**Belief Change Under Conditions of Moral Conviction**

1. Can support for highly polarized positions, with strong moral conviction be ‘demoralized’?
   1. Via a pragmatic/economic argument?
   2. Via a personal benefit/hedonic argument?
2. Can we increase the effectiveness of social consensus on changing support for highly polarized positions by reducing perceived moral conviction?

Moral convictions are attitudes that are perceived as grounded in fundamental distinctions between right and wrong. Moral conviction, while often correlated with strength of belief, is fundamentally distinct from strong but non-moral attitudes/beliefs. This is because they are perceived as a universal and objective truth, which is relatively resistant to influence from equals or superiors. Thus, morally convicted beliefs are more difficult to change as compared to beliefs grounded in preference or social convention (Skitka et al., 2021). Our research question is fundamentally, how can we change attitudes that are backed by moral conviction.

Given that morally convicted beliefs are hard to change, one plausible solution would be to first ‘demoralize’ the belief, and then attempt change. However, there is mixed evidence that this is effective. Moral conviction has been successfully reduced by framing arguments using pragmatic or economic counterarguments (Kodapanakkal et al., 2022, Kutlaca, 2013), or by emphasizing the personal or hedonic benefit of a counter position (Bastian et al, 2015; Feinberg et al, 2019). In contrast, unsuccessful reduction of moral conviction has been attributed to choosing topics that are already highly politicized and polarized (e.g., COVID-19 vaccination), framings that are dependent on belief in authority influence, or flawed psychometric measurements of moral conviction itself (Aignesberger et al., 2023; Fenzi et al., 2022; Brannon et al., 2019).

For our first study, we plan on testing several methods of moral conviction reduction (pragmatic/economic argument, and a personal/hedonic benefit argument), across an array of contemporary issues that falls along the spectrum of political belief and polarization (support for universal health care, capital punishment, and desire to exercise). Additionally, we plan on using an expanded item for measuring moral conviction that is an adaption of work by Skitka et al., (2021), which is novel insofar as it will measure perceptions of objectivity and universality in belief, which has been assumed but has not been directly assessed. Our goal is to directly reduce moral conviction on these stances, any change in behavior or belief would be a useful, but incidental benefit.

As our final goal is change in attitudes, after ‘demoralization’ we would like to leverage the effects of social influence. One of the strongest findings in psychology is that people conform towards the consensus group opinion (Asch, 1956; Deutsch M, 1955). However, one aspect of morally convicted beliefs that sets them apart from simple ‘strongly held’ beliefs, is that they appear to be independent of normative/majority influence (Skitka et al., 2005). In preliminary work, we were able to successfully manipulate social consensus by presenting false survey results that were assumed to be real, presenting either an artificially high or artificially low level of agreement with a position. We plan to directly test this interaction by attempting to reduce moral conviction and then seeing if that increases conformation towards the societal consensus.

They had problems because they didn’t measure X well, why wasn’t X measured well and what did they do? What are we going to do about X? – This can be it’s own ‘study 1’ the best construct for examining this concept.

Look into more deeply ‘construct validity’ that is the primary concern with regards to this improved psychometric. Is the interpretation of this construct correct? What is it that the study had trouble with, with regards to construct validity? “Model Operation Bias” – using a single measure to assess a construct is seen as kind of a bad thing (this is what is happening with the single item moral conviction screener). Moral conviction is related to strength of belief – how do we measure them differently?

Read a SCALE development paper.

The moral recognition/amplification bit is an extra addendum. Make sure to describe at the front end the two-part process of ‘moralization’ more broadly.

Sarstedt 2007

* Single item measures have some concerns regarding reliability and criterion validity
  + Single items do have some advantages
    - Ease of application
    - Brevity
    - Low costs
    - Higher response rates
    - Single items may be adequately reliable, however… reliable measures are NOT necessarily valid.
* Some single item measures show adequate convergent validity with multi-item measures (correlations from .58 to .68).
  + Multi item measures have their own advantages
    - Greater reliability: Multiple indicators adjusts for random error. The combination of numerous items therefore ‘averages’ out this random error, resulting in a superior measurement value.
      * More items = measurement error decreases and measurement accuracy increases.
    - Greater construct validity: larger set of indicators results in more coverage of many possible distinct ‘construct facets’ (applicable here w.r.t. universality and objectivity, which are two additional facets).
      * Additionally, many theoretically constructs are continuous (e.g., moral conviction), and thus multiple items provide a better approximation of the data.
      * Multi-items = more segmentation of the data = greater variability. This is good in that it helps prevent ceiling and floor effects, and that higher variability in multi-item measures results in higher correlation with the goal criterion.
        + Thus, multi-item measures are predicted to have higher criterion validity.
        + E.g., “A multi-item predictor correlates higher with a multi-item criterion than a single item predictor w/ the same” and “A multi-item predictor has more predictive relevance for a criterion than a single-item predictor.
  + Practical advantages: Multi-item measures are more flexible for missing values, as item-nonresponse can be addressed using data imputation.
* Reliability assessment
  + The extent to which a scale produces consistent results given repeated measurements. E.g., degree to which measurement model is free from random error.
  + For single item measures, standard approach is examining how much each measure predicts a relevant outcome.
* Results: From a strict psychometric perspective, single items have poorer properties than multi-items.
  + Multi-items generalize better with other concrete related attributes.
  + Thus, most constructs are too complex to be measured effectively with a single-item measure.
  + Lower levels of reliability and validity are traded off against the ‘practical advantages’.

Diamantoupoulos 2012:

* Predictive validity of single items can vary dramatically across different constructs.
* Simulation study identifying the influence of difference factors on predictive validity for single and multi item measures.
  + In most conditions for practical applications, multi-item scales have superior predictive validity.
  + Tests by assessing if a single item on the MI scale has comparable or superior performance to the scale in aggregate.
* With single item measures, respondents ignore aspects that are irrelevant to this particular situation, and differentially weight the attributes!
  + This results in problems when the construct is not able to be easily described or made clear to the participants (e.g.,, what the heck is moral conviction?)
  + SI measures require more ‘abstract’ thinking, and therefore can be too vague for respondents to answer ‘correctly’, whereas they could answer individual questions about the item that were smaller in scope.
* Thus, when a concept is nebulous or far from concrete, multi-item measures are the way to go!
  + This is exactly the case in our moral conviction study.

So… we DO need a multi-item thing, how do we set this up?

Scherer 2016: Development of the Medical Maximizer-Minimizer Scale

Study 1

* How did they develop this scale? What steps were important?
* Iterative process that began with generating a list of potentially relevant items.
  + More items were added by the research group and then all items were checked for conceptual clarity and literacy level.
  + Item generation erred towards over-inclusiveness.
* 27 items that were rated on a 7 point likert scale.
  + Items were assessed for sufficient variability (no floor or ceiling type issues)
  + Exploratory Factor Analysis was estimated using Maximum Likelihood.
    - Scree plot resulted in a four factor solution with 51% of variance explained.
  + Item loading on factor 1 consistent with prior definition of maximizing/minimizing (active treatment vs watching and waiting)
  + Item loading on factor 4 consistent with maximum-minimizing, but in a more abstract or extreme way.
  + Item loading on factor 2 correlated with factor 1, but reflected in general avoidance of medicine (overly specific preference)
  + Factor 3 was loaded on the belief that medical treatments did no harm (overinclusiveness).

Study 2

* Confirmatory factor analysis of the retained items from study 1, that were consistent w/ the a-priori definition.
  + Testing for convergent and discriminant validity.
* Testing the divergent validity – Is maximizing/minimizing different from hypochondria/distrust in medicine?’
  + Added a measure of hypochondria and belief in medicine directly to assess discriminant validity.
  + Hypothesized that these results would be related to, but not completely redundant w/ factors in study 1.
* Testing convergent validity – Participants reported their health care utilization and responded to hypothetical health decision scenarios (e.g., exactly what the measure is supposed to predict the behavior of!)
* Participants tested on the main 10 items, and then answered questions about health care utilization, health care access, and short form item survey for hypochondria.
  + Finally, assessed 4 medical decision scenarios related to getting more vs less health care.
* Resulted in a Bifactor Model where the items were loaded on a single factor, and 1-3 were cross-loaded on a second factor.
* Discriminant validity:
  + Maximizing is correlated to, and thus related to, but distinct from, hypochondria.
  + No association with belief in medicine being harmful and distrust in medicine
  + Maximizing/minimizing preferences are mostly independent of health care access.
    - Although this last measure had very little variability, (85% sample had health insurance) which hurts the potential correlations.
* Convergent validity: associations between scores and healthcare utilization were assessed, and maximizer scores were predicted to result in more utilization and more care. Maximizing was positively associated with 11/15 of measures of health care utilization. Medication refusal was negatively associated with the mean score (but not factor score).
  + Hypothetical scenarios were examined, and maximizers were more likely than minimizers to prefer more action (surgery, not waiting, continued chemo, etc.), but not in cases of experimental treatment.

Study 3

* Test-retest reliability, tried to assess this at two time points with no manipulation.
* Directly compared it to another behavioral/affective scale assessing a related but not entirely overlapping construct.

Study 4

* Assessed whether medical maximizing and minimizing can be affected by experimental manipulation
  + Information about the test, whether or not the test has overdiagnosis/ overtreatment, or no information about how relatively harmful it is.

Royal 2016: Face Validity Problems

* Good paper b/c in general, the original measure by Skitka relies heavily on the concept of ‘face validity’ for measures of moral conviction.
  + Prior work from 1947 validity theorists state that ‘the appearance of validity does NOT constitute scientific evidence.
  + Face validity is sloppy at best and misleading at worst.
* Plausible that many researchers are not experts in validity theory, and thus defer to narrower bands of study when developing their scale.
* Skitka 2010 states that moral conviction is great because they EXPLICITLY measure instead of ASSUME an issue is morally relevant…
  + “To avoid confounding our measure of moral conviction with other aspects of attitudes, such as attitude importance or centrality, we have generally used a single-item and face valid measure of moral conviction”
  + This doesn’t address whether or not the issue has convergent or discriminant validity???
    - Convergent validity was assessed on the single item measure, by tracking positive correlation with attitude extremity, attitude certainty, attitude importance, and political orientation.
      * Convergent validity was also tested by seeing how well it predicted a similar ‘face valid’ measure of moral conviction “the degree that participants indicated that their attitude on a given issue was connected to their fundamental beliefs about right and wrong”
      * Claimed validity b/c strength of moral conviction and belief about right/wrong were highly correlated.
    - Discriminant validity: Unique variance in moral conviction is associated with downstream consequences (intolerance, policy acceptance, peer influence resistance, and voting), when controlling for attitude extremity, importance, certainty, etc.
  + Reliability was assessed by doing test-retest reliability for 13 different issues from 1 to 3 month intervals.
    - Test-retest reliability was high! This is a good thing!

Strauss Smith 2009:

* Test validity: a test’s ability to predict practical criterion! However… tests of criterion-validity are only as good as the criteria used in the task.
  + “reliance on criterion-related validity “involves the acceptance of a set of operations as an adequate definition of whatever is to be measured [or predicted] Typically, the validity of the criterion was presumed, not evaluated independently.”
  + Test validity furthermore does not help development in basic theory, because this does not provide strong foundation for deducing likely relationships amongst variables.
* Construct Validity: Needing to articulate specific theories describing relationships between psychological processes, in order to the evaluate the performance of measures you think represents those processes (e.g., I think due to theory, X causes Y, so we can test whether or not X causes Y directly, and that’s evidence for the theory).
  + Convergent validity – associations among independent measurement procedures designed to reflect the same or similar constructs (this doesn’t hold up for skitka, as the second measure used to test conv. Validity is NOT independent!)
  + Discriminant validity – A new measure of a construct needs to be substantially less correlated with measures of conceptually unrelated constructs, than with OTHER indicators of the construct.
    - Requires the contrast of relationships of measures of constructs in the same conceptual domain (e.g., personality or symptom dimension constructs).
  + Due to method variance: validation studies require the simultaneous consideration of two or more traits measured by at least two different methods.
* Each study using a measure is simultaneously a test of the VALIDITY of a measure AND a test of the theory defining the construct.
  + Each new tests provides additional information supporting the theory/validation claims.
  + E.g. We want to determine how much someone is a ‘nerd’, construct validity requires a clear definition of what a ‘nerd’ is, and distinguish it from measures of shyness, introversion, etc.
* Construct validity is comprehensive – encompassing all sources of evidence supporting specific interpretations of a score from a measure, as well as actions based on those interpretations.
  + Failed tests of initial core hypothesis could be NOT just due to failures of theory, but failures in ‘auxiliary’ theories invoked to test a hypothesis.
* How do we measure construct validity in the field of psychology?
  + Strong programs depend on precise theory and represent an ideal
  + Weak programs stem from less fully articulated theories and construct definitions.
* Best approach is an ‘iterative’ process where tests of partially developed theories provide information, leading to theory refinement and elaboration, and so on and so forth.
  + Are these theoretical statements and tests informative?
  + Is this theory in line with what else is known in the field?
  + How does this hypothesis shed light on the validity of a theory/measure?
* Construct Representation and Nomothetic Span (e.g., meaning of a construct as established through it’s network of relationships with other constructs, i.e., convergent and discriminant validity)
  + For example, IQ has good nomothetic span, because individual differences in various measures of that construct all show similar meaningful patterns of relationship with other variables as expected!
  + Construct representation – validation of the theory of the response processes that result in a score (accuracy/reaction time for example) in cognitive task performance.
    - The psychological processes that lead to a given response on a trial, or the pattern of responses across conditions in an experiment.
    - Or to be more specific, model predictions are confirmed in the testing of the theory itself.
* Construct Homogeneity: Try to have variant items, sure… but if items are only moderately intercorrelated, it is likely that they do NOT represent the same underlying construct!
  + If one item in a heterogeneous test predicts a criterion, you can’t know which aspect of the item accounts for the covariance!
  + If you use a single score to reflect multiple dimensions, you can’t know which dimensions account for which.
  + E.g., if one person is high in X and low in Y, and another is high in Y and low in X, and the average predicts score Z… the score Z will look the same for both of them, even if X and Y have potentially different correlates.
  + For example: Use of a neuroticism score, as a summation of scores on several separable traits, is a problem b/c you have theoretical imprecision.
* Another example: People can obtain the same ‘depression’ scores with VERY different symptom patterns (e.g., a summation of x or y total behaviors), thus, depression is a useful social construct, but NOT a coherent psychological entity useful for validation studies.
  + For example, each factor predicting depression has different heritability, and thus people w/ the same score can have different heritability risks!
* Try to use cohesive unidimensional constructs! NOT multifacted complex constructs!
  + But… when is a construct measure ‘focused’ enough? If you keep cutting measures down until it’s so specific, you lose coverage of a targeted construct and you lose predictive power.
* You need attention to method variance – examination of constructs with different methods is CRUCIAL to construct validation
  + However, how different is different enough for methods? E.g., are interview and questionnaire different enough? They are both self-reported, but the operationalization of the information collection is different.
  + It’s a continuum, wherein for example, self-report and interview are closer to each other than are self-report and informant report or behavioral observation.
* The term “construct validation” refers to the process of simultaneously validating measures of psychological constructs and the theories of which the constructs are a part.

Colliver 2012: From test validity to construct validity and back!

* Proposing a shifted view from being based on ‘theoretical constructs’ to more ‘reality-based’ attributes.
* Validity is does a test/measurement measure what it says it does?
* Contemporary viewpoint shifted to absorb validation as hypothesis testing!
  + In construct validity theory – the construct (e.g., intelligence, burnout, etc.) is a theoretical concept defined by how it relates to other constructs.
  + Established by any evidence that supports this network of constructs, or laws that constrain the constructs.
  + However… this is problematic when considering that no systems of laws explicitly link constructs and observables.
* Current approach seeks to establish validity based on evidence for an interpretation of the target construct.
* Modern approach instead focuses on ‘what can be measured’ and ‘what counts as measurement’
  + Distinction between the measurement of constructs vs the measurement of ‘attributes’ – Constructs are abstract theoretical terms, given meaning by a nomological network. Attributes are thought to exist apart from theory, and measured by instruments for which outcomes are causally determined by the attribute.
  + Attributes are thought to exist independent of the measurement (e.g., height, weight, blood pressure)

Carpenter 2012: Ten Steps in Scale Development

* Best practices for developing a scale, with general concepts.
  + Scales try to capture not directly observable latent concepts with a group of concrete statements. Scales are “collections of items combined into a composite score intended to reveal levels of theoretical variables not readily observable by direct means”
  + This is exactly the case for moral conviction, no direct method to examine it.
* If the latent construct is multidimensional, the scale will need subscales that represent each dimension, summed to become one composite scale!
* Exploratory factor analysis is the most often applied approach to evaluate scales.
  + Finds grouping and guidance on factor numbers.
  + Recommended over CFA as researcher assumptions on how many dimensions are in the construct can be wrong.
* Step 1: Research the intended meaning and breadth of the theoretical construct
  + Specification of theoretical constructs can result in infinite items, but finding the optimal sample of items and dimensions to empirically represent the abstract concept is the goal.
  + Select an appropriate conceptual label!
  + Write and defend a formal conceptual definition of the construct.
  + Identify potential dimensions and items.
  + Conduct qualitative research to generate and validate dimensions and items.
    - This can be interviews, focus groups, expert feedback, etc. Participant engagement can sometimes reveal additional dimensions that were previously not considered.
  + Use feedback to refine the scale!
* Step 2: Sampling procedure
  + More participants = more stable scale, but generally at least 300 is a good shot.
  + Ideally, 20 cases per variable is also a good rule of thumb.
* Step 3: Data quality examination
  + Missing data, outliers, linearity, and extreme multicollinearity.
* Step 4: Verify the factorability of the data.
* Step 5: Conduct common factor analysis.
* Step 6: Select factor extraction method, translating the data from a variable space to a factor space.
* Step 7: Determine number of factors.
  + How many subscales??? – Again, not the case when looking at a unidimensional measure.
* Step 8: Rotate factors to determine how much overlap exists.
  + E.g., If factors are not correlated, are we measuring two constructs rather than one?
* Step 9: Retain/Delete items based on a priori criteria.
  + Each subscale should include at least 3 items!
* Step 10: Present results!

ADDENDUM

According to the domain model of attitude moralization, the process of attitude moralization is a two-part process. The first part is the shift from seeing an initial attitude as being grounded in preference or opinion (e.g., Coke vs. Pepsi) to instead being seen as having moral significance. This initial step is labeled as “Moral Recognition”. Moral recognition can occur when activities previously seen as non-moral (e.g., eating meat) become connected to already pre-existing moral beliefs (i.e., ending life is immoral, eating meat requires the ending of life, thus, eating meat is immoral). This process is defined as ‘Moral Piggybacking’. Moral recognition can also occur when previously unknown moral objections to one’s preferences are made salient (e.g., learning that a brand of luxury goods is produced by slave labor). Finally, moral recognition can occur when individuals are induced to attach strong emotions (e.g., disgust or anger) to the act or concept that is being moralized (i.e., eating meat being moralized through the viewing of videos that show animal suffering inflicted by the meat industry).

The second process is the increase in moralization of already moralized attitudes, wherein a relatively less moralized attitude becomes more moralized. One aspect in which moral amplification differs from moral recognition is that people who hold weakly moralized attitudes are likely already aware of societal norms that exist both in support and opposition to their beliefs. Persuasive arguments framed using specifically moral language, centered on perception of harm, rights, and liberties have also been shown to be effective for moral amplification. Additionally, as has already been seen in the process of moral recognition, changes in attitude-specific emotions (e.g., happiness, excitement, anger, and disgust) predict parallel changes in extremity of attitude moralization. While the two concepts of moral amplification and recognition are theoretically distinct, many of the psychological factors affecting one also affect the other.

Distinguishing these two concepts opens the door to future work that delineates which factors may be more applicable to one process or the other. Furthermore, as new technologies and societal changes develop, the process of moral recognition becomes increasingly salient. For example, moral recognition is pertinent when considering AI chatbots, as an entirely novel concept, the public in general has no preconceived notions with regards to it’s morality. This can be contrasted with contemporary topics that have been discussed for years, such as abortion, which lends itself more to studies related to further amplification or demoralization of an already morally weighty topic.

One goal of our research is to be successful at ‘demoralizing’ beliefs held with moral conviction. In the previous literature, while some researchers have been successful in ‘demoralization’, other researchers have been unsuccessful (Brannon 2019; Clifford 2017). Additionally, some research by Asadullah and colleagues (2019) indicated that there was even a null effect of moral conviction. The primary shared issue that these studies hold were an improper parameterization of the psychometric measurement of moral conviction itself.

For example, Brannon and colleagues (2019) were unable to reduce moral conviction as they defined it, on attitudes related to genetically modified organisms used as food. However, they also acknowledged that their null effect on moral conviction was likely due to the poor psychometric properties of the single-item measure they adapted from Skikta and colleagues (2005). OBJECTIONS TO SINGLE ITEM MEASURES

In comparison, Clifford and colleagues