Junior Management Science

www.jums.academy
ISSN: 2942-1861

Implicit Measurement of the Moral Self-Image Using the Go/No-Go Association Task (GNAT) - An Empirical Investigation of the Convergent Validity Between Explicit and Implicit Measures

Louisa Felicitas Bläßer

Technical University of Munich

Abstract

While people are increasingly aware of climate change, many still resist lifestyle changes. Research now focuses on understanding conscious (explicit) and unconscious (implicit) attitudes to encourage sustainable behavior. This thesis used the Go/No-Go Association Task (GNAT) to measure participants' implicit moral self-image and examine its correlation with an explicit moral self-image questionnaire, indicating convergent validity and effective application of the GNAT as an implicit measure of the moral self-image. After applying exclusion criteria, 68 participants were randomly assigned to two groups with different word lists. Results showed that repeated exposure to fewer words in group A led to little or no correlation, while group B, using more varied words, showed higher correlation and good convergent validity. This demonstrates that the GNAT effectively measures moral self-image when learning effects are avoided. The findings offer insights into implicit attitudes that influence decisions and yield practical implications for different stakeholders. This thesis contributes through its experimental design, adapted exclusion criteria, and sample correction of all perfect responses, validating the GNAT as an implicit measure and offering a foundation for future research.

Keywords: convergent validity; explicit measures; go/no-go association task (GNAT); implicit measures; moral self-image

1. Introduction

Have you ever ordered something online, used non-recyclable packaging, chosen a non-organic product, not separated food waste appropriately, or traveled by plane? The answer is likely yes, as we face many sustainable decisions daily. Unsustainable behavior is perceived as immoral, as people are increasingly aware of their impact on climate change (Sachdeva et al., 2015). But why do people engage in unsustainable and, consequently, immoral behaviors?

There has been an increasing focus on understanding the psychological drivers that motivate immoral behavior in the last decades, especially with the intensifying global climate crisis (Sachdeva et al., 2015). The ultimate goal is to use this knowledge to nudge people further into being more sustainable (Fischer et al., 2012; Sachdeva et al., 2015). Therefore, it is essential to understand why people behave immorally and how people's morality can be measured.

Traditional explicit measures often fall short of capturing the perception of people's moral selves. Social desirability biases influence the answers given in such self-report questionnaires (Crowne & Marlowe, 1960). In recent decades, various implicit measures have been developed to measure unconscious attitudes. The most famous method is the Implicit Association Test (IAT), which was further developed into the Go/No-Go Association Task (GNAT). Previous research has already implicitly assessed the moral self-image (perception

I would like to thank Konrad Kober, my Bachelor's thesis supervisor, for his dedicated support, continuous encouragement, and constructive feedback throughout the duration of this thesis. I am especially grateful for his valuable assistance and expertise in the experimental design. Additionally, I would like to thank Prof. Dr. Alwine Mohnen from the Chair of Corporate Management at the Technical University of Munich for serving as my examiner.

DOI: https://doi.org/10.5282/jums/v10i4pp966-984

of one's own morality) with the IAT, yielding promising results (Perugini & Leone, 2009). However, little research has been devoted to the GNAT, and it has only been used once to capture the moral self-image (Ferguson, 2018).

This bachelor's thesis uses the GNAT to measure the implicit moral self-image and to examine the convergent validity (the ability of two measures to capture a joint construct (Carlson & Herdman, 2012)) of this implicit measurement method by correlating it with an explicit moral self-image questionnaire. It aims to answer the following research question: "To what extent can the Go/No-Go Association Task (GNAT) be effectively applied to measure moral self-image, and is there a correlation between the outcomes of this method and the explicit moral self-image?"

The timing for this research is crucial, as it aligns with the growing interest in sustainable practices and the need to deepen the understanding of the internal, implicit motivations behind a behavior change (Mazar & Zhong, 2010; Sachdeva et al., 2015; Schlegelmilch & Simbrunner, 2019). Applying implicit psychological tests, such as the GNAT, to consumer behavior or marketing strategies opens up a new field of research. These measures improve the assessment of implicit attitudes toward products or brands because the implicit test procedures are based on less biased, unconscious answers and reactions and, therefore, are very valuable for subsequent analyses. An effective and validated tool to determine people's moral self-image could provide essential insights for different stakeholders to influence consumers toward more sustainable and moral choices.

This thesis is structured as follows:

Chapter 1 briefly introduces the topic and aim of this thesis. Chapter 2 reviews current research literature, presenting theories of moral behavior, important definitions, and measurement methods, leading into Chapter 3, which covers the methodology and detailed research design of the performed GNAT experiment. Chapters 4 and 5 present the compelling results and elaborate on the experiment's key findings, discussing its limitations, suggestions for future research, and implications for practitioners in management and policy. Chapter 6 provides a comprehensive summary of the main findings of the experiment, reflecting on the research's significance.

2. Theoretical Background

A wide range of theories are trying to explain why people behave immorally. The rational economic model expects people to behave immorally whenever their potential gain exceeds their expected punishment since it is the best choice economically (Becker, 1968). Following this reasoning, people should behave immorally every time they could potentially gain more than they would lose. In contrast, it can be observed that people intrinsically limit their immorality and avoid too much lying if it threatens their perception of their own morality (moral self-image) (Mazar et al., 2008; Sachdeva et al., 2009). It appears that the idea of an entirely rational person (e.g., homo economicus (Melé & Cantón, 2014)) does not apply to most people and situations. Instead, an internal force seems to restrict people from exploiting the potential benefits of cheating to its full extent (Cornelissen et al., 2013; Mazar et al., 2008).

It becomes evident that people face an internal conflict whenever they have an opportunity to cheat (Barkan et al., 2015; Mazar et al., 2008). This ethical dissonance is a state of tension that occurs when people are either tempted to benefit from their immoral behavior or to uphold a positive moral self-image, also known as moral self-concept[1] (Mazar et al., 2008). Festinger (1957) describes this state as cognitive dissonance and argues that its presence motivates people to subsequent action, which reduces this dissonance. People developed different strategies to engage in immoral behavior to resolve this internal conflict and distressing state, particularly without updating (and depressing) their moral self-image.

Erikson (1964) explains this motivating force as the intrinsic need for people to act according to their (moral) identity. Researchers interpret moral identity, defined as "*the use of moral values to define the self*" (Johnston et al., 2013, p. 209), as a moderator and motivational driver to act morally (Aquino & Reed, 2002; Blasi, 1993; Erikson, 1964).

The following section will introduce different theories that investigate why people behave immorally.

2.1. Moral Theories

2.1.1. Self-Concept Maintenance

Mazar et al. (2008) propose a theory of self-concept maintenance. They argue that people try to balance maintaining an honest self-concept and gaining from lying. According to Mazar et al. (2008), people would cheat to a certain extent as long as they do not need to update their moral self-concept (of being honest). This compromise allows them to benefit from cheating without negatively impacting their moral self-image. The authors also suggest that people use different techniques to decide on this motivational dilemma and to determine the degree to which cheating aligns with their moral self.

A powerful technique is, for instance, self-serving justification (Shalvi et al., 2015). It suggests that people would try to find reasons for questionable behavior to make it seem less immoral when their moral self-image is threatened. Shalvi et al. (2015) distinguish between pre-violation justification (before the immoral action) and post-violation justification (after the immoral action).

Pre-violation justification excuses the immoral action and thus reduces the threat to the moral self-concept beforehand (Shalvi et al., 2015). There are several strategies for this pre-violation justification. Examples are ambiguous actions, altruistic cheating, and moral licensing (Shalvi et al., 2015). Whenever the norms and rules for a situation are ambiguous, the actor could invent facts and reasons to justify his[2]

[1] These terms can be used interchangeably (Jordan et al., 2015)

[2] For better readability, only the pronouns "he/him/his" are used throughout this thesis

actions. If a lie does not cause harm to other people but instead would help to benefit the actor and other people, it is more likely to observe cheating (altruistic cheating) (Erat & Gneezy, 2012). Through moral licensing, people justify their bad behavior with their initial good actions (Merritt et al., 2010; Shalvi et al., 2015). In contrast, post-violation justification is a tool to justify immoral behavior after the action has already been conducted. This could be, for instance, through (partially) confessing or distancing themselves from their action by looking at others' immoral behavior (Shalvi et al., 2015).

People generally try to maintain or even enhance their moral self-image (Jordan et al., 2015; Mazar et al., 2008; Shalvi et al., 2015). They do this by behaving morally or biasing their cognitive perception with examples like self-serving justifications (Monin & Jordan, 2009). Monin and Jordan (2009) argue that people who value morality greatly pay more attention to their moral self-image, and deviations from their moral self-concept impact their self-worth more significantly compared to people with a lower importance on being moral. This moral self-image can also be influenced by previous and current situations. A deviation from their aspired level motivates people to take subsequent actions to reduce this dissonance (Cornelissen et al., 2013; Jordan et al., 2015). Monin and Jordan (2009) refer to that as behavior-generating power. To predict peoples' behavior, their individual moral self-image, which fluctuates and deviates over time (Jordan et al., 2015), must be considered.

2.1.2. Moral Balancing Model

There are two contrasting approaches when predicting peoples' moral actions after they have acted morally or immorally. Either the actor behaves consistently with his initial action, or the subsequent behavior is the opposite of his previous action (moral balancing).

Freedman and Fraser (1966) introduced the Foot-In-The-Door-Technique, which is nowadays widely used in negotiation strategies and a great example of consistent behavior. They elaborate that people who already agreed to do a small favor were more likely to agree to do a second, even larger favor.

An example of consistent moral behavior would be if a man returns a lost wallet to the owner after offering a seat in public transport to an elderly woman. The negative case, which still reflects consistent behavior, would be that the man does not offer his seat to the elderly woman and keeps the wallet as well.

Mullen and Monin (2016) argue that people show consistent behavior when they focus abstractly on values and their initial behavior. In contrast, people exhibit a balancing behavior when they think more concretely about their initial behavior and what they have accomplished with it. This alternating pattern is described as moral balancing. After a previous immoral action, the actor behaves morally in the subsequent action, or vice versa.

This moral balancing model was developed in 1990 by the psychologist Mordecai Nisan. It states that people consider previous behavior when making moral decisions. According to Nisan (1990), people try to balance their current moral self around a fixed personal moral standard (equilibrium). This personal reference point is essential for people as they constantly compare their current state with this self-set standard, which they want to maintain over time (Miller & Effron, 2010). Nisan (1990) assumes that when their moral status drops below a personal tolerable level, people will refrain from doing an immoral action. However, this satisfactory level of morality is lower than the ideal level, and those minimum requirements are determined mainly by a person's moral identity (Nisan, 1990).

Following this reasoning, a person who recently did something immoral would instead choose an altruistic action in order to compensate for the previously generated deficit in his own moral balance (moral cleansing). A person who is currently in moral surplus would be more likely to perform a subsequent selfish action (moral licensing) (Nisan, 1990). In other words, balancing happens when a moral initial behavior leads to the opposite in a subsequent behavior (Jordan et al., 2011; Mullen & Monin, 2016; Zhong et al., 2010).

When balancing a previous action, these two directions can be observed: moral cleansing and moral licensing.

Moral Cleansing

Moral cleansing (or moral compensation) happens when a previous immoral behavior causes a subsequent moral behavior (Mullen & Monin, 2016; Perkins et al., 2024). This can be explained by an analogy of a moral bank account, the moral credits model (Perkins et al., 2024). If a person's metaphorical moral bank account is in deficit, he wants to rebalance it with a subsequent moral behavior (Nisan, 1990). This could be done by performing a morally good action or refraining from immoral actions, such as cheating (Cornelissen et al., 2013). Continuing with the previous example, a man who did not offer his seat on the bus to an elderly woman would be more likely to return a lost wallet to its owner to compensate for this deficit in his moral balance. Researchers explain this effect through people's motivation and willingness to invest effort to repair their shortfalls (Jacobsen et al., 2018).

Additionally, there is strong evidence that people need to physically cleanse themselves after behaving immorally. Zhong and Liljenquist (2006) show that people who recall an immoral act would be more likely to choose antiseptic wipes compared to other products. They explain that the participants need to wash away their sins and cleanse themselves after their moral purity has been threatened.

Moral Licensing

The moral licensing effect describes the contrasting and somewhat counterintuitive observation: Good previous behavior leads to less positive or even bad behavior. In other words, people justify their bad behavior with their previous good action (Jacobsen et al., 2018; Merritt et al., 2010).

Moral licensing can be explained from two different perspectives: the moral credits model and the moral credentials model.

The moral credits model explains the licensing effect as people accumulate credits in their hypothetical moral bank account when they do something good. They can use these credits and "withdraw" them to justify subsequent negative behavior while maintaining an overall positive balance (Effron & Monin, 2010; Merritt et al., 2010; Miller & Effron, 2010). Moral licensing starts with a surplus in the moral bank account and withdraws credits to allow people to perform a negative action (Perkins et al., 2024).

The second explanation, the moral credentials model, explains the moral licensing effect with a different interpretation of the subsequent behavior. According to Monin and Miller (2001), people are less likely to interpret their subsequent behavior as immoral after they have performed an initial moral act. Instead of earning a right to perform this immoral act without punishment, the initial moral behavior has provided a lens through which the following behavior is interpreted differently (Mullen & Monin, 2016). This process is more likely when the subsequent behavior is ambiguous and can be interpreted positively (Mullen & Monin, 2016). For example, by recommending a woman for one job, people built positive credentials as being someone without prejudice and were more willing to express that a man was better suited for a second job (Monin & Miller, 2001). In this experiment by Monin and Miller (2001), the second behavior was ambiguous. It could be explained by illegitimate or legitimate motives (sexism or pragmatism). The credentials of not being sexist, e.g., established through actively recommending a woman for the first job, help to interpret the second action positively, e.g., favoring a man for the second job, due to the actor's history, without affecting the actor's moral self-image (Monin & Jordan, 2009; Monin & Miller, 2001).

Both models explain that a previous moral action can lead to immoral or questionable behavior later on. The key difference is that in the moral credits model, the actor is fully aware of the second action's immorality but decides to afford this decrease in his overall moral balance. In contrast, in the moral credentials model, the positive first action helps to disambiguate and interpret the second action differently (Monin & Jordan, 2009). To clarify this tension between the two models, Monin and Jordan (2009) suggest that the moral credits model is at work in unambiguous cases, where the meaning of the target behavior is clearly interpreted as immoral and unaffected by the previous action, while the moral credentials apply for ambiguous cases. However, both models predict the same behavior and support the importance of acknowledging a dynamic moral self-image. These models suggest that recent actions shape a person's moral self-image and influence his future moral behavior (Monin & Jordan, 2009).

Moral Self-Image in the Moral Balancing Model

The moral credits model describes a mechanism for people to balance their moral or immoral behavior with an accumulated or depleted moral bank account, reflecting the increase or decrease of the moral self-image, respectively (Merritt et al., 2010; Zhong et al., 2010). This enables people to repair their moral self-image by compensating for their selfish actions afterward (moral cleansing) (Perkins et al., 2024; Schlegelmilch & Simbrunner, 2019) or using their bolstered moral self-image (from a previous action) to perform a subsequent immoral act (moral licensing) (Cornelissen et al., 2013; Effron & Monin, 2010; Monin & Jordan, 2009; Nisan, 1990).

This emphasizes that the moral self-image plays a central role in moral decision-making. Its discrepancies from the actor's personal standard (equilibrium) motivate balancing behavior (Nisan, 1990). While the balancing could be observed, and there is empirical evidence (Cornelissen et al., 2013; Lee & Hsieh, 2013; Ploner & Regner, 2013), measuring the moral self-image is also important. Cornelissen et al. (2013) first attempted to measure the moral self-image with a scale of differences between the desired and the perceived moral self. Jordan et al. (2015) developed this scale further to provide a tool that actively and explicitly measures the moral self-image. However, there is still little empirical evidence of the deviations and fluctuations in time of the moral self-image (Perkins et al., 2024).

2.2. Definitions

As the previous section illustrated, different theories try to explain immoral behavior. Given the important role of the moral self, moral psychology increasingly shifted its focus to it to extend moral reasoning and predict behavior (Monin & Jordan, 2009). Before introducing some measurement methods, two terms must be defined accordingly in the context of the moral self: moral identity and moral self-image.

2.2.1. Moral Identity

Aquino and Reed (2002) define moral identity "*as a self-conception organized around a set of moral traits.*" (p. 1424). They suggest that moral identity is relatively stable over time and identify two dimensions: Internalization and Symbolization. Internalization describes how important it is for a person to have (nine) moral traits: "*caring, compassionate, fair, friendly, generous, helpful, hardworking, honest, and kind.*" (Aquino & Reed, 2002, p. 1426). Symbolization describes the degree to which a person wants to be seen as moral or demonstrate these traits through their actions to others (Aquino & Reed, 2002). The researchers propose that people behave morally when they assess a specific moral trait as essential for their self-concept. Moral identity should, therefore, be a motivational driver for acting consistently (Aquino & Reed, 2002) and is the basis for moral motivation (Erikson, 1964; Nisan, 1990).

To measure moral identity actively, Aquino and Reed (2002) asked participants to rate how important it is for them to possess these traits (Internalization) and if they participate in activities (e.g., hobbies), wear clothes or buy products that identify them as having these characteristics (Symbolization) (Aquino & Reed, 2002).

2.2.2. Moral Self-Image

Jordan et al. (2015) introduced the concept of the moral self-image to explain how the self-perception of the individual's morality fluctuates. They define the moral self-image as the malleable and dynamic moral self-concept.

A person's moral self-image can be described as the answer to the question "*'How moral am I?'*" (Monin & Jordan, 2009, p. 347). This reflects exactly how morally individuals see themselves at any point in time. The moral self-image is part of the dynamic working self-concept, the malleable part of the self (Jordan et al., 2015). It is completely subjective and only measures how moral persons perceive themselves (Jordan et al., 2015). Monin and Jordan (2009) highlight that individuals can constantly show differences in their moral self-image, as it can be lowered or bolstered through previous actions, which motivates subsequent behavior. The researchers agreed that the moral self-image has a behavior-generating power (Jordan et al., 2015; Monin & Jordan, 2009).

Due to the lack of previous empirical measurement methods, Jordan et al. (2015) introduced an explicit nine-point Likert scale to measure the moral self-image as highly connected to the traits of a typical moral person (based on the traits introduced by Aquino and Reed (2002)). This elaborated scale has been used as an explicit moral self-image measure in previous research (Ferguson, 2018) to investigate the convergent validity, which "*reflects the extent to which two measures capture a common construct.*" (Carlson & Herdman, 2012, p. 18).

2.3. Measurement Methods

2.3.1. Explicit vs. Implicit Measures

In order to understand, predict, and control human behavior, psychologists have been trying to measure people's cognitive processes, attitudes, and self-image (de Houwer, 2006).

A straightforward approach is to conduct a survey and actively ask participants about their opinions toward a situation or an object. This explicit method is easy to conduct, comprehensible, and easily measured (de Houwer, 2006). The most common approach for measuring the moral identity or the moral self-image is letting participants rate different personality traits on a Likert scale. This approach assesses how important these personality traits are for them (moral identity) (Aquino & Reed, 2002) or how much they are already fulfilling some characteristics compared to the person they want to be (moral self-image) (Jordan et al., 2015). However, despite their wide use (Asendorpf et al., 2002), these surveys might be subject to impression management (Paulhus, 1984), which means that participants include answers to be seen in a favorable light. When being asked, people are influenced by concerns about their self-presentation (Doherty & Schlenker, 1991; Schnabel et al., 2007) and social desirability (Crowne & Marlowe, 1960), which could incentivize them to give socially conform answers to the interviewer. Additionally, these surveys are limited to the introspective personality and might not reflect a person's entire personality (Schnabel et al., 2007).

Because of these disadvantages, new implicit measures have been developed. Initially introduced in social psychology, implicit measures are now widely applied across different disciplines and commonly used in psychology (de Houwer et al., 2009). But what is an implicit measure exactly? de Houwer (2006) suggests using the synonym automatic when explaining implicit effects. A process can be called automatic when it still operates, although the participants are unaware of results, stimulus, or procedure, do not have a specific goal, or do not invest many cognitive resources (de Houwer, 2006). Following this argumentation, the same should apply to an implicit measure. This measure intends to get an immediate (automatic) response from people without them being aware of it or involving their cognitive thinking. Such an implicit or indirect measurement method could be used to measure a person's unconscious attitude (Bartels & Schoenrade, 2022; Schnabel et al., 2007).

2.3.2. Implicit Association Test (IAT)

As a way to avoid biases of explicit measurement, Greenwald et al. developed the Implicit Association Test (IAT) in 1998. This test aims to measure the relative implicit association strength of two contrasting concepts (target categories) (e.g., FLOWER-INSECT) and PLEASANT-UNPLEASANT[3] (evaluation attribute) (Greenwald et al., 1998)).

In their initial experiment, all participants should react by pressing an assigned key on the left or right. In different blocks, a target category and an attribute are assigned to one key. For example, the left key is assigned to FLOWER + PLEASANT, whereas the right key is assigned to INSECT + UNPLEASANT. Whenever a stimulus (either a FLOWER, an INSECT, a PLEASANT, or an UNPLEASANT word) appears on the screen, the participant should press the assigned key (Greenwald et al., 1998). In other words, the stimuli should be classified into four mutually exclusive categories (FLOWER, INSECT, PLEASANT, or UNPLEASANT) (Greenwald et al., 1998; Schimmack, 2021). Participants are required to distinguish between words referring to INSECT + FLOWER and words referring to PLEASANT + UNPLEASANT words. For instance, with the assigned keys described above, the stimuli tulip or happy should be assigned to the left key (FLOWER + PLEASANT), whereas wasp or rotten should be assigned to the right key (INSECT + UNPLEASANT) (Greenwald et al., 1998). After the first combined task of discrimination between the target categories and evaluation attributes, a second block with a reversed combined

[3] The categories (target categories and evaluative attributes) are written in capital letters.

task was conducted. In this reversed combined task, one key was assigned to INSECT + PLEASANT, and the other key was assigned to FLOWER + UNPLEASANT.

It is fundamentally assumed that participants' response time is faster when the association between the target category and the evaluative attribute is stronger (de Houwer, 2006; de Houwer, 2001; Greenwald et al., 1998; Johnston et al., 2013). This means that the pairing FLOWER + PLEASANT should be easier compared to the INSECT + PLEASANT pairing if the association between FLOWER and PLEASANT words is stronger (de Houwer, 2006; de Houwer, 2001; Greenwald et al., 1998; Johnston et al., 2013). With this experiment, Greenwald et al. (1998) provided significant results demonstrating that the incompatible combination of INSECT + PLEASANT was more challenging to confirm, and participants had longer response times compared to the compatible combination of FLOWER + PLEASANT. The authors explain this effect with a stronger association and familiarity between FLOWER + PLEASANT words and than between INSECT + PLEASANT words, indicating a more positive attitude toward FLOWERS than INSECTS (Greenwald et al., 1998).

While the IAT was quite revolutionary, set new standards, and offered new opportunities, criticism about the IAT and implicit measures, in general, needs to be addressed. There are concerns about its construct validity (Schimmack, 2021), its capability to predict behavior (Bartels & Schoenrade, 2022; Brownstein et al., 2020), and its temporal instability (Brownstein et al., 2020; Schimmack, 2021).

Schimmack (2021) raises concerns that there is no consensus about what the IAT measures and that it is difficult to compare if it measures something different than explicit measures. This problem has been recognized by the dual attitudes model (Wilson et al., 2000) (also known as the double dissociation model (Perugini, 2005)), which clearly distinguishes implicit and explicit attitudes into two systems (Wilson et al., 2000). According to this model, implicit measures predict impulsive, spontaneous, and automatic behavior, while explicit measures predict controlled and conscious behavior (Wilson et al., 2000). In agreement with the double dissociation model, Johnston et al. (2013) suggest that implicit and explicit attitudes can only be measured with implicit or explicit measurement methods, respectively.

The contrasting perspective describes an additive view, where both types of attitude describe a "*different portion of variance in the same criterion*" (Perugini, 2005, p. 29). Fazio and Olson (2003) argue that both measures assess the same construct and explain a potential difference between the measurement methods with participants' deliberative control strategies.

The IAT and other implicit measurement methods can add predictive insights to self-report measures (Brownstein et al., 2020) and investigate the implicit moral self-image. Considering the concerns about the behavior predictability of the IAT, there are several studies about predictions of voting behavior with the IAT. For example, Friese et al. (2007) successfully predicted the voting behavior and attitudes for the German parliamentary elections in 2002. They used a single-target IAT, where one key was assigned to an evaluative attribute and the target category, while the second key was assigned only to the opposing attribute. This single-target IAT yielded excellent validity in predicting voting behavior (Friese et al., 2007).

2.3.3. Go/No-Go Association Task (GNAT)

To expand the use of implicit measurement methods and the Implicit Association Test (IAT), a new method was developed by Nosek and Banaji (2001). They introduced the Go/No-Go Association Task (GNAT), which mainly focuses on the error rate as the dependent variable to measure the strength of implicit associations (Greenwald et al., 1998; Nosek & Banaji, 2001).

Unlike the previously known implicit methods, in the GNAT, only a single concept (target category / e.g., ME) is evaluated considering one attribute dimension (evaluative attribute / e.g., GOOD) (Bassett & Dabbs, 2005; Ferguson, 2018; Nosek & Banaji, 2001). The GNAT does not need two contrasting concepts (two target categories); hence, it is more flexible and can reveal new aspects of social cognition. Another difference is that only one response (key) is required for the GNAT (Nosek & Banaji, 2001), simplifying the experimental setup.

During the task, a target stimulus (signal item) or a distracter stimulus (noise item) is presented on the screen for some milliseconds. Following the experimental design and example of Ferguson (2018) for the GNAT, when a stimulus word that is similar to either the attribute (e.g., GOOD) or the target category (e.g., ME) is shown, the participant should press the space bar (or any key) to give a Go response. On the contrary, when the word on the center of the screen does not match the attribute or the target category (= distractor), no response (No-Go) is required, and the participant should refrain from pressing any key.

According to Nosek and Banaji (2001), the strength of association in the GNAT is determined by how well stimuli words associated with the target category and the attribute (for example, ME + GOOD) are distinguished from distractor items unrelated to these concepts. The authors suggest that the sensitivity between the pairing conditions (in this example, ME + GOOD or ME + BAD) illustrates the strength of the association between the target category and the evaluative attribute (Greenwald et al., 1998; Nosek & Banaji, 2001). In general, the faster and/or the fewer errors (and therefore easier) the response, the stronger the association. Greenwald et al. (1998) and Nosek and Banaji (2001) argue that both error rates and average response times can provide information about task performance due to a speed-accuracy tradeoff. Nevertheless, most implicit measures focus solely on response times "*as the dependent variable and therefore may lose relevant information contained in error rates*." (Nosek & Banaji, 2001, p. 628).

Previous research in psychology has used the GNAT to investigate different implicit attitudes. For example, implicit spider fear associations (Teachman, 2007), implicit bias in phrasing drug addiction (Ashford et al., 2019), and implicit attractiveness beliefs of people who are constantly worrying about their physical appearance (Buhlmann et al., 2011). Those studies provide significant insights into the reliability and validity of the GNAT. In those applications, the researchers suggest that the GNAT is an effective tool for measuring involuntary associations and might help measure implicit associations, especially since it does not require a comparison category on a second key (Buhlmann et al., 2011; Teachman, 2007; Williams & Kaufmann, 2012). Williams and Kaufmann (2012) specifically investigated the reliability of the GNAT. They recommend a minimum of 40 trials per block for minimally acceptable reliability and at least 80 trials per block for good reliability. They argue that the GNAT is a valuable tool with many advantages and should be used in further research. (Williams & Kaufmann, 2012). By omitting a comparison concept (unlike the IAT), the GNAT can use distractor items more flexibly, allowing for a direct assessment of the attitude (Nosek & Banaji, 2001). "*In addition, the GNAT may be less susceptible to errors introduced by term valence and less biased by response criteria than reaction time-based techniques*." (Boldero et al., 2007, p. 354).

The convergence between implicit and explicit personality traits was examined by Boldero et al. (2007). They support the reliability and convergent validity of the GNAT when controlling the systematic variance of the GNAT. However, as their explicit measure was conducted before the implicit GNAT, it is possible that this inflated their associations and, thus, the correlation between both measures (Boldero et al., 2007).

Unfortunately, sufficient research in the moral domain, including the GNAT, has not yet been conducted. Previous studies focused on predictions about moral behavior with implicit measurement methods, like the IAT (Perugini & Leone, 2009). Another study tried to measure the moral identity with the IAT (Johnston et al., 2013). The only known publication of an application of the GNAT to measure the implicit moral self-image is a dissertation by Ferguson (2018), who failed to show moral balancing effects.

Implicit measures, such as the IAT and GNAT, can easily be implemented into software and be used as a portable version on mobile devices, thus overcoming the limitation of requiring a local computer (Bassett & Dabbs, 2005; Dabbs et al., 2003). This provides several advantages. Firstly, it tests participants in a more natural setting outside a laboratory (Bassett & Dabbs, 2005; Dabbs et al., 2003). This real-life setting could lower the feeling of being observed during the experiment and lead to more honest and impulsive answers. Secondly, Bassett and Dabbs (2005) argue that portable versions could be used to measure malleable attitudes at different times and important events. Thirdly, having a mobile and portable version of these tests could help to reach populations that would usually not participate in laboratory experiments. Fourthly, more people could participate because the effort needed is much less, as they are no longer required to go to the laboratory (Bassett & Dabbs, 2005).

3. Methodology

3.1. Research Design

This bachelor's thesis aims to apply an existing implicit measurement method, the Go/No-Go Association Task (GNAT), to measure the participants' implicit moral self-image and analyze these results for a correlation with their explicit moral self-image. It intends to investigate the effectiveness of the GNAT with the convergent validity between the implicit and explicit measures. As explained in Chapter 2, most of the previous research to measure the moral self-image is not based on the GNAT (e.g., Johnston et al., 2013; Perugini and Leone, 2009).

The underlying parameters of the performed experiment and the test setup were as follows:

3.1.1. Experimental Design

Objective & Material & Groups & Variables

Objective: Measure the implicit moral self-image and analyze the correlation with the explicit moral self-image (convergent validity).

Material and Groups: Inspired by Ferguson (2018), who failed to provide significant evidence for moral balancing using the GNAT, her experiment was reproduced in group A using the exact same stimuli (six words for GOOD/BAD and four words for ME/OTHER) to be consistent with her research method. In group B, the list of words was extended for potentially stronger effects.

Group A: These stimuli words were replicated from Ferguson (2018):

> ME: *me, I, my, myself*
>
> OTHER: *other, others, them, they*
>
> GOOD: *good, honest, faithful, modest, sincere, altruist*
>
> BAD: *bad, dishonest, deceptive, pretentious, arrogant, cheater*

Group B: The extended stimuli words (attributes) were selected from different scientific papers and measured in a pre-study according to their evaluative intensity (see Appendix V.I Pre-Study GOOD & BAD). This group used 20 stimuli words for each attribute and 15 for each concept category (ME & OTHER). To be consistent with Ferguson (2018) and the previous group A, the 40 attribute stimuli (57.14%) and 30 concept category stimuli (42.86%) in relation to the total stimuli of 70 in group B was almost equal to the 12 attribute stimuli (60%) and 8 concept category stimuli (40%) initially introduced by Ferguson (2018).

The used stimuli words in group B were:

> ME: *me, I, my, myself, mine, self, personally, oneself, person, intrinsic, own, individual, ego, inner essence, inner self*

Table 1: Possible outcomes of the Go/No-Go Association Task, inspired by "Trial of yes-no experiment" in Macmillan (2002)

		Response			
		Go		***No-Go***	
Stimulus	***Signal (target item)***	Hit	○	Miss	×
	Noise (distractor item)	False Alarm	×	Correct Rejection	○

OTHER: *other, others, them, they, their, themselves, theirs, his, him, her, anybody, anyone, those people, persons, the individuals*

GOOD: *caring, fair, compassionate, friendly, hardworking, generous, helpful, kind, honest, faithful, altruist, modest, sincere, genuine, joyful, patient, grateful, loyal, forgiving, respectful*

BAD: *hostile, unfair, lazy, unhelpful, ruthless, selfish, evil, brutal, hateful, angry, impatient, bad, dishonest, deceptive, pretentious, arrogant, cheater, disrespectful, disloyal, egocentric*

Variables: The stimuli word lists (group A or B) served as independent variables. The dependent variables were Hit-/False-Alarm rates.

Blocks & Trials & Stimuli

The GNAT was divided into two blocks. The target category ME was paired with the attribute GOOD in the starting block. In the second block, the same target category, ME, was paired with the opposite attribute, BAD.

Each block comprised a total of 96 trials for group A or 86 trials for group B. Both blocks started with 16 practice trials (not considered in the analysis), followed by a reminder screen, before proceeding to the 80 critical trials (considered in the analysis) for group A or 70 critical trials for group B. A trial started when a stimulus word from one of the categories (ME, OTHER, GOOD, BAD) emerged on the screen. It ended when the word disappeared. As a constant reminder of the current combination (pairing) in each block, labels for the target category (ME) and the attribute (GOOD or BAD) remained on the screen's upper left and right corners. The labels and stimuli items were displayed in black font against a white screen.

The participants were advised to either (1) give a Go response by quickly pressing the space bar if the stimulus word displayed could be categorized into one of the two labeled categories (signal item) or (2) refrain from pressing any key (No-Go response) for words that could not be categorized (noise items). The stimulus word appeared in the center of the screen and remained visible until the response deadline was reached or a key was pressed. The subsequent trial started when the participant pressed the space bar or after the response time ran out. Like Nosek and Banaji (2001), the opposing category (OTHER) or the alternate attribute served as distracter trials (noise). For example, when GOOD was the signal, BAD was the noise, and vice versa. A signal-to-noise ratio of 1:1 was held constant for all trials and both groups.

The 20 stimuli words Ferguson (2018) used for the critical trials in group A were selected randomly. Each word was repeated four times for a total of 80 trials, which was in the range of 50 to 80, yielding sufficient and good reliability, as Williams and Kaufmann (2012) recommended. For group B, the stimuli items were chosen randomly and selected without repetition from the four categories (ME, OTHER, GOOD, BAD) to reach 70 trials. Within both groups, each block (pairing) consisted of an equal number of words in order to minimize learning effects, as was again advised by Williams and Kaufmann (2012). The critical trials used the complete set of stimuli words, which were selected randomly and appeared in random order. Additionally, for the word list in group B, a small-scale pre-study (Appendix V.I Pre-Study GOOD & BAD) on the evaluative intensity of these words was conducted beforehand. This ensured that the words' evaluative intensity was a) strong enough to yield sufficient results and b) similar between the words, which was strongly suggested by Nosek and Banaji (2001).

Response Deadline & Feedback

The participants had to categorize the stimuli words as quickly and accurately as possible during the short time displayed on the screen. The response deadline was constant at 700 milliseconds (ms) across all trials and blocks, following the recommended range of 500ms to 850ms by Nosek and Banaji (2001). The interstimulus interval between two trials was held constant at 500ms.

During this interstimulus interval, immediate feedback on performance accuracy was provided. Trials where signal items were accurately identified (Hit) or noise items were correctly ignored (Correct Rejection) were recorded as correct responses, indicated by a green "O" appearing. Trials were marked as errors when noise items were mistakenly identified as signals (False Alarm) or signal items were overlooked (Miss). For these trials, a red "X" was displayed in the center of the screen after the stimulus item disappeared. The possible outcomes and corresponding feedback are visualized in Table 1.

3.1.2. Statistical Analysis

The statistical analysis of the GNAT was based on the signal detection theory, first introduced by Green and Swets (1966), as cited in Macmillan (2002), and as previous experiments using the GNAT already have done (Ferguson, 2018;

Nosek & Banaji, 2001; Teachman, 2007).

An essential part of the signal detection theory is the calculation of the sensitivity, described as d-prime (d'). It indicates the participant's ability to differentiate signal (target items) from noise (distractor items) (Green & Swets, 1966; Macmillan, 2002; Nosek & Banaji, 2001). D-prime (d'), based on the signal detection theory and applied in the GNAT by Nosek and Banaji (2001), is calculated for each block (ME + GOOD and ME + BAD) with the following formula:

$$d' = Z(\frac{Hits}{Signal\ items}) - Z(\frac{False\ Alarms}{Noise\ items})$$

It calculates the difference between the standardized Z-score of the ratio of correct Go responses (Hits) to all signal items and the standardized Z-score of the ratio of Go responses for noise items (False Alarms) to all noise items (Macmillan, 2002). D-prime (d') only considers the participant's Go responses, as only the absolute number of Hits and the absolute number of False Alarms relative to the total amount of signal or noise items in the experiment are included in the formula. Generally, a higher d' reflects a higher sensitivity. This means that the participant could discriminate signal items from noise items more easily. In other words, he had more correct Go responses (Hits) than Go responses for distractor items (False Alarms).

Implicit Moral Self-Image Score

The implicit moral self-image score or $d'(GNAT)$ determined by the GNAT was defined as the difference in associative strength (sensitivity) between the two blocks (Nosek & Banaji, 2001). The d' of the second pairing (ME + BAD) was subtracted from the d' of the first pairing (ME + GOOD), which resulted in the final implicit moral self-image score $d'(GNAT)$[4]:

$$implicit\ moral\ self-image\ score:$$
$$d'(GNAT) = d'(ME + GOOD) - d'(ME + BAD)$$

A positive implicit moral self-image score indicates a positive moral self-image since the associations between ME + GOOD are stronger than between ME + BAD. In contrast, a negative implicit moral self-image score is interpreted as a negative moral self-image because of a stronger association between ME + BAD than between ME + GOOD (Ferguson, 2018).

Explicit Moral Self-Image Score

The explicit moral self-image score was determined with the questionnaire (see Appendix V.II Explicit Moral Self-Image Questionnaire) introduced by Jordan et al. (2015) to investigate the relationship and correlation between explicit and implicit moral self-image. Participants were asked about nine moral traits on a nine-point Likert scale, where 1: [I am] much less "moral trait" (e.g., caring) than the person I want to be and 9: [I am] much more "moral trait" (e.g., caring) than the person I want to be. The explicit moral self-image score was calculated by taking the average of all responses in the explicit moral self-image questionnaire. This score ranged from 1 to 9, with 5 demonstrating a neutral explicit moral self-image, lower numbers indicating a lower explicit moral self-image, and higher numbers indicating a higher explicit moral self-image, respectively.

Pearson Correlation Coefficient

The (Pearson) correlation coefficient (r) between each group's implicit moral self-image scores, based on response rates, and the corresponding explicit moral self-image scores, was determined to evaluate the convergent validity between both measurement methods. This coefficient is standardized from -1, indicating a perfect negative linear relationship, to 1, indicating a perfect positive linear relationship, and 0, indicating no linear correlation (Fahrmeir et al., 2016). For the formula, see V.III Formula of the Pearson Correlation Coefficient.

3.2. Participants

124 out of 206 test-takers successfully finished the experiment, corresponding to a conversion rate of 60.19%. Out of these 124 participants, 56 were assigned to group A and 68 to group B at random. This sample of 124 participants consisted of 71 females (57.26%), 52 males (41.93%), and 1 person not indicating their gender (0.81%). The participants' average age was 30.08 years. 84 people (67.74%) participated on mobile devices, whereas 40 (32.26%) participated on laptops or desktops.

3.3. Procedure

The participants received a link to the oTree website, an open-source experiment platform (Chen et al., 2016). They could access the experiment through the internet using their laptop as a stationary device or mobile devices such as smartphones or tablets. Using a cookie on the oTree website ensured that participants were only allowed to participate once per device. After opening this link, the participants were forwarded to the welcome page, where they received the first information about this experiment. They were told that it was part of a bachelor's thesis and that their contribution would help assess error rates and reaction times when categorizing various English words. On the next page, they received detailed instructions about the test procedure.

For the GNAT, the participants were randomly assigned to one of the two groups and, thus, different word lists. After that, they saw an overview of the current labels in the first block (ME + GOOD) and the list of words. They started immediately with a training test, the first 16 practice trials.

[4] For further clarification: $d'(GNAT)$, also known as implicit moral self-image score based on response rates, represents the final score of the GNAT, which was calculated as the difference between the d' of the different blocks (pairings) (see formula).

Table 2: GNAT procedure in the experiment

• Welcome • Instructions: How It Works	
Group A	**Group B**
Block 1: (ME + GOOD) • Overview (labels and word list) • 16 Practice Trials • Reminder • 80 Critical Trials	***Block 1: (ME + GOOD)*** • Overview (labels and word list) • 16 Practice Trials • Reminder • 70 Critical Trials
Block 2: (ME + BAD) • Overview (labels and word list) • 16 Practice Trials • Reminder • 80 Critical Trials	***Block 1: (ME + BAD)*** • Overview (labels and word list) • 16 Practice Trials • Reminder • 70 Critical Trials
• Explicit Moral Self-Image Questionnaire • Demographic Questions	

After finishing a task, the participants saw the number of errors they had made. They were directed to a page reminding them of the current pairing and word list, that they should answer as quickly and accurately as possible, and that the next part (critical trials) counts toward their scores. For the second block with the new pairing ME + BAD, the same procedure started again, consisting of the first 16 practice trials, the reminder, and the critical trials.

After finishing the second block, participants were explicitly asked about their moral self-image using the scale developed by Jordan et al. (2015). They were asked about nine moral traits on a nine-point Likert scale, where 1: [I am] much less "moral trait" (e.g., caring) than the person I want to be and 9: [I am] much more "moral trait" (e.g., caring) than the person I want to be. Finally, some demographic questions, including gender, age, educational level, employment status, first language, and English proficiency, as well as the occurrence of technical problems, were asked. A graphical representation of the procedure for both groups is visualized in Table 2.

3.4. Hypotheses

The following hypotheses were derived to measure the effectiveness of the GNAT according to its convergent validity with an explicit moral self-image questionnaire:

> **Hypothesis 1:** The participants' implicit moral self-image scores (based on response rates) will **correlate positively** ($r > 0$) with the explicit moral self-image scores (questionnaire-based) in both groups.

This hypothesis assumes that both implicit and explicit moral self-image reflect the same construct but are measured through different methods, following Johnston et al. (2013) and Fazio and Olson (2003). Therefore, a positive correlation between both measures is expected, as the scores may vary but do not need to be congruent. A person with a positive implicit moral self-image in the GNAT should also indicate a positive explicit moral self-image in the questionnaire, and vice versa.

> **Hypothesis 2:** The **correlation** between the participants' implicit moral self-image scores and explicit moral self-image scores will be **higher in group B than in group A.**

This hypothesis assumes that the word list without repetition of group B would be more effective (and yield higher correlation) in the GNAT than the smaller, repeated word list of group A. It is expected that the repetition of words will lead to participants' better memorization during the first block, allowing for improvements and better performance in the second block, reducing the implicit moral self-image score. This would reduce the correlation coefficient compared to group B, where words are not repeated.

Table 3: Exclusion criteria for the standard sample

Exclusion Criteria
1. No recorded data
2. No response was given during the trials
3. Reaction times < 250ms
4. Reaction times > 735ms
5. $d'(ME + GOOD) \leq 0$ or $d'(ME + BAD) \leq 0$
6. Basic English proficiency
7. Indication of technical issues
8. Perfect Hit rate
9. No False Alarms

4. Results

4.1. Exclusion Criteria

Some participants had to be excluded from the analysis to ensure a valid and valuable data set (standard sample). Therefore, the data set was manually revised, especially due to technical difficulties during data collection, such as the failure to record responses and/or reaction times as well as heavy deviations from the response deadlines (e.g., speeding up the appearing words (up to 1ms) or very long loading times (up to 1,080ms)).

The following exclusion criteria for the standard sample were defined (and visualized in Table 3):

All participants whose responses were not recorded were excluded. Either the server failed to measure the reaction times for some trials, or the participants failed to respond correctly by not giving Go responses (in one block or throughout the experiment).

Furthermore, all participants with response times below 250 milliseconds (< 250ms) and above 735 milliseconds (> 735ms), to still account for some processing time of the server, were dropped. Greenwald et al. (2003) suggest excluding reaction times below 300 milliseconds (< 300ms), which could indicate random responding, but due to the advanced digitalization, frequent use of electronic devices nowadays, and the relatively young sample (average age of 30.08 years), response latencies within the range of 250ms to 735ms were included.

Additionally, following the argument by Nosek and Banaji (2001), all d-primes (d') (based on response rates) below 0 were excluded as they indicate that the participant was not following the instructions properly or was unable to distinguish any signal from noise. A value of 0 equals chance (Nosek & Banaji, 2001).

All participants had to have at least intermediate English proficiency to be considered. Therefore, all participants who indicated basic English skills were dropped. They were also excluded if they indicated technical issues in the final demographic questionnaire.

Further, due to problems with the mathematical Z-score calculation, all participants with perfect Hit rates (a participant correctly responded to all signal items) or with no False Alarms (a participant never gave a Go response for distractor items) were excluded from the primary analysis. Nevertheless, the effects of a potential correction when calculating d' were investigated.

4.2. Participants per Group

Group A (word list by Ferguson (2018)):

After the strict exclusion, 30 participants of the initial 56 people finishing the experiment in group A were considered (conversion rate of 53.57%). The sample consisted of 17 females (56.67%) and 13 males (43.33%), with an overall average age of 29.43 years.

Most people were students (56.67%). The rest was either full-time (30%) or part-time employed (13.33%). As their highest completed level of education, nine people indicated high school degrees (30%), ten had a Bachelor's degree (33.33%), and 36.67% had a Master's degree. The majority spoke German as their first language (76.67%), two spoke Portuguese (6.67%), and one person spoke English, Dutch, Spanish, and Finnish, respectively. One person stated being bilingual in German and French (3.33%). 73.33% of participants experimented on a mobile device, while 26.67% used stationary devices.

Group B (extended word list):

Following the same exclusion criteria, 38 participants of the initial 68 people finishing group B were considered (conversion rate of 55.88%). The sample consists of 21 females (55.26%) and 17 males (44.74%) with an overall average age of 30.82 years.

Half of the participants were students (50%). The rest were either full-time (36.84%), part-time (5.26%) or self-employed (5.26%), and one person had already retired (2.63%). As their highest completed level of education, eight people indicated high school degrees (21.05%), twenty people had a Bachelor's degree (52.63 %), and 26.32% had a Master's degree. Most people spoke German as their first language (76.32%), five people were native English speakers (13.16%), two people spoke Spanish (5.26%), and one person spoke Ukrainian (2.63%). One person stated to be bilingual in German and English. 69.23% of participants performed the experiment on mobile devices, and 28.25% on desktops or laptops.

4.3. Implicit Moral Self-Image Score Based on Response Rates

Signal detection theory proposes calculating participant sensitivity (d') for each block. It subtracts the standardized Z-scores of the ratio of Go responses for noise items (False Alarms) to all noise items from the standardized Z-score of the ratio of correct Go responses (Hits) to all signal items (Macmillan, 2002). Applying this theory to the GNAT, it is assumed that participants are more sensitive and can discriminate signal items from noise items more easily when the association of the two signal components is more positive than a negative or neutral association (Nosek & Banaji, 2001). "*Greater sensitivity indicates a stronger association between the target category and attribute. This association is defined to be a measure of automatic attitude toward the target category.*" (Nosek & Banaji, 2001, p. 635).

To calculate the implicit moral self-image score with the GNAT, the approach of Ferguson (2018) and Nosek and Banaji (2001) was followed by comparing the d' of both pairings for each participant:

$$Implicit\ moral\ self-image\ score:$$
$$d'(GNAT) = d'(ME + GOOD) - d'(ME + BAD)$$

According to Ferguson (2018), a positive implicit moral self-image score indicates a positive moral self-image (the associations between ME + GOOD are stronger than between ME + BAD), whereas a negative implicit moral self-image score is interpreted as a negative moral self-image (stronger association between ME + BAD than between ME + GOOD).

4.3.1. Results of Standard Sample

Table 4 shows all characteristics of the obtained implicit moral self-image scores for both groups and compares the Pearson correlation coefficients of both groups.

Table 4: Characteristics of implicit moral self-image scores, based on response rates, and Pearson correlation coefficients for group A and group B with strict exclusion criteria

	Group A	**Group B**
Max	1.3804	1.5475
Min	-1.4816	-0.6629
Mean	0.2360	0.4154
Median	0.2783	0.3588
Pearson correlation coefficient	-0.0078	0.4870

Group A

In group A, the final implicit moral self-image scores, calculated as the participants' difference in association strength (d') of both blocks, ranged from a score of -1.4816 to 1.3804, with a mean of 0.2360 and a median of 0.2783. Eleven participants (36.67%) performed better in the ME + BAD block, indicating a negative implicit moral self-image. The questionnaire-based average explicit moral self-image score was obtained as 5.2852 in group A.

Correlating the implicit moral self-image scores with the explicit moral self-image scores conveyed through the questionnaire after the GNAT yielded a Pearson correlation coefficient of r = -0.0078. This coefficient does not indicate a linear correlation between the explicit and implicit moral self-image scores. Counterintuitively, it may suggest a very slight negative tendency. This means that people who performed better in the ME + GOOD pairing of the GNAT, relative to the ME + BAD pairing, provided answers in the questionnaire indicating a lower explicit moral self-image score and vice versa. This does not support hypothesis 1 of the correlation coefficient being positive for all groups ($r > 0$). Figure 1 depicts the distribution of values and ultimately demonstrates the correlation (dotted line).

Group B

The implicit moral self-image scores conducted through the GNAT for group B had a mean of 0.4154, a median of 0.4154, and ranged from -0.6629 to 1.5475. Only nine participants (23.68%) had a negative implicit moral self-image score and thus a higher association of ME + BAD. The average explicit moral self-image score in this group was 5.2281.

The Pearson correlation coefficient between implicit moral self-image scores and explicit moral self-image scores was calculated as r = 0.4870, which indicates a moderate positive linear relationship. This finding suggests that people who performed better in ME + GOOD pairing had a higher average score for the explicit moral self-image. This supports the first hypothesis of a positive correlation ($r > 0$) and the effectiveness of the GNAT. Figure 2 depicts the distribution of values and ultimately demonstrates the correlation (dotted line) for group B. A similar pattern but different ranges on the x-axis can be observed in Figure 1 and Figure 2. When comparing both correlation coefficients, it becomes evident that group B correlates higher than group A, which supports hypothesis 2.

4.3.2. Correction of 0.5 for all Perfect Responses

As previously stated, all participants with perfect Hit rates and no False Alarms (perfect responses) were excluded from the data set due to problems when calculating the Z-scores and d' of each block. Nonetheless, the effects of a correction suggested by Kadlec (1999) and Macmillan (2002) on the correlation coefficient for both groups were investigated. This approach involves correcting the absolute numbers of Hits or False Alarms by either adding 0.5 to no False Alarms or subtracting 0.5 from perfect Hit rates, which yielded 39.5 Hits in group A and 34.5 Hits for group B. Correcting these perfect response rates allowed four more participants to be included in group A and two more in group B.

This resulted in Pearson correlation coefficients of r = 0.0106 for group A and r = 0.4384 for group B. In this case, both groups demonstrate a positive correlation supporting hypothesis 1. The coefficient increases for group A and decreases for group B compared to the standard sample, as shown in Table 5. However, in support of hypothesis 2, the correlation coefficient of group B is still higher than group A's.

Table 5: Comparison of different Pearson correlation coefficients before and after sample correction for groups A and B

Pearson correlation coefficient	**Group A**	**Group B**
Standard sample (before correction)	-0.0078	0.4870
New sample, including participants with perfect responses (after correction)	0.0106	0.4384

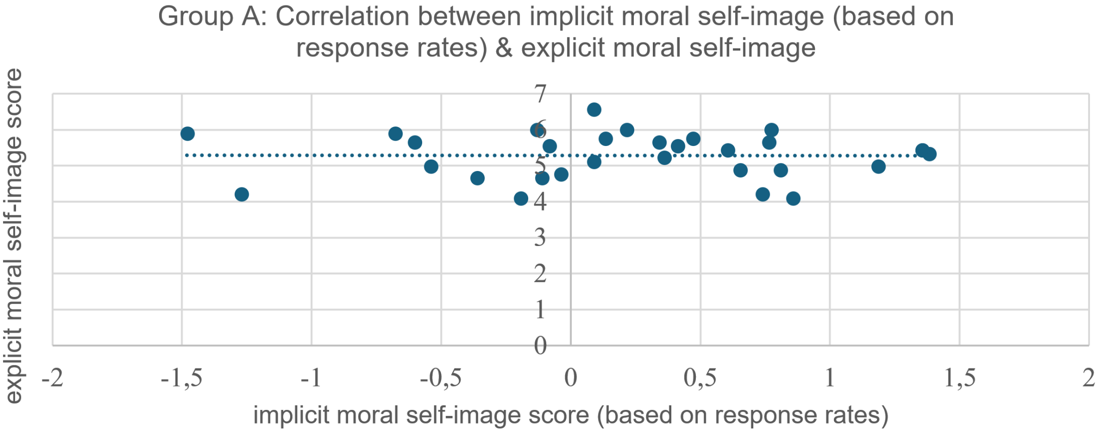

Figure 1: Correlation (dotted line) between the implicit moral self-image scores, based on response rates, and the explicit moral self-image scores for group A

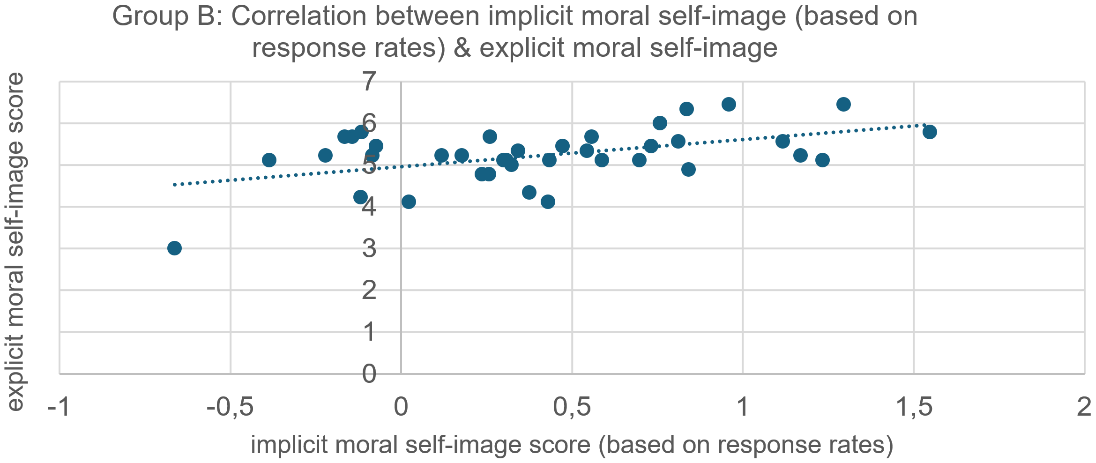

Figure 2: Correlation (dotted line) between the implicit moral self-image scores, based on response rates, and the explicit moral self-image scores for group B

5. Discussion

5.1. Key Findings

This study was designed to investigate the effectiveness of the GNAT for measuring the moral self-image with its convergent validity by relating it to the well-established explicit moral self-image questionnaire by Jordan et al. (2015). The indicator considered in this analysis is the Pearson correlation coefficient (r), where a high correlation indicates a high convergent validity. The implicit moral self-image was analyzed with a focus on the response rates using the signal detection theory by Green and Swets (1966), as cited in Macmillan (2002).

This bachelor's thesis is the second known attempt to measure the implicit moral self-image with the GNAT. Given that the first attempt by Ferguson (2018) was unsuccessful in showing significant results for moral balancing and was somewhat opaque and vague about the detailed research design, this was replicated in group A but extended in group B. Based on literature, Ferguson's (2018) list of words was extended with more synonyms for group B (see Appendix V.I Pre-Study GOOD & BAD) to further explore the effect of different stimuli words in the GNAT. Another crucial difference between both groups was the repetition of stimuli. Group A only used 20 stimuli words (10 signals and 10 noises), which were each repeated four times to yield a total of 80 critical trials, whereas group B used 70 stimuli words (35 signals and 35 noises) for 70 trials without repetition.

In hypothesis 1, both groups were expected to have a positive correlation between the implicit and explicit moral self-image scores ($r > 0$). This assessment was based on the idea that implicit and explicit moral self-image explain the same construct and are only obtained by different measurement methods, following Johnston et al. (2013) and Fazio and Olson (2003). Consequently, this would lead to a positive correlation, even though the scores allow for some deviation and do not necessarily need to be completely congruent.

This hypothesis 1 has been proven wrong for group A, as the first group with fewer stimuli had a negative correlation coefficient of r = -0.0078 in the standard sample. This negative relationship could be explained, as it seems reasonable that participants who performed better in the second block (ME + BAD) had already seen all stimuli at least four times before and memorized the stimuli used in the first block. This prior "training" in the first block might have caused a higher d' in the second block (ME + BAD), which led to a negative implicit moral self-image score in the GNAT. With this improved personal performance in the second block (ME + BAD), participants could have bolstered their moral self-image and could have reported a higher score in the questionnaire, which led to a negative correlation. This appears plausible since they received immediate feedback and saw the total number of errors after each block. The same mechanism could be applied to participants who performed even worse in the second block. Participants not fulfilling their target in the second block, e.g., not outperforming their error rate from the first block, might be disappointed and state a negative explicit moral self-image. As the negative tendency was extremely small, this influence of personal performance on the explicit questionnaire was insignificant and could not be observed after the correction. However, with a convergent validity of r = -0.0078 for group A or even after the sample correction of r = 0.0106, both coefficients were very close to zero, indicating no significant correlation and, thus, no convergent validity. The absence of (sufficient) convergent validity might imply that both measures do not assess the same construct or that either of the two measures is flawed (Cunningham et al., 2001).

In contrast, group B had a Pearson coefficient of r = 0.4870 (even after correction r = 0.4384), which makes a particular relationship and a moderate convergent validity between both measures evident for this group. This further proves that the GNAT, as an implicit measure, is not flawed and holds good convergent validity, which can be interpreted as proof that implicit and explicit measures capture the same construct (to a certain extent). This promising relationship could be understood as even more substantial and already as significant since it is much higher than the previous standard for implicit measures (usually around 0.2 to 0.3 (Perugini, 2005)). Cunningham et al. (2001) noted that the extent of measurement error in any measure establishes a natural upper limit for correlations. They interpreted their correlation of r = 0.35 between the implicit IAT and an explicit scale already as significant and argued that "*the two sets of measures are correlated, yet distinct*." (Cunningham et al., 2001, p. 167).

For hypothesis 2, it was expected that the correlation between the implicit moral self-image scores and the explicit moral self-image scores (based on the questionnaire) for group B would be higher than that of group A. This hypothesis has been proven right due to a significantly higher correlation coefficient of r = 0.4870 for group B compared to r = -0.0078 for group A. Even when considering the correction for all perfect responses, the same effect could be observed with a correlation coefficient of r = 0.4384 and r = 0.0106 for group B and group A, respectively.

This could also be explained by stronger learning effects in group A, as the same 20 stimuli were repeated four times in each block. This was not the case for group B. Each stimulus was only shown once within a block due to the extended word list. Additionally, in relative terms, more participants with a negative implicit moral self-image score were observed for group A (36.67%) than for group B (23.68%). This could be because participants in group A had fewer stimuli and more time to learn, which gave them an advantage for the second block, where they performed better in 36.67% of all cases. This percentage is lower in group B, at 23.68%. Therefore, it can be assumed that learning effects in group B were not that strong, which led to a more significant correlation between implicit and explicit measures. In addition, learning effects due to unequal block length were avoided in the experimental design, as both blocks in each group had the same length; hence, this should not cause this significant difference in correlation between both groups (Williams & Kaufmann, 2012).

An alternative explanation for the higher correlation could be the higher percentage of native English speakers in group B (13.16%) compared to group A (3.33%). Indeed, a native English speaker would react faster and more accurately than a native German speaker. However, assuming that the characteristics of each participant, for instance, language or age, were the same in both blocks, this should not affect the overall results. This is because the response rate of each participant was calculated into an individual d' for each block by only comparing the response rates within each participant (identical person). The d' of a native English speaker could then be generally higher in each block compared to the d' of a German speaker, but after calculating the difference between both blocks (the implicit moral self-image score), this effect can be neglected. Therefore, language, age, and educational level should not affect the overall results in both groups.

As mentioned, previous literature usually obtained lower correlation coefficients (generally between 0.2 and 0.3) and explained this as the natural difference between explicit (relying on self-reporting) and implicit (relying on the associative strength and measured with reaction times or error rates) measurement methods (Cunningham et al., 2001; Perugini, 2005). Assuming the correlation is a perfect indicator of the underlying system, one could argue that a lower correlation would prove a double dissociation system, where explicit and implicit attitudes exist and describe different evaluations. This only applies if the assumption holds that explicit attitudes are only measured by explicit measures and implicit attitudes only by implicit measures. In contrast, a higher correlation would support the existence of an additive pattern, implying that explicit and implicit measures describe the same attitude, as both measures yield similar results. As there is no clear consensus among researchers, it is difficult to come to a conclusion. However, one could believe that "*implicit and explicit attitudes can be best understood as implicit or explicit measures of the same attitude*." (Perugini, 2005, p. 31), given the promising correlation in group B.

The initial research question: "To what extent can the Go/No-Go Association Task (GNAT) be effectively applied to measure moral self-image, and is there a correlation between the outcomes of this method and the explicit moral self-image?" should be answered separately for both groups. Group A has shown little, even negative, to no correlation between the two measures, which indicates an ineffective application of the GNAT in measuring the moral self-image. For group B, there is, in fact, a significant positive correlation between the implicit and explicit measures, indicating an effective application because learning effects were avoided by using various stimuli without repetitions.

This thesis has closed the gap in previous research, as it has proven that the GNAT is an effective instrument for measuring the moral self-image if learning effects are considered in the research design. It suggests utilizing various stimuli words without repetition instead of repeating fewer stimuli words. This finding has highlighted the extraordinary importance of the experimental design for yielding significant results. However, the GNAT still has some limitations, and further research investigating the GNAT's ability to predict behavior and convergent validity is highly recommended, whereas the first results were already quite promising. This thesis was the first study conducted to assess the convergent validity of the GNAT for capturing the moral self-image.

5.2. Limitation of Results & Suggestions for Future Research

It is important to emphasize the limitations of this research, especially to enable future research. First of all, investigating the correlation between explicit and implicit measures could be misleading and ambiguous due to various explanations for both attitudes and missing consensus about its significance. According to Perugini (2005), a low correlation could suggest insufficient convergence validity between the two measurement methods. However, it could also be seen as evidence for the double dissociation, supporting the theory that a dual system of attitudes exists and that those attitudes are unrelated. Instead, it might be more expressive to separately investigate the capability of both measures to predict behavior (Perugini, 2005) and explore this in further experiments. Validating the GNAT as a moral self-image measure according to its ability to predict behavior would provide further insights and could support its usefulness (Perugini, 2005).

Nevertheless, Carlson and Herdman (2012) generally advised using measures with convergent validities above 0.7 ($r > 0.7$) and avoiding those below 0.5 ($r < 0.5$). In the context of implicit measures and especially regarding the moral self-image, one might have to deviate from this suggestion, as there are various reasons for the low correlation. More research should be conducted to review these recommendations.

Future research could extend the experiment over a longer timeframe and ask participants about their moral self-image before and after specific moral or immoral actions. Since the measured moral self-image is a snapshot (in daily life) and can deviate in time (Jordan et al., 2015; Monin & Jordan, 2009), tests could be conducted at regular intervals with the identical test group to analyze the ranges of variation around their fixed personal reference, potentially initiating moral balancing behavior (Jordan et al., 2015; Nisan, 1990).

Regarding the experimental design, further investigation is needed about the effect of the block order on the moral self-image, as previous research has already suggested that it might have an impact (Greenwald et al., 1998, 2003; Nosek & Banaji, 2001). It would be interesting to see if a reversed order of the blocks yields the same results and convergent validity. An extensive experiment should investigate the occurrence of learning effects due to increased familiarity with the procedure and stimuli words in the second block. This could be done by changing the order of the blocks ME + GOOD and ME + BAD in different groups.

When investigating the convergent validity, an additional aspect is changing the order of the explicit questionnaire. Asking participants explicitly about their moral self-image before and after the GNAT could examine potential priming effects (of both explicit questionnaire and implicit GNAT), eventually inflating the correlation (Boldero et al., 2007). Further, including the questionnaire twice (before and after the GNAT) could provide insights into how the individual's performance in the GNAT changes the explicit moral self-image.

Additionally, it would make sense to expand this sample and conduct the experiment with more participants. Simple deviations already greatly impacted the coefficients obtained in the current sample of 30 (group A) to 38 participants (group B). Additionally, the uneven size of both groups raised concerns about the comparison. Most participants were native German speakers (76.67% in group A and 76.32% in group B), which differentiated from previous experiments, where only native English speakers were asked to perform the GNAT in English (Ferguson, 2018; Nosek & Banaji, 2001; Williams & Kaufmann, 2012). Repeating the same experiment in the first language of each participant could eventually provide more robust results, as the implicit association could be much stronger, investigating more "natural" and more accessible associations.

In addition, the portable nature of the GNAT (Bassett & Dabbs, 2005), as conducted in this online experiment, allowed for measuring the moral self-image of participants in everyday life. This setting enabled participants to perform the experiment in a familiar environment, without mandatory seminars, away from an unpleasant and stressful (graded) laboratory atmosphere, and whenever they felt ready and comfortable to process the test. Participants were only instructed to perform the test to support a bachelor's thesis in assessing error rates and reaction times when categorizing English words without telling them the true purpose of this experiment. This more relaxing environment might have supported less biased, unconscious responses and yielded real-life insights into participants' moral self-image without them being aware of it. However, this could also have led to participants not taking the test seriously and giving less concentrated responses than in a controlled laboratory environment, which made it harder to compare the results with previous studies.

Furthermore, the differences in results obtained through mobile and stationary devices were not investigated due to the limited scope of this bachelor's thesis and the small sample size. Examining this further with a larger sample is highly recommended, as the experimental Go response differed for both versions. Participants using touch screens had to tap exactly on the word, whereas participants using laptops had to press the space bar.

Moreover, due to the limited scope of this bachelor's thesis and the partially experienced technical discrepancies, a detailed analysis of the reaction times (response latencies), according to Cohen (2013), was omitted. An analysis applying the signal detection theory based on reaction times is proposed in Appendix V.IV Implicit Moral Self-Image Score Based on Reaction Times. However, an analysis based on Cohen (2013) and the often-used algorithm developed by Greenwald et al. (2003), relying on mean latencies and standard deviation of all reaction times, is strongly recommended in the future as it could hold important insights and support the GNAT's validity by confirming the results and correlations obtained.

5.3. Future Implications

After introducing the IAT and the GNAT, a new field of research opened up for assessing people's implicit attitudes. This could not only be used for psychological counseling (Asendorpf et al., 2002) but can also provide valuable and hidden insights into behavioral economics and for different economic stakeholders, e.g., companies, employees, investors, and policymakers.

Companies could use implicit measures to gain deeper insights into their stakeholders and to predict consumer behavior, especially when bringing this into the context of the moral balancing theory. They could exploit the knowledge about the compensatory balancing behavior of people after an initial immoral act. For example, companies could offer customers the option to donate to a prosocial charity after a self-centered purchase (e.g., a carbon-intensive flight for vacation), as adapted from Schlegelmilch and Simbrunner (2019). This allows customers to compensate for a decreased moral account with a subsequent good action (moral cleansing). Moral licensing effects could be utilized, for instance, by allowing people to demonstrate moral behavior (e.g., donating to a charity) before offering them a carbon-intensive flight (adapted from Schlegelmilch and Simbrunner (2019)). Both effects influence the customer's purchase decision, giving them a better feeling after balancing or justifying their self-centered purchase, potentially leading to a higher demand in the future. Additionally, "*[i]mplementing donation options in a web shop is an easy way for a company to signal that it is socially responsible.*" (Schlegelmilch & Simbrunner, 2019, p. 551). In Marketing, companies could cluster their target group, as people with higher importance on their morality have a higher incentive to maintain their moral self-image and are more likely to be influenced by (negative) deviations (Aquino & Reed, 2002; Monin & Jordan, 2009). If a person with a high importance on morality performs an immoral act, it has a larger negative impact on his moral self-image, and he is more likely to compensate for it afterward (Monin & Jordan, 2009). This could be utilized for personalized advertisements, pricing, and influencing people. As the GNAT reveals deeper, unconscious attitudes and feelings toward brands or products, it could be applied to more generic and strategic decisions, such as product development process, brand positioning, pricing, and improved market research.

Assessing the implicit moral self-image may serve as an efficient tool in the recruitment process (Asendorpf et al., 2002), potentially revealing the implicit biases and attitudes of both recruiters and prospective employees. This could be a helpful instrument when selecting employees as it, for example, could examine the accordance with company values. Kim (2003) discovered that participants could successfully develop strategies to fake the results in an IAT when instructed to react slower. Remarkably, this was not done spontaneously. This can affect the argument for using implicit measures, such as the IAT or GNAT, when selecting employees, as they could improve their performance if they understood the underlying mechanism and did this test more than once (Kim, 2003). Asendorpf et al. (2002) raised concerns about the ethical use of implicit measures as they reveal participants' involuntary answers and are not under their own control. This data, therefore, needs to be handled carefully, and sufficient data security should be implemented.

Investors could use this measure to uncover the implicit attitudes and morality traits of founders or companies' top management (e.g., CEO). They could assess the morality of these managers and draw conclusions for their investment decisions. For example, investing only in honest (or moral) founder teams potentially implies a transparent and more honest environment, exchange of information, and prospective company success.

Policymakers could benefit from an effective tool to measure peoples' moral self-image and implicit attitudes to understand their concerns and reasonings. This could be applied to effectively nudge them toward more sustainable behavior by implementing reasonable restrictions and policies. Additionally, this could affect elections and politics, as voters feel better understood by politicians.

6. Conclusion

An effective tool to measure people's moral self-image provides valuable insights, extending the research in behavioral economics and moral balancing (Mazar & Zhong, 2010).

This thesis explored the effectiveness of the GNAT by examining the convergent validity between the GNAT and the explicit moral self-image questionnaire. It discovered that a lower correlation between implicit and explicit measures, thus a lower convergent validity, was obtained if a set of only 20 stimuli words were repeated four times in each block (group A). This was explained by the participants' learning effects throughout the experiment, decreasing the convergent validity and effectiveness of the GNAT. For this reason, this thesis postulates using different stimuli without repetition instead, to avoid learning effects. This approach yielded a higher convergent validity (correlation) between implicit and explicit measures of the moral self-image, indicating higher effectiveness of the GNAT (group B).

These findings are crucial to effectively measuring the moral self-image, which is essential later on to investigate and understand its changes when aiming to predict moral behavior. This knowledge can be used for practical applications in behavioral economics, such as corporate management or policymaking.

Furthermore, this thesis contributes to research with its experimental design and analysis of the GNAT, including the adapted exclusion criteria, a sample correction (of all perfect response rates), and a first attempt to analyze the results based on reaction times (see Appendix V.IV Implicit Moral Self-Image Score Based on Reaction Times). This offers promising approaches for future research, which should particularly focus on a joint analysis of reaction times and response rates, experimental adjustments of the GNAT (e.g., changing the block order), and measuring the participant's moral self-image over a longer timeframe. Further research is necessary to validate the power of the GNAT to predict (e.g., consistent or balancing) behavior before bringing it into practical use in consumer behavior, for instance, by nudging people toward purchasing a more expensive but sustainable product.

However, the GNAT yields promising validity and insights into understanding people's implicit moral self-image and actions. This can be a powerful tool in the future, especially since it can be conducted entirely online, which helps assess the moral self-image in daily life (outside a laboratory environment).

Next time you order something online, use non-recyclable packaging, choose non-organic products, do not separate food waste appropriately, or book a plane ticket, remember the effect of these decisions on the environment and ultimately on your moral self-image, which influences future behavior.

References

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, *83*(2), 380–393. https://doi.org/10.1037/0022-3514.83.2.380

Ashford, R. D., Brown, A. M., & Curtis, B. (2019). The Language of Substance Use and Recovery: Novel Use of the Go/No–Go Association Task to Measure Implicit Bias. *Health Communication*, *34*(11), 1296–1302. https://doi.org/10.1080/10410236.2018.1481709

Asuero, A. G., Sayago, A., & González, A. G. (2006). The Correlation Coefficient: An Overview. *Critical Reviews in Analytical Chemistry*, *36*(1), 41–59. https://doi.org/10.1080/10408340500526766

Barkan, R., Ayal, S., & Ariely, D. (2015). Ethical dissonance, justifications, and moral behavior. *Current Opinion in Psychology*, *6*, 157–161. https://doi.org/10.1016/j.copsyc.2015.08.001

Bartels, J. M., & Schoenrade, P. (2022). The Implicit Association Test in Introductory Psychology Textbooks: Blind Spot for Controversy. *Psychology Learning & Teaching*, *21*(2), 113–125. https://doi.org/10.1177/14757257211055200

Bassett, J. F., & Dabbs, J. M. (2005). A portable version of the go/no-go association task (GNAT). *Behavior Research Methods*, *37*(3), 506–512. https://doi.org/10.3758/BF03192721

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, *76*(2), 169–217. https://doi.org/10.1086/259394

Blasi, A. (1993). The Development of Identity: Some Implications for Moral Functioning. In G. G. Noam & T. E. Wren (Eds.), *Studies in contemporary German social thought. Moral self: Working conference : Revised papers* (pp. 99–122). MIT Press.

Boldero, J. M., Rawlings, D., & Haslam, N. (2007). Convergence between GNAT-assessed implicit and explicit personality. *European Journal of Personality*, *21*(3), 341–358. https://doi.org/10.1002/per.622

Brownstein, M., Madva, A., & Gawronski, B. (2020). Understanding Implicit Bias: Putting the Criticism into Perspective. *Pacific Philosophical Quarterly*, *101*(2), 276–307. https://doi.org/10.1111/papq.12302

Buhlmann, U., Teachman, B. A., & Kathmann, N. (2011). Evaluating implicit attractiveness beliefs in body dysmorphic disorder using the Go/No-go Association Task. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(2), 192–197. https://doi.org/10.1016/j.jbtep.2010.10.003

Carlson, K. D., & Herdman, A. O. (2012). Understanding the Impact of Convergent Validity on Research Results. *Organizational Research Methods*, *15*(1), 17–32. https://doi.org/10.1177/1094428110392383

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cornelissen, G., Bashshur, M. R., Rode, J., & Le Menestrel, M. (2013). Rules or consequences? The role of ethical mind-sets in moral dynamics. *Psychological Science*, *24*(4), 482–488. https://doi.org/10.1177/0956797612457376

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349–354. https://doi.org/10.1037/h0047358

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*(2), 163–170. https://doi.org/10.1111/1467-9280.00328

Dabbs, J. M., Bassett, J. F., & Dyomina, N. V. (2003). The palm IAT: A portable version of the implicit association task. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *35*(1), 90–95. https://doi.org/10.3758/BF03195500

de Houwer, J. (2006). What are Implicit Measures and Why are We Using Them? In R. Wiers & A. Stacy (Eds.), *Handbook of Implicit Cognition and Addiction* (pp. 11–28). SAGE Publications, Inc. https://doi.org/10.4135/9781412976237.n2

de Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

de Houwer, J. (2001). A Structural and Process Analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*(6), 443–451. https://doi.org/10.1006/jesp.2000.1464

Doherty, K., & Schlenker, B. R. (1991). Self-Consciousness and Strategic Self-Presentation. *Journal of Personality*, *59*(1), 1–18. https://doi.org/10.1111/j.1467-6494.1991.tb00765.x

Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? *Personality & Social Psychology Bulletin*, *36*(12), 1618–1634. https://doi.org/10.1177/0146167210385922

Erat, S., & Gneezy, U. (2012). White Lies. *Management Science*, *58*(4), 723–733. https://doi.org/10.1287/mnsc.1110.1449

Erikson, E. H. (1964). *Insight and responsibility: Lectures on the ethical implications of psychoanalytic insight* ([Norton pbk. ed.]). W.W. Norton.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-50372-0

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition. Research: Their meaning and use. *Annual Review of Psychology*, *54*, 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Ferguson, R. (2018). *How flexible is morality? A test of the moral credits model of moral balancing* [Doctoral dissertation, Australian Catholic University].

Festinger, L. (1957). *A theory of cognitive dissonance* ([Renewed 1985 by author]). Stanford University Press.

Fischer, J., Dyball, R., Fazey, I., Gross, C., Dovers, S., Ehrlich, P. R., Brulle, R. J., Christensen, C., & Borden, R. J. (2012). Human behavior and sustainability. *Frontiers in Ecology and the Environment*, *10*(3), 153–160. https://doi.org/10.1890/110079

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, *4*(2), 195–202. https://doi.org/10.1037/h0023552

Friese, M., Bluemke, M., & Wänke, M. (2007). Predicting voting behavior with implicit attitude measures: The 2002 German parliamentary election. *Experimental Psychology*, *54*(4), 247–255. https://doi.org/10.1027/1618-3169.54.4.247

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley New York.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Jacobsen, C., Fosgaard, T. R., & Pascual-Ezama, D. (2018). WHY DO WE LIE? A PRACTICAL GUIDE TO THE DISHONESTY LITERATURE. *Journal of Economic Surveys*, *32*(2), 357–387. https://doi.org/10.1111/joes.12204

Johnston, M. E., Sherman, A., & Grusec, J. E. (2013). Predicting moral outrage and religiosity with an implicit measure of moral identity. *Journal of Research in Personality*, *47*(3), 209–217. https://doi.org/10.1016/j.jrp.2013.01.006

Jordan, J., Leliveld, M. C., & Tenbrunsel, A. E. (2015). The Moral Self-Image Scale: Measuring and Understanding the Malleability of the Moral Self. *Frontiers in Psychology*, *6*, 1878. https://doi.org/10.3389/fpsyg.2015.01878

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality & Social Psychology Bulletin*, *37*(5), 701–713. https://doi.org/10.1177/0146167211400208

Kadlec, H. (1999). Statistical properties of d' and β estimates of signal detection theory. *Psychological Methods*, *4*(1), 22–43. https://doi.org/10.1037/1082-989X.4.1.22

Kim, D. Y. (2003). Voluntary Controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly*, *66*(1), 83. https://doi.org/10.2307/3090143

Lee, Y. H., & Hsieh, G. (2013). Does slacktivism hurt activism? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 811–820. https://doi.org/10.1145/2470654.2470770

Macmillan, N. A. (2002). Signal Detection Theory. In H. Pashler (Ed.), *Stevens' Handbook of Experimental Psychology: Volume 4: Methodology in Experimental Psychology. THIRD EDITION* (pp. 43–90). John Wiley & Sons, Inc. https://doi.org/10.1002/0471214426.pas0402

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

Mazar, N., & Zhong, C. B. (2010). Do green products make us better people? *Psychological Science*, *21*(4), 494–498. https://doi.org/10.1177/0956797610363538

Melé, D., & Cantón, C. G. (2014). The Homo Economicus Model. In *Human Foundations of Management* (pp. 9–29). Palgrave Macmillan UK. https://doi.org/10.1057/9781137462619_2

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral Self-Licensing: When Being Good Frees Us to Be Bad. *Social and Personality Psychology Compass*, *4*(5), 344–357. https://doi.org/10.1111/j.1751-9004.2010.00263.x

Miller, D. T., & Effron, D. A. (2010). Chapter Three - Psychological License: When it is Needed and How it Functions. In *Advances in Experimental Social Psychology* (pp. 115–155, Vol. 43). Academic Press. https://doi.org/10.1016/S0065-2601(10)43003-8

Monin, B., & Jordan, A. H. (2009). The Dynamic Moral Self: A Social Psychological Perspective. In *Personality, Identity, and Character* (pp. 341–354). Cambridge University Press. https://doi.org/10.1017/CBO9780511627125.016

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), 33–43. https://doi.org/10.1037/0022-3514.81.1.33

Mullen, E., & Monin, B. (2016). Consistency Versus Licensing Effects of Past Moral Behavior. *Annual Review of Psychology*, *67*, 363–385. https://doi.org/10.1146/annurev-psych-010213-115120

Nisan, M. (1990). Moral Balance: A Model of How People Arrive at Moral Decisions. In T. E. Wren (Ed.), *The Moral Domain: Essays in the Ongoing Discussion between Philosophy and the Social Sciences* (pp. 283–314). The MIT Press.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, *19*(6), 625–666. https://doi.org/10.1521/soco.19.6.625.20886

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Perkins, B. G., Podsakoff, N. P., & Welsh, D. T. (2024). Variance in Virtue: An Integrative Review of Intraindividual (Un)Ethical Behavior Research. *Academy of Management Annals*, *18*(1), 210–250. https://doi.org/10.5465/annals.2022.0057

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *The British Journal of Social Psychology*, *44*(Pt 1), 29–45. https://doi.org/10.1348/014466604X23491

Perugini, M., & Leone, L. (2009). Implicit self-concept and moral action. *Journal of Research in Personality*, *43*(5), 747–754. https://doi.org/10.1016/j.jrp.2009.03.015

Ploner, M., & Regner, T. (2013). Self-image and moral balancing: An experimental analysis. *Journal of Economic Behavior & Organization*, *93*, 374–383. https://doi.org/10.1016/j.jebo.2013.03.030

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science*, *20*(4), 523–528. https://doi.org/10.1111/j.1467-9280.2009.02326.x

Sachdeva, S., Jordan, J., & Mazar, N. (2015). Green consumerism: moral motivations to a sustainable future. *Current Opinion in Psychology*, *6*, 60–65. https://doi.org/10.1016/j.copsyc.2015.03.029

Schimmack, U. (2021). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *16*(2), 396–414. https://doi.org/10.1177/1745691619863798

Schlegelmilch, B. B., & Simbrunner, P. (2019). Moral licensing and moral cleansing applied to company-NGO collaborations in an online context. *Journal of Business Research*, *95*, 544–552. https://doi.org/10.1016/j.jbusres.2018.07.040

Schnabel, K., Asendorpf, J., & Greenwald, A. (2007). Using Implicit Association Tests for the Assessment of Implicit Personality Self-Concept. In G. J. Boyle, G. Matthews, & H. Saklofske (Eds.), *Handbook of Personality Theory and Testing*. SAGE Publications, Inc. https://doi.org/10.4135/9781849200479.n24

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-Serving Justifications. *Current Directions in Psychological Science*, *24*(2), 125–130. https://doi.org/10.1177/0963721414553264

Teachman, B. A. (2007). Evaluating implicit spider fear associations using the Go/No-go Association Task. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(2), 156–167. https://doi.org/10.1016/j.jbtep.2006.10.006

Williams, B. J., & Kaufmann, L. M. (2012). Reliability of the Go/No Go Association Task. *Journal of Experimental Social Psychology*, *48*(4), 879–891. https://doi.org/10.1016/j.jesp.2012.03.001

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. https://doi.org/10.1037/0033-295X.107.1.101

Zhong, C. B., Ku, G., Lount, R. B., & Murnighan, J. K. (2010). Compensatory Ethics. *Journal of Business Ethics*, *92*(3), 323–339. https://doi.org/10.1007/s10551-009-0161-6

Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science (New York, N.Y.)*, *313*(5792), 1451–1452. https://doi.org/10.1126/science.1130726