Testing Measurement Invariance of the Moral Foundations Questionnaire Across 27 Countries

Kathryn Iurino[1] [iD] and Gerard Saucier[1]

Abstract

It has become clear that there are multiple "moralities": diverse bases that guide people's judgments of right and wrong. The widely known Moral Foundations Theory stipulates that there are at least five such moralities, measurable via questionnaire, and tends to assume that these distinct foundations are rooted deep in humanity's evolutionary past. Were this true, we should find that the structure of five foundations is cross-culturally generalizable. Such assumptions are best tested in a diverse range of global populations with no built-in Western bias. Here, we test the measurement invariance of the short-form Moral Foundations Questionnaire across 27 countries spanning the five largest continents. We find that it is difficult to specify Moral Foundations Questionnaire items in a quantitative five-factor model that will converge nonproblematically across a wide variety of populations.

What considerations are relevant to moral judgment? Defining the scope of moral considerations is the first step to understanding the psychological nature of moral judgment—how moral judgments are made, the role of moral judgments in human functioning, and how moral judgments vary across individuals and societies. Within psychology, Kohlberg (1971) and Gilligan (1977) focused on the most basic, uncontroversial moral obligations we have toward all people. These include not harming others, caring for others (Gilligan, 1977), and acting in a just manner—for example, by treating people fairly and not deceiving them (Kohlberg, 1971). However, concerns for individual well-being and justice are not the only considerations that inform moral judgment for many people. Drawing on Shweder, Much, Mahapatra, and Park's (1997) work, Haidt and colleagues incorporated insights from psychology, anthropology, and evolutionary theory to expand the domain of moral concerns studied in psychology (Haidt & Graham, 2009; Haidt & Joseph, 2007).

Using a social intuitionist framework, Graham et al. (2011; Graham et al., 2013) defined what is necessary for a domain to be considered a "moral foundation" by several criteria. First, the domain must show up as a common concern in third-party normative judgments (i.e., show up in gossip) and must be culturally widespread. Moreover, the domain should involve automatic affective evaluations for which there is evidence of innate preparedness (i.e., some form of it should be found prior to experience, as evident by its presence in nonhuman primates and in infants) rather than being merely a "social convention," and it should have a basis in evolutionary psychology. That is, it should be plausible that the domain's associated psychological features served an adaptive function in our evolutionary past.

The Moral Foundations Questionnaire

Based on these criteria, Graham et al. (2011) proposed five distinct domains of moral judgment. They developed the Moral Foundations Questionnaire (MFQ) to assess the extent to which an individual's moral judgments arise from the activation of "modules" (Haidt & Joseph, 2007) in these domains: harm/care, fairness/justice, in-group/loyalty, authority, and purity. There is good evidence for the internal validity, external validity, and predictive utility for scales measuring the five foundations within relatively homogeneous English-speaking samples. In the key Graham et al. (2011) study, 80% were from the United States, 87% were White, 63% were male, all spoke English, and the mean level of education was in-between "completed college" and "some graduate school." These

[1]University of Oregon, Eugene, OR, USA

Corresponding Author:
Gerard Saucier, Department of Psychology, University of Oregon, 1227 University of Oregon, Eugene, OR 97403, USA.
Email: gsaucier@uoregon.edu

participants appear to have been highly motivated; they navigated themselves to YourMorals.org and filled out questionnaires purely out of personal interest in "exploring their morals" and receiving feedback.

The MFQ has been theoretically generative, revealing interesting relationships between gender, culture, personality, political orientation, and moral concerns with more resolution than was previously possible (Graham et al., 2011; Graham et al., 2013). However, more work is needed to evaluate the measurement properties of the scales across cultural contexts. Although studies have found supportive evidence for a five-factor model in Western countries such as France, Sweden, and New Zealand (Davies, Sibley, & Liu, 2014; Métayer & Pahlavan, 2014; Nilsson & Erlandsson, 2015) as well as in Turkey (Yilmaz, Harma, Bahçekapili, & Cesur, 2016), Davis et al. (2016) found poor fit for a five-factor model in a U.S. sample with a higher proportion of Black participants than in previous research. Furthermore, when examining each foundation separately, Davis et al. (2016) found evidence for factorial but not scalar invariance across White and Black college students from the United States.

More tests of measurement invariance are needed to gauge the cross-cultural applicability of the MFQ's five-foundation model, as there have been no previous investigations of its measurement invariance in non-U.S. countries. Furthermore, since Moral Foundations Theory aims to measure those sources of moral judgment most characteristic of human functioning, highly relevant to evaluating Moral Foundations Theory is testing the cross-cultural generalizability of the MFQ. A finding of measurement invariance would provide more confidence that use of the MFQ across cultures can shed light on meaningful differences between cultures rather than merely reflecting the measurement properties of the MFQ.

Here, the goal was to evaluate the MFQ across more culturally diverse samples, with world regions more fairly represented than in the Graham et al. (2011) study. We planned to test for four increasingly stringent levels of measurement invariance following Muthén and Muthén (2012). First, we planned to test for configural invariance; this would suggest that the same five-factor structure is found across countries, and that the same indicators can be used to define these factors across countries. Second, we planned to test for factorial/metric invariance (factor loadings invariant across countries); this would suggest that the content of the factors is the same across countries, and that relationships between variables can justifiably be compared across countries. Third, we planned to test for scalar invariance (intercepts invariant across countries); this would suggest that means across countries can be compared. If scalar invariance is not met, observed differences in means cannot be assumed to reflect actual differences in the underlying latent construct, even though a few studies have already (prematurely) made inferences from mean differences (Graham et al., 2011; Van

Leeuwen, Park, Koenig, & Graham, 2012). Finally, we planned to test for full invariance (error variances invariant across countries).

We chose to test for measurement invariance within a confirmatory factor analysis (CFA) rather than an item response theory (IRT) framework for a few reasons. First, CFA allows for testing of the equivalence of the factor structure, whereas IRT can only test for equivalence between isolated scales of items. To lay the groundwork for the cross-cultural applicability of the MFQ as it relates to Moral Foundations Theory, we needed first to establish that the same five-factor structure exists across countries. Second, CFA is more amenable to testing the equivalence between many groups; IRT methods generally require pairwise comparisons, which would have been less practical than a CFA approach given the number of countries included in this study. Third, a CFA approach is more consistent with how the MFQ scales have predominantly been modeled and used in past research (as simple aggregated sums of individual items). Since we were interested in replicating the results from previous studies using a more culturally diverse selection of countries, we chose to model the five foundations in the same way as past research.

Method

Participants

Data for this study come from the Survey of World Views (SWV), collected online from institutions of higher education in 33 different countries in 2012. For this project, countries were selected so as to overcome a frequent limitation in psychological research: an overrepresentation of Western industrialized democracies (Henrich, Heine, & Norenzayan, 2010) and East Asian countries, while underrepresenting the rest of the world (Saucier et al., 2015). Although Graham et al.'s (2011) data included a wide range of world regions, 94.7% of participants were from Western industrialized democracies (U.S., Canada, Western Europe, United Kingdom, Australia), while Eastern Europe, Latin America, South Asia, Middle East, Southeast Asia, East Asia, and Africa combined constituted just 5.3%. For the SWV, countries were selected so as to be more representative of the world's population. The 33 SWV countries add up to 67.3% of the world's population and 76.2% of the aggregate gross domestic products of all countries in the world (Central Intelligence Agency, 2012). (For more details on the SWV, see Saucier et al., 2015.)

Here, we excluded countries with less than 100 participants, leaving 27 countries. The mean age for samples ranged from about 20 years (Ukraine, Philippines) to 26 years (Mexico); percentage of females varied from 23% (Bangladesh) to 88% (Poland). Table 1 provides

Table 1. Demographic Characteristics for 27 Countries, Grouped by Region.

Country/territory/ region	N	% Female	Mean age, years	Mean-%-missing[a]
Africa (sub-Saharan)				
Tanzania	240	32	24.9	3.60
Kenya	275	34	24.6	1.85
Ethiopia	359	28	24.0	2.49
North Africa/Middle East				
Morocco	411	52	25.5	2.87
Turkey	402	55	21.1	2.45
South Asia				
Bangladesh	247	23	21.8	2.47
India	359	63	21.0	5.17
Nepal	337	59	21.0	3.59
Southeast Asia				
Malaysia	311	67	20.5	2.89
Philippines	393	69	20.0	5.99
Thailand	343	72	21.6	2.77
Singapore	287	56	21.6	4.56
East Asia				
China (mainland)	307	74	20.8	5.20
Taiwan	380	64	22.5	2.03
Japan	375	63	20.9	2.48
East/southeast Europe				
Ukraine	236	63	20.2	2.97
Poland	221	88	21.2	0.32
Greece	233	72	21.8	2.08
Western Europe				
Spain	333	65	22.8	2.30
Germany	338	52	23.5	0.71
England	218	64	22.7	1.19
North America				
Canada	215	63	21.8	0.98
United States	408	58	21.9	2.24
Latin America				
Mexico	139	66	26.1	4.96
Peru	276	62	21.8	7.03
Argentina	220	56	24.4	3.61
Brazil	192	79	22.2	0.91

[a]Mean-%-missing is the mean percentage of missing responses across 20 Moral Foundations Questionnaire items.

demographic characteristics for each country's participants, including its *N* and mean-%-missing-data.

Materials

Due to space considerations, the SWV used the 20-item version of the MFQ, consistent with advice on moralfoundations.org/questionnaires. This version has four items for each of the five foundations; each set of four items is equally divided into two question formats. One format asks participants to rate "How relevant is the following to

your moral concerns?" on a 6-point scale (*not at all relevant* to *highly relevant*). The second format asks participants to rate "How strongly do you agree with each of the following statements?" on a 6-point scale (*strongly disagree* to *strongly agree*). Items were administered in the country's official language (or one of them, if multiple). Six countries—United States, Canada, England, Kenya, India, Singapore—used English. Items were translated and checked/revised by a back-translation method. Supplementary materials include a list of English-language items (see Supplemental Appendix A, available with the online version of the article); survey materials in 32 different languages can be found at the following URL: http://psychometriglossia.uoregon.edu/.

Procedure

Moral Foundations Theory is presented as a model that should be globally applicable rather than regionally specific, so our tests took a pan-cultural approach. Because we wanted to cross-validate (i.e., derive/revise in one set, then test in another) any model specification, we randomly split 27 countries into three sets of nine countries each. A constraint was that each set, where feasible, represent all major world regions found in the data set: North America, South America, East Asia, Southeast Asia, South Asia, Western Europe, Eastern Europe, sub-Saharan Africa, and North Africa/Middle East. A split into three sets enabled us to make modifications to the models in the first set of countries, then test how well these models generalized to an independent, comparable set of countries, following Wiggins (1973). If modifications developed in the first set of countries did not generalize well in the second set, we could respecify the model there, then see how well this respecified model generalized in a third set. Such "modification and cross-validation" methods incorporate replication tests into the research design.

Set 1 included the United States, Argentina, Taiwan, Malaysia, India, Germany, Greece, Kenya, and Turkey. Set 2 included Canada, Peru, China, Thailand, Nepal, Spain, Ukraine, Tanzania, and Morocco. Set 3 included Mexico, Brazil, Japan, Philippines, Bangladesh, England, Poland, Ethiopia, and Singapore. Each set had one country from each of nine regions listed above, except that Set 3 had only the first eight, with Singapore being the ninth sample.

Analyses

Data were analyzed using M*plus* Version 7. Missing data were handled using the full-information maximum-likelihood method. Prior to all analyses, three selection criteria were applied in all countries. Excluded were cases with all data missing, or judged to be possible random responders (based on near-zero correlations between their responses

and the rest of the sample, as one would expect from random data), or judged to be nonindependent (the same person participating twice, or perhaps two respondents apparently coordinating answers) by the study coordinator who monitored incoming data. Across the original set of 33 countries, this eliminated 545 cases (6%) of the data.

Latent variables were scaled by setting their variances to 1; all factor loadings were freely estimated. Giving credence to past research supporting the validity of five factors, we sought a way to specify the five factors that might be recommended for the future for other MFQ-20 data. Seeking to stay as true as possible to the five-factor model specified in Graham et al. (2011), we followed the procedure (specified in advance, before running all analyses) described next.

First, we tested the five-factor model in each Set-1 country individually. If this failed (resulted in nonconvergence or nonpositive-definite matrices in most countries), we planned to adopt a less demanding modeling approach: We would standardize responses within each Set-1 country based on the country mean and standard deviation, so that factors could not reflect between-country differences, then pool the data from the Set-1 countries and perform a TARGET rotation using as targets loadings from the five-factor model in Graham et al. (2011, p. 376). Target loadings would inform specification of a model in which each item can load not only on the intended factor but also on any additional factors with a cross-loading above .30. This alternative specification for a five-factor model could be tested in Set-2 countries, examining generalizability to a new, independent set of countries. If the model ran in more than half of the Set-2 countries (i.e., without nonconvergence or nonpositive-definite latent-variable covariance matrices), the plan was to test measurement invariance for this model across those countries it ran in. If the model ran in less than half of the countries, we would repeat the same procedure that we followed in Set 1: Set-2 countries' data would be standardized by country and pooled, with a new target rotation to arrive at another alternative five-factor model specification, to then be tested using CFA in Set 3. In case this procedure did not result in any viable model across countries, the plan was to follow the same procedure but with a less restrictive approach for deriving an alternative model specification: exploratory factor analysis using geomin rotation.

Following Hu and Bentler (1999), we examined varied fit indices that evaluate the degree of misspecification: standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), and comparative fit index (CFI). SRMR measures absolute fit: how well (on average) the model reproduces the correlation matrix. RMSEA indicates absolute fit adjusting for model parsimony; here the magnitudes of the covariance residuals are adjusted for degrees of freedom. CFI reflects the proportion of improvement in fit relative to the null (or independence) model. Hu and Bentler (1999) suggest that to confirm adequate fit for a model, CFI should ideally be greater than .95, RMSEA should be less than .06, and SRMR should be less than .08.

For any model that ran successfully in more than half of the countries in a set, we planned to test for four progressively more stringent levels of measurement invariance, following Muthén and Muthén (2012, p. 483). The first would test configural invariance, discerning whether the factor structure was invariant across countries. The second would test for factorial invariance, holding factor loadings equal across countries. In both the configural and factorial invariance models, all factor means were set to 0, all intercepts freely estimated for all countries, and factors were all scaled by setting their variances to 1.

Third, we would test for scalar invariance, seeing whether holding the intercepts constant across groups resulted in substantial decrement in fit. For the scalar-invariance model, a first country's factor mean is set to 0 and all other means are freely estimated (again, with all factor variances set to 1). The final test would be full invariance. Except for this additional equality constraint, specification for the full-invariance model was the same as for scalar invariance.

Results

In each of the Set-1 countries, the latent-variable covariance matrix was nonpositive-definite, with at least one correlation between factors estimated to exceed 1.00. Since the regular five-factor model had estimation problems in all Set-1 countries, we performed a target rotation on their pooled data.[1] Overall, just 10 of 20 items had their highest loading on the factor targeted, indicating that a somewhat different combination of items emerged measuring each factor (see Table 2; items in boldface are indicators used for each factor in the CFA). In order, the factors (1 through 5) could be interpreted as follows: moral relevance of suffering and differential treatment, fairness as crucial to morality, moral relevance of loving one's country (and traditions), judgments emphasizing "binding" moral foundations, and moral relevance of purity concerns. Only the second factor drew substantially from both relevance and judgment items.

We then attempted to cross-validate this structure using a CFA in Set-2 countries, with indicators being all items loading above .3 in the Set-1 Target Rotation. This model successfully ran in six of nine countries (excluding Canada, Nepal, and Tanzania). The model failed to converge in Canada; the latent-variable covariance matrix (psi) was nonpositive-definite in Tanzania and Nepal—in both countries the correlation between the second and fourth factors was estimated to exceed 1.00. Table 3 displays the fit statistics of the model for each of the six countries the model successfully ran in.

Table 2. Results of Target Rotation in Set-1 Countries.

Item	1	2	3	4	5
MF12 (EMOTIONALLY)	**.766**^{T}	.036	−.088	−.035	.099
MF13 (TREATED)	**.593**	.226^T	.023	−.112	.096
MF02 (FAIRLY)	.167	**.592**^{T}	.000	.130	−.150
MF19 (UNFAIRLY)	.182	**.559**^{T}	.071	**−.314**	**.322**
MF08 (JUSTICE)	.162	**.542**^{T}	.069	.202	−.138
MF01 (COMPASSION)	.248^T	**.371**	.057	.213	−.114
MF07 (ANIMAL)	.167^T	**.337**	.041	.083	−.088
MF18 (WEAK)	**.311**^{T}	**.323**	.221	−.121	.045
MF14 (LOVECOUNTRY)	.139	−.147	**.88**^{T}	.177	−.198
MF21 (TRADITIONS)	−.006	−.125	**.346**	.265^T	**.330**
MF04 (KIDRESPECT)	.035	.092	.058	**.543**^{T}	.197
MF11 (UNNATURAL)	−.020	−.085	.025	**.466**	**.332**^{T}
MF10 (SEXROLES)	−.028	−.056	.016	**.465**^{T}	.251
MF05 (HARMLESSDG)	.026	.134	.035	**.433**	.209^T
MF03 (HISTORY)	.030	.089	.205^T	**.403**	.004
MF09 (FAMILY)	.060	−.029	.057^T	**.364**	.115
MF16 (DECENCY)	.052	.041	.193	.197	**.541**^{T}
MF22 (DISGUSTING)	−.011	.107	.094	.116	**.538**^{T}
MF20 (BETRAY)	.022	.255	.197^T	−.038	**.471**
MF15 (AUTHORITY)	.045	−.045	**.305**	.260^T	**.424**

Notes. In boldface are items that were salient on each factor and items with cross-loadings that were selected to load on each factor in the Model 1 confirmatory factor analysis. Loadings marked ^{T} are those that were originally specified to have nonzero targets, the exact value of which was equal to the factor loadings that Graham et al. (2011) found for the five-factor model. All other loadings had a target loading of 0.

Since the model developed with Set-1 data ran in more than half the Set-2 countries, we used it to test measurement invariance in Set-2 countries (see Table 4 for results), also looking for minor modifications to improve the model fit, which might be cross-validated in Set-3 countries.

A main issue with this model was with items/indicators that were allowed to cross-load: These items tended to have very large standard errors, and standardized factor loadings that were estimated to be much larger than loadings for other items. The worst performing cross-loading item was MF19, and after that, MF18 and MF21. Therefore, we created two modified models. In one, MF19 was not allowed to cross-load on any factors other than the factor it was most salient on (Model 2); in another, MF19, MF18, and MF21 were not allowed to cross-load on any factors other than their most salient one (Model 3).

Only Model 2 met the criterion of running in more than half of Set-2 countries, so we then tested Model 1 (the original model) and Model 2 in Set-3 countries. Model 2 ran in three of nine Set-3 countries, and Model 1 ran in two of nine (see Table 5). Neither model converged in Mexico; for Bangladesh, England, Japan, and Poland, both models resulted in a latent-variable covariance matrix that was nonpositive-definite. This recurring tendency to produce

Table 3. Results From fitting CFA Models in Set-2 Countries.

Country	Model	χ^2	df	CFI	RMSEA	SRMR
Canada	Model 1 (nc)	—	—	—	—	—
	Model 2 (npd)	—	—	—	—	—
	Model 3 (npd)	—	—	—	—	—
China	Model 1	428.091	154	.805	.076	.073
	Model 2 (npd)	—	—	—	—	—
	Model 3 (npd)	—	—	—	—	—
Morocco	Model 1	291.832	154	.951	.047	.041
	Model 2	565.633	156	.855	.080	.085
	Model 3	568.962	158	.855	.080	.086
Nepal	Model 1 (npd)	—	—	—	—	—
	Model 2	402.775	156	.832	.069	.073
	Model 3	410.882	158	.828	.069	.074
Peru	Model 1	298.888	154	.87	.058	.059
	Model 2	472.781	156	.715	.086	.086
	Model 3	475.138	158	.714	.085	.087
Spain	Model 1	462.354	154	.788	.078	.084
	Model 2	569.047	156	.716	.089	.089
	Model 3	590.285	158	.702	.091	.092
Tanzania	Model 1 (npd)	—	—	—	—	—
	Model 2 (npd)	—	—	—	—	—
	Model 3 (npd)	—	—	—	—	—
Thailand	Model 1	397.385	154	.881	.068	.061
	Model 2	566.137	156	.8	.088	.093
	Model 3 (npd)	—	—	—	—	—
Ukraine	Model 1	283.247	154	.812	.060	.071
	Model 2 (nc)	—	—	—	—	—
	Model 3 (npd)	—	—	—	—	—

Notes. CFA = confirmatory factor analysis; df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; nc = no convergence, npd = not positive-definite latent-variable covariance matrix. Model 1 is the original model derived from target rotation in Set-1 countries. Model 2 removes the cross-loadings from MF19 to the authority and purity factors. Model 3 additionally removes the cross-loading from MF18 to fairness and the cross-loading from MF21 to purity.

Table 4. Results From Measurement Invariance Tests for Model 1 in Set-2 Countries That the Model Ran in (N = 1,906).

	χ^2	df	CFI	RMSEA	SRMR
Configural	2161.795	924	.870	.065	.065
Factorial	2784.588	1,054	.819	.072	.131
Scalar	4743.841	1,129	.621	.100	.161
Full	6073.29	1,229	.492	.111	.215

Note. df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual. For the scalar invariance test, the latent-variable covariance matrix in China and Ukraine were not positive-definite, and for the full invariance test, the latent-variable covariance matrix in China, Morocco, and Thailand were not positive-definite.

nonpositive-definite matrices indicates overextraction (five factors overly highly correlated, too many factors for the data).

Table 5. Results From Fitting CFA Models in Set-3 Countries.

Country	Model	χ^2	df	CFI	RMSEA	SRMR
Brazil	Model 1	284.876	154	.853	.067	.070
	Model 2	293.687	156	.846	.068	.073
Ethiopia	Model 1 (npd)	—	—	—	—	—
	Model 2	435.107	156	.786	.071	.082
Singapore	Model 1	433.726	154	.793	.080	.078
	Model 2	468.686	156	.768	.084	.081

Note. CFA = confirmatory factor analysis; df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual. The models displayed are the only countries these models successfully ran in. Model 1 and Model 2 are the same as those tested in Set 2.

Discussion

Moral Foundations Theory is cross-cultural in its scope, and much interesting research has been done using its associated questionnaires across different populations. The current study set out with the goal of establishing the measurement invariance of the short-form MFQ in order to establish whether cross-cultural comparisons using this questionnaire are justified. However, we were not able to replicate Graham et al.'s (2011) results indicating that a five-factor model is a suitable approach to modeling the moral foundations.

The difficulty arriving at a five-factor model that converged nonproblematically across diverse populations has a few possible explanations. First, our sample was different from Graham et al.'s (2011) sample in a few notable ways. SWV respondents were compensated monetarily rather than by delivery of feedback, which might tap into different strata of the population with differing response-motivations than in previous data (the majority of which were from the United States, all of which were English-speaking, and all of which were intrinsically motivated enough to navigate to YourMorals.org in order to receive feedback). It is possible that a five-factor model is most appropriate for highly motivated, English-speaking participants.

Second, our administration of the MFQ differed from Graham et al. (2011): Due to logistical limitations, we were not able to randomize the order of the items in the MFQ for each participant and thus may have obtained inflated correlations between adjacent items of the same method, thereby leading to inflated correlations between factors. As much as feasible, the order of MFQ items should be randomized across respondents; in this way the adjacency of items will have fewer effects on responses and results.

Third, we used the short-form MFQ, and models might have run more successfully with the long-form measure. Though Graham et al. (2011) found that the two-item combinations in the short-form MFQ were almost as good as the three-item combinations in the long-form MFQ in terms of

internal and external validity, the longer version would have made it possible to run a model that accounted for method variance, as the substantive factors could have been split according to question format without being left to only two indicators per factor. We attempted to run a model with five substantive factors and two method factors, but this model was empirically underidentified; more items may have resolved this problem.

Fourth, using larger samples (e.g., >1,000) as in Graham et al. (2011) within each country might have helped.

Overall, it may be that MFQ structure is not very resilient to smaller samples, short-form versions of the measure, fixed item-order, or less motivated participants. Since our results suggested five factors are too many for the data, it may be worth considering whether other factor structures would be more robust across cultures and methodologies. It is possible that a structure where each domain mapped onto a distinct emotion would be more cross-culturally generalizable, or one where each domain clearly mapped onto a specific function. For example, some moral concerns may function primarily to protect individuals from harm, whereas others may function primarily to preserve the status quo. Future research should investigate whether a common structure can be derived across a diverse set of countries using a comprehensive item pool.

Though this is a demanding test, other studies using the same sample have examined the cross-cultural generalizability of instruments such as the Social Axioms Survey (Bou Malham & Saucier, 2014) and the Questionnaire Big Six (Thalmayer & Saucier, 2014). Both have found better evidence for measurement invariance using the same data set, suggesting the current findings are not an artifact of this particular data set. In the case of the Social Axioms Survey, only two countries did not run due to nonpositive-definite matrices; compared to here, where all of the countries in the first set did not run without nonpositive-definite matrices. This difference in "ease-of-fit" cross-culturally likely arises because the social-axioms constructs and measures were developed from the outset via international collaboration (Leung et al., 2002), rather than (as here) developed in a Western context and then exported to non-Western settings. To minimize regional bias and to enhance downstream cross-cultural applicability, we can recommend that psychological models, where possible, be developed initially from international collaborations.

Development of a sound measure of the psychological bases for moral judgment requires consideration of how well the measure generalizes across cultures; it is important that the phenomena captured are indeed culturally widespread. Generalizability is an important consideration in evaluating any measure, but even more so when the associated theory suggests something about innate human predispositions. The present study addresses the question of whether the five moral foundations, as measured by the

short-form MFQ, generalize to samples more representative of the world (with more non-Western representation) than have been previously studied. Though we do not find support for the generalizability of five moral foundations, our findings provide guidance regarding the use of the short-form MFQ in cross-cultural research and highlight the need for more research addressing the generalizability of longer versions of this questionnaire.

Note

1. Multigroup exploratory structural equation modeling (ESEM) was not part of our prespecified procedure sequence, but comparing results from using ESEM is relevant. When developing a model using multigroup ESEM in Set-1 countries, the fifth factor did not have any salients. Given our a priori preference for five factors, the ESEM was a less effective approach than target rotation (see Supplemental Appendix B for the factor loadings we obtained using ESEM, available with the online version of the article).

Supplemental Material

Supplemental material for this article is available online.

ORCID iD

Kathryn Iurino [iD] https://orcid.org/0000-0001-6461-7032

References

Bou Malham, P., & Saucier, G. (2014). Measurement invariance of social axioms in 23 countries. *Journal of Cross-Cultural Psychology*, *45*, 1046-1060.

Central Intelligence Agency. (2012). *The world factbook*. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html

Davies, C., Sibley, C., & Liu, J. (2014). Confirmatory factor analysis of the Moral Foundations Questionnaire: Independent scale validation in a New Zealand sample. *Social Psychology*, *45*, 431-436.

Davis, D. E., Rice, K., Van Tongeren, D. R., Hook, J. N., DeBlaere, C., Worthington, E. L., Jr., & Choe, E. (2016). The moral foundations hypothesis does not replicate well in Black samples. *Journal of Personality and Social Psychology*, *110*(4), e23.

Gilligan, C. (1977). In a different voice: Women's conceptions of the self and morality. *Harvard Educational Review*, *47*, 481-517.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55-130.

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

Haidt, J., & Graham, J. (2009). Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality. In J. Jost, A. C. Kay & H. Thorisdottir (Eds.), *Social and psychological bases of ideology and system justification* (pp. 371-401). New York, NY: Oxford University Press.

Haidt, J., & Joseph, C. (2007). The moral mind: How 5 sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence & S. Stich (Eds.), *The innate mind: Vol. 3. Foundations and the future* (3rd ed., pp. 367-391). New York, NY: Oxford University Press.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*, 61-83.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Kohlberg, L. (1971). From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In L. Mischel (Ed.), *Cognitive development and epistemology* (pp. 151-284). New York, NY: Academic Press.

Leung, K., Bond, M. H., de Carrasquel, S. R., Muñoz, C., Hernández, M., Murakami, F., . . . Singelis, T. M. (2002). Social axioms: The search for universal dimensions of general beliefs about how the world functions. *Journal of Cross-Cultural Psychology*, *33*, 286-302.

Métayer, S., & Pahlavan, F. (2014). Validation of the Moral Foundations Questionnaire in French. *Revue Internationale de Psychologie Sociale*, *27*, 79-107.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nilsson, A., & Erlandsson, A. (2015). The Moral Foundations taxonomy: Structural validity and relation to political ideology in Sweden. *Personality and Individual Differences*, *76*, 28-32.

Saucier, G., Kenner, J., Iurino, K., Bou Malham, P., Chen, Z., Thalmayer, A. G., . . . Altschul, C. Cross-cultural differences in a global "Survey of World Views." (2015). *Journal of Cross-Cultural Psychology*, *46*, 53-70.

Shweder, R., Much, N., Mahapatra, M., & Park, L. (1997). The big three of morality (autonomy, community, and divinity), and the "big three" explanations of suffering. In A. Brandit & P. Rozin (Eds.), *Morality and health* (pp. 119-169). New York, NY: Routledge.

Thalmayer, A., & Saucier, G. (2014). The Questionnaire Big Six in 26 nations: Developing cross-culturally applicable Big

Six, Big Five, and Big Two Inventories. *European Journal of Personality*, *28*, 482-496. doi:10.1002/per.1969

Van Leeuwen, F., Park, J. H., Koenig, B. L., & Graham, J. (2012). Regional variation in pathogen prevalence predicts endorsement of group-focused moral concerns. *Evolution and Human Behavior*, *33*, 429-437.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Yilmaz, O., Harma, M., Bahçekapili, H. G., & Cesur, S. (2016). Validation of the moral foundations questionnaire in Turkey and its relation to cultural schemas of individualism and collectivism. *Personality and Individual Differences*, *99*, 149-154.