

# A brief assessment tool for investigating facets of moral judgment from realistic vignettes

Michael Kruepke<sup>1</sup> · Erin K. Molloy<sup>2</sup> · Konrad Bresin<sup>1</sup> · Aron K. Barbey<sup>3</sup> · Edelyn Verona<sup>4</sup>

© Psychonomic Society, Inc. 2017

**Abstract** Humans make moral judgments every day, and research demonstrates that these evaluations are based on a host of related event features (e.g., harm, legality). In order to acquire systematic data on how moral judgments are made, our assessments need to be expanded to include real-life, ecologically valid stimuli that take into account the numerous event features that are known to influence moral judgment. To facilitate this, Knutson et al. (in *Social Cognitive and Affective Neuroscience*, 5(4), 378–384, 2010) developed vignettes based on real-life episodic memories rated concurrently on key moral features; however, the method is time intensive (~1.4–3.4 h) and the stimuli and ratings require further validation and characterization. The present study addresses these limitations by: (i) validating three short subsets of these vignettes (39 per subset) that are time-efficient (10–25 min per subset) yet representative of the ratings and factor structure of the full set, (ii) norming ratings of moral features in a larger sample (total N = 661, each subset N = ~220 vs. Knutson et al. N = 30), (iii) examining the generalizability of the original factor structure by replicating it in a larger sample across vignette subsets, sex, and political ideology, and (iv) using

latent profile analysis to empirically characterize vignette groupings based on event feature ratings profiles and vignette content. This study therefore provides researchers with a core battery of well-characterized and realistic vignettes, concurrently rated on key moral features that can be administered in a brief, time-efficient manner to advance research on the nature of moral judgment.

**Keywords** Moral judgment · Vignette · Moral · Assessment · Sex

## Introduction

Morality, the capacity to discern between right and wrong, is a hallmark of human nature and an important component of personal identity (Strohlinger & Nichols, 2014). Moral judgments are evaluations based on this sense of morality and compare the actions or opinions of an individual or group to specific norms and values (e.g., religious, cultural). Through these evaluations, moral judgments are thought to provide a mechanism for morality to guide or constrain people's thoughts and behavior (Greene et al., 2009; Graham, Haidt, & Nosek, 2009). Although moral judgment (MJ) often guides human behavior, the interrelations between specific event features (e.g., harm, legality, self-benefit) and how these features influence MJ, and in turn particular forms of behavior, remain to be well characterized, potentially due to the methods typically used to study MJ (Blasi, 1980; Thoma 1994).

Investigations of MJ often employ scenarios from philosophy to assess judgments of moral appropriateness (Greene et al., 2001). For example, the “Trolley Problem” asks if it is morally appropriate to change the course of a runaway trolley so that it kills one person instead of five. Such thought experiments provide key insights into variables that influence MJ (e.g., personal force and intention; Greene et al. 2009).

---

✉ Michael Kruepke  
mdkruepke@gmail.com

<sup>1</sup> Psychology Department, University of Illinois at Urbana-Champaign, Urbana-Champaign, Champaign, IL 61820, USA

<sup>2</sup> Computer Science Department, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

<sup>3</sup> Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

<sup>4</sup> Psychology Department, University of South Florida, Tampa, FL, USA

However, scholars have questioned whether moral dilemmas from philosophy provide a sound foundation for understanding everyday MJ and behavior (Kahane, 2015). More recently, researchers have emphasized the importance of administering more realistic, ecologically-valid materials, such as pictures of moral content (Moll, de Oliveira-Souza, Eslinger et al., 2002) and sentences about moral norms (Moll, de Oliveira-Souza, Bramati et al., 2002) to investigate how specific, often isolated, event features (e.g., emotional content, legality of an action, harm severity) impact MJ. This research indicates that MJ is not motivated by a single feature but instead depends on numerous sources of information (Graham et al., 2009). Converging evidence from neuroscience indicates that MJ engages multiple brain systems involved in cognitive, social, and emotional processing (e.g., executive function, social cognition, affect, and motivation; Greene & Haidt, 2002). While this work has been critical to understanding the processes involved in MJ, it does not provide precise information about how specific event features are integrated or weighed against each other to form MJ (e.g., is harm severity or legality of an action more important in making a MJ?). To provide a more precise assessment of MJ under ecologically-valid conditions, methods for the concurrent examination of moral event features using realistic stimuli have been developed (Knutson et al., 2010).

Knutson et al. (2010) conducted a normative study of real-life narratives based on episodic memories of moral events drawn from the work of Escobedo (2009). In their work, Escobedo et al. collected 758 first-person moral vignettes from a sample of 100 healthy, English-speaking adults from Southern California (47 males; age 40–60 years; mean age 48.8, SD = 5.9 years). Vignettes were solicited using three types of cue words: emotions, actions, and superlatives, representing both positive and negative moral experiences (e.g., emotion cue: compassionate, guilty; action cue: honest, unfaithful; superlative cue: best life event, worst life event). To perform their normative study, Knutson et al. condensed Escobedo et al.'s 758 first-person moral narratives (average word count = 218) into 312 shorter moral vignettes (average word count = 43) and had this set of vignettes rated on 13 event features previously implicated in MJ (Rusbult & Van Lange, 1996; Shweder, Much, Mahapatra, & Park, 1997; Haidt, 2007). The 13 event features, along with their anchor points (as well as three new event features added in an exploratory manner in this study), are listed in Table 1.

To further characterize the vignettes, and determine whether a smaller number of moral dimensions underlie the event feature ratings, Knutson et al. conducted a factor analysis of participant ratings using the 10 main event features and the set of 312 vignettes (Table 1; ratings of frequency, personal familiarity, and general familiarity were not included). This analysis revealed three components: (i) norm violation, (ii) social affect, and (iii) intention (Table 2). Norm violation represents moral features emphasized by Shweder et al. (1997)

and Haidt (2007), and is based on ratings of adherence to conventional norms, harm to others, benefit to others, legality, and moral appropriateness. Social affect reflects emotional components of MJ (Moll, de Oliveira-Souza, Bramati et al., 2002; Moll, de Oliveira-Souza, Eslinger et al., 2002), and is represented by ratings of emotional intensity, emotional aversion, and whether others are involved in or impacted by the events depicted in the vignette. Intention reflects instrumentality of the action (Koster-Hale, Saxe, Dungan, & Young, 2013), and relates to premeditation or planning and the level of benefit obtained by the actor in the vignette. These components parallel and expand on previous models of MJ. For example, the importance of emotional content is clear in Haidt's Moral Foundations Theory (MFT; Haidt, 2007), although MFT does not directly address intention, which has been shown to impact MJ (Koster-Hale et al., 2013). By concurrently assessing the relative impact of several moral event features, Knutson et al.'s stimuli bring together, within the same paradigm, moral features differentially emphasized across theories of MJ. More broadly, Knutson et al.'s vignettes provide researchers with a set of ecologically-valid vignettes to elucidate how specific event features are integrated or weighed against each other to form MJ.

Despite the importance of Knutson et al.'s work, several obstacles have limited the impact and utility of their assessment battery for research on MJ. Indeed, although numerous articles have referenced and encouraged researchers to use Knutson's stimuli (Gold, Pulford, & Colman 2014; Feldmanhall et al., 2012; Bzdok et al., 2012; Kahane, 2015; Ugazio, Lamm, & Singer, 2012), only two published studies have utilized them (Vranka & Bahník, 2016; Simpson & Laham, 2015). To begin with, the method put forth by Knutson et al. is time intensive due to the large number of vignettes employed and the fact that each subject was asked to rate each vignette. A full administration of Knutson et al.'s vignettes requires individuals to read 312 vignettes and make 4,056 ratings, which take at least an hour to complete. Specifically, given typical reading rates for comprehension (200–300 words per minute; Carver 1990) and an average word length of 43 words across all 312 vignettes, reading the vignettes takes between 45 and 67 min. Completing ratings for each vignette adds an additional 34–135 minutes, given 13 event feature scales per vignette and assuming a window of 0.5–2.0 s per rating. Assessments of this length can negatively impact data collection and quality by increasing participant drop-out, inattentiveness, and random responding (Herzog & Bachman, 1981). Additionally, long and inflexible assessments are unlikely to be broadly adopted and may be difficult or impossible to implement in some methodologies (e.g., imaging, psychophysiological). Of note, Knutson et al. indicated that researchers are not required to administer the full 312 vignette assessment. The authors recommend investigators select vignettes to fit their unique needs based on vignette content, event feature ratings, and/or factor scores. Although using this top-down approach in selecting

**Table 1** Summary table of administered event features scales

Event feature scale	Question administered	Anchors
Emotional intensity	How emotionally intense was this event?	1 = Not at all emotionally intense 7 = Extremely emotionally intense
Emotional aversion	How emotionally unpleasant was this event?	1 = Not at all aversive or unpleasant 7 = Extremely aversive or unpleasant
Harm	How much harm did this action do to others?	1 = No harm to others 7 = Extreme harm to others
Other-benefit	How much did this action benefit others?	1 = No benefit for others 7 = Extreme benefit for others
Self-benefit	How much did this action benefit the main actor (YOU)?	1 = No benefit to main actor (YOU) 7 = Extreme benefit to main actor (YOU)
Premeditation	How much planning went into this action?	1 = The action was completely unplanned 7 = The action was completely planned
Legality	How legal was this action?	1 = The action was extremely illegal 7 = The action was extremely legal
Social norms	Does this action follow social rules?	1 = This action breaks social rules 7 = This action follows social rules
Socialness	Are other people involved in this action?	1 = No other people are involved in the action 7 = Other people are extremely involved in the action
Moral appropriateness	Was this action morally appropriate?	1 = Extremely morally inappropriate 7 = Extremely morally appropriate
Frequency	How often do you think this type of event actually happens?	1 = This type of event rarely occurs 7 = This type of event occurs all the time
Personal familiarity	Have you ever experienced this type of event?	1 = Never experienced this type of event 7 = Frequently experienced this type of event
General familiarity	Have you thought about this type of event?	1 = Never thought about this type of event 7 = Frequently think about this type of event
Self-harm	How much harm did this action do to the main actor (YOU)?	1 = No self-harm towards main actor (YOU) 7 = Extreme self-harm towards main actor (YOU)
Once vs. Repeated Event	Was this action a one-time event or something that the main actor (YOU) did frequently?	1 = One-time event 7 = Frequently
Acted differently	How likely is it that the main actor (YOU) would have acted differently in this specific event?	1 = Extremely unlikely 7 = Extremely likely

*Note:* Feature scales in block one were included in Knutson et al.'s (2010) analyses as well as in the factor and LPA analyses here. Feature scales in block two were in Knutson et al.'s initial ratings but were not included in their factor analysis. These are included in our 13-scale factor analyses. Feature scales in block three are new to this study and are explored in the 16-scale factor analyses. See Supplemental Materials – Appendix F and G, respectively, for 13- and 16-scale factor analyses

vignettes can prove fruitful (Green, 2009), it also introduces several challenges for the field that interfere with efforts to systematically study MJ.

First, selecting a smaller set of vignettes based on specific needs may lead to variation in methods, and thus results. Specifically, studies may look at different sets of vignettes, such that there is no, or very little, overlap across studies. This could yield results that are study- and stimuli-specific. This concern could be further compounded by not reporting which vignettes were used, including in the two published studies that have used Knutson's stimuli (Vranka & Bahník,

2016; Simpson & Laham, 2015). This can make it difficult or even impossible to compare findings across studies, which can directly hinder progress. While researchers will likely continue to have good reasons to select specific vignettes (e.g., a desire to focus on particular ratings or content), our subsets will provide a standardized methodology that is brief and easy to administer. Indeed, a key reason MJ research using the Moral Foundations Theory (MFT) has been so successful is because researchers use a standardized measurement, which facilitates comparisons and the steady development of a nomological net around constructs related to this theory.

**Table 2** Factor analyses results of Knutson et al. (2010), of each of the three vignette subsets, and of the three subsets combined

Event feature scale	Knutson et al. 2010														
	Subset 1			Subset 3			Subset 4			Subsets 1 + 3 + 4					
	Norm violation	Social affect	Intention	Norm violation	Social affect	Intention	Norm violation	Social affect	Intention	Norm violation	Social affect	Intention			
Social norms	<b>0.947</b>	0.154	0.144	<b>0.962</b>	-0.119	0.030	<b>0.956</b>	-0.179	-0.023	<b>0.975</b>	-0.006	-0.050	<b>0.966</b>	-0.093	-0.012
Harm	<b>0.803</b>	0.473	0.009	<b>-0.798</b>	0.522	-0.161	<b>-0.814</b>	0.460	-0.048	<b>-0.778</b>	0.506	-0.063	<b>-0.790</b>	0.502	-0.110
Legality	<b>0.737</b>	-0.288	0.115	<b>0.793</b>	0.295	-0.136	<b>0.785</b>	0.335	-0.046	<b>0.743</b>	0.400	-0.233	<b>0.757</b>	0.369	-0.158
Other benefit	<b>-0.883</b>	0.046	0.051	<b>0.882</b>	-0.041	0.119	<b>0.898</b>	-0.111	-0.054	<b>0.879</b>	0.111	0.229	<b>0.893</b>	-0.009	0.093
Moral appropriateness	<b>-0.956</b>	-0.102	-0.120	<b>0.957</b>	-0.158	0.068	<b>0.951</b>	-0.193	-0.042	<b>0.984</b>	-0.034	-0.032	<b>0.968</b>	-0.129	-0.003
Emotional intensity	0.024	<b>0.896</b>	-0.066	-0.321	<b>0.859</b>	0.050	-0.213	<b>0.896</b>	0.067	0.067	<b>0.920</b>	0.004	-0.140	<b>0.903</b>	0.030
Socialness	-0.115	<b>0.763</b>	0.154	0.287	<b>0.773</b>	-0.016	0.087	<b>0.712</b>	-0.116	0.309	<b>0.756</b>	0.185	0.243	<b>0.774</b>	0.044
Emotional aversion	0.336	<b>0.762</b>	-0.258	-0.618	<b>0.741</b>	-0.078	-0.521	<b>0.788</b>	-0.012	-0.338	<b>0.874</b>	-0.155	-0.488	<b>0.809</b>	-0.110
Premeditation	-0.002	0.175	<b>0.859</b>	0.166	0.372	<b>0.827</b>	-0.001	0.201	<b>0.914</b>	0.039	0.381	<b>0.867</b>	0.068	0.320	<b>0.867</b>
Self-benefit	0.244	-0.304	<b>0.772</b>	-0.044	-0.323	<b>0.841</b>	-0.069	-0.371	<b>0.813</b>	-0.038	-0.291	<b>0.850</b>	-0.050	-0.302	<b>0.855</b>
Variance explained	40%	24%	15%	49%	22%	14%	48%	21%	14%	41%	29%	16%	44%	25%	15%

Note: Knutson et al.'s (2010) results are based on all 312 vignettes. Each of the subsets in the present study contains a unique set of 39 vignettes. Subsets 1 + 3 + 4 contains all the vignettes from the three subsets validated here and thus has 117 unique vignettes. Differences in loading direction for the “norm violation” component compared to Knutson et al.'s results reflect our reversals of the scales for “social norm violation” (higher scores mean higher social norm adherence in our study) and “illegality” (higher scores mean more legality in our study); however, the results here reflect the same factor structure and interpretation. Main loadings are bolded

Second, little work has been done to develop well-tested methods to investigate, in a realistic and concurrent fashion, the multiple features known to influence MJ (e.g., harm, legality, intention). Thus, few studies have been able to investigate the links between the wide array of features implicated in MJ, with perhaps the exception of the self-report MFT measure. Selecting a smaller set of vignettes based solely on certain content or event features may perpetuate a focus on isolated event features (e.g., emotional intensity) and the often extreme examples of these features (e.g., abortion). The challenges posed by continuing this focus are likely to be compounded by the difficulties introduced through potential methodological variation noted above.

Third, if vignettes are selected with a narrow focus on certain content or event feature ratings, it is unclear whether the three dimensions extracted by Knutson would be reflected in the smaller set. This structure provides one of the first published assessments exploring the underlying relationships between the wide range of features implicated in MJ. The utility of understanding MJ through such a dimensional structure is clearly demonstrated by research using similar methods to investigate the cognitive and neural foundations of religious belief (Kapogiannis et al., 2009). Maintaining this dimensional structure across studies would thus directly facilitate integration of results as well as the investigation of MJ as the multi-faceted construct it is known to be. As it stands, to fully leverage the contributions of Knutson et al.'s work, researchers are required to either administer the full 312 vignette battery or independently curate and validate smaller, and thus shorter, subsets of vignettes.

Another obstacle that may have limited the use of Knutson's original assessment battery is their study's relatively limited sample size. Knutson et al.'s validation and factor results were based on 30 participants (15 males, 15 females; mean age 26.7 years,  $SD = 5.1$ ; mean years of education 17,  $SD = 2.4$ ). Given the recent focus on replication and the historic importance of generalizability, researchers may be reluctant to use methods developed in a single, relatively small sample (Open Science Collaboration, 2015; Shavelson et al., 1989). Moreover, due to the sample size, Knutson et al. were unable to assess the impact of raters' demographic features, such as sex (male, female) and political affiliation (liberal, moderate, conservative). Historically, MJ research has investigated these differences in attempts to more fully explain the influence of personal identity on MJ (Gilligan 1982; Jaffee & Hyde, 2000; Graham et al., 2009). Exploring whether sex or political affiliation impacts how these vignettes are rated would more fully characterize these vignettes and allow these vignettes to be couched within the context of broader and classic research on MJ.

Lastly, Knutson et al.'s analyses were limited to factor analysis of event feature ratings. In this regard, Knutson et al.'s suggestion that researchers select vignettes that are relevant to their research questions follows a top-down or conceptual approach. Some researchers may favor a more bottom-up or

empirical approach, where vignette content is characterized and unique vignette groupings are identified through statistical classification methods, such as latent profile analysis (LPA). Indeed, such an approach can provide researchers with a set of empirically valid vignette groupings that vary on content and event feature ratings which can be used to assess MJ in a common manner across labs. Classification through LPA can therefore offer a valuable complement to factor analysis, with factor analysis focusing on the underlying nature of moral event features and LPA focusing on characterizing vignette content.

The present study directly addresses existing limitations by: (i) validating three short subsets of vignettes (39 vignettes per subset) that are time-efficient (10–25 min per subset) yet representative of the ratings and factor structure of the full set, (ii) norming vignette ratings of several moral event features in a larger and more diverse sample (total sample  $N = 661$ , each subset sample  $N = \sim 220$ ), (iii) examining the generalizability of the original factor structure by replicating it in a large sample across subsets of vignettes, sex, and political ideology, and (iv) using LPA to empirically characterize vignette groupings based on specific event feature profiles and vignette content.

## Method

### Participants

Results are based on 661 students (58% female) at a large Midwestern university who completed the study for extra credit. Most participants were 18–25 years old (99.7%), considered English their primary language (88.7%), and identified as non-Hispanic Caucasians (62.6%), with 11.3% identifying as Hispanic, 9.1% as Chinese, 7.6% as Korean, 7.3% as Asian Indian, 7% as African-American, and 7% as Mexican (all other identifications were <3%; participants were encouraged to select all that applied). Participants were more evenly varied in their political affiliations (38% liberal, 25.9% moderate, 15% conservative, 4.7% libertarian, 1.8% socialist, 0.3% tea party, 14.4% rather not say). Regarding religious affiliations, 43% identified with Christianity, 12.3% as agnostic, 8.5% as atheist, 3.2% with Judaism, and 3.1% with Islam (all other identifications were <3%).

Although 802 individuals participated, 141 individuals were removed for our analyses, leaving us with a sample of 661 participants. Following guidelines to ensure analysis of data from attentive responders (Huang et al., 2012), we removed those who did not complete the full survey ( $N = 134$ , 16.7% of total sample), and a small number of participants who took between 15–17 min to complete the survey ( $N = 7$ , 0.8% of total sample). The latter were removed based on pilot testing indicating that conscientious completion of the study (i.e., going through the consent, demographics, instructions,



and all 16 feature scale ratings for all 39 vignettes via Qualtrics) took between 20–30 min. Demographics for those removed did not differ from those included and results presented do not differ when excluded participants are included in analyses (see Supplemental Materials – Appendix A; Supplemental Material and data are available online via the Open Science Framework: [osf.io/26tyy](https://osf.io/26tyy)). Institutional review board (IRB)-approved protocols were followed throughout the study, and all participants included in our analyses provided informed consent (IRB #14386).

### Vignette subsets

To limit the length of the battery without reducing its scope, we developed three subsets, each with 39 vignettes, taken from Knutson et al.'s original 312 vignettes. These subsets were each designed to take 10–25 min. This reflects a 1- to 3.2-h decrease in completion time (68–95%) from Knutson et al.'s full set of 312 vignettes. To ensure all subsets adequately represented the features of the full set, subsets were created using a *vignette selection heuristic* developed by the authors (full details and code available via Molloy & Kruepke, 2017) and implemented in MATLAB (The MathWorks, Inc., 2014). The heuristic utilizes Knutson et al.'s mean rating data to select representative vignettes endorsed at “low,” “neutral,” and “high” levels on each of the original 13 (i.e., Knutson et al.'s) event feature rating scales. These levels are defined by the intervals [a, b], [b, c], [c, d], respectively, such that a is the minimum rating for the scale, b is the mean scale rating minus the standard deviation of scale's rating, c is the mean scale rating plus the standard deviation of the scale's rating, and d is the maximum rating for the scale. This approach is a common example of data binning, which takes continuous values (e.g., event feature ratings) and bins them into discrete categories (e.g., “low,” “neutral,” “high”). Binning methods such as this are routinely used to improve machine learning algorithms (Kotsiantis & Kanellopoulos 2006), are at the heart of many neuroimaging techniques (Ince et al., 2016), and provide a well-validated method to characterize a broad range of data in a smaller, yet representative, fashion. Seeking to produce a time-efficient method, we selected only one vignette per level for each of the 13 event feature-rating scales. This provided 39 vignettes per subset. Utilizing this heuristic allowed us to produce pseudo-random subsets of vignettes such that all 13 event features have at least one vignette rated at each of the three levels of endorsement. This heuristic was used to generate five subsets of 39 unique vignettes.

Given our expected sample size (~600) and to ensure an adequate number of participants completed ratings on each set, only three subsets were administered in this study. Exploratory factor analyses (EFA) using Knutson's original data indicated that all five subsets demonstrated similar factor structures and loadings to the full set. Based on these results,

we selected subsets 1, 3, and 4 to administer to participants in the current study, as they demonstrated lower cross-loadings, using Knutson's data, compared to subsets 2 and 5. Tucker's Index of Factor Congruence (Tucker Index), which is a well validated tool used to determine the degree of similarity between factor loadings, was also assessed post-hoc (Lorenzo-Seva & Ten Berge, 2006; R Core Team, 2014; Revelle, 2015). Tucker's Index values range from -1 to +1. Values in the range of 0.85 to 0.94 indicate acceptable similarity between factors, whereas values over .95 indicate near equality between factors (Lorenzo-Seva & Ten Berge, 2006). The factor structures based on Knutson's data obtained for subsets 1, 3 and 4 (the ones chosen in this study), as well as the ones not chosen (subsets 2 and 5), all showed similarity to Knutson's full set, with average Tucker's Index ranging from .86 to .98 (Subset 1 = .98, Subset 2 = .97, Subset 3 = .86, Subset 4 = .98, Subset 5 = .97; for full results see Supplementary Materials – Appendix B). Tucker's Index values indicated that all five subsets created by our heuristic have factor structures similar or equivalent to the full set. Thus, our selection of specific subsets to increase sample size per subset did not bias our results.

Vignettes were administered to participants online via Qualtrics®. Individuals were randomly assigned to one of the three subsets such that subsets had a roughly equal number of participants (i.e., subset 1: N = 220; subset 3: N = 225; subset 4: N = 216). Vignettes were rated on 16 event features using a fixed-point sliding scale (1–7, starting point at 4). These included the 10 main scales used in Knutson's analyses, the three additional scales found in Knutson et al. but not used in their analyses, as well as three new scales: self-harm; regularity of the actor's actions (“once vs. repeated event”); and likelihood that the participant would have acted differently than the actor in the vignette (“act differently”) (Table 1). The latter scales were included in our study to broaden the assessment of event features implicated in MJ (Cresswell & Karimova, 2010; Reynolds & Ceranic, 2007; Phillips et al., 2015), but were only peripheral to our main goal.

To enhance the clarity and standardize the interpretation of the scales across participants, we modified some aspects of Knutson et al.'s scales (see Supplementary Materials – Scale Changes). Specifically, we changed scale presentation from a label (e.g. other-benefit) to a question (e.g., How much did this action benefit others?); slightly adjusted the wording of 11 scales; and maintained consistent wording and grammar for anchors across scales by reversing anchors for legality and social norms.

## Data analysis and results

### Factor analyses

Data were analyzed in two complementary ways. First, following Knutson et al., we conducted EFA using the 10-event

feature scales assessed in Knutson et al. (Table 1), with ratings averaged across participants. We focused on these scales due to their direct measurement of features previously implicated in MJ (e.g., harm) and to assess the reproducibility of Knutson et al.'s factor structure in this larger sample. This event feature-centered approach identifies whether underlying factors influenced feature ratings, and EFA allowed us to examine the factor structure without being confined to a specific solution (i.e., Knutson et al.'s original structure). Given that our aim was to replicate a factor structure with a smaller subset of items as a step to improve assessment feasibility and not necessarily confirm a theoretical model, EFA was considered the appropriate choice over confirmatory factor analysis (CFA), which is often theory-driven (Church & Burke, 1994; McCrae et al., 1996; Ferrando & Lorenzo, 2000).

EFA analyses were conducted in the same manner as Knutson et al., using SPSS's (20.0.0) principal component extraction with a Varimax rotation and Kaiser normalization (Kaiser, 1958). Oblique rotations (i.e., Promax, Direct Oblimin) produced structures similar to Knutson's results and to the Varimax results reported here (i.e., similar factor structures and acceptable Tucker Index values) as well as weak correlations between components (mean  $r = .105$ ,  $SD = .088$ , range = .001–.360; see Supplementary Material – Appendix C). Analyses were conducted separately for each of the three subsets as well as with the subsets combined (i.e., 117 vignettes). For all EFA analyses, Tucker's Index of Factor Congruence was calculated to determine the similarity of factor loadings across samples (i.e., Knutson et al. results vs. our results), across subsets (i.e., comparing subsets 1, 3, and 4) and across grouping variables (i.e., sex and political affiliation) (Lorenzo-Seva & Ten Berge, 2006).

Factor analyses on the 10 event feature scales in the three subsets separately, as well as together, produced a three-factor solution similar to Knutson et al.'s analyses (Table 2) and evidenced high Tucker Index values (Table 3), indicating similarity in factor loadings of these subsets relative to the Knutson sample as well as similarity across the subsets themselves. Specifically, we identified a norm violation component, a social affect component, and an intentionality component all with eigen values greater than one. Norm violation consistently accounted for the most variance and involved positive loadings from social norms, legality, benefit to others, and moral appropriateness, as well as negative loadings from harm to others. Social affect accounted for the second most variance, with positive loadings from emotional intensity, socialness, and emotional aversion. Lastly, intention accounted for the smallest amount of variance, with positive loadings from premeditation and self-benefit.

Based on previous research and theory suggesting that MJ may vary according to an individual's sex (Jaffee & Hyde, 2000) or political affiliation (Graham et al., 2009), we examined whether factor structures differed as a function of sex

(male, female) or political affiliation (liberal, moderate, conservative). Religious affiliation (Graham et al., 2009; Galen, 2012) was not directly investigated with the current sample since our sample was predominantly Christian (i.e.,  $N = 284$  Christianity,  $N = 81$  Agnostic,  $N = 56$  Atheist,  $N = 21$  Judaism,  $N = 20$  Islam, all remaining affiliations  $N < 20$ ). As indicated by high Tucker Index values, analogous results were found across sex (male, female; Table 4) and political affiliation (liberal, moderate, conservative; Table 5) (see Supplemental Materials – Appendix D and E, respectively for full factor structures and additional Tucker Index comparisons). Based on these results, the norm violation, social affect, and intention components appear to generalize well across samples, subsets, sex, and political affiliation, providing increased confidence in their ability to assess core and key aspects of MJ.

Similar factor structures and Tucker Index values were seen when conducting EFA on the 13 original event feature scales and including our three new scales (i.e., 16-scale analysis; see Table 1 for scale identification), with results again indicating norm violation, social affect, and intention components. In addition, a fourth component, event familiarity and likelihood, was identified in both the 13-scale and 16-scale analyses (Supplementary Material – Appendix F and see Supplementary Material – Appendix G, respectively). Mirroring findings for the 10 event features, these results held across sex and political affiliation, each of the three vignette subsets, and in the full set of 117 vignettes used here (see Supplementary Material – Appendix D for sex results, Appendix E for political affiliation results).

To further inform the utility of these subsets for use across a range of samples (e.g., those where reading ability may be a limitation; Greenberg et al., 2007), we also calculated the readability and comprehensibility of the vignettes via the Flesch-Kincaid reading ease and grade level indices (Flesch, 1948). Across all vignettes, reading ease and grade level ranged from 46.4 to 98.9 (mean = 82.82,  $SD = 9.97$ ) and from 2.6 to 11.1 (mean = 5.22,  $SD = 1.53$ ), respectively (specific indices for each vignette are found in the Supplementary Materials – Vignette Information). Despite the wide range in readability indices across vignettes, the subsets did not differ significantly in reading ease ( $F(5, 111) = 0.391$ ,  $p = .854$ ) or grade level ( $F(5, 111) = 0.447$ ,  $p = .815$ ), with subset 1 showing a minimum reading ease of 62 and grade level of 2.7, subset 3 a minimum reading ease of 62 and grade level of 2.6, and subset 4 a minimum reading ease of 46 and grade level of 2.9.

### Latent profile analysis

Our second analytic approach, LPA, used the vignette as the unit of analysis (averaged across participants) and provides an empirically derived characterization of the vignettes' content. LPA identifies unique patterns (i.e., profiles) of responding across a set of items. These profiles can be used to identify

**Table 3** Tucker's Index of Factor Congruence was computed between the factors from Knutson et al. (2010), the three vignette subsets individually, and the subsets combined. Tucker's Index values range from -1 to

+1. Values in the range of 0.85 to 0.94 indicate acceptable similarity between factors, whereas values over .95 indicate near equality between factors (Lorenzo-Seva &amp; ten Berge, 2006)

	Knutson et al. 2010			Subset 1			Subset 3			Subset 4		
	Norm violation	Social affect	Intention	Norm violation	Social affect	Intention	Norm violation	Social affect	Intention	Norm violation	Social affect	Intention
vs. Subset 1	0.97	0.99	0.94	-	-	-						
vs. Subset 3	0.99	0.99	0.93	0.99	0.99	0.97	-	-	-			
vs. Subset 4	0.99	0.98	0.97	0.97	0.99	0.97	0.98	0.97	0.93	-	-	-
vs. Subsets 1 + 3 + 4	0.99	0.99	0.97	0.99	1.00	0.99	1.00	0.99	0.97	0.99	0.99	0.98

*Note:* Knutson et al.'s (2010) results are based on all 312 vignettes. Each subset in the present study contains a unique set of 39 vignettes. Tucker's Index values were calculated with social norms and legality in the current sample reverse scored to match anchor points in Knutson et al.'s original event feature scale structure

empirically derived categories of vignette content. Here, nine of the original event feature scales (i.e., all those used in the 10-event feature factor analysis, save for moral appropriateness) were used to identify groupings of similar vignettes via profiles of event feature ratings. We excluded moral appropriateness in clustering the vignettes to examine which vignette groups were rated as more or less morally appropriate. LPA was conducted in Mplus 5.0 (Muthén & Muthén, 2008) and used all 117 vignettes tested here. In line with the literature (Clark et al., 2013; Nylund et al., 2007), we fit models starting with two profiles, adding profiles until the Bayesian Information Criterion (BIC; Schwarz, 1978) achieved its first minimum and then increased. Simulation studies have shown that the model with the lowest BIC is most likely to be the correct model, and the BIC outperforms other methods of model selection (Nylund et al., 2007). In addition to the BIC, we considered the number of vignettes per profile grouping and profile interpretability to determine best fit (Clark et al., 2013).

After identifying the best fitting solution, we characterized vignette groups based on profiling features and content and assigned names to each profile. We then compared the groupings on judgments of moral appropriateness. We applied three different indices to describe mean differences in ratings of moral appropriateness across vignette groupings. First, we report the point estimate and 95% confidence interval (CI) for the mean. CIs that do not overlap are generally different

at  $p < .01$ , while CIs that overlap by less than half (.5) a margin of error are generally different at  $p < .05$  (Cumming, 2013). Second, we report Cohen's  $d$  as a measure of effect size. Based on Cohen's recommendations (1992), we interpreted effect sizes of 0.2 and below as "small" effects, effects near 0.5 as "medium" effects, and effects larger than 0.8 as "large" effects. Nonetheless, we also report  $p$ -values from pairwise comparisons. Lastly, we examined whether ratings of moral appropriateness for groupings differed based on sex (i.e., male, female; Jaffee & Hyde, 2000) or political affiliation (i.e., liberal, moderate, conservative; Graham et al., 2009).

For the LPA, a seven-profile solution evidenced the lowest BIC ( $BIC = 2,665.3$ ); however, there were convergence errors (e.g., best likelihood score not replicated), suggesting issues in interpreting this model. A six-profile solution had the next lowest BIC ( $BIC = 2,668.77$ ) and produced clear and meaningful vignette groupings. Therefore, based on BIC and interpretability, we determined the six-profile solution to be the best fitting model (Figs. 1 and 2). The resultant groupings did not differ on Flesch-Kincaid indices of reading ease ( $F(5, 114) = 1.933, p = .149$ ) or grade level ( $F(5, 114) = 2.251, p = .110$ ).

To help interpret the six profiles identified, we plotted the means and confidence intervals for the event feature ratings used to define the profiles in two complementary ways. Figure 1 uses box plots to visualize the values for each profile for each event feature, allowing for comparisons across

**Table 4** Tucker's Index of Factor Congruence between males and females for subsets 1, 3, and 4 and for the subsets combined

Comparisons	Norm violation	Social affect	Intention
Female 1 + 3 + 4 vs. Male 1 + 3 + 4	1.00	.99	.99
Female Subset 1 vs. Male Subset 1	1.00	1.00	1.00
Female Subset 3 vs. Male Subset 3	1.00	.99	.97
Female Subset 4 vs. Male Subset 4	1.00	.99	.99



**Table 5** Tucker’s Index of Factor Congruence between political affiliations for subsets 1, 3, and 4 and for the subsets combined

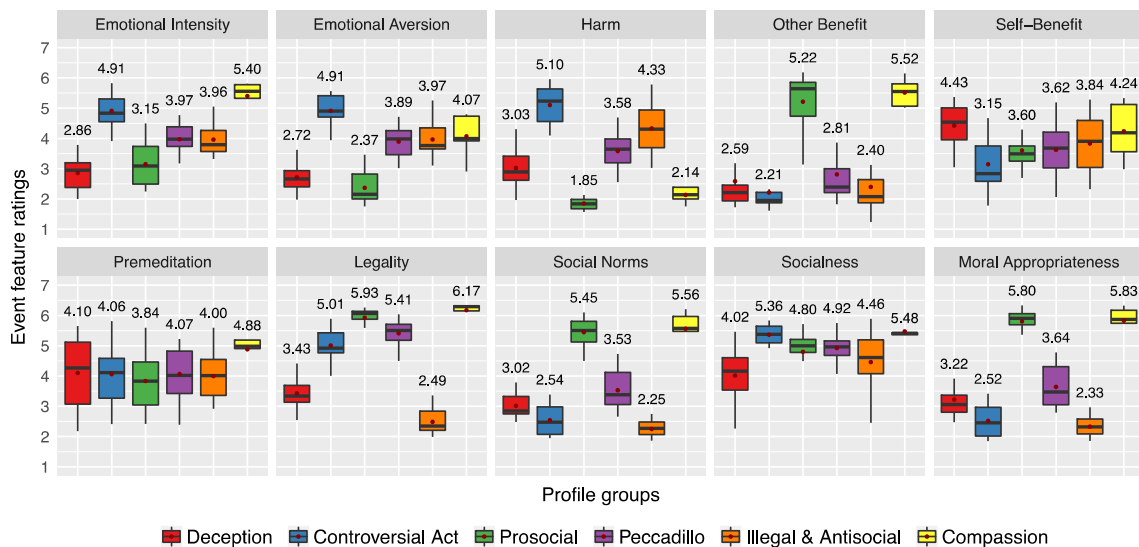
Comparisons	Norm violation	Social affect	Intention
Liberal 1 + 3 + 4 vs. Conservative 1 + 3 + 4	1.00	1.00	1.00
Liberal Subset 1 vs. Conservative Subset 1	.99	1.00	.98
Liberal Subset 3 vs. Conservative Subset 3	1.00	.99	.99
Liberal Subset 4 vs. Conservative Subset 4	1.00	1.00	.99
Liberal 1 + 3 + 4 vs. Moderate 1 + 3 + 4	1.00	1.00	1.00
Liberal Subset 1 vs. Moderate Subset 1	1.00	1.00	.99
Liberal Subset 3 vs. Moderate Subset 3	.99	.99	.99
Liberal Subset 4 vs. Moderate Subset 4	1.00	1.00	1.00
Conservative 1 + 3 + 4 vs. Moderate 1 + 3 + 4	1.00	1.00	1.00
Conservative Subset 1 vs. Moderate Subset 1	1.00	1.00	.99
Conservative Subset 3 vs. Moderate Subset 3	.99	1.00	.99
Conservative Subset 4 vs. Moderate Subset 4	1.00	1.00	.99

profiles (e.g., Prosocial profile is rated significantly lower on harm than the Controversial Act profile,  $t(44) = -25.02$ ,  $p < .001$ ,  $d = -7.19$ ). Figure 2 shows the results of the six-factor solution by plotting each event feature on the x-axis and separate lines for each profile. This allows for the characterization of the profile in relation to itself (e.g., Prosocial profile is rated higher on legality than harm).

The first profile, labeled the Deception profile ( $n$  = number of vignettes, % = percentage of total vignettes administered [ $n/117$ ];  $n = 22$ , 19%), involves vignettes with low ratings on all event feature scales in absolute terms and in relation to other profiles. For example, the Deception profile was significantly lower on emotional intensity than all other profiles – except for the Prosocial profile ( $d$ ’s ranged from  $-2.09$  to  $-4.88$ , all  $p$ ’s  $< .001$ ), and showed elevations on benefit to self. Actions

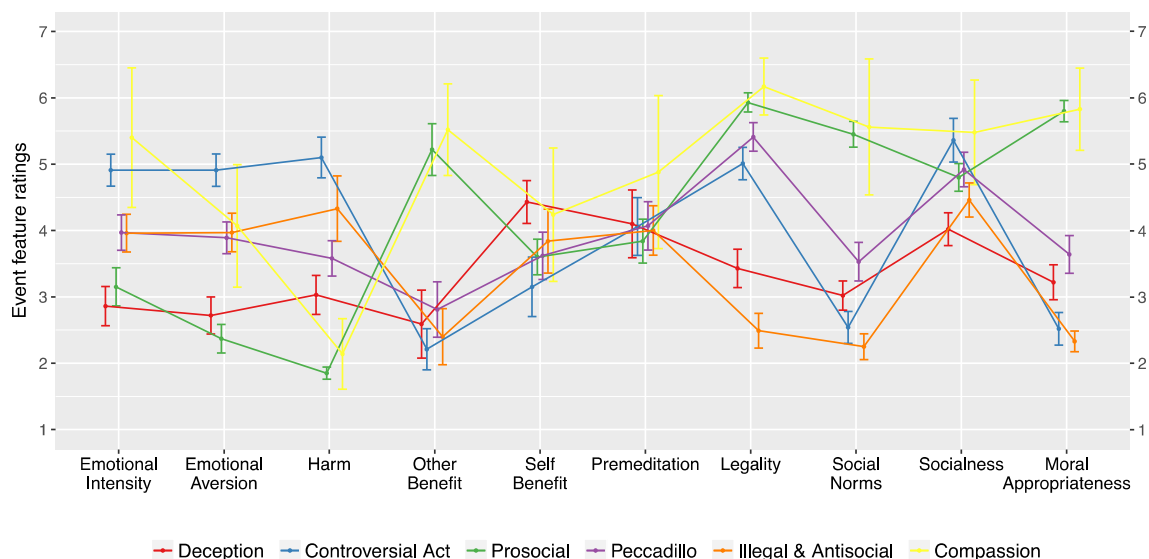
in this group typically break social norms, especially cheating, lying, or stealing, while benefiting the self. Example vignettes are, “Back in high school I kind of had an agreement with the guy sitting next to me. We would show each other our papers whenever we were taking a test. Both of us were pretty good students, we just would make sure we shared test answers if we needed to” and “I applied for a position at this company. As it turned out the pay was not very good at all for the amount of work that I was doing. So I lied and told the manager that I had another job offer and I would take it unless I got a raise.”

The Controversial Act profile of vignettes ( $n = 21$ , 18%) was characterized by high ratings on emotional intensity (e.g., significantly greater than Prosocial, Peccadillo, Illegal & Antisocial, and Deception profiles,  $d$ ’s range from 1.77 to 3.94, all  $p$ ’s  $< .001$ ), emotional aversion, harm to others, and



Note: Red dots and numbers above each box plot represents mean event feature rating. Horizontal line in each box plot represent median. Moral appropriateness ratings were not employed in the latent profile analyses, and were thus not used to identify the six-profile solution. Moral appropriateness ratings are displayed here to provide a complete picture of the quantitative and qualitative differences between profiles.

**Fig. 1** Box Plot of latent profile analysis of nine event feature scales with a six-profile solution



Note: Bars represent 95% confidence intervals for the mean. Moral appropriateness ratings were not employed in the latent profile analyses, and were thus not used to identify the six-profile solution. Moral appropriateness ratings are displayed here to provide a complete picture of the quantitative and qualitative differences between profiles.

**Fig. 2** Profile plots of latent profile analysis of nine event feature scales with a six-profile solution

legality and low ratings on benefit to others and social norms. These center on actions that are mostly legal, yet generate negative emotions, likely due to the violation of social norms causing harm to others and potentially controversial behaviors in the vignettes. Example vignettes are, “I left my second marriage and I left my step-kids there too. My youngest stepson has some disabilities, but I left him there. I could not cope with his druggie, drinking father and so I decided to leave everything behind” and “One night I was having sex with my boyfriend. He said that he had a condom on but at the end I found out that he didn’t. I became pregnant and since I just had had a baby recently, I decided to have an abortion.”

The Peccadillo profile of vignettes ( $n = 24$ , 21%) was characterized by elevation on legality, slightly lower than neutral ratings on social norms, and neutral ratings on emotional intensity and unpleasantness. Actions in these vignettes are typically legal yet break some, often more minor, social norms (e.g., lies or minor sins). They are similar to vignettes in the Controversial Act profile but are much less emotionally charged (i.e., significantly lower than Controversial Act on emotional intensity,  $t(43) = -5.96$ ,  $p < .001$ ,  $d = -1.77$ , and emotional aversion,  $t(43) = -6.88$ ,  $p < .001$ ,  $d = -2.08$ ), partly because there is less overt other-harm involved (i.e., significantly lower than Controversial Act profile on harm,  $t(43) = -8.41$ ,  $p < .001$ ,  $d = -2.53$ ). Example vignettes are, “While I was in college, I was in a long distance relationship with a girl. We talked every night on the phone and really tried to make it work. Meanwhile, I was having study sessions with an attractive girl in my class and very tempted to cheat on my

girlfriend” and “One night I was going out late and I didn’t want my son to know about it. I was a single mom and at that point in time he always seemed to want to act like the parent. So I snuck out of the house to go out.”

The Illegal and Antisocial profile ( $n = 20$ , 18%) contained vignettes low on legality (i.e., significantly lower than all other profiles,  $d$ 's range from  $-1.74$  to  $-9.56$ , all  $p$ 's  $< .001$ ) and social norms and neutral on other scales. This profile was similar to the Deception profile, but differed in that the acts were rated as more illegal, emotional (i.e., emotional intensity and emotional aversion), and harmful (i.e., significantly higher on harm;  $d$ 's range from  $1.74$  to  $2.59$ , all  $p$ 's  $< .001$ ). Vignettes in this group are the most clearly illegal and involve antisocial behavior. Example vignettes are, “As I was backing out of a parking lot I bumped a parked car and left a minor dent. I didn’t even feel the impact when I hit the car but it left a little bit of damage. I drove away without leaving a message or trying to contact the person” and “I was thirteen years old and I went into the grocery store where I lived. There was a comb that I wanted in the store, so I just took it. I didn’t really need it but I just wanted the thrill of stealing it and nobody catching me.”

The Prosocial profile ( $n = 25$ , 21%) evidenced high scores on benefit to others, legality, and social norms and low scores on emotional intensity, emotional aversion, and harm to others. Events in this group represent prosocial actions, such as being charitable and honest. Example vignettes are, “I found a wallet with a fifty-dollar bill in it. I found a phone number to call and contacted the woman whose wallet it was.

She was very appreciative and came to my house to pick it up” and “I was at the pharmacy buying something and I noticed a man who was sitting outside selling trinkets. He was homeless and it was freezing out. So I went next door to a store and bought him some food and new clothes.”

Lastly, the Compassion profile ( $n = 5$ , 4%) exhibited high scores on emotional intensity, benefit to others, planning, legality, and social norms and low scores on harm to others. Like the Prosocial set, these vignettes represent prosocial acts but involve more emotional events than the former group, as evidenced by significantly higher emotional intensity ( $t(28) = 6.97$ ,  $p < .001$ ,  $d = 3.79$ ) and emotional aversion ratings ( $t(28) = 6.32$ ,  $p < .001$ ,  $d = 2.63$ ). Examples are, “My wife was diagnosed with cancer. I was there with her every step of the way even though it was extremely emotionally and mentally demanding. I helped her through all her appointments and emotional distress” and “I had a bad relationship with my father and had not talked to him for years. He had left my mother. When my mother died I gathered the strength to call him and tell him that she died and that I loved him.”

To help further differentiate vignette groupings, we conducted an ANOVA with grouping profile as a between-subject factor and judgments of moral appropriateness as the dependent measure. There was a significant main effect of profile,  $F(5, 111) = 142.51$ ,  $p < .001$ , which highlighted their distinctness. The Illegal and Antisocial profile was rated the least morally appropriate ( $M = 2.33$ ,  $SD = .44$ , 95% CI [2.18–2.48]) and significantly lower than all other profiles ( $d$ 's range from  $-1.66$  to  $-6.12$ , all  $p < .001$ ) aside from Controversial Acts, ( $d = -.35$ ,  $p = .260$ ). The Controversial Act profile ( $M = 2.52$ ,  $SD = .53$ , 95% CI [2.27–2.76]) was the next lowest and was also significantly different from all other profiles ( $d$ 's range from  $-2.09$  to  $-6.16$ , all  $p$ 's  $< .001$ ), aside from the Illegal and Antisocial profile. The Deception ( $M = 3.22$ ,  $SD = .60$ , 95% CI [2.95–3.48]) and Peccadillo profiles ( $M = 3.64$ ,  $SD = .66$ , 95% CI [3.35–3.92]) were both rated as slightly more morally inappropriate and were significantly different from all others [ $d$ 's] range from  $.71$  to  $4.90$ , all  $p < .001$ ). Finally, the Prosocial and Compassion profiles were rated as the most morally appropriate (Prosocial:  $M = 5.80$ ,  $SD = .39$ , 95% CI [5.64–5.96]; Compassion:  $M = 5.83$ ,  $SD = .49$ , 95% CI [5.21–6.45]). They did not differ from each other ( $d = .04$ ,  $p = .927$ ) and were rated significantly higher on appropriateness relative to all other profiles ( $d$ 's range from  $4.02$  to  $6.47$ , all  $p < .001$ ).

To examine whether sex or political affiliation moderated the relationship between grouping profile and ratings of moral appropriateness, we used multi-level modeling, nesting vignette profile within subjects. Although not all participants saw the same vignettes, each subset of vignettes contained a fair spread of vignette profiles, allowing us to examine whether between-person differences affected morality ratings for different vignette profiles. Vignette profile (i.e., Deception, Controversial Acts, Peccadillo, Illegal and

Antisocial, Prosocial, and Compassion) was a within-subject factor and sex (male, female)/political affiliation (liberal, moderate, conservative) were between-subject factors in separate analyses. Although there was no significant interaction with political affiliation,  $F(10, 21000) = 1.59$ ,  $p = .101$ , there was for sex,  $F(5, 26000) = 33.39$ ,  $p < .001$ . To follow-up on this interaction, we compared males and females on their ratings of moral appropriateness within vignette profile. Males and females significantly differed in their ratings for the Deception, Controversial Act, Illegal and Antisocial, and Prosocial profiles (all  $p$ 's  $< .001$ ). Females rated the Deception, Controversial Act, and the Illegal and Antisocial profiles as less morally appropriate than males (Deception:  $d = .34$ , Controversial Act  $d = .56$ , Illegal and Antisocial  $d = .87$ ), and the Prosocial profile as more morally appropriate than males ( $d = -.71$  [males–females]),  $p < .001$ . The other profiles were similarly rated across the sexes, with the Peccadillo and Compassion profiles demonstrating no significant differences between males and females (Peccadillo:  $d = .14$ ,  $p = .051$ ; Compassion:  $d = -.04$ ,  $p = .792$ ). Thus, females, relative to males, appeared harsher in their moral judgments of negative actions, and more morally approving of prosocial actions.

## Discussion

Our study delivers three major advances in the assessment and study of MJ. First, we created and validated three brief subsets of vignettes that reliably capture the underlying structure (e.g., factor structure) of Knutson et al.'s full set of 312 vignettes (i.e., all subsets evidence acceptable to high Tucker Index values, showing congruence with Knutson's original structure). Each subset contains 39 unique vignettes and reduces the necessary time investment for participants from 1.4–3.4 h to 10–25 min (per subset). These subsets provide researchers empirically-verified sets of vignettes that can be used to concurrently assess several main event features known to influence MJ, without burdening participants with a lengthy protocol. Furthermore, these subsets can be combined to increase reliability through repetitions of characteristically similar stimuli. In this way, researchers can choose to administer 39, 78, or 117 vignettes with corresponding time investments of 10–25 min, 20–50 min, and 30–75 min. This will allow the development of assessment protocols to fit specific studies and methodologies (e.g., fMRI, psychophysiology, behavioral). Moreover, given the similarity of factor loadings across subsets (i.e., high Tucker Index values; Table 3) these subsets may be suitable as parallel forms, which would directly facilitate experimental (e.g., pre-post) and longitudinal (time-point 1 vs. time-point 2) testing of MJ. Future studies will be needed to further and directly test this possibility.

Second, we provide event feature rating data and clarify factor structure results using a large number of raters relative to Knutson et al.'s original paper (Supplementary Material – Vignette Information). Specifically, in comparison to Knutson et al.'s ratings of 312 vignettes by 30 individuals, this study involved a total of 661 participants across 117 vignettes, with groups of around 220 unique participants rating subsets of 39 unique vignettes. Using the rating data from this larger sample, we replicated Knutson et al.'s three-component solution for event feature ratings (i.e., norm violation, social affect, intention) in each of our subsets (Tables 2 and 3). As such, we demonstrated the generalizability of Knutson et al.'s results. Further, we revealed the stability of these factor solutions (via Tucker Index values) across sex (males, females; Table 4) and political affiliation (i.e., liberal, moderate, conservative; Table 5; see Supplemental Materials – Appendix D and E, respectively for full factor structures and additional Tucker Index comparisons). Demonstrating the stability of these factor structures provides increased confidence in these solutions. Overall, these results demonstrate the importance of previously identified moral features, with norm violation representing features emphasized by Shweder et al. (1997) and Haidt (2007), social affect reflecting the emotional components of MJ (Moll, de Oliveira-Souza, Bramati et al., 2002; Moll, de Oliveira-Souza, Eslinger et al., 2002), and the intention component indicating the importance of instrumentality and intent (Koster-Hale et al., 2013). Moreover, since different event features are typically studied in isolation, these results provide important insights into the multi-dimensional nature of MJ, including that moral processes may be best described by three linked, yet independent, domains – social norms, social emotions, and intentions. Based on the variance explained by these factors (Table 2), we also see that social norms may play the largest role in our moral judgments, followed by social emotions, and then intention. This information speaks to, and our methods provide novel ways of exploring, many of the core discussions in MJ research around the influence of cultural/social norms (Haidt & Joseph, 2007), the weight of emotion versus reasoning (e.g., social intuitionist model: Haidt, 2001; dual process model: Greene, 2001), and the importance of intent (Young et al., 2010, Cushman, Sheketoff, Wharton, & Carey, 2013).

Third, our study provides empirical characterization of different vignette types, identifying six distinct groups that vary according to vignette content (Figs. 1 and 2). These groupings expand the empirical characterization of the 117 vignettes administered here to specific aspects of vignette content, providing six distinct groupings that can be used to assess the impact of certain moral features or content types on behavior and decision making. For example, comparing responses to vignettes in the Prosocial set to those in the Compassion set can help elucidate the role of emotion (i.e., emotional intensity and emotional aversion) in judgments of prosocial behavior,

whereas contrasting responses in the Illegal and Antisocial set to responses in the Peccadillo set can evaluate the impact of legality. In addition, we demonstrated that these groupings vary in ratings of moral appropriateness in ways that would be expected given their content. Specifically, we observed that vignettes in the Compassion and Prosocial groups are rated as the most morally appropriate, whereas those in the Peccadillo, Deception, Controversial Act, and Illegal and Antisocial groups are rated as increasingly morally inappropriate.

Although ratings of these groupings did not differ across political affiliation, we did find evidence that females, relative to males, rated vignettes in the Deception, Controversial Act, and Illegal and Antisocial groups as less morally appropriate and vignettes in the Prosocial group as more morally appropriate, with effects ranging in size from small to large (Deception  $d = .34$ , Controversial Act  $d = .56$ , Prosocial  $d = -.71$ , Illegal and Antisocial  $d = .87$ ). Put another way, females appear to be harsher in their MJ of negative actions and more morally approving of prosocial actions. Previous research using the MFT framework indicates that females, relative to males, attend more to the Harm ( $d = .58$ ), Fairness ( $d = .22$ ), and Purity (Sanctity;  $d = .15$ ) foundations, whereas males attend slightly more to the In-group (Loyalty) and Authority foundations ( $d_s < .06$ ; Graham et al., 2011). These data suggest that female's judgments of moral appropriateness may focus on features of harm, self-benefit (fairness), and other-benefit (fairness) – suggestions further supported by previous research on the impact of sex on empathy (Davis, 1983) and egalitarianism (Arts and Gellissen, 2001). In contrast, and also in line with the MFT research, males may focus more on aspects related to social norms (e.g., loyalty) and legality (e.g., authority). While this provides a potential path to link our results with other established findings, given our current data and event feature scales, we are unable to speak at this level of specificity. Particularly, our current event feature scales do not map directly onto MFT's foundations in a one-to-one fashion. Moreover, as seen in Figs. 1 and 2, and by the very nature of LPA, groupings are not solely differentiated on any specific event feature. We note potential avenues for working through these limitations below.

### Directions for future research and development

We also note areas for further study and development. First, the moral content covered by the current stimuli should be expanded upon. The original real-life narratives (Escobedo et al., 2009) adapted by Knutson et al., focused on morality defined by personal code and did not seek to address religious morality. Religious belief and events are found in many conceptualizations of MJ (e.g., Shweder's, 1997, domain of divinity; Haidt's, 2007, foundation of purity/sanctity) and investigating these features more directly (e.g., religious vignettes and/or event feature scales) will likely foster a more inclusive



and accurate picture of MJ. In addition, future developers of these stimuli may wish to more directly incorporate aspects of prominent MJ theories or stimuli. For example, a recent publication (Clifford et al., 2015) provides researchers with a set of moral vignettes and scales (e.g., This action violates norms of loyalty) to assess the MFT. These scales could be directly and easily incorporated into the stimuli studied here and would provide researchers a more direct link between the key event-features explicated here and MFT. Indeed, this path would help further explicate our findings regarding sex differences noted above.

In line with future expansions, our extended analyses (see Supplementary Materials) indicate that the three event features excluded from Knutson et al.'s original analysis (frequency, personal familiarity, and general familiarity), as well as three new event feature scales measuring self-harm (self-harm scale), behavioral consistency (i.e., regularity of the actor's actions; once vs. repeated scale), and counterfactual thought (i.e., likelihood that the participant would have acted differently than the actor in the vignette; act differently scale) may provide additional information on how individuals make MJ. The addition of the three latter scales indicates a potential fourth factor, event familiarity and likelihood. At the same time, this additional factor increases the variance explained from 83% in the 10-event feature three-factor structure to 86% for the 13-event feature four-factor structure, and 85% for the 16-feature four-factor structure. Thus, this additional scale may not significantly increase the variance explained, but may provide a fuller and more detailed map of the moral landscape. Future studies should seek to validate our findings which may provide a more encompassing assessment of MJ. Overall, improved coverage would allow for greater clarification on how specific event features and concerns are integrated or weighed against each other to form MJ.

Second, future studies should refine measurement and exploration of potential individual difference moderators. For example, gender can more accurately be differentiated from biological sex, which we relied on here, through measures of masculinity, femininity, and gender identity (Palan et al., 1999). In a similar vein, our measure of political affiliation could not distinguish between socially conservative and economically conservative ideologies, with the former potentially impacting MJ to a greater degree (Graham et al., 2009). This may partly account for why we did not detect differences in ratings of MJ vignettes based on political affiliation, particularly since other studies have reported differential moral processing across political ideologies (Graham et al., 2009). Furthermore, although we were unable to investigate the impact of religious affiliation due to limited religious diversity in our sample, future studies may wish to attend more to religiosity (e.g., belief in a god, engagement in religious activity, dedication to religious doctrine) than religious affiliation (e.g., whether you identify with Christianity, Judaism, Islam, etc), as

the former is increasingly seen as more important (Graham & Haidt, 2010). Thus, in general, future studies could further explicate these areas by including more granular measures of individual differences and by potentially incorporating the MFT vignettes/scales created by Clifford et al. (2015).

Third, given that this study relied on a convenience sample of US college students, studying more diverse populations will be necessary to facilitate discussions on the impact of culture and whether specific aspects of morality are culturally dependent or universal. Fourth, investigating the test-retest reliability of the feature ratings within individuals, especially across the lifespan, will be crucial to bolstering confidence in these stimuli for different study designs (e.g., experimental, longitudinal) and in elucidating developmental processes involved in MJ formation.

Lastly, future studies should consider or aim to assess these stimuli within a CFA framework. CFA is a traditional choice for confirming a factor structure based on theory to justify the constraints in the model. Here, our aim was to replicate a factor structure with a subset of items as a step to improve assessment feasibility and not necessarily confirm a theoretical model. In this regard, it was not clear that Knutson's factor results, derived from a sample of 30 people, provided a well-established framework of the factors underlying moral judgment from which to draw strong a priori constraints. As such, EFA that imposes no a priori constraints in the model was an appropriate choice for our analyses (Church & Burke, 1994; Ferrando & Lorenzo, 2000; McCrae et al., 1996). Given issues around the assumption of independent observations (which our study violates, as the event feature scales were the units of analysis, collapsed across participants; Benlter & Chou, 1987), one way to move forward with a CFA would be to conduct a CFA in a latent model framework such as structural equation modeling (SEM), wherein the level of analysis would be at the individual participant level, and the complex structure of multiple scales and multiple vignettes could be modeled. This approach, however, would address an entirely different question than Knutson et al. Nonetheless, the goal and success of our study was to provide researchers with a more time-efficient method for investigating MJ within the framework developed by Knutson et al., which can be validated further in CFA.

## Conclusion

This study advances research on the nature of moral judgment and its links to behavior by using a large sample to develop and characterize a core battery of realistic vignettes concurrently rated on key moral event features that can be administered in a brief, time-efficient manner. In doing so, we provide investigators with effective and flexible tools to fit their unique and specific needs while ensuring broad coverage of



the many factors implicated in MJ. For example, the stimuli and methods advanced here are adaptable to various methods of inquiry, such as behavioral and imaging studies, which can provide converging lines of insight into MJ. Moreover, because the stimuli are fully characterized in numerous ways, such as event feature ratings, factor loadings and structure, and vignette content, researchers can choose to focus on data or areas of specific interest to them while still being able to provide insights in other domains. For example, researchers initially focused on vignette content will still be able to analyze their data in reference to event feature ratings or factor loadings and structure. This will not only foster a richer understanding of the factors implicated in MJ but will provide a more economical trade-off between time investment and data-collection. Overall, these methodological advances help to broaden our assessment of MJ to include real-life, ecologically valid stimuli that take into account, in a concurrent fashion, the numerous features implicated in MJ. In addition, our preliminary results concerning the impact of sex and political affiliation on MJ offer novel insights into the importance of these individual difference variables and provide clear avenues for further investigation and stimuli development to explicate the nature of MJ.

**Author Note** EKM was supported by the NSF IGERT Fellowship, Grant No. 0903622.

We would like to thank the reviewers for their comments and assistance in improving the manuscript.

## References

- Arts, W., & Gelissen, J. (2001). Welfare states, solidarity and justice principles: Does the type really matter? *Acta Sociologica*, *44*(4), 283–299.
- Bentler, P. M., & Chou, C. H. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*, 78–117.
- Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, *88*(1), 1.
- Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, *217*(4), 783–796.
- Carver, R. P. (1990). *Reading rate: A review of research and theory*. San Diego, CA: Academic Press.
- Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three-and-four-dimensional models. *Journal of Personality and Social Psychology*, *66*, 93–114.
- Clark, S. L., Muthén, B., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 681–703. doi:10.1080/10705511.2013.824786
- Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, *47*(4), 1178–1198.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.
- Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. doi:10.1126/science.aac4716
- Cresswell, M., & Karimova, Z. (2010). Self-Harm and Medicine's Moral Code: A Historical Perspective, 1950–2000. *Ethical Human Psychology and Psychiatry*, *12*(2), 158–175.
- Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21.
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113–126.
- Escobedo, J. R. (2009). *Investigating Moral Events: Characterization and Structure of Autobiographical Moral Memories. Unpublished Dissertation*. Pasadena, California: California Institute of Technology.
- FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441.
- Ferrando, P. J., & Lorenzo, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica*, *21*, 301–323.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.
- Galen, L. W. (2012). Does religious belief promote prosociality? A critical examination. *Psychological Bulletin*, *138*(5), 876.
- Gilligan, C. (1982). *In a different voice*. Harvard University Press.
- Gold, N., Pulford, B. D., & Colman, A. M. (2014). The outlandish, the realistic, and the real: Contextual manipulation and agent role effects in trolley problems. *Frontiers in Psychology* *5*, 35. doi:10.3389/fpsyg.2014.00035
- Graham, J., & Haidt, J. (2010). Beyond beliefs: Religions bind individuals into moral communities. *Personality and Social Psychology Review*, *14*(1), 140–150.
- 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.
- Greenberg, E., Dunleavy, E., and Kutner, M. (2007). Literacy Behind Bars: Results From the 2003 National Assessment of Adult Literacy Prison Survey (NCES 2007-473). U.S. Department of Education. Washington, DC: National Center for Education Statistics
- Greene, J. D., Cushman, F. a., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. doi:10.1016/j.cognition.2009.02.001
- Greene, J., & J. H. (2002). How (and where) does moral judgement work? *Trends in Cognitive Sciences*, *6*(12), 517–523.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science (New York, N.Y.)*, *293*(5537), 2105–2108. doi:10.1126/science.1062872
- 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). The new synthesis in moral psychology. *Science*, *316*, 998–1002.
- Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *The Public Opinion Quarterly*, *45*(4), 549–559. doi:10.1086/268687

- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114. doi:10.1007/s10869-011-9231-8
- Ince, R. A. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2016). A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. *Human Brain Mapping, doi:10.1002/hbm.23471*
- Jaffee, S., & Hyde, J. S. (2000). Gender differences in moral orientation: A meta-analysis. *Psychological Bulletin, 126*(5), 703–726. doi:10.1037/0033-2909.126.5.703
- Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience, 00*(00), 1–10.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187–200.
- Kapogiannis, D., Barbey, A. K., Su, M., Zamboni, G., Krueger, F., & Grafman, J. (2009). Cognitive and neural foundations of religious belief. *Proceedings of the National Academy of Sciences, 106*(12), 4876–4881.
- Knutson, K. M., Krueger, F., Koenigs, M., Hawley, A., Escobedo, J. R., Vasudeva, V., Adolphs, R., Grafman, J. (2010). Behavioral norms for condensed moral vignettes. *Social Cognitive and Affective Neuroscience, 5*(4), 378–384. doi:10.1093/scan/nsq005
- Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America, 110*(14), 5648–53. doi:10.1073/pnas.120799211
- Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization Techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 47–58.
- Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57–64.
- MATLAB (2014). Natick, Massachusetts. The MathWorks Inc.
- McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Pauonen, S. V. (1996). Evaluating replicability of factors in the revised NEO personality inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552–566.
- Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *Neuro Image, 16*(3 Pt 1), 696–703.
- Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 22*(7), 2730–2736.
- Molloy, E. K. and Kruepke, M. D., (2017). Selecting representative subsets of vignettes for investigating multiple facets of moral judgement: Documentation and MATLAB Code. GitHub repository, <https://github.com/ekmolloy/select-vignette-subsets>
- Muthén, L. K., & Muthén, B. O. (2008). *Mplus (Version 5.1)*. Los Angeles, CA: Muthén & Muthén.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569.
- Palan, K. M., Areni, C. S., & Kiecker, P. (1999). Reexamining masculinity, femininity, and gender identity scales. *Marketing Letters, 10*(4), 357–371.
- Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition, 145*, 30–42. doi:10.1016/j.cognition.2015.08.001
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>
- Revelle, W. (2015) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <https://CRAN.R-project.org/package=psych> Version = 1.5.8.
- Reynolds, S. J., & Ceranic, T. L. (2007). The effects of moral judgment and moral identity on moral behavior: An empirical examination of the moral individual. *The Journal of Applied Psychology, 92*(6), 1610–1624. doi:10.1037/0021-9010.92.6.1610
- Rusbult, C. E., & Van Lange, P. A. M. (1996). Interdependence processes. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 564–596). New York, NY: Guilford Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922.
- Shweder, R., Much, N., Mahapatra, M., & Park, L. (1997). Divinity and the “Big Three” Explanations of Suffering. *Morality and Health, 119*.
- Simpson, A., & Laham, S. M. (2015). Individual differences in relational construal are associated with variability in moral judgment. *Personality and Individual Differences, 74*, 49–54.
- Strohming, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*(1), 159–171. doi:10.1016/j.cognition.2013.12.005
- Thoma, S. (1994). Moral judgments and moral action. *Moral development in the professions: Psychology and applied ethics*, 199–21.
- Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion, 12*(3), 579.
- Vranka, M. A., & Bahnik, Š. (2016). Is the Emotional Dog Blind to Its Choices? *Experimental Psychology, 63*(3), 180–188.
- Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron, 65*(6), 845–851.