

Stockholm School of Economics  
Department of Economics  
5350 Master's thesis in economics  
Academic year 2022-2023

## **Can Psychological Priming Affect Self-Rated Product Desirability? Preregistered Experimental Evidence from 1274 Individuals**

Shahin Eidinejad (42124)

### **Abstract**

Conventional economic theory often assumes that preferences are exogenously fixed and remain constant over time. A growing literature is trying to relax this assumption by endogenizing preferences and shedding light on how they are determined and potentially affected by external factors such as culture, environment, or identity. In this paper, I contribute to this strand of literature by using insights from psychology to examine whether psychological priming affects preferences. In particular, I test whether reading an unethical text affects preferences, insofar as they manifest through a self-rated desirability rating, for hygiene products. To do this, I conducted two large-scale, identical, preregistered (<https://osf.io/jphmb/>) experiments on separate samples with N=458 and N=816, respectively. Contrary to previous findings, I find no reliable evidence that reading an unethical text affects self-rated desirability for hygiene products, suggesting that this priming likely has no effect on preferences and that past findings may be false positives.

Keywords: Preferences, Priming, Reliability  
JEL: D9, D91

Supervisor: Anna Dreber Almenberg  
Date submitted: December 5, 2022  
Date examined: December 15, 2022  
Discussant: Emma Hamre

Acknowledgements: I thank Anna for great supervision and the Knut and Alice Wallenberg Foundation for financing.

# Table of Contents

1. Introduction .....	1
2. Past research.....	5
2.1. Priming in Economics.....	5
2.2. Moral Transgressions and Cleanliness .....	8
2.3. Moral purity and increased desirability for cleansing products.....	10
3. The Experiment .....	14
3.1. General information .....	14
3.2. Experimental Design.....	14
3.3. Participants and Power Calculations.....	21
3.3.1. First Experiment.....	21
3.3.2. Second Experiment .....	24
3.4. Statistical Tests .....	24
4. Results .....	25
4.1. Experiment 1 .....	25
4.1.1. Descriptive Statistics .....	25
4.1.2. Findings.....	27
4.2. Experiment 2.....	29
4.2.1. Descriptive statistics.....	29
4.2.2. Findings.....	31
5. Discussion .....	32
5.1. Endogenous preferences .....	32
5.2. Relation to past findings .....	33
5.3. Different findings in the experiments .....	34
5.4. Participant attentiveness .....	35
6. Conclusion.....	36
Bibliograhya .....	37
Appendix .....	42
Treatment Survey Prolific.....	42

# 1. Introduction

One of the fundamental concepts employed in economics is rational choice theory, which offers a framework for analysis in which agents make rational choices in order to maximize utility based on their tastes or preferences. While this framework is a powerful tool for accurately analyzing a wide range of phenomena, it has a limitation in that it sheds little light on what determines preferences or how they might change with time and context. Traditional economics often assumes, in a simplifying manner, that agents are endowed with exogenously fixed preferences that remain constant, and that behavioral changes are driven by changes in factors such as incentive structures or beliefs about probabilities (Stigler and Becker, 1977)

Economics has a growing body of literature exploring the determinants of preferences (see Bernheim et al., 2021; Dietrich and List, 2013; Cohn and Maréchal, 2016). With some evidence suggesting that economic preferences may be influenced by factors such as culture (Fehr and Hoff, 2011; Henrich et al., 2001) or environmental cues (Cohn et al., 2015). Improving our understanding of preference determinants can contribute to the advancement of economic theory and a more thorough understanding of general decision-making and human behavior. As a result, further research in this field can yield results with fruitful implications for economics and other disciplines studying behavior and decision-making.

In this paper, I use insights derived from psychology to study how preferences, insofar as they manifest through a self-rated measure of product desirability, can be influenced by an exogenous psychological trigger. Previous work has found suggestive evidence that psychological priming, the activation of mental concepts through external cues, can

increase self-rated product desirability (see e.g., Trakulpipat et al., 2021). I focus on one particular finding in the priming literature where there is suggestive evidence that reading an unethical text can increase self-rated desirability for cleansing products, in order to study how preferences can be affected by priming.

In their seminal paper, Zhong and Liljenquist (2006) hypothesized that there is a physio-psychological link between physical and moral disgust, such that a threat to moral purity would compel a need for physical cleansing (Macbeth effect). In one of their experiments (study 2), they found that participants (N=27) who experienced an *implicit* threat to moral purity by hand-copying an unethical story rated cleansing products as more desirable than participants who copied an ethical one. This provides suggestive evidence that reading an unethical text might elicit feelings of physical disgust that necessitate physical cleansing, and this manifests as higher self-rated desirability for cleansing products. However, replications of Zhong and Liljenquist (2006) (study 2) with larger sample sizes (N=153; N=156; N=286) failed to replicate their findings (Earp et al., 2014), casting some doubt on the reliability of their results.

Similar studies have provided suggestive evidence for a link between an *explicit* threat to moral purity through the enactment of immoral behavior and increased desirability for cleansing products directly related to the morally “dirty” body part. Lee and Schwarz (2010) found that participants (N=87) who experienced an explicit threat to moral purity by lying through a voicemail (mouth) rated mouthwash as more desirable compared to participants who transgressed by lying in writing through an email (hands). Similarly, participants who lied in an email rated hand sanitizer as more desirable than subjects who lied through voicemail. Suggesting that the potentially existing link between threats to

moral purity and increased desirability for physical cleansing products might be specific to the sensory-motor modality involved.

Despite the past efforts to disentangle the potential physio-psychological link between physical and moral disgust and its relation to cleansing product desirability, there are remaining gaps in the literature. Firstly, to my knowledge, no paper in the extant literature has investigated whether there is a link between an *implicit* threat to moral purity and an increased preference for cleansing products *directly* related to the morally “dirty” body part. Neither Zhong and Liljenquist (2006) nor the replication attempts (Earp et al., 2014) include cleansing products directly related to the morally “dirty” body part. Thus, it is still unclear whether an implicit threat to moral purity by, for example, reading a short immoral story will increase preferences for cleansing products used to cleanse the morally “tainted” body part.

Secondly, a pressing drawback with both Zhong and Liljenquist's (2006) and Lee and Schwarz's (2010) work is that they do not use neutral control groups in their experiments. Thus, their experimental designs do not allow for accurate inference regarding the hypothesized link between moral and physical disgust. We cannot, for example, accurately infer whether the effect found in Zhong and Liljenquist (2006) is driven by the unethical treatment, the ethical one, or a combination of both.

In this paper, I address these two gaps in the literature. The primary purpose of this paper is thus to examine whether an *implicit* threat to moral purity by reading an immoral story increases self-rated desirability for cleansing products used to cleanse the morally “dirty” body part. To this end, I run two identical experiments where I employ an improved experimental design with a neutral prompt, unlike in past research, to eliminate potential bias induced by having participants read an ethical prompt. I will also use the

improved experimental design to serve a secondary purpose, which is to provide an accurate reexamination of whether an implicit threat to moral purity increases preferences for cleansing products *not* directly related to the morally “dirty” body part, as suggested by the findings of (Zhong and Liljenquist, 2006).

The contributions of my paper are fourfold. First, the current paper contributes to the growing behavioral economics literature studying preferences through priming techniques (see e.g., Cohn and Maréchal, 2016; Callen et. al., 2014). My experiments contribute to the behavioral economics literature by giving insights into how preferences, insofar as they exhibit through self-rated desirability for products, can be affected by an exogenous psychological trigger. The findings in my paper can potentially guide further development of economic models and be used to examine the accuracy of conventional economic theory.

Second, this is the first paper to explore whether an implicit threat to moral purity increases preferences for cleansing products used to cleanse the morally “dirty” body part. Thus, this study contributes to the psychology literature by providing novel evidence exploring the potential link between moral and physical disgust. My work can also elucidate whether cleansing-related behaviors seen in religious rituals, common culture, and natural language use as “dirty mouth” reflect a link between moral and physical disgust that cause the manifestation of such phenomena.

Third, I contribute to the literature by further developing past experimental designs and removing a potential source of bias that could stem from not having a neutral control group. A more precise understanding of potential cleansing effects is of interest to researchers trying to develop a better understanding of preference formation, cognition, and the psyche in general. The experimental design in this paper can also be extended to

examine (and reexamine) other cleansing effects and, with a slight modification, be applied to the case where there is an enactment of immoral behavior, like that of Lee and Schwarz (2010).

Fourth, to my knowledge, the current paper provides evidence on cleansing effects and their relation to increased product desirability from the largest sample up to date. The combined number of participants in my experiments is more than 45 times as many as the participants in Zhong and Liljenquist (2006). Thus, the findings can be used to assess the universality of cleansing effects, the reliability of previous findings, and guide potential future power calculations. To my knowledge, I am also the first to explore cleansing effects with a preregistered analysis plan, which lowers “researcher degrees of freedom” (Simmons et al., 2011), i.e., undisclosed flexibility in data collection and analysis, and potentially contributes to the development of improved scientific practices.

The remainder of my paper is structured as follows. First, I review some of the past literature related to my paper and outline my hypotheses. Second, I describe the experimental design and statistical procedures. Last, I present the results and discuss the implications of my findings.

## **2. Past research**

### **2.1. Priming in Economics**

A growing body of literature explores the effects psychological priming can have on economic preferences. In this section, I outline some of the past work on this topic.

Cohn et al. (2015) studied countercyclical risk aversion by priming financial professionals (N=162) with either a stock market “boom” or a market “bust”. The subjects were randomly assigned to the two treatments where those in the “boom” prime were

exposed to an upward trending stock market chart whereas those in the “bust” group faced the opposite situation. Participants were then to decide how much of their endowment they wanted to invest in a risky asset. They found a statistically significant difference in the portion invested in the risky asset between the two groups, where those in the “boom” prime had a larger risky share. Cohn et al.’s (2015) findings thus provide some suggestive evidence, elicited through priming techniques, for the existence of countercyclical risk aversion among financial professionals.

In a similar vein, Callen et al. (2014) studied the relationship between violence and risk preferences in a large sample of Afghan individuals (N=816) who had been exposed to violence. The participants were primed with either a “Fear”, “Happy” or “Neutral” prime. The priming involved the participants describing one event in the past that caused them fear, happiness, or a neutral event. They found that participants primed with the “Fear” condition exhibited an increased preference for certainty in an economic game compared to the other groups. Their findings provide suggestive evidence that trauma exposure priming can impact risk preferences.

Others have studied how religious or identity priming can affect cheating and altruism. Shariff and Norenzayan (2007) examined whether religious priming affects dictator game-giving. The subjects (N=50) in their study 1 were split into two groups where one of the groups was primed by unscrambling ten five-word sentences where five of them contained religious target words such as “God” or “spirit” and the other half received no prime. Then the subjects played a standard dictator game where they were endowed with ten one-dollar coins and could decide to either keep all coins or donate a portion or all of them to an anonymous receiver. Shariff and Norenzayan (2007) found that those in the religious prime allocated more money to the anonymous receivers than those in the



control group. This suggests that religious priming could potentially impact dictator game giving, which can be a potential proxy for altruism.

Cohn, Maréchal, and Noll (2014) studied how identity-priming inmates from a maximum-security prison (N=182) by making their criminal identity more salient affected their cheating rates in an economic game. They found that the identity prime made the subjects more prone to cheating than those who had not been primed with their criminal identity. This provides suggestive evidence that personal identity perception can affect levels of honesty.

In general, priming in economics has been used to study the impact it can have on preferences. It allows for a simple way to introduce exogenous variation while, on average, keeping all else constant. Therefore, it is an efficient tool for studying preference formation. More broadly, priming has been a popular tool in past research, particularly in psychology, and have thought to have produced very reliable results. However, after several large-scale replication projects (e.g., Open Science Collaboration, 2015), the reliability of many, but far from all, priming studies has become a contested issue (e.g., Nature, 2019).

Nonetheless, the literature above represents a small sample of the studies in economics that use priming to study its effects on preferences. I highlighted these particular papers as they represent past work published in some of the most influential economics and psychology journals (e.g., American Economic Review and Psychological Science). For a more comprehensive review, I point the reader to the work of Cohn and Maréchal (2016).

In my paper, I use priming to study how it can affect self-rated product desirability preferences. My treatment primes are based on the findings in a particular strand of

literature in psychology which examines the relationship between moral transgressions and physical cleanliness. In the sections below, I review the related literature in psychology.

## **2.2. Moral Transgressions and Cleanliness**

Considerable amounts of research have explored the potential links between moral and physical disgust. Schnall et al. (2008b) analyzed whether being exposed to physical dirtiness impacts the judgment of the severity of others' moral transgressions. They explored whether participants (N=127) who had experienced a mild odor (treatment group 1) or a strong odor (treatment group 2) had a more severe moral judgment than participants who had not experienced a bad smell (experiment 1). They found that both treatment groups had a more severe moral judgment than those in the control group, but there was no difference in judgment between the two treatments.

In a second experiment, Schnall et al. (2008b) assessed whether participants (N = 43) who were led into and seated in an unkempt room (treatment) had more severe moral judgment than participants who experienced a clean room (control). Similar to the findings in their first experiment, they found that extraneously induced disgust made moral judgments more severe.

Gollwitzer and Melzer (2012) studied the link between moral transgressions and physical cleanliness in a different context. They assessed whether participants (N=70) that had played video games involving violence subsequently rated cleansing products as more desirable than a separate control group (N=55) who only rated product desirability. They found that subjects in the video game treatment rated cleansing products more desirable than those in the control group.

Others have explored the psychological consequences of physical cleansing. For example, Schnall et al. (2008a) (experiment 2) found that participants (N=44) who physically cleansed after watching a 3-minute-long clip, which elicited strong feelings of disgust, were more likely to judge moral actions as less ethically wrong than participants who had not cleansed physically (control group). In a similar vein, Xu et al. (2012), using a 2 (good vs. bad luck)  $\times$  2 (wash vs. not washing hands) between participants design, found that participants (N=59) perceived the influence of one's good or bad luck as a consequence of e.g., a good or bad financial decision was removed by cleansing one's hands.

The above findings present a small sample of the plethora of research effort that explores the vast array of potential psychological effects of cleanliness. I outline these particular papers because they represent some of the most closely related literature to my work. For a more comprehensive literature review, I refer the reader to Trakulpipat et al. (2021). However, the reader should note that a majority of past research exploring cleanliness effects has small sample sizes, which can make the results largely unreliable even when statistically significant at conventional levels (Gelman and Carlin, 2014).

Moreover, to my knowledge, no past work on this topic has been preregistered, increasing the probability of false positive results due to “researcher degrees of freedom” (Simmons et al., 2011) such as “garden of the forking paths” (Gelman and Loken, 2013). Forking is a largely unintentional process on behalf of researchers that can happen in cases where there are degrees of freedom in, for example, choices of covariates, statistical tests, and subgroup analyses, which can lead the researchers to make choices that favor the tested hypotheses. This problem is sometimes referred to as “p-hacking” when researchers consciously seek to find statistically significant results.

Given the two problems outlined and the ongoing debate regarding the strength of the evidence for general cleansing effects (e.g., Ropovik et al., 2021), I caution the reader to not over-interpret the papers presented in the section above or in the literature review by Trakulpipat et al. (2021) as definitive. I argue a more sober view of them is as mildly suggestive of the various ways that cleanliness effects might manifest.

My paper mainly focuses on one specific way in which this phenomenon potentially manifests: how implicit threats to morality affect self-rated desirability for cleansing products. In the next section, I review the main articles that my work builds on.

### **2.3. Moral purity and increased desirability for cleansing products**

My paper primarily builds on the work of Zhong and Liljenquist (2006) (study 2) and Lee and Schwarz (2010). In this section, I outline their work and replication attempts of it (Earp et al., 2014; Schaefer et al., 2015) in detail.

Zhong and Liljenquist (2006) hypothesized that a threat to moral purity would elicit feelings similar to physical disgust and necessitate physical cleansing (Macbeth effect). To test their hypothesis, they conducted a two-level, single factor (ethical vs. unethical) between participants' experiment. The participants (N=27) were randomly assigned to one of the two groups and told that they were engaging in two unrelated tasks. First, they hand-copied an ethical or unethical story. Second, the participants rated the desirability of 10 products, 5 of which were cleansing-related; however, the authors did not specify exactly how the rating procedures were conducted. They found that participants who hand-copied the unethical story rated the desirability of the cleansing products statistically significantly higher than those who copied the ethical one. Therefore, Zhong and Liljenquist (2006) concluded that their initial hypothesis, that a threat to moral purity

requires physical cleansing as measured by increased desirability for cleansing products, was supported by their findings

However, their paper is not without its drawbacks. Given the small sample size of 27 participants, the associated statistical power and the strikingly large magnitude their results (Cohen's  $d=1.08$ ), and that they are the first to provide evidence for the Macbeth effect, there are strong reasons to suspect that their findings might be a false positive. This suspicion is further strengthened by the failure of their results to replicate in several replications with higher statistical power (Earp et al., 2014). Their results failed to replicate in samples of participants from the United Kingdom ( $N=153$ ), the United States ( $N=156$ ), the same country in which they conducted their original experiment, and India ( $N=286$ ) (Earp et al., 2014). A second drawback with Zhong and Liljenquist's (2006) (study 2) and the subsequent replication attempts of it is that none of the cleansing products included in the desirability rating were related to the morally "dirty" body part. This is a particularly pressing drawback as later research has found evidence suggesting that the potentially increased preference for cleansing products after a threat to moral purity might be specific to the sensory-motor modality involved in the transgression (Lee and Schwarz, 2010; Schaefer et al., 2015).

In their paper, Lee and Schwarz (2010) built upon the work of Zhong and Liljenquist (2006). From natural language use such as a "dirty mouth" or "dirty hands", Lee and Schwarz (2010) hypothesized that the sensory-motor modality involved in an ethical transgression might prominently figure in the embodiment of moral purity. They thus tested the hypothesis of whether an explicit threat to moral purity would affect preferences for cleansing products directly related to the morally "dirty" body part. To test their hypothesis, they used a 2 (modality: mouth vs. hands)  $\times$  2 (ethical vs. unethical)

experiment where the participants (N=87) were to either perform an immoral act with their mouths, by lying through a voicemail or with their hands by lying in an email they send. The primary factor that differentiates their study from that of Zhong and Liljenquist (2006) (study 2) is that the participants enact immoral behavior, which poses an explicit threat to moral purity, whereas the participants in Zhong and Liljenquist (2006) (study 2) hand-copied an unethical story and only experienced an implicit threat to moral purity. After the treatment, the participants rated the desirability of several products as a part of an ostensible marketing survey. However, the authors did not specify they how conducted the rating procedures. In the list of products, they included mouthwash and hand sanitizer to test their hypothesis that the link between moral and physical disgust is specific to the sensory-motor modality involved in the transgression.

Lee and Schwarz (2010) conclude that their hypothesis was confirmed as they found evidence that participants who enacted an immoral behavior by lying through phone (mouth) evaluated mouthwash more positively than those who lied in writing, whereas participants that lied in writing by sending an email (hands) evaluated hand sanitizer more positively than those who lied through phone. While their results provide some suggestive evidence that confirms their hypothesis, their paper is not without its drawbacks. In particular, given their sample size and that the study was not preregistered, giving many researcher degrees of freedom, there is a rather high likelihood that their findings, if evaluated on their own, are a false positive. However, a replication of their results by Schaefer et al. (2015) with 35 participants has successfully replicated the findings of Lee and Schwarz (2010) for the desirability of mouthwash. Thus, even though both papers have rather small sample sizes, given that both of their findings present similar results,

there is less of a reason to suspect that their findings are false positives. Although, the current evidence is still rather far from being conclusive.

A particularly pressing drawback with Zhong and Liljenquist (2006) and Lee and Schwarz (2010) is that they do not use neutral control groups in their experiments. The control group in Zhong and Liljenquist (2006) read an ethical story that could potentially affect their preferences for cleansing products. Similarly, Lee and Schwarz (2010) examined whether preferences for cleansing products differed between those who lied in writing, with their hands, compared to those who lied with their mouth. Thus, in both papers, it is unclear whether the statistically significant findings are driven by the unethical treatment or by the prompts used for the 'control' groups. While I am aware that both their papers present bar charts attempting to visualize what drives the effect, I argue that much weight cannot be attached to the content of the graphs for two reasons. First, the sample sizes in both their papers are comparatively small causing the evidence in general to be inconclusive. Secondly, neither of their studies was preregistered, allowing for many "researcher degrees" of freedom in constructing the graphs and conducting the analysis.

In light of the past findings, I set out with the primary hypothesis that an implicit threat to moral purity elicited by reading an immoral story will increase self-rated desirability for cleansing products used to clean the morally "dirty" body part. As a corollary, I set out with the secondary hypothesis that an implicit threat to moral purity will not cause a change in self-rated desirability for other products.

In the subsequent section, I outline the experiment to test my hypotheses.

## **3. The Experiment**

### **3.1. General information**

To reliably test my hypotheses, I conduct two identical preregistered (<https://osf.io/jphmb/>) experiments on separate populations. I conduct the first experiment on students at the Stockholm School of Economics and Stockholm University, then once the first experiment is finished, I replicate it with participants from Prolific. The only differences between the two experiments are (1) that the participants in Prolific receive a one Pound (£) payment for completing the study, whereas the students in Stockholm receive no financial compensation for GDPR compliance reasons, and (2) some minor changes in wording in the introductory page.

In the coming sections, I describe the experimental design and procedures in more detail.

### **3.2. Experimental Design**

I conduct a two-level single factor (amoral vs. immoral) between participants experiments randomized (50/50) at the subject level. The participants receive a link to an online survey<sup>1</sup>, powered by Qualtrics, where they engage in 2 seemingly unrelated tasks. Firstly, a short reading task, and secondly, a product desirability rating. Before starting the reading and desirability survey tasks, the participants fill out basic demographic information about themselves. At this stage, the participants are also informed about the approximate completion time for the survey and that all answers are anonymous to avoid conformity bias.

---

<sup>1</sup> Full survey is attached in the appendix.



After filling in the basic information, participants engage in the reading task. In the immoral prime, participants read a short immoral story of someone lying to their colleague through the phone in a voicemail and are asked to answer three questions about the story. Participants in the amoral prime read a short amoral story and are asked to answer three questions about it.

Participants in the immoral prime conditions read the following story:

“Two years ago, when I was an associate at a law firm, I was coming up for promotion against another hard-working associate. For several months, my colleague had been working on a major case that would ultimately make or break my colleague’s career at the firm. However, my colleague could not find a very important document, without which it was highly unlikely that my colleague would have sufficient evidence to win the major case. The night before the trial, as I was walking through the office, I found the very important document that my colleague was desperately in need of. I called my colleague by phone and left a voicemail where I lied and told my colleague that the document was nowhere to be found in the office, knowing that my promotion would be secured.”

The morally “dirty” body part is the mouth since the protagonist lies verbally through a voicemail.

Participants in the amoral group read a truncated version of the prompt used in the immoral prime, where the two last sentences have been removed. They thus read the following story:

“Two years ago, when I was an associate at a law firm, I was coming up for promotion against another hard-working associate. For several months, my colleague had been working on a major case that would ultimately make or break my colleague’s career at the firm. However, my colleague could not find a very important document, without which it was highly unlikely that my colleague would have sufficient evidence to win the major case.”

The reading prompts are motivated by the prompts used by Zhong and Liljenquist (2006) and Lee and Schwarz (2010) but have two key differences from them. First, I use immoral vs. amoral prompts instead of unethical vs. ethical ones. The rationale for this choice is that it mitigates potential bias induced by the possible impact reading an ethical prompt has on participants’ preferences for the products included in the desirability rating. If I were to detect an effect using ethical vs. unethical prompts, I would not be able to accurately infer whether it was caused by the unethical prompt, the ethical one, or a combination of both. Thus, I argue that an amoral vs. immoral prime is better suited to make inferences regarding my hypotheses.

Second, the prompts do not reveal the gender of the colleague to reduce potential bias that could be induced by this information, thus allowing for a potentially more accurate estimation of the causal effect. Except for the two differences specified, the prompts in my paper are broadly similar to those used by Zhong and Liljenquist (2006) and Lee and Schwarz (2010).

During the reading task, participants will also be asked to answer three questions related to the story. These questions are of dual purpose. One question serves as a form

of attention check. I have formulated the attention-check question so it probes whether the participants have attentively read the critical part of the prompt: the discovery of the very important document. Thus, increasing the probability that (a) I identify participants who skimmed over the critical part of the passage and (b) potentially increase the likelihood of participants attentively reading it. Only participants who correctly answer the attention-check question will be included in the statistical analysis. Although there is a chance that this could lead to selection bias as those who answer the question correctly might be systematically different from those who answer incorrectly, I argue that this possibility is improbable since all participants (N=18) in a pilot survey conducted during Spring 2022 before launching the experiment answered the question correctly, and it is not a trick question or the like. As an additional attention-check, I have enabled the forced responses option in Qualtrics for all questions in the survey. This also makes it so e.g., half-completed survey responses cannot be sent in.

The other two questions are formulated as standard personality survey questions to serve as decoys to lower the probability that the participants figure out the actual purpose of the experiment. Moreover, for the same reason, none of the three questions directly relate to the purpose of the experiment to mitigate potential social desirability bias.

The three questions asked are:

1) “From the information presented above, was the important document found?”

(Attention-check)

- Answer alternatives: yes, no
- Correct answers: yes (immoral - treatment); no (amoral - control)

2) “Who do you think deserves the promotion the most?” (Decoy 1)

- Answer alternatives: the main character, the colleague, I don’t know

3) “Hard-workers should be rewarded for their work even if they are severely disliked by their colleagues” (Decoy 2)

- Answer alternatives: yes, no, I don’t know

After the reading task, participants will be asked to rate the desirability of eight products from 1 (very undesirable) to 5 (very desirable), where 5 is the highest, as a part of the product desirability rating survey. The list of products includes mouthwash, toothpaste, hand sanitizer, shampoo, wireless headphones, eco-friendly water bottle, a vegan chocolate bar, and a reusable grocery bag. The primary hypothesis is that those in the immoral treatment group will have an increased preference for oral hygiene products. I test this conjecture by including mouthwash and toothpaste in the list of products as they are used to cleanse the morally “dirty” body part (mouth). Thus, I expect an increase in the average desirability for mouthwash and toothpaste in the immoral treatment compared to the amoral group. Hand sanitizer and shampoo are included to reexamine whether an implicit threat to morality affects the desirability of cleansing products not directly related to the morally “dirty” body part, similar to the findings of Zhong and Liljenquist (2006). I include two cleansing products to assess the effect of the treatment both in the general and sensory-motor-specific case to lower the probability of observing an effect driven by random noise in the data.

The wireless headphone, eco-friendly water bottle, vegan chocolate bar, and reusable grocery bag were included primarily as decoys to decrease the possibility that participants suspect that the purpose of the experiment is to assess how the treatment affects preferences for cleansing products. The order of appearance for the products in the desirability rating is randomized to minimize potential bias induced by order of appearance. Moreover, the desirability rating will also serve as a proxy for the attention

paid by participants in the second part of the experiment, as I exclude responses that have given the same desirability rating to all products, as this is highly suggestive of the respondent simply rushing through the rating. While there is a slight chance of this exclusion criteria leading to selection bias, I argue that it is improbable since it is highly unlikely that participants find all eight products in the rating equally desirable.

The rationale for conducting online experiments rather than an in-person lab experiment is threefold. Firstly, since online experiments are less costly and time-consuming than lab experiments, it will allow me to gather a larger sample size (Eynon et al., 2017; Reips, 2000). One of the primary drawbacks of a large portion of the past literature (Zhong and Liljenquist, 2006; Lee and Schwarz, 2010; Schaefer et al., 2015) is that they have small sample sizes, which casts doubt on the statistical power and the general validity of their results. Thus, since I am conducting online experiments, I can combat this issue by having a higher number of participants than that in past work.

Secondly, partly from having a higher number of participants, my findings will likely have higher external validity than previous literature. This is because I will likely be able to have a more diverse group of participants, instead of just recruiting some undergraduate students from one particular course at a university, as in e.g., Lee and Schwarz (2010). Moreover, since participants will be able to complete the tasks in an environment that is not as highly controlled and artificial as a lab setting, this might lead to higher external and ecological validity.

Thirdly, conducting an online experiment, where all the experimental procedures and processes will be made available on the web, will increase the transparency of my work compared to that of lab experiments. In lab experiments, many, or at least some, of the procedures are not or cannot be accurately recorded (Reips, 2000). An example of this

could be any potential non-verbal communication between the participant and experimental instructor, which are likely not documented in the instruction protocol.

Naturally, there are also drawbacks to online experiments (Eynon, 2017; Reips, 2000). The two most prominent disadvantages of online experiments are that the response rate tends to be low and attrition high. These drawbacks are particularly pressing in cases where the experiment is time-consuming and complex. I try to combat these problems by making my experimental survey short and easy to complete. It should not take more than approximately three minutes to complete and it does not include any cognitively demanding tasks.

Another drawback of online experiments is that there is no possibility for the participants to ask clarifying questions in case the instructions are unclear. To decrease the chances of the instructions being unclear and assess how long it takes for participants to complete the survey, I conducted a pilot study during Spring 2022 where no respondents (N=18) indicated that the instructions were unclear. The average self-reported completion time was 3 minutes. Thus, I conclude that the survey instructions are clear and that it takes approximately 3 minutes to complete it, on average.

Another challenge with online experiments is the probability of multiple submissions by the same participant. It seems to be quite a rare phenomenon in general (Eynon, 2017; Reips, 2000), and the chances of it happening seem to be higher in cases where participants have strong opinions about the topic and if there is financial compensation for completing the task (for an example, see e.g. Konstan et al., 2005). Regarding my experiments, I argue the chances of multiple submissions are low. Firstly, there is no reason to suspect that my experiment will raise strong opinions. Secondly, for the first experiment, there is no monetary compensation and for the second experiment,

participants in Prolific cannot submit responses to the same survey twice with their user ID. Thus, I argue that there is no pressing reason to suspect that there will be issues with multiple submissions with my experiments. Nevertheless, I have employed the “prevent multiple submission/prevent ballot-box stuffing” option in Qualtrics. While this cannot guarantee that no multiple submissions will be made, it offers one further layer of protection against it.

Lastly, there is a possibility of participants filling in half the survey, leaving and doing something else, then finishing it. Thus, potentially losing the effect from the treatment prime. I try to combat this by not allowing the participants to close down the survey and return to it.

In light of past research, I also argue that there are no strong reasons to suspect that I am sacrificing internal validity by conducting an online experiment instead of an in-lab one. Past research has consistently found that there are no stark differences when comparing results from in-person experiments to online (see e.g., Prissé and Jorrot, 2022; Arechar et al., 2018; Germine et al., 2012; Gosling et al., 2004; Krantz and Dalal, 2000), which is the most commonly used method when comparing the validity of online vs. in-person experiments. The consistency in results seems to hold particularly for short and simple experiments. Thus, I argue that the decision to run the experiment online will likely not affect the internal validity of my findings.

### **3.3. Participants and Power Calculations**

#### **3.3.1. First Experiment**

Using the software G\*power 3.1 (Faul et al., 2007), I calculated that the required sample size to achieve 80% statistical power for a standard two-tailed t-test with two independent groups, an assumed effect size of Cohen’s  $d = 0.3$  and  $\alpha = 0.005$  is 596. Thus, given the

power calculations, I aim to recruit 650 participants to account for observations that I drop from the statistical analysis.

I used the effect size for mouthwash from Lee and Schwarz (2010) as the guide for my calculations. Using their exact effect size, mean treatment  $\mu_{\text{treatment}} = 0.21$  with  $SD = 0.72$ , and mean control  $\mu_{\text{control}} = -0.26$  with  $SD = 0.94$ , yields a Cohen's  $d = 0.56$  with an  $\alpha = 0.05$ . However, given that they have a small sample size, they provide the initial findings for this particular effect, and well-known publication bias, their effect size is likely to be positively biased. Thus, I assumed a 'true' effect size of approximately 50% of the magnitude of the original findings. My assumption of the 'true' effect size is guided by (i) the findings of large-scale replication projects that have documented that the effect sizes of well-powered direct replications are approximately 50% of the original effect size (Open Science Collaboration, 2015), and (ii) a meta-analysis of the Macbeth effect which concluded that if unethical primes generate an effect on desirability for cleansing products, that it is likely to be categorized as a small one (Siev et al., 2018).

Participants consist of two groups of students: (1) all students enrolled at the Stockholm School of Economics (SSE), excluding my classmates in the MSc Economics program (2021 intake) and others who might know the topic of my work (to avoid social desirability bias), and (2) doctoral students at Stockholm University. I distribute the survey directly to their student emails in a message that includes a link to it. All SSE student email addresses are readily accessible to all students and staff at SSE, and the email addresses of Ph.D. students at Stockholm University are publicly available. Thus, I do not need to gather personal data to distribute the survey.

I gather data for the first experiment in two rounds: during Spring 2022 and Fall 2022. During Spring, I distribute the survey to all students at SSE and to Ph.D students at



Stockholm University. I distribute the survey from the beginning of April and it will remain open until 2022-05-31. I have decided to implement a sharp cutoff for data collection at a pre-specified date to limit “researcher degrees of freedom” and avoid the reliability problems associated with it. If I receive the desired number of responses during Spring, I will not conduct a second round of data collection for this experiment.

If I have not attained the desired number of responses in the Spring, I will conduct a second round of data collection in the Fall of 2022. I will then send out the survey to all newly enrolled students at SSE (first-year BSc or first-year MSc, excluding BSc students who enrolled in an MSc). I will send out the survey at the beginning of the Fall semester (mid-September) and keep it open until a preset date.

The subject pool was chosen largely for convenience reasons and to allow me to gather data while not violating GDPR laws. Nonetheless, there are several benefits to the study population. Firstly, all email addresses and contact information are readily available to me. Thus, the survey can be distributed to the participants in a time- and cost-effective manner. Secondly, since the survey can be disseminated efficiently through email, it provides a way of reaching out to a more diverse sample of students than some of the past work that only includes a small number of students from one course at one particular university, as in e.g., Zhong and Liljenquist (2006). Thirdly, having a student subject pool decreases the probability of potential problems with, for example, language barriers since all procedures are in English.

Naturally, however, a drawback with the chosen subject pool is that the results will not be generalizable to the general population. Nonetheless, since the primary purpose of my paper is to make causal inferences, I value internal validity as my primary concern, which is likely not negatively affected by the chosen subject pool.

### **3.3.2. Second Experiment**

In the second experiment, I aimed to recruit 880 participants to achieve 80% power to detect an assumed effect size of Cohen's  $d = 0.26$  at a 0.005% significance level. The power calculations were guided by the findings in the first experiment, where I found an effect size of Cohen's  $d = -0.26$  for mouthwash (see the results section 4.1.2 for more information), which was the only product with a statistically significant effect.

I used Prolific to gather responses for the second experiment and had a pre-set data collection cut-off at 880. In Prolific, I required that participants speak fluent English to access the survey. I used Prolific to gather responses for the second experiment as it provides an efficient way to gather a large number of rather reliable responses (Palan and Schitter, 2018) from a population quite different than the first experiments. Thus, providing an efficient platform to test the reliability of the findings in the first experiment.

### **3.4. Statistical Tests**

I employ a standard two-tailed t-test to analyze the data. As a robustness check, I also run a Mann-Whitney u-test, which is the non-parametric counterpart of the t-test.

To test the primary hypothesis, I construct an 'oral hygiene rating' by taking the average desirability rating for mouthwash and toothpaste for each participant. For example, if mouthwash receives a rating of 3 and toothpaste a rating of 5, the oral hygiene rating for the participant will be  $\frac{(3+5)}{2} = 4$ . I use the t- and u-test to examine if there is a statistically significant difference in the mean between the immoral treatment group and the amoral control group. I will not run any interaction controls to check if e.g., men and women react differently to the treatment. It would deviate from the primary focus of the

paper, and given the desired sample size, such tests might be underpowered and not provide reliable results.

To test the secondary hypothesis, I create a 'general hygiene rating', by taking the average desirability rating for hand sanitizer and shampoo, and an 'other desirability rating' by taking the average rating for wireless headphones, water bottle, reusable grocery bag, and vegan chocolate bar. Both ratings will be created in the same fashion as the oral hygiene rating. I use the t- and u-test to test if there is any statistically significant difference in the means of the two ratings from the treatment.

Following Benjamin et al. (2018), I use a significance threshold of  $0.005 < P$  for findings to be interpreted as “statistically significant”, and a threshold of  $0.005 < P < 0.05$  for results to be interpreted as providing “suggestive evidence”. I use this interpretive framework to lower the probability of drawing unreliable or false conclusions. The framework is of particular importance for my findings since, from a Bayesian perspective, the “priors”, i.e., the expectations from past evidence, of the main hypothesis being true are arguably rather low.

## **4. Results**

### **4.1. Experiment 1**

#### **4.1.1. Descriptive Statistics**

Table 1 gives an overview of the descriptive statistics for experiment 1. The survey had a response rate of 14.5% and 90% of the respondents passed the attention-check tests, resulting in a total usable sample of  $N=458$ . The mean age for the participants is 25.7 with a range from 17 to 63. 55% of the usable sample are men and 72% of the participants study at the Stockholm School of Economics, with a majority pursuing degrees within business and economics.

**Table 1: Descriptive Statistics**

<b>Responses</b>	
Emails sent	3500 <sup>2</sup>
Total Responses	509
Response rate	14.5%
Usable responses	458
Percent of responses who passed attention-check	90%
<b>Age</b>	
Mean	25.7
Min	17
Max	63
<b><sup>3</sup>Gender Identity</b>	
Man	251
Woman	202
Other	5
<b>University</b>	
Karolinska Institute	2
Royal Institute of Technology	3
Stockholm School of Economics	331
Stockholm University	122
<b>Field of Study</b>	
Arts/humanities	13
Business/Economics	349
Engineering/Computer Science	12
Natural Sciences/Medicine	43
Social Sciences/Law	34
Other	7

Table 2 shows baseline balance between treatment and control group in the first experiment. The table reports results from t- and u-tests on all observable covariates and I find no statistically significant difference between the two groups, suggesting that the randomization on observables was successful.

---

<sup>2</sup> Approximate number of emails sent.

<sup>3</sup> Data for Gender Identity coded as Man="1", Woman="2", Other="3"; University is coded as Karolinska Institute="1", Royal Institute of Technology="2", etc. Same pattern applies for the coding of Field of Study. This information is relevant for the tests presented in Table 2 where I examine the balance in covariates.

**Table 2 Baseline Balance in Covariates**

	Obs. Treatment	Obs. Control	Mean Treatment	Mean Control	Diff.	P-value (t-test)	P-value (u-test)
Gender	237	221	1.489	1.443	.046	.367	.4735
Age	237	221	25.43	26.045	-.615	.365	.3193
University	237	221	3.240	3.262	-.022	.624	.7597
Field of Study	237	221	2.456	2.484	-.028	.774	.7051

### 4.1.2. Findings

Tables 3 and 4 report the main results from experiment 1 and figure 1 visualizes the results in box plots. In table 3, I find robust suggestive evidence (t-test:  $p=0.028$ ; u-test:  $p=0.0315$ ) that the treatment group had a lower oral hygiene desirability rating than the control group (Cohen's  $d=-0.21$ ). When looking at the individual product ratings in table 4, I find that the effect on the oral hygiene rating is solely driven by an observed effect on mouthwash, where I find strongly suggestive evidence (t-test:  $p=0.005$ ; u-test:  $p=0.0109$ ) that the self-rated desirability for mouthwash is lower in the treatment group than in the control group (Cohen's  $d = -0.26$ ). My findings in the first experiment run contrary to the primary hypothesis of the paper which postulates that the treatment, on the average, will increase the oral hygiene rating.

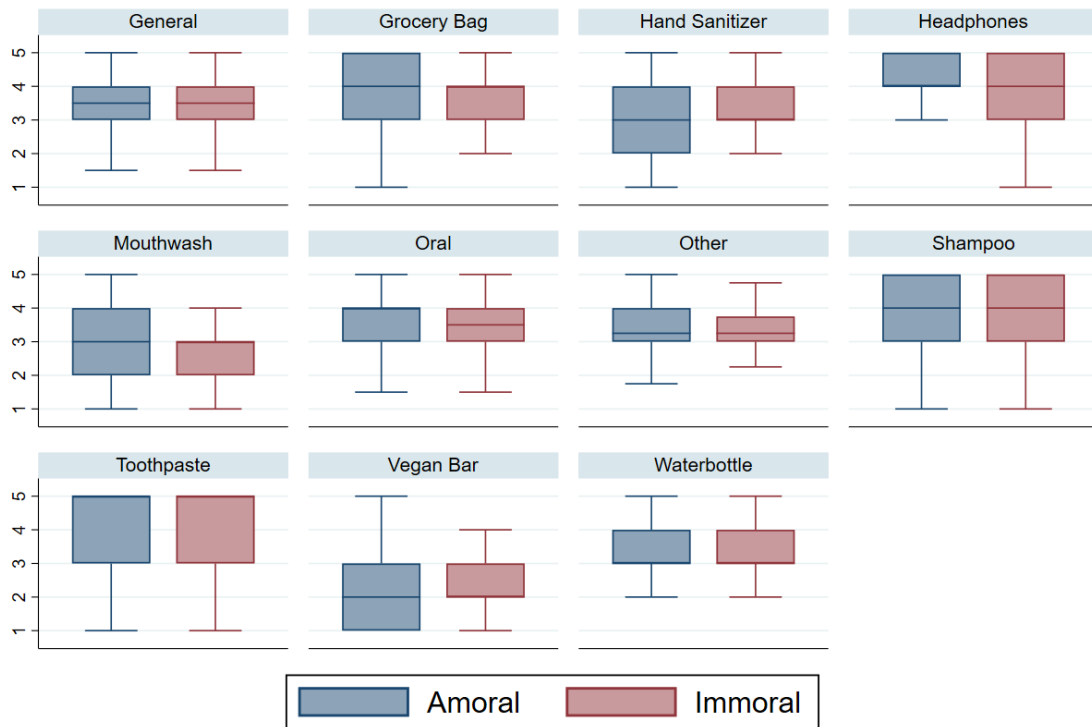
I find no evidence of the treatment having an effect on any of the other product desirability ratings, which confirms the secondary hypothesis of the paper that the treatment will not have any effect for the general hygiene or other product rating. Moreover, from figure 1, I find that the distribution of all the ratings look largely similar.

**Table 3: Impact of Treatment on Product Desirability Ratings**

	Obs. Treatment	Obs. Control	Mean Treatment	Mean Control	Diff.	P-value (t-test)	P-value (u-test)
Oral hygiene rating	237	221	3.424	3.604	-.180	.028	.0315
General hygiene rating	237	221	3.538	3.566	-.028	.734	.7974
Other rating	237	221	3.355	3.359	-.004	.962	.8982

**Table 4: Impact of Treatment on Individual Product Desirability Ratings**

	Obs. Treatment	Obs. Control	Mean Treatment	Mean Control	Diff.	P-value (t-test)	P-value (u-test)
Mouthwash	237	221	2.755	3.031	-.276	.005	.0109
Toothpaste	237	221	4.093	4.176	-.083	.418	.3079
Shampoo	237	221	3.975	3.977	-.002	.979	.9159
Hand Sanitizer	237	221	3.101	3.154	-.053	.602	.8514
Headphones	237	221	4.025	4.109	-.084	.418	.3275
Water bottle	237	221	3.329	3.366	-.037	.738	.6377
Reusable Grocery Bag	237	221	3.569	3.611	-.042	.695	.6293
Vegan chocolate bar	237	221	2.498	2.348	.150	.161	.0943

**Figure 1: Desirability Ratings**

The box spans the 25th to the 75th percentile with the interior horizontal line representing the median (50th percentile). The whiskers span the full range of the observations.

## **4.2. Experiment 2**

### **4.2.1. Descriptive statistics**

Table 5 gives an overview of the descriptive statistics for experiment 2, conducted with participants from Prolific. The demographic information for certain variables available through Prolific was incomplete. Thus, I have split table 5 into one section that presents information on the variables I have complete information on for the usable sample and one section where I present the incomplete demographic information for the total sample.

I received N=881 responses, with an attention-check pass rate of 93%, which resulted in a total usable sample of N=816. The mean age was 29.7, ranging from 18 to 69. There was an approximately 50/50 split between men and women. Out of the participants that I have demographic information about educational status and nationality, 59% are not students and the major nationalities represented are South Africa, Portugal, and Poland.

**Table 5: Descriptive Statistics**

<b>Responses</b>	
Total Responses	881
Usable responses	816
Percent of responses who passed attention-check	93%
<b>Age</b>	
Mean	29.7
Min	18
Max	69
<b>Gender Identity</b>	
Man	406
Woman	402
Other	8
<i>Incomplete Demographics Info</i>	
<i>Total Sample</i>	
<b>Nationality</b>	
Greece	43
Hungary	20
Poland	106
Portugal	154
South Africa	234
Spain	32
United Kingdom	66
Zimbabwe	23
<b>Total</b>	678
<b>Student Status</b>	
Yes	319
No	467
<b>Total</b>	786

Table 6 presents information on baseline balance in the observable covariates that I have complete information on. I find no statistically significant difference in the gender distribution between the two groups. However, there is suggestive evidence (t-test:  $p=0.025$ ; u-test:  $p=0.036$ ) of an age difference where those in the control group are, on average, 1.36 years older than those in the treatment group. While this indicates a slight imbalance between the age distribution in the two groups, there is no particular reason to believe that it has any impact on the final results, and, by definition, the difference occurred randomly.

**Table 6: Baseline Balance in Covariates**

	Obs. Treatment	Obs. Control	Mean Treatment	Mean Control	Diff.	P-value (t-test)	P-value (u-test)
Gender	419	397	1.511	1.514	-.003	.947	.8006
Age	419	397	29.01	30.37	-1.36	.025	.0360



### 4.2.2. Findings

Tables 7 and 8 report the main results of experiment 2 and figure 2 visualizes the results in box plots. In figure 2, I find that the distributions for all the ratings are largely similar. Moreover, I find no evidence of a statistically significant difference in self-rated product desirability rating for either toothpaste or mouthwash between the two groups. Thus, from the data gathered in both my experiments, I find no reliable evidence that the treatment has an effect on oral hygiene product desirability ratings. Therefore, I reject the primary hypothesis of the paper and conclude that the treatment has no reliable effect on self-rated product desirability for oral hygiene products.

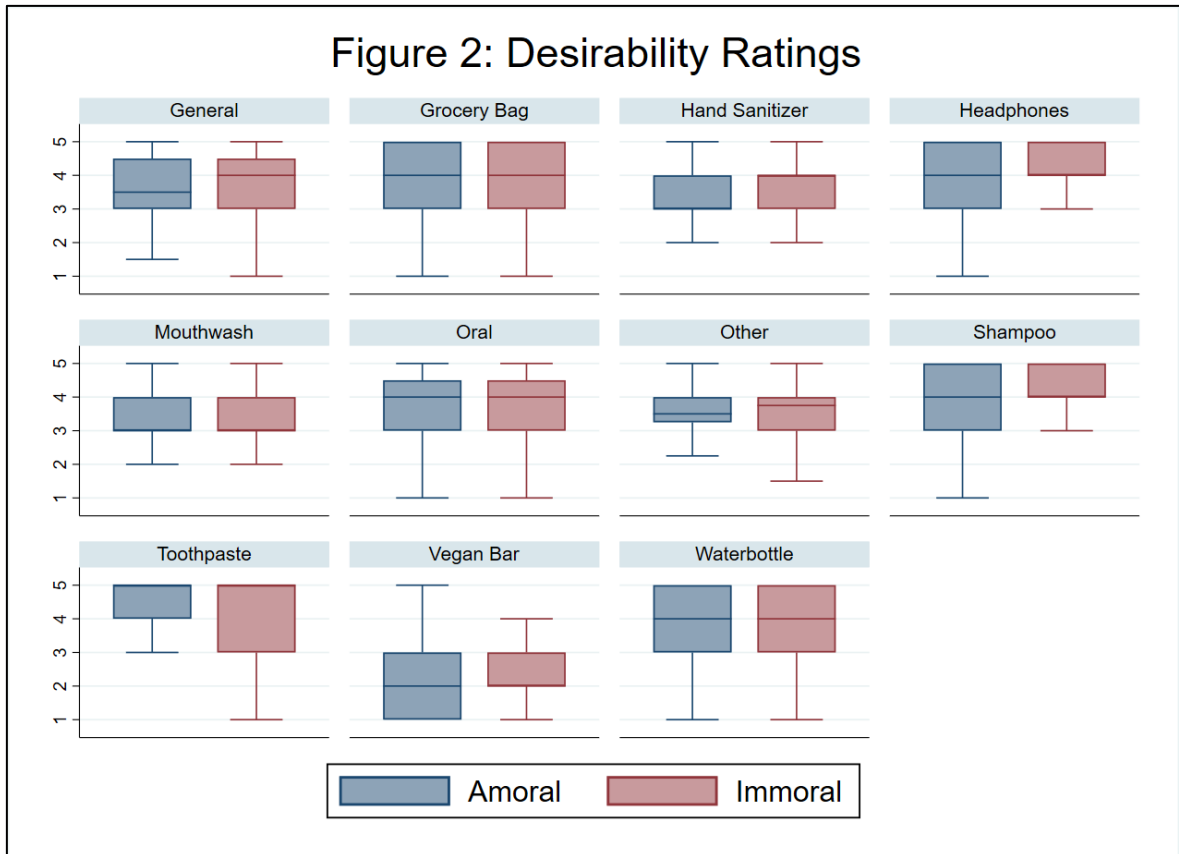
Moreover, I find no evidence of a statistically significant effect on any of the other products included in the product desirability rating, which confirms the findings in the first experiment. Thus, I conclude that the treatment had no effect on self-rated desirability for general consumer products, which reliably confirms the secondary hypothesis of the paper.

**Table 7: Impact of Treatment on Product Desirability Ratings**

	Obs. Treatment	Obs. Control	Mean Treatment	Mean Control	Diff.	P-value (t-test)	P-value (u-test)
Oral hygiene rating	419	397	3.819	3.816	.003	.969	.8851
General hygiene rating	419	397	3.801	3.709	.092	.123	.0892
Other rating	419	397	3.597	3.582	.015	.763	.7146

**Table 8: Impact of Treatment on Individual Product Desirability Ratings**

	Obs. Treatment	Obs. Control	Mean Treatment	Mean Control	Diff.	P-value (t-test)	P-value (u-test)
Mouthwash	419	397	3.461	3.441	.020	.798	.8474
Toothpaste	419	397	4.177	4.191	-.015	.831	.9383
Shampoo	419	397	4.124	4.003	.121	.069	.0606
Hand Sanitizer	419	397	3.477	3.416	.061	.438	.4698
Headphones	419	397	4.098	3.992	.106	.157	.1046
Water bottle	419	397	3.905	3.882	.023	.764	.9057
Reusable Grocery Bag	419	397	3.969	4.071	-.102	.162	.1675
Vegan chocolate bar	419	397	2.418	2.385	.033	.694	.6419



The box spans the 25<sup>th</sup> to the 75<sup>th</sup> percentile with the interior horizontal line representing the median (50<sup>th</sup> percentile). The whiskers span the full range of the observations.

## 5. Discussion

### 5.1. Endogenous preferences

I find no reliable evidence that reading an unethical text impacts self-rated desirability for cleansing products. My findings are in line with the predictions of standard economic theory, which generally holds that tastes and preferences remain constant through time. Nonetheless, while my findings do not necessarily contradict the assumptions of conventional economics, I argue that they do not yield particular support to it. Rather, they show that it is unlikely that an exogenous trigger by reading an unethical text produces effects that are not in line with standard economic theory but it does not exclude the possibility that other effects can produce opposite results. For some examples of

suggestive evidence that run contrary to the predictions of theory see e.g., Cohn and Maréchal (2016) and Callen et al. (2014).

If I had found statistically significant and reliable effects, it would have contradicted standard theory and implied that there is likely another dimension in which economic models could be potentially improved. Nonetheless, even if I had found statistically significant effects, one of the limitations of my experiments is that the participants did not make choices that had monetary consequences. Experiments in economics traditionally feature monetary incentives, where decisions affect payoffs, to simulate a real-world environment and study “revealed preferences” rather than hypothetical choices. However, due to the institutional setting with regard to GDPR laws where I conducted my experiments, it was not feasible to feature monetary payoffs in my experiments since it required some form of personal data gathering. Therefore, it remains somewhat unclear how exactly statistically significant results in my experiments would impact actual market behavior. Although, I argue it is not unreasonable to assume that there exists a positive relationship between findings in a hypothetical case and in a case with incentives.

## **5.2. Relation to past findings**

From the cumulative evidence from my experiments and in the replications by Earp et al. (2014) of Zhong and Liljenquist’s (2006)(study 2) original experiment, I argue there is very little evidence for the existence of a Macbeth effect. It is unlikely that a threat to moral purity by reading an unethical text compels a need for physical cleansing that manifests through higher desirability for products used to cleanse oneself. Thus, I conclude that the Macbeth effect does not exist and that it is likely that Zhong and Liljenquist’s (2006) findings were false positives caused by e.g., statistical noise.

Given the findings in both my experiments, that of Earp et al. (2014), and the meta-analysis by Siev et al. (2018), if general cleansing effects exist, I argue that they are likely to be of small magnitude. Thus, it is also highly likely that experiments exploring similar cleansing effects such as the work presented in the literature review (section 2.2) by e.g., Xu et al. (2012) or Gollwitzer and Melzer (2012), are likely markedly underpowered with largely unreliable findings. Therefore, we need to carefully reassess previous results and their reliability before we can draw any valid conclusions.

While my experiment does not conceptually replicate the findings of Lee and Schwarz (2010) since their experiment includes actual enactment of unethical behavior, I would still caution against building theories or concepts on their findings. I argue that there is no strong reason to suspect that enacting an unethical behavior rather than just reading about one would generate an effect almost twice the magnitude of the effect size of Cohen's  $d = 0.3$  used for the power calculations in my experiment 1. Thus, their work is likely markedly underpowered, even if combined with the observations in the replication of their work by Schaefer et al. (2015), and would need some form of replication with larger sample sizes before we can ascertain that their findings are reliable.

### **5.3. Different findings in the experiments**

My findings for the potential effect of the treatment on the self-rated desirability rating for mouthwash are different between the first and the second experiment. In the former, I find suggestive evidence for an effect in the opposite direction of what was hypothesized and this finding does not replicate in the latter experiment. What caused this difference in findings is unclear but there are several potential explanations.

Firstly, there are three observable differences between the experimental populations: (a) the mean age, (b) the nationalities of participants, where it is highly likely that a

majority of participants in the first experiment were Swedish whereas this is not the case in the second experiment, and (c) student status. Either of these differences might have contributed to the difference in findings. Although, the hypotheses do not postulate any confounding effect of, for example, student status or nationality so there is no strong reason to believe that these differences contributed to the difference in findings.

Secondly, the one-euro payment to participants in Prolific may have affected their general level of attention, which could have impacted the final results. While this might be the case, given that the attention-check pass rates between the two experiments are similar (90 and 93%), I would argue that this is an unlikely reason for the difference in results.

Thirdly, it is possible that the effect does not exist and that the findings in the first experiment are just statistical noise. This might be likely (i) given the state of the past evidence, (ii) that the hypothesis is in the opposite direction, and (iii) since the first experiment only has 65% statistical power to detect an effect of Cohen's  $d=0.3$  at a 0.005 significance level. But to confirm that the evidence in the first experiment was caused by statistical noise we would need a replication on a very similar population (ideally the same population, with repeated sampling) with even more participants. However, I think the non-existence of the effect is the most probable reason for the difference in findings between my two experiments because of the weak priors, that the effect was in the opposite direction than what was hypothesized, and the statistical power of the first experiment.

## **5.4. Participant attentiveness**

The attention-check pass rates for my experiments were 90 and 93 percent, respectively, which indicates that approximately 10 percent of the participants did not carefully pay

attention to the entire survey. I rule out the possibility that the attention checks are unclear or too demanding since a majority passed them. Thus, the pass rate highlights the importance of attention checks to ascertain or at least increase the probability that all participants included in the statistical analysis have been attentive. A fruitful avenue for future research could be to assess the optimal way of ensuring that participants are attentive when participating in online experiments or surveys. Interestingly, the monetary incentive in the second experiment did not markedly change the attention-check pass rate. This might suggest that small financial incentives do not necessarily improve participants' attentiveness when filling out surveys.

## **6. Conclusion**

There is a growing body of research in economics studying how preferences are determined and what factors can affect them. In this paper, I studied if psychological priming by reading a short unethical text can affect preferences, insofar as they manifest through a self-rated desirability rating, for cleansing products. To this end, I conducted two preregistered experiments with identical experimental designs and gathered evidence from 1247 individuals. I find no reliable evidence that reading an unethical text affects self-rated desirability for cleansing products and conclude that this prime does not affect preferences. My findings suggest that past work exploring this effect are likely false positives and have experiments that are markedly underpowered.

Fruitful avenues for future research could be to examine other primes' potential effect on preferences or to replicate some of the past work using larger sample sizes to assess their reliability.

# Bibliography

- Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21:99-131
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature human behaviour*, 2(1):6–10.
- Bernheim, B. D., Braghieri, L., Martínez-Marquina, A., and Zuckerman, D. (2021). A theory of chosen preferences. *American Economic Review*, 111(2):720–54.
- Callen, M., Isaqzadeh, M., Long, J. D., and Sprenger, C. (2014). Violence and risk preference: Experimental evidence from afghanistan. *American Economic Review*, 104(1):123–48.
- Cohn, A., Engelmann, J., Fehr, E., and Maréchal, M. A. (2015). Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105(2):860–85.
- Cohn, A. and Maréchal, M. A. (2016). Priming in economics. *Current Opinion in Psychology*, 12:17–21.
- Cohn, A., Maréchal, M. A., and Noll, T. (2015b). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies*, 82(4):1289–1308
- Dietrich, F. and List, C. (2013). Where do preferences come from? *International Journal of Game Theory*, 42(3):613–637.

- Earp, B. D., Everett, J. A., Madva, E. N., and Hamlin, J. K. (2014). Out, damned spot: Can the “macbeth effect” be replicated? *Basic and Applied Social Psychology*, 36(1):91–98.
- Eynon, R., Fry, J., and Schroeder, R. (2017). The sage handbook of online research methods.
- Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G\* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191.
- Gelman, A. and Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6):641-651.
- Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348.
- Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review*, 19(5):847–857.
- Gollwitzer, M. and Melzer, A. (2012). Macbeth and the joystick: Evidence for moral cleansing after playing a violent video game. *Journal of Experimental Social Psychology*, 48(6):1356–1360.




- Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American psychologist*, 59(2):93.
- Konstan, J. A., Simon Rosser, B., Ross, M. W., Stanton, J., and Edwards, W. M. (2005). The story of subject naught: A cautionary but optimistic tale of internet survey research. *Journal of Computer-Mediated Communication*, 10(2):00–00.
- Krantz, J. H. and Dalal, R. (2000). Validity of web-based psychological research. In *Psychological experiments on the Internet*, pages 35–60. Elsevier.
- Henrich, J., Boyd, R., Bowes, S., Colin, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review*, 91(2):73-78.
- Lee, S. W. and Schwarz, N. (2010). Dirty hands and dirty mouths: Embodiment of the moral-purity metaphor is specific to the motor modality involved in moral transgression. *Psychological science*, 21(10):1423–1425.
- Nature.com. (2019). What’s next for psychology’s embattled field of social priming. Accessed 2022-11-18. URL: <https://www.nature.com/articles/d41586-019-03755-2>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.
- Prissé, B. and Jorrat, D. (2022). Lab vs online experiments: No differences. *Journal of Behavioral and Experimental Economics*, 100

- Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In *Psychological experiments on the Internet*, pages 89–117. Elsevier.
- Ropovik, I., Sparacio, A., and IJzerman, H. (2021). The lack of robust evidence for cleansing effects. *Behavioral and Brain Sciences*, 44.
- Schaefer, M., Rotte, M., Heinze, H.-J., and Denke, C. (2015). Dirty deeds and dirty bodies: Embodiment of the macbeth effect is mapped topographically onto the somatosensory cortex. *Scientific reports*, 5(1):1–11.
- Schnall, S., Benton, J., and Harvey, S. (2008a). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological science*, 19(12):1219–1222.
- Schnall, S., Haidt, J., Clore, G. L., and Jordan, A. H. (2008b). Disgust as embodied moral judgment. *Personality and social psychology bulletin*, 34(8):1096–1109.
- Shariff, A. F. and Norenzayan, A. (2007). God is watching you: Priming god concepts increases prosocial behavior in an anonymous economic game. *Psychological science*, 18(9):803–809.
- Siev, J., Zuckerman, S. E., and Siev, J. J. (2018). The relationship between immorality and cleansing. *Social Psychology*.
- Simmons, P. J., Nelson, D. L., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359-1366.
- Stigler, J. G. and Becker, S. G. (1977). De Gustibus Non Est Disputandum. *The American Economic Review*, 67(2):76-90.

- Trakulpipat, C., Wiwattanapantuwong, J., Dhammapeera, P., and Tuicomepee, A. (2021). “macbeth effect”: The link between physical cleanliness and moral judgement. *Journal of Social Sciences*, 42(4):779–786.
- Xu, A. J., Zwick, R., and Schwarz, N. (2012). Washing away your (good or bad) luck: Physical cleansing affects risk-taking behavior. *Journal of Experimental Psychology: General*, 141(1):26.
- Zhong, C.-B. and Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, 313(5792):1451–1452.

# Appendix

## Treatment Survey Prolific



Hi! Please see the consent form below:

Purpose of research: To examine product desirability as a part of an MSc Thesis at the Stockholm School of Economics.

What you will do in this research: You will participate in a simple product desirability study by answering a short survey consisting of (1) a short reading task and (2) a product desirability ranking task.

Time required: Participation will take about 3 minutes.

Risks: I believe there are no known risks associated with this research study. Your participation in this study will remain anonymous. I will not store your identity. Your responses will be assigned an arbitrary code number.

Benefits: You will receive £1 for participating.

Participation and withdrawal: Your participation in this study is completely voluntary and you can withdraw at any time. You are free to skip any question that you choose.

To contact the researcher: This study is being done by Shahin Eidinejad, student at the Stockholm School of Economics. If you have questions about this research, please contact 42124@student.hhs.se

By clicking the link to the survey you are indicating that you are at least 18 years old, have read and understood this consent form and agree to participate in this research study. At the end of the study, you will be redirected to a completion URL to receive credit for participating.

---

Please enter your Prolific ID

---

Please state your gender.

☐ Man

☐ Woman

☐ Other

---

Please state your age in numbers (e.g. 32)



Please read the story below and answer the questions about it.

"Two years ago, when I was an associate at a law firm, I was coming up for promotion against another hard-working associate. For several months, my colleague had been working on a major case that would ultimately make or break my colleague's career at the firm. However, my colleague could not find a very important document, without which it was highly unlikely that my colleague would have sufficient evidence to win the major case. The night before the trial, as I was walking through the office, I found the very important document that my colleague was desperately in need of. I called my colleague by phone and left a voicemail where I lied and told my colleague that the document was nowhere to be found in the office, knowing that my promotion would be secured."

From the information presented above, was the important document found?

Yes

No

Who do you think deserves the promotion the most?

The main character

The colleague

I don't know

"Hard-working employees should be rewarded for their work even if they are severely disliked by their colleagues"

Do you agree with this statement?

Yes

No

I don't know





Please rate how desirable you consider the consumer products below.

	Very Undesirable (1)	(2)	Moderately Desirable (3)	(4)	Very Desirable (5)
Vegan Chocolate Bar	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Eco-Friendly Water Bottle	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hand Sanitizer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mouthwash	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Wireless Headphones	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Toothpaste	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shampoo	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reusable Grocery Bag	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



We thank you for your time spent taking this survey.  
Your response has been recorded.