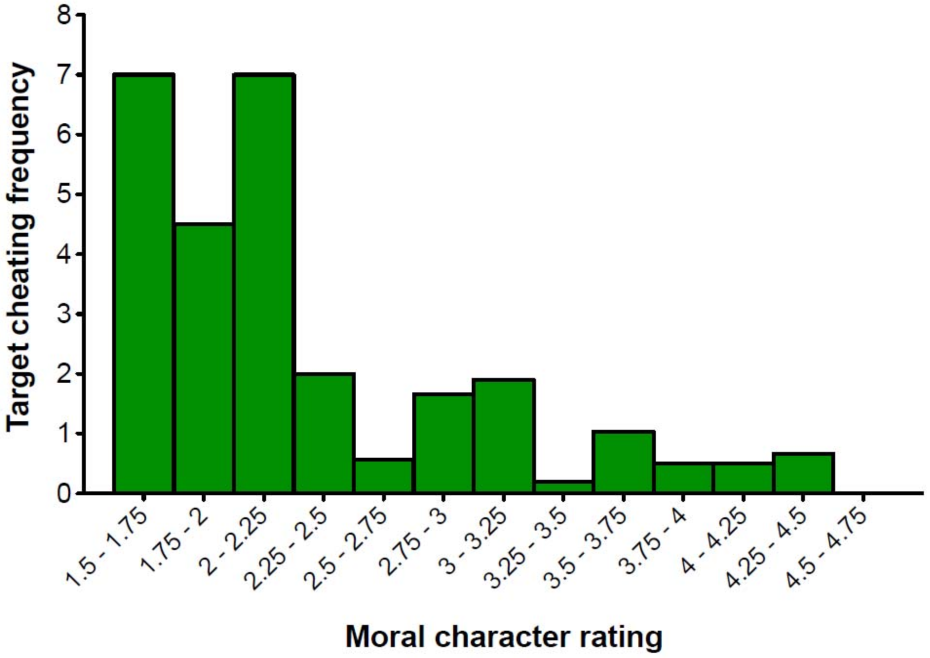


Figure 7. Study 3: Cheating in the Problem-Solving Task as a function of Judges' Ratings of Moral Character for Mistake Questions

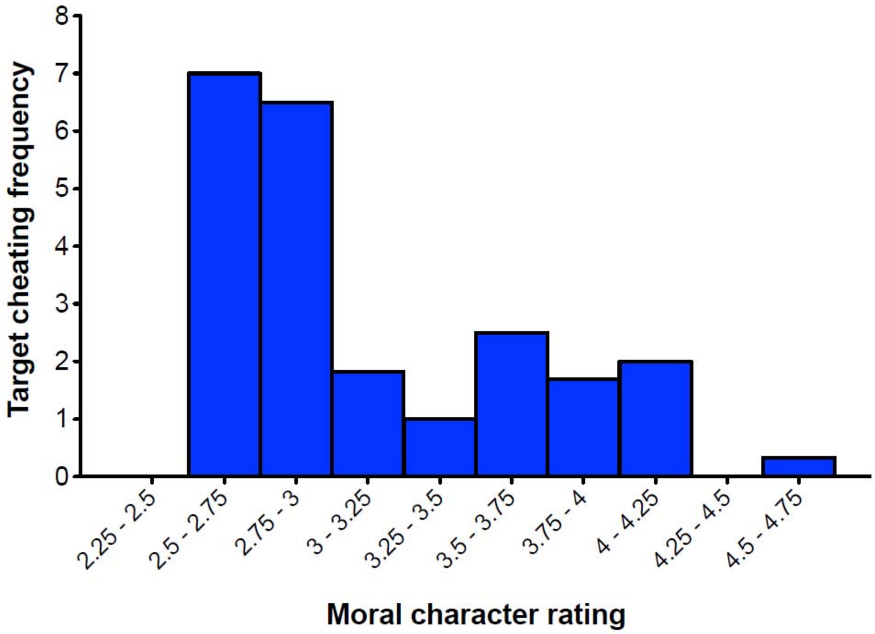


Figure 8. Study 3: Cheating in the Problem-Solving Task as a function of Judges' Ratings of Moral Character for Dilemma Questions

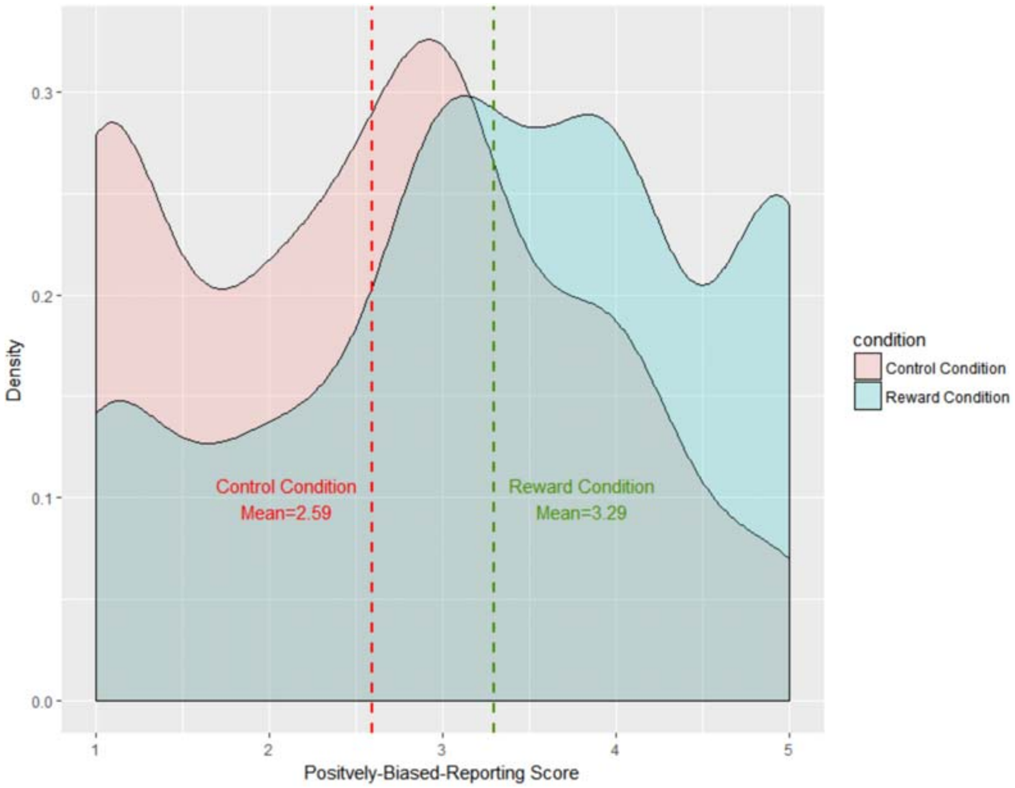


Figure 9. Study 7: Distributions of the Positively-Biased-Reporting Score for the Control and Reward Conditions

Appendix 1. Interview Questions

1. [Mistake] Please tell us about a time when you made a mistake at work. How did you feel when this occurred? What did you do? What, if anything, did you learn from this experience?

2. [Dilemma] Please describe an experience in which you were faced with a difficult dilemma at your job--a situation where you found it hard to decide what to do. What factors did you consider? What did you do? What, if anything, did you learn from this experience?

3. [Employer] How would your current or last employer describe you?

Appendix 2. Instructions for the Number Task in Study 7

The Number Task Instructions

In the Number Task there are SENDERS and RECEIVERS. You will be assigned to one of these roles.

In total, you will complete the Number Task five times with five different partners, and you will remain in the same role for all five iterations of the task.

You will be asked to make a series of decisions and your decisions will be matched with the decisions of other participants in this experiment to determine your payoffs. Neither you nor your partners will know the identity of one another.

The amount you earn will depend on the decisions you and your partners make. For each decision, you can earn a $0.25 bonus payment (for a maximum bonus payment of $1.25 for all five decisions).

Please press "Next" to be assigned to the role of "SENDER" or "RECEIVER."

SENDER INSTRUCTIONS

You have been assigned to the role of SENDER. You are paired with an anonymous RECEIVER.

We will use a random number generator to select a whole number between 1 and 5. You (the SENDER) will learn what number was selected. The RECEIVER will not learn what number was selected.

You will then choose one of the five following messages to send to the RECEIVER:

Message 1: “The number picked was 1.”
Message 2: “The number picked was 2.”
Message 3: “The number picked was 3.”
Message 4: “The number picked was 4.”
Message 5: “The number picked was 5.”

After you send the RECEIVER a message, the RECEIVER will choose a number between 1 and 5.

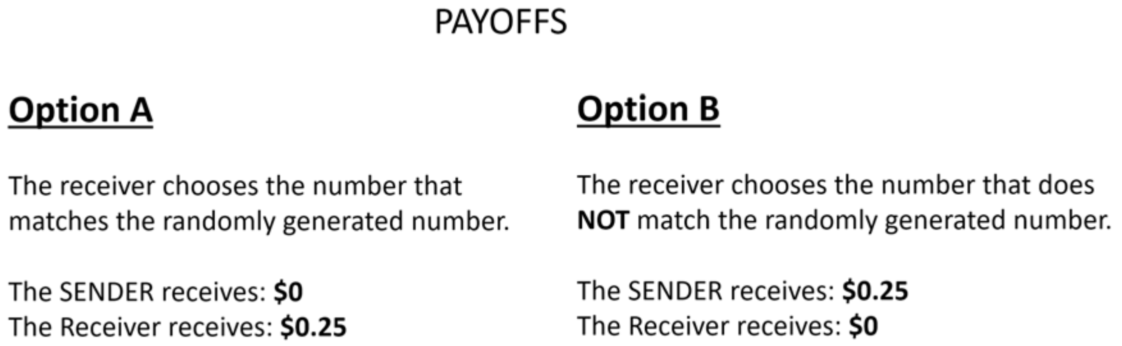
The only information the RECEIVER will have is the message the SENDER sends.

The choice the RECEIVER makes will determine the bonus payments that both the RECEIVER and SENDER earn.

Here is how the payment system works:

If the RECEIVER chooses the number that matches the randomly generated number, the payouts are given according to Option **A**

If the RECEIVER chooses a number that does **not** match the randomly generated number, the payouts are given according to Option **B**



PAYOFFS

Option A	**Option B**
The receiver chooses the number that matches the randomly generated number.	The receiver chooses the number that does **NOT** match the randomly generated number.
The SENDER receives: **$0** The Receiver receives: **$0.25**	The SENDER receives: **$0.25** The Receiver receives: **$0**

Appendix 3. Instructions for the Problem-Solving Task in Study 7

The Problem-Solving Task

In this part of the study, you will complete a problem-solving task under time pressure. To encourage you to do well on this task, you will be rewarded for good performance.

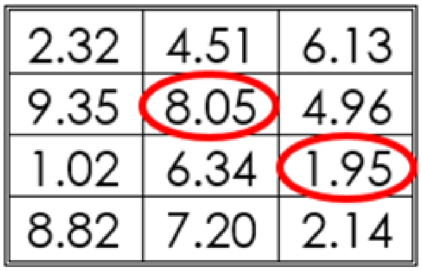
Your earnings will be based on your performance. The better you do on the task, the more money you will earn.

The Problem-Solving Task Instructions

You will be shown matrices, consisting of 12, three-digit numbers. In each matrix, there are two numbers that add up to 10 (for example, 8.05 and 1.95 in the example matrix below).

You will have 7 seconds to find the two numbers in each matrix that add up to 10.

Example Matrix



2.32	4.51	6.13
9.35	8.05	4.96
1.02	6.34	1.95
8.82	7.20	2.14

Once the 7 seconds are over, you will indicate whether you identified a pair that adds up to 10.

In total, you will be given 5 matrices to solve.

For each correctly solved matrix, you will receive $0.25, for a maximum possible bonus payment of $1.25 (for all 5 matrices).

Press "Next" when you are ready to start The Problem-Solving Task.

Appendix 3. Complementary Tables

Study 1

Table S1A. Study 1: Descriptive Statistics of Self-Reported Traits and Correlation with Targets' Cheating

	N	Min	Max	Mean	SD	Correlation with Judge-reported Moral Character	Correlation with Cheating
Cheating	195	0.00	16	1.87	3.06		
Honesty-Humility	195	1.50	5.00	3.40	.63	.15*	-.05
Conscientiousness	195	2.10	5.00	3.60	.55	.26*	-.17*
Guilt Proneness	195	1.00	5.00	3.91	.79	.23*	-.10

* p<.05

Table S1B. Study 1: Descriptive Statistics of Online Judges' Average-Moral-Character-Rating and Correlations with Targets' Cheating

	Total Target N	Total Judge N	Average Judge N	Word Count Mean	Word Count SD	Min	Max	Mean	SD	Correlation with Cheating
Mistake	99	76	15.35	67.71	37.40	1.62	4.33	3.24	.57	-.41***
Dilemma	96	76	15.83	84.98	55.33	2.33	4.63	3.37	.48	-.24*
Total	152	152	15.59	76.21	47.53	1.62	4.63	3.31	.53	-.28***

*** p<.001, ** p<.01, * p<.05

Study 2

Table S2A. Study 2: Descriptive Statistics of Self-Reported Traits and Correlation with Targets' Counterproductive Work Behaviors

	N	Min	Max	Mean	SD	Correlation with Judge-reported Moral Character	Correlation with CWB
CWB	495	.00	114	6.54	13.98		
Honesty-Humility	495	1.60	5.00	3.64	.59	.19***	-.26***
Conscientiousness	495	2.20	5.00	3.92	.56	.07	-.30***
Guilt Proneness	495	1.00	5.00	4.31	.77	.24***	-.21***

*** p<.01

Table S2B. Study 2: Descriptive Statistics of Online Judges' Average-Moral-Character-Rating and Correlations with Targets' CWB

	Total Target N	Average Judge N	Word Count Mean	Word Count SD	Min	Max	Mean	SD	Correlation with CWB
Mistake	159	16.79	68.29	40.69	2.27	4.59	3.49	.46	-.18*
Dilemma	168	15.78	75.18	43.89	1.94	4.76	3.48	.51	-.16*
Employer	168	16.20	50.20	20.81	1.79	4.43	3.58	.42	-.23*

** p<.01, * p<.05, +<.10

Study 3

Table S3A. Study 3: Descriptive Statistics of Six Judges' Average-Moral-Character-Rating and Correlations with Targets' Cheating

	Min	Max	Mean	SD	Correlation with Cheating
Mistake	1.50	4.33	3.25	.58	-.41**
Dilemma	2.00	4.67	3.45	.49	-.28**

*** p<.001, ** p<.01, * p<.05

Table S3B. Study 3: Correlations between Judge's Moral Character Judgments and Self-Reported Honesty-Humility, Conscientiousness, and Guilt Proneness

	Cheating	Average-Moral-Character-Rating	HH-Self	C-Self
Average-Moral-Character-Rating	-.29**			
Self-reported Honesty-Humility (HH-Self)	-.05	.19**		
Self-reported Conscientiousness (C-Self)	-.16*	.20**	.24**	
Self-reported Guilt proneness	-.10	.30**	.45**	.22**

** p<.01, * p<.05

Study 4

Table S4A. Study 4: Descriptive Statistics of Self-Reported Traits

	N	Min	Max	Mean	S.D.	H	C	GP
Honesty-Humility (H)	296	2.00	7.00	6.21	.95			
Conscientiousness (C)	296	3.00	7.00	6.00	1.08	.53***		
Guilt Proneness (GP)	296	1.00	5.00	4.25	.78	.47***	.30***	
Agreeableness	296	1.50	7.00	5.67	1.19	.49***	.41***	.40***

***: p < .001

Table S4B. Study 4: Target Size and Word Count Descriptive Statistics

	Total Target N	Total Judge N	Word Count Mean	Word Count SD
Mistake	96	5	68.49	41.77
Dilemma	96	5	79.00	51.48
Employer	44	5	34.66	19.07

Table S4C. Study 4: Descriptive Statistics of Judge-Rating-Average

	Min	Max	Mean	SD	H	C	GP
Mistake							
Honesty-Humility (H)	2.60	4.60	3.40	.39			
Conscientiousness (C)	2.20	4.80	3.23	.51	.56***		
Guilt Proneness (GP)	2.20	4.60	3.45	.52	.54***	.53***	
Agreeableness	1.80	4.20	3.25	.38	.57***	.59***	.49***
Dilemma							
Honesty-Humility	2.20	4.60	3.44	.51			
Conscientiousness (C)	1.80	4.60	3.51	.56	.70***		
Guilt Proneness (GP)	2.00	4.40	3.28	.52.	.77***	.73***	
Agreeableness	1.40	4.60	3.30	.52	.26*	.23**	.36**
Employer							
Honesty-Humility	2.00	4.00	3.16	.31			
Conscientiousness (C)	1.60	4.60	3.84	.52	.39*		
Guilt Proneness (GP)	1.60	4.00	3.13	.35	.63***	.70***	
Agreeableness	2.00	4.60	3.40	.58	.37*	.47**	.60***

*** p < .001, ** p < .01, * p < .05

Study 5

Table S5A. Study 5: Descriptive Statistics of Judge-Rating-Average

	Average-Judge N	Min	Max	Mean	SD	Correlation with Self-Reports	Correlation with Cheating
Mistake							
Honesty-Humility	15.56	1.40	4.29	3.25	.63	.23*	-.43***
Conscientiousness	20.81	1.07	4.42	3.06	.65	.29**	-.51***
Guilt Proneness	16.16	1.32	4.65	3.10	.68	.24*	-.27**
Dilemma							
Honesty-Humility	16.25	2.10	4.44	3.31	.46	.02	-.16
Conscientiousness	16.04	1.30	4.23	3.15	.64	.27*	-.42***
Guilt Proneness	15.83	1.71	4.50	3.32	.53	.30**	-.34***

**: p<.001, **: p<.01

Study 6

S6A. Study 6: Target Size and Word Count Descriptive Statistics

	Total Target N	Total Judge N	Word Count Mean	Word Count SD
Mistake	59	6	61.90	32.95
Dilemma	59	6	76.22	46.19
Employer	56	4	29.89	14.29

Table S6B. Studies 6: Descriptive Statistics of Self- and Peer-Reported Traits

	N	Min	Max	Mean	S.D.	H	C	GP
Self-Reports								
Honesty-Humility (H)	171	1.10	4.80	3.31	.64			
Conscientiousness (C)	172	2.00	5.00	3.72	.53	.17*		
Guilt Proneness (GP)	172	1.00	5.00	3.92	.87	.36***	.31***	
Agreeableness	171	1.30	4.60	3.20	.60	.25**	.05	.18*
Peer-Reports								
Honesty-Humility (HH)	87	1.70	4.70	3.37	.52			
Conscientiousness (C)	87	1.80	4.70	3.83	.55	.44***		
Guilt Proneness (GP)	87	1.40	5.00	4.19	.80	.25*	.08	
Agreeableness	87	1.40	4.70	3.33	.61	.47***	.11	.00

**: p<.001, *: p<.01

Table S6C. Study 6: Descriptive Statistics of Judges' Rating Average

	Min	Max	Mean	SD	H	C	GP
Mistake							
Honesty-Humility (H)	2.00	4.33	3.40	.40			
Conscientiousness (C)	2.00	4.17	3.23	.37	.39**		
Guilt Proneness (GP)	2.50	4.83	3.52	.53	.36**	.40**	
Agreeableness	2.67	4.00	3.21	.26	.67***	.38**	.49***
Dilemma							
Honesty-Humility	2.00	4.67	3.42	.49			
Conscientiousness (C)	2.17	4.50	3.56	.49	.61***		
Guilt Proneness (GP)	2.17	4.67	3.25	.46	.81***	.72***	
Agreeableness	2.17	4.50	3.37	.47	.56***	.57***	.68***
Employer							
Honesty-Humility	2.25	4.00	3.18	.36			
Conscientiousness (C)	2.25	4.50	3.71	.52	.15		
Guilt Proneness (GP)	2.75	3.50	3.10	.17	.49***	.24+	
Agreeableness	2.75	4.50	3.62	.40	.33*	.28*	.29*

*** p < .001, ** p < .01, * p < .05, +: p ≤.10

Table S6D. Study 6: Self-Judge Correlations and Judge-Peer Correlations Across Interview Question Conditions

	Self-Judge Correlations				Judge-Peer Correlations			
	HH	C	GP	A	HH	C	GP	A
Mistake	-.05	.20	.08	-.10	-.18	.28	.13	.15
Dilemmas	.25+	.05	.17	.21	.22	-.05	.48**	.20
Employer	.12	.31*	-.16	.01	-.30	.01	-.32+	-.23

**:p<.01, +: p ≤.10

Study 7

Table S7A. Study 7: Descriptive Statistics of Self-Reported Traits

	N	Mean	S.D.	Conscientiousness	Guilt Proneness
Mistake					
Honesty-Humility	201	3.47	.70		
Conscientiousness	201	3.82	.69	.51***	
Guilt Proneness	201	4.01	.85	.56***	.37***
Dilemma					
Honesty-Humility	195	3.57	.72		
Conscientiousness	195	3.97	.63	.43***	
Guilt Proneness	195	4.05	.79	.55***	.49***
Employer					
Honesty-Humility	210	3.67	.69		
Conscientiousness	210	4.02	.60	.49***	
Guilt Proneness	210	4.18	.78	.57***	.46***

***: p < .001, **: p < .01, *: p < .05

Table S7B. Study 7: Descriptive Statistics of Responses to Interview Questions

	Target N	Word Count Mean	Word Count SD	Correlation of Word Count and Positively Biased Reporting
Mistake				
Combined Sample	201	63.06	46.07	-.06
Reward Condition	104	86.62	42.55	-.04
Control Condition	97	104.27	68.32	-.27**
Dilemma				
Combined Sample	195	95.13	57.02	-.04
Reward Condition	99	142.36	77.54	-.04
Control Condition	96	113.51	62.82	-.13
Employer				
Combined Sample	210	128.16	71.96	.09
Reward Condition	108	57.12	46.71	.04
Control Condition	102	69.34	44.75	.06

**: p < .01

Table S7C. Studies 7: Descriptive Statistics of Judges' Average-Ratings across Evaluation Conditions

	Mistake			Dilemma			Employer		
	Average Judge N	Mean	SD	Average Judge N	Mean	SD	Average Judge N	Mean	SD
Moral Character	7.09	3.18	.55	11.03	3.42	.53	11.70	3.46	.47
Honesty-Humility	10.07	3.26	.58	9.87	3.39	.53	9.94	3.36	.45
Conscientiousness	10.70	2.93	.61	10.38	3.27	.68	8.31	3.43	.63
Guilt Proneness	10.57	3.14	.70	9.62	3.17	.54	9.38	3.02	.43
Fairness	9.08	3.26	.59	10.51	3.38	.63	9.69	3.33	.44
Sincerity	10.70	3.27	.62	9.87	3.38	.55	9.59	3.35	.44
Modesty	10.32	3.15	.44	10.90	3.15	.47	9.86	2.81	.52
Greed	10.95	3.16	.63	8.72	3.24	.50	9.11	3.05	.48

Table S7D. Study 8: Correlations of Self-Reports and Judge Reports across Interview Question Conditions

	Mistake	Dilemma	Employer
Honesty-Humility	.24**	.23**	.08
Conscientiousness	.40***	.36***	.28***
Guilt Proneness	.16*	.17*	.11
Fairness	.21**	.23**	.13+
Sincerity	.03	.12+	.13+
Modesty	.17*	.11	.19**
Greed Avoidance	-.00	-.06	.05

*** p < .001, ** p < .01, * p < .05, +: p ≤.10

Table S7E. Study 7: Two-Sided Censored Regression of Targets' Average Levels of Cheating and Lying on Judges' Average Rating and Word Count

	Mistake (Conscientiousness)		Dilemma (Guilt proneness)		Employer (Conscientiousness)		Employer (Guilt Proneness)	
	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value	B (SE)	p-value
Intercept	2.35 (.16)	.00	2.87 (.16)	.00	2.40 (.18)	.00	2.36(.17)	.00
Word Count	-.07 (.16)	.67	-.01 (.17)	.93	-.04 (.18)	.84	-.17(.24)	.49
Average Judge Rating	-.28 (.18)	.11	-.15 (.18)	.41	-.16 (.19)	.40	.17(24)	.49
Positively Biased Reporting (PBR)	.19 (.17)	.25	.17 (.15)	.26	.36 (.18)	.05	.16(.17)	.35
PBR × Average Judge Rating	-.24 (.16)	.13	-.11 (.16)	.48	-.37 (.21)	.07	-.40(.22)	.07

Table S7F. Study 8. Correlations of Judges' Rating of Targets' and Linguistic Cues of Achievement Across Interview Question Conditions

	Greed	Honesty-Humility	Fairness	Sincerity	Modesty
Mistake Questions	.11	-.01	.03	-.05	.10
Dilemma Questions	.12	.19***	.08	.05	.12
Employer Question	.04	-.05	-.17*	-.15*	-.01

***: p < .001, *: p < .05

Table S7G. Study 7. Correlations among Judges' Evaluations on Targets' Moral character, Honesty-Humility, Conscientiousness, Guilt Proneness, and Four Elements of Honesty-Humility

	Moral character	HH	C	GP	Fairness	Sincerity	Modesty
Mistake							
Honesty-Humility (HH)	.81						
Conscientiousness (C)	.73	.67					
Guilt Proneness (GP)	.67	.68	.47				
Fairness	.78	.73	.74	.62			
Sincerity	.72	.67	.67	.65	.72		
Modesty	.76	.74	.55	.74	.68	.65	
Greed	.23**	.26***	.37***	.13+	.28***	.29***	.17*
Dilemma							
Honesty-Humility (HH)	.77						
Conscientiousness (C)	.70	.69					
Guilt Proneness (GP)	.48	.41	.31				
Fairness	.78	.68	.67	.38			
Sincerity	.69	.67	.67	.49	.66		
Modesty	.62	.60	.59	.39	.51	.48	
Greed	.00	.04	.22**	-.13+	.14+	.12+	-.05 (n.s.)
Employer							
Honesty-Humility (HH)	.46						
Conscientiousness (C)	.65	.34					
Guilt Proneness (GP)	.50	.26	.44				
Fairness	.71	.31	.60	.43			
Sincerity	.21	.23	.17	.25	.25		
Modesty	.21	.45	.08	.13	.15	.26	
Greed	.35***	-.05	.43***	.13+	.32***	.01	-.34***

***: p<.001, **: p<.01, *: p<.05

Note. Correlations among HH, C, GP, Fairness, Sincerity and Modesty are all significant with α=.001.

Study 8

Table S8A. Study 8: Text Cue Correlations with Targets' Unethical Behavior and Judges' Moral Character Evaluation

	Unethical Behavior			Moral Character Judgments		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Mistake						
Pronouns						
First Person Pronoun Ratio	-.14	-.19*	.02	-.03	-.02	-.06
Third Person Pronoun Ratio	.34**	.04	-.08	.02	.05	.05
Negative Emotions						
Anger	**.35****	**.01**	**.14***	**-.33****	**-.15+**	**-.16***
Anxious	-.06*	.10	.11+	.17	.09	-.03
Sad	**.32****	**.20***	**.07**	**-.04**	**-.03**	**-.04**
Affiliation vs. Achievement						
Affiliation	-.05	.02	.01	.01	.14+	.03
Achievement	.17+	.08	.01	.03	-.06	.03
Social Words						
Prosocial Words	-.06	.05	-.10+	.18+	.12+	.11+
Social Words	.09	.11	.06	-.06	.04	-.03
Dilemma						
Pronouns						
First Person Pronoun Ratio	**-.08**	**-.08**	**.04**	**-.29****	**-.20****	**-.19****
Third Person Pronoun Ratio	**.10**	**.19***	**-.19****	**.31****	**.19***	**.11**
Negative Emotions						
Anger	-.07	.03	-.02	.18+	.10	.02
Anxious	-.12	.16*	.09	-.10	-.02	-.05
Sad	.07	.02	-.04	-.11	-.05	.02
Affiliation vs. Achievement						
Affiliation	-.08	-.07	.10+	.24*	.07	.05
Achievement	-.14	.08	.06	.16	-.11+	.06
Social Words						
Prosocial Words	.09	-.08	-.07	.10	.08	.08
Social Words	**-.04**	**-.16***	**-.08**	**.26****	**.22****	**.09+**
Employer						
Pronouns						
First Person Pronoun Ratio	.00	.03	.02	.11+	-.21**	.11
Third Person Pronoun Ratio	-.02	.00	-.08	.01	.18*	.01
Negative Emotions						
Anger	.01	-.02	.14*	-.18*	-.02	-.18**
Anxious	.04	-.07	.07	.06	.06	.01
Sad	-.06	.01	.11	.01	.01	.06
Affiliation vs. Achievement						
Affiliation	**-.02**	**-.19***	**.01**	**.28*****	**.15***	**.28****
Achievement	**.03**	**-.12+**	**.01**	**-.06**	**.07**	**-.06**
Social Words						
Prosocial Words	**-.10+**	**-.09**	**-.10**	**.24*****	**.31****	**.24****
Social Words	-.03	.02	.06	.11+	.09	.11

*** p < .001, ** p < .01, * p < .05, +: p ≤.10