2014-01-01

Developing and Testing an Implicit Measure of Moral Foundation Accessibility

Scott D. Frankowski
University of Texas at El Paso, sdfrankowski@miners.utep.edu

DEVELOPING AND TESTING AN IMPLICIT MEASURE OF MORAL FOUNDATION ACCESSIBILITY

SCOTT DAVID FRANKOWSKI

Department of Psychology

APPROVED:

Michael Zárate, Ph.D., Chair

James Wood, Ph.D.

Daniel Jones, Ph.D.

Penelope Espinoza, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

DEVELOPING AND TESTING AN IMPLICIT MEASURE OF MORAL FOUNDATION ACCESSIBILITY

by

SCOTT DAVID FRANKOWSKI, B.S.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF ARTS

Department of Psychology

THE UNIVERSITY OF TEXAS AT EL PASO

December 2014

Acknowledgements

Thank you friends and family for your support throughout my academic career. Thank you, Jessica, your support and friendship since I started college means the world to me. Thank you, Nazy, you're one of the best, gurl. Thank you, Melissa, for all of your help on research this past year, and for being such a great friend. Thank you, Michael, for your input and help throughout this project.

Abstract

Moral Foundations Theory provides a framework for understanding moral judgments and behavior. With the present research, I developed a word fragment task as an implicit measure of moral foundation accessibility. In an experiment, I used this measure as a predictor of moral attitudes and behaviors toward two moral violations. Responses on this implicit measure predicted moral attitudes; however, the priming conditions did not affect responses. The failure of the primes are discussed, as well as future directions.

Table of Contents

List of Tables

Introduction

Every day we are confronted with making decisions about what is right and wrong. We develop a moral code based on our upbringing, genetics, culture, religious practices, personal philosophies, and educational background. We typically believe that our moral decisions are well thought out and correct. When we judge the actions or beliefs of others, we often do so from the basis of our own moral code. When our morality is tested, many of us can reason why we are right. Recent developments in moral psychology challenge common notions of a reasoned moral code. There is a growing body of evidence that moral judgments are often gut-level affective reactions and that moral reasoning is used to explain our moral intuitions.

With the present research, I explore the implicit nature of moral intuitions within Haidt's (Haidt & Graham, 2007) moral foundations framework. In previous research reported in this thesis, I developed a word fragment completion task as an implicit measure of moral foundation accessibility. The primary aim of the current research is to provide initial validation of this measure by testing how responses to these word fragments predict disapproval of, and moral action in response to, a moral violation.

1.1 Historical and Current Research on Morality

The study of morality has traditionally focused on acknowledging harm, providing care, and reasoning fairness and justice (e.g. Piaget, 1932; Kohlberg, 1963). A protégé of Kohlberg's, Turiel (1983), defines morality as "prescriptive judgments of justice, rights, and welfare pertaining to how people ought to relate to each other" (p. 3). Haidt and Graham (2007) contest that this is a predominantly modern and Western view of morality. Through much of human history, groups' survival and ability to thrive has *not* depended on modern notions of categorical individual rights. Whereas Piaget and Kohlberg view morality based on group loyalty and

submission to authority as underdeveloped forms of morality, these socially binding forms of morality have arguably been necessary for human survival.

Rokeach (1973) proposed that moral values extend beyond acknowledging harm and striving for fairness, and that people could strive for values such as obedience and security – i.e. such values are not necessarily a sign of an underdeveloped moral compass as Kohlberg or Piaget would suggest. Moral foundations theory builds on the work of Rokeach by emphasizing the automaticity of moral judgments.

Kohlberg's theory of moral development (1963) states that people employ moral reasoning as a means of addressing moral dilemmas. In Kohlberg's research, participants are given dilemmas to sort through. For example, one scenario is about a man whose wife is dying. A pharmacist has a drug that will cure the wife's condition; however, the pharmacist is charging an exorbitant price for the drug that the man cannot afford. The man steals the drug and the participants are tasked with making the moral judgment of whether the act was right or wrong and provide reasoning for their judgment. Participants are assigned to a stage of morality based upon how they *reason* that stealing the drug is right or wrong.

Whereas in Kohlberg's theory moral judgments are consciously derived, the social intuitionist model of morality (Haidt, 2001) states moral judgments are "quick moral intuitions followed (when needed) by slow, ex post facto moral reasoning" (p. 817). Importantly, the social intuitionist model of morality does not eliminate moral reasoning, but instead states that moral reasoning can affect automatic reactions, and that this process happens slowly overtime, such that our morals can be malleable *but* our moral judgments are automatic. In the social intuitionist model of morality, Haidt (2001) discusses moral judgments in the context of Zajonc's (1980) theory on the primacy of affect which states that affective reactions precede cognitive

reasoning. Zajonc found that affective reactions occur quicker than cognitive reasoning, and occur in the absence of cognitive encoding. Furthermore, people's reasoning is often only an explanation for an initial reaction that is often unreliable (Nisbett & Wilson, 1977). Thus, automatic moral reactions should, at times, conflict with moral reasoning.

Haidt, Bjorklund, and Murphy (2000) demonstrated the primacy of moral judgments over moral reasoning in an experiment in which participants read about an incestuous encounter between siblings. The scenario stated that the siblings were both single stable adults, that the women was on birth control and that the man used a condom, and that they both had agreed it would be a one-time occurrence. The scenario goes on to state that the siblings keep the night as a special secret that made them closer as brother and sister. Participants were asked what they thought of the encounter and whether it was okay for the siblings to do what they did. Predictably and overwhelmingly, participants condemned the actions of the siblings and produced reasons why the encounter was wrong. When dangers of inbreeding were raised, participants were reminded that the siblings used multiple forms of contraception. When participants reasoned that the siblings would be emotionally damaged, they were reminded that the siblings had actually grown closer with no negative emotional repercussions. Eventually participants would just state that they cannot explain why it is wrong but that they just *know* it was wrong. Haidt (2001) poses the question, "What model of moral judgment allows a person to know that something is wrong without knowing why?" (p. 814). Indeed, if moral judgments are based solely on reasoning and reflection such as Piaget and Kohlberg propose, people would be able to explain why the scenario of sibling incest is wrong; or alternatively, they would be able to overcome their initial reaction and state that the siblings did nothing immoral. Most people, however, do not reach the conclusion that the incestuous act was morally acceptable.

1.2 Moral Foundations Theory

1.2.1 The foundations. Haidt and Graham (2007) propose moral foundations theory which consists of five moral foundations: Harm/Care — acknowledging harm and caring for others; Fairness/Reciprocity — expecting to be treated and to treat others fairly; Ingroup/Loyalty — actions that benefit one's group are considered moral; Authority/Respect — respecting and submitting to established authority figures is moral; and, Purity/Sanctity — perceiving the body, mind, and spirit to be pure and sacred.

1.2.2 Virtues and vices. Much of the research on moral foundations theory further splits the five foundations into virtues and vices. Virtues are the aspects of the moral foundation that are held in esteem and which people strive to achieve. Vices are themes associated with violations of the foundation virtues. For example, a virtue of the fairness foundation would be striving for racial equality; while conversely, a vice of the fairness foundation would be racial discrimination.

1.2.3 Individualizing and binding foundations. Haidt and Graham (2007) define the foundations of ingroup/loyalty, authority/respect, and purity/sanctity as "binding foundations" because their function is to "to bind people together into hierarchically organized interdependent social groups" (Haidt, 2007, editorial on www.edge.org). The foundations of harm/care and fairness/reciprocity are considered "individualizing foundations" in that they "protect individuals from each other and allow them to live in harmony as autonomous agents who can focus on their own goals" (Haidt, 2007).

1.3 Moral Foundations Research

Many moral foundations studies have examined how endorsements of different foundations predict other attitudes and behaviors. Moral foundation endorsement is measured as

an individual difference trait with the Moral Foundations Questionnaire which has five correlated subscales corresponding to each foundation (Haidt, Graham, & Joseph, 2009; Graham et al., 2011). Much of the research has focused on the relation between foundation endorsement and political ideology. It has been shown that people with liberal ideological leanings endorse the individualizing foundations more than the binding foundations; whereas, people with conservative leanings tend to endorse all five of the foundations equally relative to one another (Haidt & Graham, 2007). Liberals tend to not endorse the binding foundations, and conservatives tend to endorse the individualizing foundations significantly less than liberals do. Dissecting ideology further, Haidt et al. (2009) found that secular liberals tend to strongly endorse individualizing foundations while showing very little endorsement of the binding foundations. Further, those on the religious left strongly endorse the individualizing foundations while also endorsing the binding foundations. Conversely, social conservatives strongly endorse the binding foundations while showing much less endorsement of the individualizing foundations.

Beyond political ideology, Graham et al. (2011) tested convergent and discriminant validity of the moral foundations questionnaire in relation to other established scales. Dispositional empathy (Davis, 1983) was significantly related to the harm/care foundation and was not related to the purity/sanctity foundation. Schwartz's (1992) values of self-discipline, cleanliness, and devotion were significantly associated with the purity/sanctity foundation and were weakly associated with the harm/care foundation. These findings provide an opportunity to test discriminant validity of an implicit measure with these explicit scales.

1.3.1 Associations with other individual differences. Authoritarianism is a person's propensity to submit to authority, aggress in the name of one's authorities, and hold conventional

values as espoused by one's authorities (Altemeyer, 1996). Social dominance orientation is a person's preference for social hierarchy and group-based dominance over low status group (Pratto, Sidanius, Stalleworth, & Malle, 1994). Both authoritarianism and social dominance correlate positively and moderately-to-strongly with conservative political ideology (Altemeyer, 2004). Previous studies also show that authoritarianism and social dominance orientation correlate positively and moderately-to-strongly with endorsing the binding moral foundations while correlating negatively with the individualizing foundations (e.g. Graham et al., 2011, Koleva, Graham, Iyer, Ditto, & Haidt, 2012). In Graham et al.'s study (2011), authoritarianism correlated more strongly with the binding foundations than did social dominance orientation. Furthermore, social dominance correlated more strongly and negatively with the individualizing foundations than did authoritarianism. Recent research shows that social dominance orientation causally predicts lower levels of empathy (Sidanius et al., 2013), suggesting that it may also lead to lower levels of moral identification with individualizing foundations. While authoritarianism and social dominance are similar in that they promote group hierarchies and punish deviations from conventional social systems, there are distinctions between the two variables that are important in regards to morality. High authoritarians tend to be more religious and dogmatic than high social dominators (Altemeyer, 1998); thus, authoritarianism ought to predict moral disapproval of purity violations whereas social dominance should not. These associations suggest that political ideology, authoritarianism, and social dominance should be taken into consideration in moral foundations research. In the pilot study presented, these variables were included when testing an implicit measure of moral foundation accessibility.

Expanding on the research examining associations between ideology and moral foundations, a series of large scale studies explored how explicit moral foundation endorsement predicts attitudes toward culture-war topics (Koleva et al., 2012). These studies showed that explicit endorsement of the foundations predicts moral disapproval of culture-war issues such as same-sex marriage, the death penalty, embryonic stem cell research, flag burning, and using pornography among many other issues. In the current research, the death penalty and using pornography were used as primes of a moral violation of harm/care (death penalty) and purity/sanctity (pornography). These two topics were chosen because moral foundation endorsement was the strongest predictor of moral disapproval for each topic. Endorsement of the purity/sanctity foundation strongly predicted moral disapproval of using pornography, and endorsement of harm/care foundation predicted moral disapproval of the death penalty. Furthermore, purity/sanctity endorsement did not predict death penalty disapproval and harm/care endorsement did not predict disapproval of pornography. While Koleva et al., examined the relation between explicit foundation endorsement and moral disapproval, there has not yet been any research on the relationship between implicit morality and explicit moral disapproval. Furthermore, some of the existing research on implicit moral foundations suggests that implicit evaluations and accessibility may operate differently than explicit foundation endorsement.

1.3.2 Implicit moral foundation studies. The study of implicit accessibility of moral foundations is in its infancy. A search of PsycInfo abstracts for the words "implicit" and "moral foundations" returns only two results in which researchers used an implicit measure of moral foundations. In his dissertation research, Graham (2010) found mixed results with the associations between implicit moral foundation measures and political ideology. In one study,

participants were shown two different moral violations (e.g. “Treating people unequally,” and, “Disobeying an authority,” p. 22) and asked to choose which one is worse as quickly as possible with their gut reaction (a measure that Graham acknowledges is the least implicit of the measures he uses in his dissertation on implicit moral foundations). Graham found that conservatism was a strong negative predictor for choosing the individualizing violations. However, this finding was the same whether participants were told to choose quickly with their gut reaction or to deliberate and carefully think through their response. Perhaps, in line with Nisbett et al. (1977), when participants were instructed to deliberate and carefully think through their response, they simply searched out support for their initial gut-level reaction.

In another study for his dissertation, Graham (2010) used an evaluative priming procedure in which participants would see a moral foundation virtue or vice related word (or a control prime) followed by a word that had a positive or negative valance. The participants were tasked with judging whether the second word was positive or negative as quickly as possible. The results showed that participants more quickly categorized foundation vice words as negative than virtue words as positive. This finding suggests that vices of moral foundations may be more salient than foundation virtues. Political ideology, however, was not a predictor of automatic negativity toward either the individualizing or binding foundations. Thus, while conservatives tend to explicitly endorse the binding foundations more than liberals, this experiment suggests they do not show increased automatic responding to the foundations.

Graham’s (2010) dissertation research provides some evidence that differences exist in implicit moral foundations tasks depending on one’s ideology. Graham, however, found no consistent effect, and his results, with the exception of one of the studies (which used the least implicit measure), did not align with the explicit endorsement of moral foundations. A series of

studies by Leidner and Castano (2012) shows implicit moral foundation accessibility can be prone to experimental manipulation.

In Leidner et al.'s (2012) studies people were led to believe that either their ingroup or an outgroup committed an atrocity. In the ingroup condition participants read a vignette about U.S. soldiers committing war crimes against civilians in Iraq. In the outgroup condition, participants read about Australian soldiers committing war crimes against Iraqi civilians. As a dependent measure, the researchers used a modified version of the Lexical Decision Task (Meyer & Schvaneveldt, 1971) in which participants had to make a quick decision about whether a word presented was a word or a non-word. Reaction times were recorded and faster decisions for moral foundation related words were evidence of increased accessibility of the associated moral foundation. The researchers found in the ingroup condition people showed greater implicit accessibility of the ingroup/loyalty and authority/respect moral foundations; whereas, in the outgroup condition people showed greater accessibility of the harm/care and fairness/reciprocity moral foundations. This research shows that there is implicit activation of the moral foundations in response to violations of morals.

In Leidner et al.'s (2012) studies, implicit accessibility of moral foundations matches neatly with the expectations of social identity theory (Tajfel & Turner, 1979) – the idea that we strive for a positive ingroup identity. In Graham's (2010) dissertation research, however, implicit morality is more ambiguous. These lines of research differ in that Graham was examining implicit *evaluations* of foundation relevant stimuli; whereas, Leidner et al. were examining implicit *accessibility* of moral foundation constructs. Whereas Leidner et al. showed that implicit accessibility of moral foundations can shift depending on context, there is currently no published research on how implicit foundation accessibility can predict explicit moral

attitudes. Graham's research showed how ideology is related to implicit moral foundation evaluations; however, there is currently no published research on implicit dispositional moral foundations. Considering the infancy of moral foundations theory, more novel approaches to implicit moral foundation accessibility ought to provide a clearer picture of the mechanisms that underlie our moral intuitions. With the following pilot study, I created a word fragment completion task as an implicit measure of moral foundation activation that has potential to be used both as a predictor and outcome variable.

Pilot Study

The primary aim of this pilot study was to identify word fragments associated with each moral foundation's virtue and vice. A secondary aim was to examine how this measure is associated with individual difference measures of ideology and explicit moral foundation endorsement. There are several reasons why a word fragment completion task serves as a good implicit measure to test for moral foundations accessibility. First, a word fragment completion task can be used both as a paper-and-pencil measure and can easily be adapted for computer survey programs. Second, this task is low cost in terms of researcher time and ease of data collection. This measure can easily be administered to a room full of participants or online. Unlike many reaction-time measures, a word fragment completion task is less confounded by variability introduced through physical attributes of participants and their computers (e.g. participant handedness, computer screen refresh rates, etc.). Thus, the task is well suited for between-subjects studies and studies conducted online. Finally, this measure is intended to be a modular measure of moral foundation accessibility. Researchers can choose the foundations' virtues and vices that are appropriate for their research questions.

In developing this word fragment completion task, we tested 194 moral foundation related word fragments. The aim of this study was to retain 12 or more words for each moral foundation's virtue and vice based on recommendations for creating word fragment tasks (Koopman, Howe, Johnson, Tan, & Change, 2013). While Koopman et al. recommend retaining fragments that participants complete as the target word 25-75% of the time, we chose to retain words that were completed as target congruent words 20-80% of the time. This more liberal approach allowed 10 or more fragments to be retained for each moral foundation virtue and vice.

In selecting words for this task, we used the moral foundations dictionary provided by

Haidt and Graham at www.moralfoundations.org. This dictionary was designed for use with the Linguistic Inquiry Word Count software, a text analysis program used in qualitative research (Pennebaker, Francis, & Booth, 2007). Most of the fragments tested were words and synonyms that appeared in this dictionary. Fragments were created using a crossword puzzle solver (www.crosswordsolver.org, as recommended by Koopman et al., 2013). Fragments were used that could be completed congruent with only one moral foundation. This pilot study included the moral foundations questionnaire (Graham et al., 2011), a measure of authoritarianism (McFarland, 2010) and social dominance orientation (Pratto et al., 1994). Authoritarianism and social dominance were chosen for inclusion because of their relation to explicit endorsement of the moral foundations. We predicted that our implicit measure would be associated with explicit endorsement of the foundations such that implicit activation of a foundation would be related to explicit endorsement of the corresponding foundation. We did not have specific predictions regarding the foundation virtues and vices.

2.1 Procedure

Participants were recruited through Amazon's Mechanical Turk (M-Turk). Participants signed an informed consent form with their M-Turk ID and were then provided a link to the study. The study began by informing participants of the task, showing an example of the task, and specifying to complete each fragment as quickly as possible with only one word. Further, participants were instructed to skip a fragment if they could not think of a word within a few seconds (as recommended by Koopman et al., 2013).

Participants then completed two practice fragments. After the practice fragments, participants continued on to the task in which they completed 194 fragments along with 40 neutral fragments. All fragments were presented in random order to participants. After

completing the word fragment task, participants completed an explicit measure of moral foundation endorsement, an authoritarianism scale, the social dominance scale, and a demographics questionnaire that included a measure of political ideology. The scales before the demographics were presented in random order.

2.2 Participants

Participants ($N = 258$) were paid $0.70 for their participation. We decided a priori that data from about 250 participants would be collected in order to detect small to medium effects. Six participants were excluded because they did not follow instructions on the word fragment completion task. These six participants either put more than one word for each fragment or they entered only the missing letters instead of the full word, contrary to the bolded instructions prior to beginning the task. Four participants completed the study more than once. The time-stamps were examined and the earliest completion time was retained. After excluding these ten, there were 248 participants included in analyses. Women comprised 68.83% ($n = 170$). One participant did not give an answer to the gender item. The mean age of participants was 38.95 years old, $SD = 13.95$, *median* = 36. One participant did not provide their age.

2.3 Individual Difference Measures

2.3.1 Moral Foundations Questionnaire.

The 30 item moral foundations questionnaire (MFQ-30; Graham et al., 2011) was used to measure explicit endorsement of each moral foundation. The MFQ-30 uses 6 items to measure each moral foundation. The foundations tend to be correlated with one another (e.g. see Table 4). For the first 15 items of the MFQ-30 participants are asked, "To what extent are the following considerations relevant to your thinking?" (0, not at all relevant to 6, extremely relevant) with items including "Whether or not someone suffered emotionally" (harm/care foundation) and "Whether or not someone

showed a lack of respect for authority" (authority/respect foundation). The next 15 items asked participants to indicate their level of agreement with items on a 6 point scale (strongly disagree to strongly agree with no midpoint). Items on this half include, "Compassion for those who are suffering is the most crucial virtue," and, "Respect for authority is something all children need to learn." Missing values on the MFQ-30 were imputed with a single imputation using SAS' Proc MI. Descriptive statistics for each foundation are as follows: Harm/Care, $\alpha = .72$, $M = 4.61$, $SD = 0.82$; Fairness/Reciprocity, $\alpha = .67$, $M = 4.59$, $SD = 0.73$; Ingroup/Loyalty, $\alpha = .81$, $M = 3.42$, $SD = 0.97$; Authority/Respect, $\alpha = .80$, $M = 3.71$, $SD = 0.99$; Purity/Sanctity, $\alpha = .88$, $M = 3.34$, $SD = 1.35$.

2.3.2 Authoritarianism. Authoritarianism was measured using a shortened version of Altemeyer's Right-wing Authoritarianism scale (McFarland, 2010). This 11 item measure is scored on a 9 point scale (strongly disagree to strongly agree), $\alpha = .92$, $M = 3.82$, $SD = 1.81$. A parallel analysis comparing principal component eigenvalues with random data eigenvalues confirms a one factor solution. Missing values were imputed with a single imputation using SAS' Proc MI. Scale items include: "It is always better to trust the judgment of the proper authorities in government and religion than to listen to the noisy rabble-rousers in our society who are trying to create doubt in people's minds," and, "It is best to treat dissenters with leniency and an open mind, since new ideas are the lifeblood of progressive change" (reverse-scored).

2.3.4 Social Dominance Orientation. Social dominance was measured with Sidanius and Pratto's (2004) Social Dominance Orientation Scale. This 16 item measure was scored on a 5 point scale (Strongly disagree to strongly agree), $\alpha = .93$, $M = 1.98$, $SD = 0.77$. Social dominance consists of two related factors, group based dominance and support for inequality. In

this study, the factors were not analyzed separately. A parallel analysis comparing principal component eigenvalues with random data eigenvalues confirms a two factor solution. Sample items include, "Sometimes other groups must be kept in their place," and, "It would be good if groups could be equal" (reverse-scored). Missing values were imputed using SAS' Proc MI with a single imputation.

2.3.5 Political Orientation. Political orientation was measured by taking the mean of political party preference (1 – 7, Strong Democrat to Strong Republican, *Other* option excluded), economic ideology (1 – 7, Strongly Liberal to Strongly Conservative, *Other* option excluded), and social ideology (1 – 7 , Strongly Liberal to Strongly Conservative, *Other* option excluded), $N = 226$, $\alpha = .86$, $M = 3.53$, $SD = 1.55$. There were fewer participants included in this variable because people who chose the *other* option for any of the three items were excluded rather than imputing values or computing a mean based on fewer than all three of the ideology variables. Thus, this measure captures partisan political orientation.

2.4 Results

For each moral foundation, a minimum of 12 word fragments were retained that participants created as moral foundation congruent words 20 - 80% of the time, and which were completed with any word at least 80% of the time. (see supplemental material at scottfrankowski.wordpress.com for a list of all fragments tested and retained, and the SAS syntax for coding fragments as congruent or incongruent). For each foundation virtue and vice, the number of items retained, the means for completing the fragments foundation congruently, standard deviations, and ranges can be found in Table 1. After assessing each foundation's virtues and vices, I also collapsed foundations to create variables that were a composite of the individualizing foundation's virtues and vices and the binding foundation's virtues and vices.

Thus, the new variables were as follows:

Individualizing Virtue — Composite of the harm/care virtues and fairness/reciprocity virtues (completed fragments would include words such as care, defend, fair, honest);

Binding Virtue — Composite of the ingroup/loyalty virtues, authority/respect virtues, and purity/sanctity virtues (e.g. conform, leader, holy, virgin);

Individualizing Vice — Composite of the harm/care vices and the fairness/reciprocity vices (e.g. exploit, harm, dishonest, rude);

Binding Vice — Composite of the ingroup/loyalty vices, authority/respect vices, and purity/sanctity vices (e.g. traitor, resist, guilt, slut).

Collapsing across foundations is in-line with previous research on implicit moral foundation accessibility (see Leidner et al., 2012; Graham, 2010). Furthermore, the subscales on the MFQ-30 were collapsed to create variables for explicit endorsement of the individualizing foundations and the binding foundations.

Interestingly, the ideology variables consistently correlated negatively with the implicit binding vices. Greater political conservatism correlated with completing the binding vice fragments as moral foundation congruent less often, $r = -.21$, $p = .002$. The more conservative participants were, they were less likely to complete word fragments such as sex, sin, disgusted, defy (i.e. binding vice words). Similarly, authoritarianism and social dominance orientation negatively correlated with the binding vices, $r = -.20$, $p = .001$ $r = -.18$, $p = .004$, respectively. Moreover, endorsement of the binding foundations as measured by the MFQ-30 correlated negatively with implicit binding vices, $r = -.22$, $p < .001$. Endorsement of the binding foundations also correlated negatively with implicit individualizing virtues ($r = -.14$, $p = .03$) *and* individualizing vices ($r = -.15$, $p = .02$); although, it was not associated with implicit accessibility

of the binding virtues, $r = .01$, *ns*. Explicit individualizing moral foundations only correlated positively with implicit individualizing vices, $r = .15$, $p = .02$. See tables 2 and 3.

2.5 Test-Retest Reliability

We assessed test-retest reliability on these data, in-line with recommendations by Koopman et al. (2013). Five to six weeks after participating, 230 of the 258 participants were contacted to take part in a follow-up study. The follow-up study consisted of only the word fragments that were tested in the first study (i.e. participants did not complete the individual differences measures again). Participants were compensated $.50 for their time. Of those contacted, 120 (52%) participated in this follow-up study. Of these 120, eight were excluded for not following instructions or not completing the task. Thus, test-retest reliability was analyzed with data from 112 participants.

2.5.1 Test-Retest Results. The word-fragment measure had fair test-retest reliability, $r = .57$, $p < .0001$. Comparably, the widely used Implicit Association Test has an average test-retest reliability of $r = .56$ (Greenwald, Poehlman, Uhlman, & Benaji, 2009). The open-ended response options for the word fragment completion task and the instructions to complete each fragment as quickly as possible restrict the reliability of the measure compared to explicit measures that use interval scales.

The descriptive statistics of this test-retest reliability study can be found in Table 1. For each moral foundation virtue and vice the means for the percent of fragments completed congruently were greater in the follow-up study than in the first study. In turn, the correlations between the foundations' word fragments at time two and the individual difference measures from time one strengthened (See Table 2). These results indicate that there may be carryover effects from the first study.

2.6 Discussion

We achieved the primary aim of this pilot study by retaining 10 or more word fragments for each moral foundation's virtues and vices. From this point on, the word fragment task will be referred to the Implicit Moral Foundation Task (IMFT). Future research can further develop the IMFT to assess how susceptible to manipulation each fragment is. Some fragments may always have the same level of activation aggregated across a sample regardless of primes or individual differences. Such fragments may not be useful and may later be discarded as a way to increase the power the measure.

Although the association between IMFT scores and individual differences were not strong in this study, the data suggest there may be dispositional states of automatic moral foundation accessibility. Note, on the IMFT conservatives differed from liberals, high authoritarians differed from low authoritarians, those high in social dominance differed from those who were low on the measure, and those who explicitly endorsed the binding foundations differed from those who did not. The patterns found in this study differ from those found in Graham's (2010) dissertation studies and in Leidner et al.'s (2012) morality shifting studies.

An interesting pattern emerges in that those who explicitly endorse the binding foundations (i.e. various forms of ideological conservatives – authoritarianism, social dominance, and political conservatism) showed lower scores on the IMFT binding vices. Moreover, those who explicitly endorse the individualizing foundations show increased scores on the IMFT individualizing vices. Thus, there is a paradoxical effect in that those who tend to endorse the binding foundations are *less* likely to see violations (vices) of the binding foundation. Comparatively, those who endorse individualizing foundations are *more* apt to see violations (vices) of harm/care and fairness. To put into context with examples: Social

conservatives ought to be less likely to see vices of authority such as tyranny, vices of ingroup/loyalty such as racism, and less likely to see vices of purity such as rape. Conversely, secular liberals ought to be more likely to see vices of harm such as abuse, and vices of fairness such as cheating.

This paradoxical effect found with the binding vices could can be related to how Haidt (2013) talks of how morality can bind and blind. Morality can bind us to our social groups while also blinding us to the plights and injustices experienced by those outside of our groups and to the injustices perpetuated by our own group. The results of this pilot study provide literal evidence of the blinding of morality. Those who are more apt to see group cohesion as virtuous (i.e. those who endorse the binding foundations), are more likely to be blind to vices of the binding moral foundations.

The relationships found in this pilot study between the IMFT and individual differences, although not predicted in regards to the foundation virtues and vices, were taken into consideration in the thesis experiment. If the results of the pilot study replicate, *lower* binding vices scores on the IMFT ought to predict greater moral disapproval of a violation of a binding foundation; and conversely, higher individualizing vice scores on the IMFT ought to predict greater moral disapproval of a violation of an individualizing foundation. In the following experiment, this model of implicit moral accessibility is tested.

Study 2

In the first study we found, as hypothesized, that scores on the IMFT correlated with explicit foundation endorsement and with other measures that have been used in moral foundations research. The purpose of this second study is to test the IMFT in an experiment. Furthermore, in study two, we present a violation of both an individualizing foundation and binding foundation, which provides a test of the previously found results in which endorsement of the binding foundation was associated with lower binding vice IMFT scores, and endorsement of the individualizing foundation was associated with higher individualizing vice IMFT scores.

The primary aim of this study was to partially validate the IMFT by showing its susceptibility to a moral violation prime as well as its ability to predict moral disapproval and moral action. This word fragment measure ought to predict moral disapproval of the culture-war topics chosen (death penalty representing a harm/care violation and pornography representing a purity/sanctity violation), and show appropriate convergent and discriminant validity with other measures.

The secondary aim of this study is to replicate the pilot study results by testing whether inhibition of a binding foundation vice (purity) leads to increased moral disapproval of pornography (a purity violation); whereas, increased activation of a individualizing foundation vice (harm/care) leads to increased moral disapproval of the death penalty (a harm/care violation). I chose pornography and the death penalty because purity was a strong predictor of moral disproval of pornography and harm/care was a strong predictor of the death penalty, while purity and harm/care did not predict disapproval of the incongruent moral violation (Koleva et al., 2011).

3.1 Hypotheses

1. IMFT scores should be susceptible to manipulation, such that in each condition participants will complete more fragments associated with the moral violation than in the other condition.

2. After priming a moral violation, IMFT scores will predict moral disapproval of the associated culture-war topic and increased donations to a related organization.

3. The IMFT will show appropriate discriminant validity. In the purity/sanctity condition, moral disapproval of pornography will *not* be associated with the harm/care foundation of the IMFT. Furthermore, in the harm/care condition, moral disapproval of the death penalty will *not* be associated with the purity/sanctity foundation of the IMFT.

4. Extending Graham et al.'s results (2011), and to further show appropriate convergent validity, the harm/care foundation of the IMFT should correlate with Davis' (1983) measure of dispositional empathy, and purity/sanctity foundation of the IMFT should correlate with Schwartz's (1992) values measure.

5. In the purity/sanctity condition (i.e. a binding moral foundation), lower purity vice IMFT scores will predict greater moral disapproval of using pornography.

6. In the harm/care condition (i.e. an individualizing moral foundation), higher harm/care vice IMFT scores will predict greater moral disapproval of the death penalty.

7. Greater harm/care vice IMFT scores ought to correlate with greater dispositional empathy (measured on the Davis, 1983 scale); whereas, purity/sanctity vice IMFT scores ought to correlate with greater endorsement of the self-discipline, cleanliness, and devotion components of the values scale (measured with the Schwartz, 1992 value items). Furthermore, replicating the pilot study, lower purity vice IMFT scores should correlate with higher scores on

a scale measuring authoritarianism.

8. There ought to be a condition by explicit moral foundation endorsement interaction such that greater explicit endorsement of the purity/sanctity foundation will predict lower purity vices IMFT scores in the purity condition compared to the harm/care condition; and, greater explicit endorsement of the harm/care foundation ought to predict higher harm/care vice IMFT scores in the harm/care condition compared to the purity/sanctity condition.

9. Members of the thesis committee suggested that I also investigate how the implicit measure interacts with the explicit moral foundations measure (MFQ-30) to predict moral attitudes and action. To this end, I also hypothesized that harm/care vice IMFT scores with greater explicit harm/care endorsement ought to predict greater disapproval of the death penalty and increased donations to an anti-death-penalty organization. Conversely, lower purity vice IMFT scores with greater explicit endorsement of the purity foundation ought to predict greater disapproval of pornography and increased donations to an anti-pornography organization.

3.2 Methods

3.2.1 Participants.

Data were collected from 315 participants. Three participants were excluded for failing to summarize the vignette or randomly responding. One participant was excluded for being a non-U.S. citizen, one other was excluded for not answering the citizenship question. Fourteen participants were excluded for spending less than 20 seconds reading the priming vignettes (of 400 words). Thus, data were analyzed from 296 participants. Fifty-seven percent of the sample were women (n = 168), 42 percent were men (n = 124), and four individuals identified as gender-variant or transgender. The mean age was 37.60 (SD = 13.09, *range* = 18 - 76). Eighty-four percent stated they were White (n = 249; Black, n = 25;

Asian, *n* = 17; Latino, *n* = 12*;* one Native American, one Pacific Islander, and one *other*; *note, participants could choose more than one option*).

Because the IMFT requires that participants be completely fluent in English, a question was asked to determine whether participants should be excluded for fluency. Nearly all of the sample indicated that English was their first language (98.3%, *n* = 291). Of four participants who stated that English was not their first language, when asked to rate on a scale of 1 - 7 their fluency (1 = Not fluent at all, to 7 = just as fluent as anyone whose first language is English), one participant chose 6, and the three others chose 7, thus no participants were excluded based on their fluency of English.

3.2.2 Procedure. Participants were recruited through M-Turk. The study was listed as an attitudes and linguistic study. Participants were paid $1.00 for their time. After agreeing to participate, participants were given a link an informed consent form in UTEP's Qualtrics system. After signing the informed consent with their M-Turk ID number, they were forwarded to the study. Participants were given a brief overview of the tasks in the study and before the priming vignettes they were given instructions and examples of the word fragment task. After these practice fragments, participants were randomly assigned to read a vignette of approximately 400 words about the death penalty or pornography. These vignettes were framed with debate points both for and against the topic. The death penalty is a violation of the harm/care moral foundation; pornography is a violation of the purity foundation. These violations were chosen because in previous research it was shown that harm/care endorsement predicted death penalty disapproval and purity endorsement predicted disapproval of pornography (Koleva et al., 2012); furthermore, in Koleva et al.'s research, harm/care endorsement did not predict moral attitudes toward pornography and purity endorsement did not predict moral attitudes toward the death

penalty. Thus, these two moral violations allow for a test of discriminant validity of the word fragment measure.

After reading the vignette, participants were asked to summarize what they read. This served as a manipulation check. Immediately after this check participants completed 73 word fragments: Twelve fragments represented the harm/care virtues, 15 harm/care vices, 12 purity virtues, 10 purity vices, and 24 neutral word fragments. The fragments used can be found in the appendix. Many of the neutral words served as a manipulation check in that they were easy to make into many different words (i.e. '_at' can be cat, hat, bat, pat, etc.). All fragments were randomized. Participants were shown one fragment on each screen. Using JavaScript code in Qualtrics, the cursor would always be in the text area when presented with each fragment and participants could hit the 'Enter' button to advance to the next fragment; thus, participants never had to take their hands off the keyboard. A timer, not visible to participants, captured the time it took them to complete each fragment. To my knowledge, the current experiment is the first to use a timer on individual word fragments. In pretesting, fragments were excluded if enough participants did not correctly fill in a word for a given fragment. With the current experiment, timing variables were used on individual word fragments to allow for analyses of fragments that participants are completing but which may be taking a long time to generate responses. Participants were instructed to fill in a word as quickly as possible; therefore, if there are fragments for which participants are taking time to ponder, those fragments can be excluded from this type of measure.

After completing the word fragments, participants answered six items about their moral disapproval of both pornography and the death penalty. There were three items for each moral violation, answered on a seven-point scale (1 = strongly disagree to 7 = strongly agree): "I

believe that using pornography (the death penalty) is morally wrong;" "I believe that other people should be opposed to using pornography (the death penalty);" and, "I believe that the government should create and enforce laws that restrict the use of pornography (the death penalty)." Both the death penalty disapproval (M = 4.12, SD = 1.86, α = .87) and pornography disapproval scales (M = 3.21, SD = 1.89, α = .90) had strong inter-item reliability.

After the moral disapproval items, participants were able to make a donation of up to $.50 to a charity against the death penalty (The National Coalition to Abolish the Death Penalty), a charity against pornography (Morality in the Media), or they were able to allocate the money to themselves as a bonus. Participants could allocate the $.50 any way they wanted, as long as the total added up to $.50. Sixty-six percent of participants kept all of the bonus for themselves (n = 195). Participants overwhelmingly allocated money to themselves, M = $.40, SD = $.16. Fifty-eight participants (18.71%) allocated money to the anti-porn charity, M = $.21, SD = $.14. Seventy-four participants (23.79%) allocated money to the anti-death-penalty charity, M = $.24, SD = $.16.

After the moral disapproval items, participants completed the MFQ-30, a measure of authoritarianism, the Schwartz Values Scale, a scale of trait empathy, and demographics that included political orientation questions.

3.2.3 Measures. ***MFQ-30***. A description of this measure and sample items can be found in the pilot study section of this paper. The subscales of the MFQ-30 had good reliability: Harm/care, M = 4.76, SD = .82, α = .75; Fairness/reciprocity, M = 4.70, SD = .77, α = .74; Ingroup/loyalty, M = 3.56, SD = .99, α = .77; Authority/respect, M = 3.40, SD = .96, α = .76; and, Purity/sanctity, M = 3.16, SD = 1.39, α = .89.

Authoritarianism. A description of this measure and sample items can be found in the

pilot study section of this paper. This single factor scale showed good reliability, $M = 3.40$, $SD = 1.81$, $\alpha = .92$.

Values scale. The Schwartz Value scale consisted of three items asking participants to rate the importance of each value (Self-discipline, Devout, and Clean) as a guiding principle in life (1 = Not important at all, to 7 = Of supreme importance). These three items showed fairly poor inter-item reliability, $M = 4.68$, $SD = 1.26$, $\alpha = .58$.

Empathy. Davis' (1983) measure of trait empathy consists of 14 items. Items, scored on a seven point scale (1 = strongly disagree to 7 = strongly agree), include, "Seeing warm, emotional scenes melts my heart and makes me teary-eyed." And, "Occasionally I am not very sympathetic to my friends when they are depressed" (reverse scored). This single factor scale had good reliability, $M = 5.41$, $SD = 1.06$, $\alpha = .93$.

3.3 Results

3.3.1 Timing of word fragments. I first looked at the mean response times of word fragments. There were fragments with long mean times (greater than ten seconds) and with large standard deviations. However, with some of these fragments, it was clear that only a few responders skewed the mean by taking minutes to respond (i.e. they may have strayed away from the task). Thus, I also looked at quantiles of the timing distribution. I chose to exclude fragments in which 25% or more of the sample completed the fragment in more than ten seconds. This ten second cut-off seems appropriate given the mean time that participants completed retained fragments, $M_{seconds} = 5.58$, $SD = 3.49$. In total, nine word fragments were excluded from analyses based on how long participants took to complete them: From the care virtues, healthy, and unharmed, were excluded (10 were retained); from the care vices, misery, torture, and violent, were excluded (12 were retained); from the purity virtues, chaste, maiden,

modest, and purity, were excluded (8 were retained); and from the pure vices, disgusted was excluded (9 were retained). See Table 5 for the descriptive timing statistics of these excluded fragments. For all of the retained fragments, individual response times that exceeded 20 seconds were coded as a fragment that was completed target-incongruently. Table 6 shows the percentages of fragments completed target congruently for the care and purity virtues and vices.

3.3.2 Condition priming affecting implicit accessibility. Unless otherwise specified, analyses were conducted with a general linear model. The reported effects are the type III sums of squares analyses in which the effect is reported holding all other effects in the model constant. IMFT scores did not differ between conditions, all Fs < 1.00. Moreover, there were no condition interactions when analyzing IMFT scores within-subjects, all ps > .20. Thus, the first hypothesis, which predicted that responses on the measure would differ as a function of priming a moral violation, was not supported.

3.3.3 Implicit moral accessibility predicting moral attitudes. Hypothesis two states that the IMFT will predict moral disapproval after priming a moral violation. The model including the condition primes, IMFT care scores (composite of virtues and vices), and the interaction term, was significant, $F(3, 292) = 3.26$, $p = .022$, $\eta^2 = .032$. Analyzing the effects in this model, harm/care IMFT scores predicted disapproval of the death penalty $F(1, 292) = 7.69$, $p = .01$, $\eta^2_p = .026$; however, there was no significant interaction with condition, $F(1, 292) = 2.38$, $p = .12$. In a hierarchal regression, the care scores on the IMFT marginally predicted death penalty disapproval beyond the explicit the harm/care foundation of the MFQ-30, $b = 1.27$, $SE = .74$, $t(293) = 1.73$, $p = .085$, $\Delta R^2 = .009$. In support of hypothesis five, which states that implicit care *vices* would predict moral disapproval whereas the virtues would not, the model including the primes, care vices, and the interaction term as predictors was significant, $F(3, 292) = 3.32$, p

= .02, η^2 = .033. Deconstructing the effects in this model, greater harm/care vice IMFT scores predicted death penalty disapproval, $F(1, 292) = 7.52$, $p = .007$, $\eta^2_p = .025$; however, there was not a significant interaction with condition, $F(1, 292) = 2.28$, $p = .13$. In a hierarchal regression, the care vice scores on the IMFT marginally predicted death penalty disapproval beyond the explicit the harm/care foundation of the MFQ-30, $b = 1.03$, $SE = .55$, $t(293) = 1.90$, $p = .062$, $\Delta R^2 = .011$. Care virtue IMFT scores did not predict moral disapproval, $F(3, 292) = 0.89$, $p = .45$. Care IMFT scores did not predict disapproval of pornography either as a main effect or after a prime, $F(3, 292) = 1.93$, $p = .13$, thus partially supporting hypothesis three by showing discriminant validity of the harm/care word fragments.

The IMFT purity scores (composite of virtues and vices) did not predict disapproval of pornography either alone or after the priming manipulation, $F(3, 292) = 1.77$ $p = .15$. While the composite of the purity fragments did not predict disapproval of pornography, the model including the priming manipulations, IMFT purity *virtues*, and the interaction term was significant, $F(1, 292) = 3.32$, $p = .02$, $\eta^2 = .033$. Deconstructing the effects in this model, greater implicit purity virtues predicted greater disapproval of pornography, $F(1, 292) = 6.17$, $p = .014$, $\eta^2_p = .021$; however, there was no interaction with the prime, $F(1, 292) = 0.55$, $p = . 46$. In a hierarchal regression, the purity virtue scores on the IMFT did not predict pornography disapproval beyond the explicit purity foundation of the MFQ-30, $p > .50$. Contrary to predictions (hypothesis six), increased purity *vice* IMFT scores did not predict greater disapproval of pornography, $F(3, 292) = 1.43$, $p = .23$. Unpredicted, and contrary to hypothesis three of the primary aims which hypothesizes discriminant validity of the measure, greater IMFT purity scores predicted increased death penalty disapproval when primed with the death penalty and decreased death penalty disapproval when primed with the pornography, $F(1, 292) = 5.63$, p

= .018, η^2_p = .019.

3.3.4 Implicit accessibility predicting moral action. Per hypothesis two, the implicit measure should predict increased donations in the corresponding condition. IMFT harm/care (both virtues and vices) did not predict donations to an organization opposed to the death penalty, $F(3, 293) = 1.77$, $p = .15$. IMFT purity (both virtues and vices) did not predict donations to an organization opposed to pornography, $F(3, 292) = 1.01$, $p = .39$. IMFT purity virtues, however, predicted increased donations to an anti-pornography organization, $F(1, 290) = 4.23$, $p = .041$, $\eta^2_p = .014$. In a hierarchal regression, however, increased purity virtue IMFT scores did not predict donations beyond the explicit purity foundation of the MFQ-30, $p > .40$. The interaction between condition and purity virtues was not significant, $F(1, 289) = 2.37$, $p = .12$. IMFT purity vices did not predict donations, $F(3, 292) = 1.42$, $p = .24$.

3.3.5 Explicit moral foundations predicting implicit moral foundation accessibility. For hypothesis eight I predicted that the explicit measure of moral foundation endorsement ought to interact with a moral violation prime to predict IMFT scores. Specifically, in the death penalty condition, endorsement of the harm/care foundation ought to predict increased IMFT care vice scores; and in the pornography condition, explicit purity endorsement ought to predict lower IMFT purity vice scores. The model including the priming conditions, explicit harm/care endorsement, and the interaction term as predictors of IMFT care vices, was marginally significant, $F(3, 292) = 2.56$, $p = .056$, $\eta^2 = .026$. Analyzing the effects in this model, increased explicit harm/care endorsement predicted increased IMFT care vice scores, $F(1, 292) = 7.58$, $p = .006$, $\eta^2_p = .025$, however, there was no interaction with the prime. The full model including explicit harm/care endorsement, the primes, and the interaction, in predicting IMFT care virtue scores, was not significant $F(3, 292) = 1.74$, $p = .16$. Unpredicted, and counter to predictions

about the discriminant validity of the word fragment measure, explicit endorsement of the harm/care foundation also predicted purity virtue accessibility, $F(1, 292) = 5.26$, $p = .023$, $\eta^2_p = .018$.

A model predicting IMFT purity virtue scores from explicit endorsement of purity, condition primes, and the interaction term, was significant, $F(3, 292) = 5.06$, $p = .002$, $\eta^2 = .049$. Analyzing the effects in this model, explicit endorsement of the purity foundation predicted greater IMFT purity virtues scores, $F(1, 292) = 12.60$, $p < .001$, $\eta^2_p = .041$, but there was no interaction with the condition primes. Explicit purity endorsement did not predict IMFT purity vice scores alone nor interacting with the primes, $F(3, 292) = 0.39$, $p = .76$.

3.3.6 Implicit accessibility interacting with explicit endorsement to predict moral disapproval. Per recommendations from committee members, I also ran analyses with the implicit measures interacting with the explicit measures of moral foundations. A model including explicit harm/care endorsement, IMFT harm/care, priming condition, and interaction terms, was significant, $F(7, 288) = 7.23$, $p < .001$, $\eta^2 = .149$. Deconstructing this interaction, IMFT harm/care interacted with explicit harm/care endorsement in an unpredicted manner such that greater IMFT harm/care with greater explicit endorsement predicted greater death penalty disapproval in the pornography condition compared to the death penalty condition, $F(1, 288) = 8.01$, $p = .005$, $\eta^2_p = .027$. Analyzing the care virtues and vices, IMFT care virtues with greater explicit harm/care endorsement predicted death penalty disapproval in the pornography condition, but not the death penalty condition, contrary to hypotheses of discriminant validity, $F(1, 141) = 4.66$, $p = .033$, $\eta^2_p = .032$. As predicted in hypothesis nine, which pertain to the implicit vices, increased IMFT care vices but *not* virtues with greater explicit harm/care endorsement predicted greater death penalty disapproval in the death penalty condition, $F(1,$

147) = 6.33, p = .013, η^2_p = .041. Implicit purity interacting with explicit purity endorsement did not predict moral disapproval.

3.3.7 Implicit accessibility interacting with explicit endorsement to predict moral action.

IMFT care interacting with explicit harm/care endorsement did not predict increased donations to an organization that is against the death penalty. There was, however, an unpredicted IMFT care virtue, by explicit harm/care, by condition interaction such that increased IMFT care virtues with increased explicit care endorsement in the death penalty condition predicted greater anti-pornography donations, F(1, 146) = 7.60, p = .007, η^2_p = .049. Purity IMFT scores (both virtues and vices), with increased explicit purity endorsement, predicted increased donations to the anti-pornography organization in the pornography condition, F(1, 140) = 4.37, p = .039, η^2_p = .03.

3.3.8 Correlations among explicit measures. Replicating prior research (Koleva et al., 2011), the harm/care foundation on the MFQ-30 correlated strongly with empathy, r = .49, p < .001. Similarly, the purity foundation on the MFQ30 correlated strongly with Schwartz' Values Questionnaire, r = .66, p < .001, and with a measure of authoritarianism, r = .77, p < .001.

3.3.9 Word fragment correlations with other measures. I also examined correlations to assess whether the IMFT showed similar convergent and discriminant validity as the explicit moral foundations measure. Hypothesis four states that the implicit measure would be associated with these explicit measures; and furthermore, hypothesis seven states that the explicit measures will be associated with the vices, rather than the virtues. Examining the implicit measure, the IMFT care virtues weakly correlated with the harm/care foundation of the

MFQ30, $r = .12$, $p = .039$, and with dispositional empathy, $r = .13$, $p = .021$. The IMFT care vices also weakly correlated with the harm/care foundation of the MFQ-30, $r = .16$, $p = .006$, and dispositional empathy, $r = .12$, $p = .009$. The IMFT purity virtues weakly correlated with the harm foundation of the MFQ-30, $r = .14$, $p = .017$; the purity foundation of the MFQ-30, $r = .20$, $p < .001$; empathy, $r = .14$, $p = .015$; values, $r = .14$, $p = .016$; and authoritarianism, $r = .17$, $p = .002$. These associations were not present in the pilot study. Contrary to the pilot study in which IMFT purity vices were negatively correlated with the purity foundation of the MFQ-30 and authoritarianism, IMFT purity vices did not significantly correlate with any explicit measures. See Table 7.

3.4 Discussion

Although I hypothesized that priming people with the moral violations of the death penalty or pornography would lead to differences in implicit accessibility as measured with the IMFT, this was not the case. There were no differences in accessibility between conditions. Furthermore, participants within conditions did not show greater accessibility of the corresponding moral foundation. While vignettes used in previous research were able to prime implicit moral foundation accessibility (e.g. Leidner et al., 2012), the vignettes used in the present study, for the most part, did not affect responses on the IMFT. Perhaps the vignettes were not strong enough to prime a moral violation. The tone of a debate with well-reasoned positions on both sides of the argument may suppress the gut-level affective reactions proposed by Haidt (2001) in the social intuitionist framework of moral judgments. Future experiments ought to use normed affective pictures to prime moral violations (e.g. International Affective Picture System; Lang, Bradley, & Cuthbert, 2008).

While the condition primes did not affect IMFT scores, in some instances the IMFT did predict moral disapproval in line with hypotheses. As hypothesized, the IMFT harm/care foundation predicted disapproval of the death penalty. I predicted that IMFT care vices would predict moral disapproval more so than the virtues. This prediction was supported. Furthermore, for the harm/care foundation, the IMFT showed discriminant validity in that it did not predict disapproval of pornography, as hypothesized. For the purity foundation, IMFT purity virtues, and not the vices, predicted pornography disapproval. This effect partially supports hypotheses, however, I had predicted based on pretesting this measure, that lower IMFT purity vice scores would predict moral disapproval rather than accessibility of the purity virtues. Pre-testing the measure provided evidence that lower IMFT purity vice scores (and the binding foundation vices in general) are associated with explicit endorsement the purity foundation and explicit ideological attitudes that are known to be positively associated with the purity foundation. Namely, in pretesting, IMFT purity vice scores were associated with explicit purity endorsement, political conservatism, and authoritarianism (r = -.17, -.16, -.18, respectively, all ps < .05). Those associations did not replicate in the current experiment (See Tables 2 and 7).

There was some support that the implicit measure predicted moral action in the form of donations. IMFT purity virtue scores predicted moral action whereas IMFT harm/care scores did not. IMFT purity virtues, but not the vices, predicted greater donations to an anti-pornography organization. Furthermore, there was a priming effect in analyzing the interaction of explicit and implicit measures such that greater explicit purity endorsement with increased IMFT purity virtues predicted greater donations to an anti-pornography organization in the pornography condition. Saliency of a moral violation in those who are most likely to view pornography as a moral violation then prompts the automatic responding on the implicit measure which predicts

moral action. This interaction occurred only with the purity violation, though, not the harm/care violation. With the current study, accessibility of a virtue predicted moral action, and accessibility of the vices predicted moral disapproval. While not a predicted effect, this could be an interesting route for future research.

While the prime alone did not produce changes in word fragment completions, the prime interacting with the IMFT responses predicted disapproval of the moral violations while controlling for the variance of the explicit measure and implicit measure on their own. Supporting secondary hypotheses regarding the foundation vices, greater explicit harm/care endorsement with increased IMFT care vice scores predicted greater death penalty disapproval in the death penalty condition. When looking at the composite of the care virtue and vices, however, increased implicit accessibility with increased explicit endorsement predicted disapproval of the death penalty in the pornography condition. While this effect may suggest that the word fragment measure lacks discriminant validity, it could also be that there is not implicit compartmentalization of the moral foundations. Priming with pornography could activate accessibility of the harm/care foundation as many people may have an automatic reaction of pornography violating that moral foundation.

There were other instances of the cross-talk between moral foundations. After a death penalty prime, increased IMFT purity virtue scores predicted death penalty disapproval; whereas, after a pornography prime, increased IMFT purity virtue scores predicted *decreased* disapproval of the death penalty. These could be spurious relationships, or it is possible that the gut level reactions to moral violations are complex and that violations occur across multiple moral foundations in unintuitive ways. Indeed, in Koleva et al.'s (2012) research, they found that increased endorsement of the purity foundation was a strong predictor of negative attitudes

toward immigrants, flag burning, and stem-cell research – issues that may not have immediate connections to the purity foundation.

Another theory of morality proposes that moral transgressions can be viewed as an exchange between an agent committing a moral transgression and the victim, or moral patient, who perceives the moral violation (Gray, Young, & Waytz, 2012). This dyadic theory proposes that morality, in moral foundations theory terms, is essentially harm – real, imagined, or perceived. In this theory of morality, pornography as a violation of purity, can be thought of as a harm to one's self, one's relationships, and one's spirituality, all of which are moral patients of such a moral violation. Morality in this dyadic framework could explain how a pornography prime may activate harm/care themes which may in turn affect moral disapproval of a harm/care violation.

3.4 Future Directions

Going forward, I will be looking closer at the word fragments used in this measure to find fragments which may work and be susceptible to priming. The vignettes used in this study were mostly not able to prime activation of the corresponding moral foundation. For a follow-up study, I will use pictures from the IAPS (Lang et al., 2008) that are associated with the moral foundations and which have high affective arousal ratings. After using these affective primes, I will develop the word fragment measure further by identifying word fragments that were more activated after their associated prime.

Per committee suggestions, I will also further examine the psychometric properties of the IMFT. Each foundation of the IMFT correlated moderately with each other foundation, suggesting that word fluency may be the strongest common factor in the measure. Partialling out

the first factor in an exploratory factor analysis may provide a clearer picture of the factor structure of the IMFT.

An interesting effect found in the pilot study and partially replicated in study two, was that greater implicit individualizing vice accessibility was associated with explicit endorsement of individualizing foundations; whereas decreased implicit binding vice accessibility was associated with explicit endorsement of the binding foundations. With future research we ought to explore this relationship further with other methods. Are secular liberals who strongly endorse individualizing foundations more apt to see violations of those foundations? Are social conservatives who strongly endorse the binding foundations less likely to see violations of aspects of the binding foundations? One way to test this would be a categorization task of moral virtue and violation pictures into moral versus immoral categories in which response times are recorded.

Conclusion

Moral foundations research is in its infancy. There has been a wealth of research to support the idea of multiple foundations that people rely on when making decisions about what is right or wrong. Central to moral foundations theory is the automaticity of moral judgments. There is very little research, however, that shows that morality is an automatic process. Examining the interplay between reasoning and automatic processes will fill a critical gap in knowledge about how people make decisions about right and wrong.

References

Altemeyer, B. (1996). *The authoritarian specter* (p. 374). Cambridge, MA: Harvard University Press.

Altemeyer, B. (2004). Highly dominating, highly authoritarian personalities. *The Journal of Social Psychology*, *144*(4), 421-448.

Davis, M. H. (1983). The effects of dispositional empathy on emotional reaction and helping: A multidimensional approach. *Journal of Personality*, *51*(2), 167-184.

Graham, J. (2010). Left Gut Right Gut: Ideology and Automatic Moral Reactions. (Unpublished doctoral dissertation). University of Virginia, Charlottesville, VA.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029-1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*,*101*(2), 366-385.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17-41.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814-834.

Haidt, J. (2007, September 21). Moral psychology and the misunderstanding of religion. Retrieved from http://www.edge.org/conversation/moral-psychology-and-the-misunderstanding-of-religion.

Haidt, J. (2013). *The righteous mind: Why good people are divided by politics and religion*. Random House LLC.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), 110-119.

Kohlberg, L. (1963). The development of children's orientations toward a moral order. *Human Development*, *6*(1-2), 11-33.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality*, *46*(2), 184-194.

Koopman, J., Howe, M., Johnson, R. E., Tan, J. A., & Chang, C. H. (2013). A framework for developing word fragment completion tasks. *Human Resource Management Review, 23*, 242-253.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainsville, FL.

Leidner, B., & Castano, E. (2012). Morality shifting in the context of intergroup violence. *European Journal of Social Psychology*, *42*(1), 82-91.

McFarland, S. (2010). Authoritarianism, social dominance, and other roots of generalized prejudice. *Political Psychology*, *31*(3), 453-477.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227-234.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277-293.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2007). Linguistic inquiry and word count: LIWC [Computer software]. *Austin, TX: liwc.net*.

Piaget, J. (1932). *The Moral Judgment of the Child.* Translated by Marjorie Gabain. New York, Harcourt, Brace & Co., 1932.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741-763.

Rokeach, M. (1973). *The nature of human values*. Free press.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, *25*(1), 1-65.

Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Ho, A.K., Sibley, C., & Duriez, B. (2013). You're inferior and not worth our concern: The interface between empathy and social dominance orientation. *Journal of Perosnality. 81*(3), 313 – 323.

Sidanius, J., & Pratto, F. (2004). *Social Dominance Theory: A New Synthesis*. Psychology Press.

Tajfel, H., & Turner. JC (1979). An integrative theory of intergroup conflict. *Social psychology of intergroup relations*, 33-47.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge University Press.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *.American Psychologist*, *35*(2), 151 - 175.

Appendix

Table 1. Descriptive statistics for the word fragment measure.

Word Fragments	# of Fragments	Mean % Congruent	*Std. Dev.*	Min. %	Max. %
All Fragments	153	46 (51)	8 (8)	21 (31)	66 (70)
Individualizing Virtue	28	53 (57)	12 (11)	15 (32)	82 (83)
Individualizing Vice	35	47 (49)	12 (12)	10 (19)	76 (76)
Binding Virtue	47	45 (48)	10 (11)	13 (25)	71 (79)
Binding Vice	43	43 (51)	10 (09)	13 (19)	67 (73)
Harm/Care Virtue	12	55 (57)	16 (14)	8 (25)	92 (92)
Harm/Care Vice	17	46 (49)	16 (15)	0 (15)	92 (86)
Fairness Virtue	16	50 (56)	16 (14)	13 (20	81 (100)
Fairness Vice	18	48 (50)	13 (14)	12 (18)	76 (83)
Ingroup/Loyalty Virtue	15	40 (43)	14 (14)	7 (7)	79 (71)
Ingroup/Loyalty Vice	16	50 (57)	14 (12)	12 (19)	81 (88)
Authority Virtue	18	46 (51)	13 (16)	11 (6)	78 (88)
Authority Vice	15	40 (51)	13 (14)	6 (13)	73 (80)
Purity Virtue	14	48 (51)	14 (14)	14 (14)	86 (86)
Purity Vice	12	38 (45)	16 (16)	0 (17)	83 (92)

Note. $N = 248$. Values in parentheses are from the follow-up study in which test-retest reliability was assessed, $N = 112$.

Table 2. Correlations between word fragment individualizing and binding foundations and explicit measures.

Word Fragments	Political Orientation	RWA	SDO	MFQ-Indv	MFQ-Binding
Virtue	.00 (-.05)	-.02 (.00)	-.10 (-.13)	.07 (.04)	-.06 (-.05)
Vice	-.19** (-.24) *	-.20** (.30)**	-.19** (-.25)**	.13* (.13)	-.22*** (-.30)**
Individualizing	-.04 (-.20)*	-.11 (-.21)*	-.14* (-.21)*	.12* (.08)	-.17** (-.26)*
Binding	-.14* (-.13)	-.11 (-.12)	-.15* (-.21)*	.08 (.10)	-.13* (-.13)
Individuating Virtue	.04 (-.01)	-.06 (-.09)	-.12 (-.06)	.06 (-.05)	-.14* (-.20)*
Individualizing Vice	-.10 (-.29)**	-.13* (-.25)**	-.11 (-.26)**	.15* (.17)	-.15* (-.23)*
Binding Virtue	-.02 (-.06)	.02 (.06)	-.06 (-.15)	.06 (.10)	.01 (.06)
Binding Vice	-.21** (-.30)**	-.20** (-.27)**	-.18** (-.18)	.09 (.06)	-.22*** (-.29)**

*$p < .05$, **$p < .01$, ***$p < .001$. $N = 248$. Values in parentheses are the test-retest correlations, $N = 112$. RWA = Right-wing Authoritarianism, SDO = Social Dominance Orientation, MFQ-Indv = composite of the explicit endorsement of the harm/care and fairness/reciprocity moral foundations, MFQ-Binding = composite of the ingroup/loyalty, authority/respect, and purity/sanctity moral foundations.

Table 3. Correlations between word fragments and explicit measures.

Word Fragments	Political Orientation	Authoritarianism	Social Dominance	MFQ-Indvidualizing	MFQ-Binding
Care Virtue	.07	.04	-.04	.09	-.04
Care Vice	-.08	-.13*	-.11	.10	-.16*
Fair Virtue	-.01	-.10	-.14*	.01	-.17**
Fair Vice	-.08	-.09	-.08	.14*	-.08
Ingroup Virtue	-.07	-.19**	-.09	.08	-.17**
Ingroup Vice	-.11	-.12	-.22***	.06	-.16*
Authority Virtue	-.01	.16*	-.05	.04	.12
Authority Vice	-.17*	-.13*	-.07	.04	-.13*
Purity Virtue	.06	.08	.01	.00	.07
Purity Vice	-.17*	-.18**	-.10	.08	-.18**

*$p < .05$, **$p < .01$, ***$p < .001$. *Note:* This table shows the correlations of the ideology variables and explicit moral foundation endorsement with the means of the word fragment categories. *MFQ-Individualizing* is the mean of the Harm and Fairness dimensions of the Moral Foundations Questionnaire. *MFQ-Binding* is the mean of the Loyalty, Authority, and Purity dimensions on the Moral Foundations Questionnaire. $N = 248$ for Authoritarianism, Social Dominance, MFQ-Individualizing, and MFQ-Binding. $N = 226$ for Political Orientation.

Table 4. Correlations among explicit measures.

Variable	1	2	3	4	5	6	7	8	9	10
1. Harm	--									
2. Fairness	.64*	--								
3. Loyalty	.24*	.23*	--							
4. Authority	.13[a]	.09	.75*	--						
5. Purity	.09	.00	.59*	.71*	--					
6. Political Party	-.17[a]	-.26*	.23*	.30*	.28*	--				
7. Political – Economy	-.15[a]	-.23*	.19[a]	.34*	.32*	.65*	--			
8. Political – Social	-.15[a]	-.27*	.30*	.46*	.58*	.71*	.65*	--		
9. RWA	-.09	-.23*	.53*	.69*	.75*	.44*	.44*	.66*	--	
10. SDO	-.40*	-.47*	.17[a]	.27*	.19[a]	.40*	.40*	.42*	.42*	--
11. Gender	.18[a]	.00	.13[a]	.16[a]	.22*	-.03	.01	.05	.17[a]	-.17[a]

[a] $p < .05$, *$p < .001$. . *Note:* RWA = Right-wing Authoritarianism, SDO = Social Dominance Orientation. Harm, Fairness, Loyalty, Authority, and Purity were all measured with the MFQ-30. $N = 248$ for Harm, Fairness, Loyalty, Authority, Purity, RWA, and SDO. $N = 232$ for Political Party; $N = 239$ for Political-Social; and, $N = 238$ for Political-Economic – higher scores correspond to greater conservatism. Gender ($N = 247$, $n_{male} = 77$, $n_{female} = 170$) coded 1 = Male, 2 = Female.

Table 5. Times to complete excluded fragments

	Mean Time (seconds)	*SD*	75th Percentile
Composite of all retained Fragments	5.58	3.49	6.64
Healthy (Hea_ _ _y)	19.10	25.18	21.72
Unharmed (Un_ _ rmed)	17.21	32.04	13.87
M_ser_ (M_ser_)	10.64	14.80	10.73
Torture (T_ _t_re)	19.05	39.07	15.35
Violent (Viol_n_)	9.36	10.76	11.07
Chaste (Ch_st_)	15.50	20.40	16.60
Maiden (M_ _den)	17.37	67.36	14.38
Modest (_ _dest)	11.55	16.95	11.14
Purity (P_ _ity)	14.72	22.83	15.22
Disgusted (Dis_us_ed)	12.48	18.90	12.10

Note. Fragments in which 25% of the sample took more than 10 seconds to complete as a word were excluded from analyses.

Table 6. Percent of fragments completed target congruently.

	Death Penalty		Pornography	
	M	*SD*	*M*	*SD*
Care Virtues	61.12%	16.91%	59.45	17.43
Care Vices	42.50	18.91	43.01	18.56
Purity Virtues	51.66	20.44	53.88	17.78
Purity Vices	38.26	19.10	38.08	19.19

Note. This table shows the percentage of word fragments completed condition congruently in each condition. All *t*s between condition < 1.00, *ns.*

Table 7. Correlations in experiment collapsed across conditions.

Variable	1	2	3	4	5	6	7	8	9
1. Care Virt	-								
2. Care Vice	.23***	-							
3. Pure Virt	.22***	.30***	-						
4. Pure Vice	.23**	.40***	.27***	-					
5. MFQCare	.12*	.16**	.14*	.02	-				
6. MFQPure	-.04	.08	.20***	-.06	.12*	-			
7. Empathy	.13*	.12*	.14*	-.07	.49***	.16**	-		
8. Values	.02	.02	.14**	-.01	.06	.66***	.21***	-	
9. Auth	-.01	-.01	.17**	-.10	-.07	.77***	.07	.57***	-

*$p < .05$, **$p < .01$, *** $p < .001$. *Note.* Variables 1-4 are the implicit variables as measured by the word fragment completion task. MFQ = Moral Foundations Questionnaire. Auth = Right wing Authoritarianism. Values = Schwartz' Values scale. Empathy = Davis' dispositional empathy scale.

Vita

Scott Frankowski is a doctoral student in the Social Cognition Lab in the Social, Cognitive, and Neurosciences program at the University of Texas at El Paso. Scott earned his Bachelor of Science *cum laude* with a major in psychology and a minor in neuroscience from the University of Wisconsin Oshkosh in 2011. Scott's research focuses on perceptions of established norms, authority, and morality. Scott received honorable mentions from the National Science Foundation in 2012 and 2013 for his graduate research program fellowship applications. Scott has been a teaching assistant for many statistics courses including a graduate level stats course for which he was awarded department TA of the year in 2014. Scott has published on group and gender identity and has presented on numerous topics at conferences.

Scott Frankowski
805 Upson Dr.
El Paso, TX 7990