# Theorizing and Measuring Religiosity Across Cultures

Personality and Social
Psychology Bulletin
2017, Vol. 43(12) 1724–1736
© 2017 by the Society for Personality
and Social Psychology, Inc
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146167217727732
journals.sagepub.com/home/pspb

**Adam B. Cohen[1], Gina L. Mazza[1], Kathryn A. Johnson[1], Craig K. Enders[2], Carolyn M. Warner[1], Michael H. Pasek[3], and Jonathan E. Cook[3]**

## Abstract

For almost 50 years, psychologists have been theorizing about and measuring religiosity essentially the way Gordon Allport did, when he distinguished between intrinsic and extrinsic religiosity. However, there is a historical debate regarding what this scale actually measures, which items should be included, and how many factors or subscales exist. To provide more definitive answers, we estimated a series of confirmatory factor analysis models comparing four competing theories for how to score Gorsuch and McPherson's commonly used measure of intrinsic and extrinsic religiosity. We then formally investigated measurement invariance across U.S. Protestants, Irish Catholics, and Turkish Muslims and across U.S. Protestants, Catholics, and Muslims. We provide evidence that a five-item version of intrinsic religiosity is invariant across the U.S. samples and predicts less warmth toward atheists and gay men/lesbians, validating the scale. Our results suggest that a variation of Gorsuch and McPherson's measure may be appropriate for some but not all uses in cross-cultural research.

## Keywords

intrinsic religiosity, extrinsic religiosity, religious orientation, Muslim, Catholic, Christian, confirmatory factor analysis, measurement invariance

Received December 23, 2016; revision accepted August 1, 2017

Gordon Allport's theorizing about religious motivations has defined the psychological study of religion since the middle of the 20th century. His theory and measurement of religious orientation came from his work on religion and prejudice. Allport was puzzled when he found that people who attended church more often showed higher levels of ethnic or racial prejudice despite the messages of universal love contained in many religions. He came to theorize that those with a mature or intrinsic religious orientation (hereafter "IR") would be less prejudiced, while those who were religious for immature, social, or extrinsic reasons (hereafter "ER") would be more prejudiced. He found some empirical support for this distinction when it came to predicting prejudice (Allport & Ross, 1967).

For Allport, being religious for intrinsic reasons meant endorsing items assessing the subjective importance of religion as a master motive in one's life (e.g., "I try hard to carry my religion over into all my other dealings in life"), feeling a divine presence (e.g., "Quite often I have been keenly aware of the presence of God or the Divine Being"), and being privately religious (e.g., "It is important to me to spend periods of time in private religious thought and meditation"). Extrinsic motivations were also defined broadly and included some social motivations (e.g., "One reason for my being a church member is that such membership helps to establish a person in the community"), some personal motivations (e.g., "The purpose of prayer is to secure a happy and peaceful life"), and still others which may or may not simply be reverse-keyed intrinsic items (e.g., "It doesn't matter so much what I believe so long as I lead a moral life").

Allport's view of religious motivations, even some 50 years after his work on this topic, continues to be the gold standard for both theory and measurement in the social and personality psychology of religion (a July 2017 Google Scholar search of "intrinsic religiosity" returned "about 41,000 results"). IR and ER are of course not the only dimensions of religion studied—one immediately also thinks of spirituality, quest, and fundamentalism (Hill & Pargament, 2003)—but Allport's view of intrinsic religiosity remains the *sine qua non* theory about what it means to be appropriately

[1] Arizona State University, Tempe, USA
[2] University of California, Los Angeles, USA
[3] The Pennsylvania State University, University Park, USA

**Corresponding Author:**
Adam B. Cohen, Department of Psychology, Arizona State University, P.O. Box 871104, Tempe, AZ 85287-1104, USA.
Email: adamcohen@asu.edu

and truly religious, and extrinsic religiosity remains the most common theoretical viewpoint of what it means to use religion to achieve instrumental goals, such as comfort or social connection, usually immaturely or insincerely. Although many updated versions of Allport's measures exist, the items are substantially similar in conceptual content and wording to Allport and Ross's original items. There are in fact 11 religious orientation measures in Hill and Hood's (1999) *Measures of Religiosity*, and several other influential scales, such as the Religious Commitment Inventory (Worthington et al., 2003) and the Duke Religious Index (Koenig & Büssing, 2010), all of which are to a large extent reliant on the theoretical meaning and the wording of the Allport and Ross's items.

## Theory and Factor Structure of IR and ER

A great deal of work in the psychology of religion has been devoted to refining Allport's measures to uncover the factor structure of the scales and the ideal wording of items (e.g., Gorsuch & Venable, 1983; Kirkpatrick, 1989), so much so that the psychology of religion has even been criticized as focused on measurement more so than theory (Kirkpatrick & Hood, 1990). Work in the field has primarily tried to uncover whether intrinsic and extrinsic religiosity are negatively correlated or orthogonal; how many subscales of extrinsic religiosity might exist; and whether items like "It doesn't much matter what I believe so long as I am good" are reverse-keyed intrinsic items (Trimble, 1997) or should be omitted (Maltby, 1999; Momtaz, Hamid, Yahaya, & Ibrahim, 2010; Smither & Walker, 2015). There have also been more theoretical questions raised about what the scales are conceptually measuring (Kirkpatrick & Hood, 1990), whether it is appropriate to juxtapose the means and ends of religion (Pargament, 1992) and whether the concepts of and distinctions between intrinsic and extrinsic are culturally universal or better represent an American, Christian, individualistic understanding of proper religious motivation (A. B. Cohen, Hall, Koenig, & Meador, 2005; Flere & Lavric, 2008; Ji & Ibrahim, 2007; Khan, Watson, & Chen, 2015).

It is important to note that theory and measurement need to inform each other. Without a clear theory, we do not know whether we are measuring a construct appropriately. Without appropriate measures, we do not know whether we are measuring the theoretical constructs of interest. Just what is intrinsic religiosity measuring? Is it simply how religiously committed someone is? Is it how much they have internalized their religion? Is it whether they believe in God? Is it whether religion affects other domains of their life? Are behaviors such as reading religious texts central in evaluating intrinsic religiosity? These all might be aspects of the same motivations for religion (requiring one concept of intrinsic religiosity) or they might be different motivations (requiring more than one concept of intrinsic religiosity).

Indeed, it is not clear on the basis of face validity whether we should expect one or several scales and subscales. If someone is intrinsically motivated to be religious in the sense that religion is the master motive of their life, is this to say that they are also intrinsically motivated in the sense that they have a connection with the divine? If someone is intrinsically oriented, is this indeed orthogonal to extrinsic motivations, such as pursuing comfort or relief from religion? And is seeking comfort or relief part of the same (extrinsic) construct as seeking social connections through religion, or should theories be developed which have separate explanations for these two approaches to religion, and posit different effects of them? And, as we have argued elsewhere, affiliating with one's religious group, which might be reflected in extrinsic religiosity items, can also be highly valued among many religious individuals and groups (A. B. Cohen et al., 2005). Only by testing theoretically derived models of the factor structure can we answer questions about which motivations are related.

We seek here to clarify our theoretical understanding of intrinsic and extrinsic religiosity by more thoroughly investigating the measurement properties of scales used to assess these constructs. Unfortunately, this process has not come as far as it could despite decades of work that have already been devoted to such efforts. Often, researchers primarily rely on an exploratory strategy to understand the measurement of intrinsic and extrinsic religiosity. Although an exploratory strategy is recommended initially when developing multifactor scales or after rejecting an existing theory, the factor structure should then be validated using a confirmatory strategy on a new sample.

When using an exploratory strategy such as exploratory factor analysis (EFA) or principal components analysis (PCA), we can identify items that appear to cluster based on the data at hand without *a priori* specifications about the constructs under investigation. Because PCA assumes that the items are measured without error (which is rarely, if ever, true), we focus on EFA, which does not rely on this assumption, throughout the remainder of this article. With EFA, all items are free to load on all factors, such that some items may load highly on a single factor with secondary loadings on the other factors and some items may load about equally on two factors.

When using a confirmatory strategy, however, we can examine different theorized factor structures. With confirmatory factor analysis (CFA), we freely estimate some parameters while constraining others to be a certain value (typically zero) or equal to other parameters. In this way, we can specify which items ought or ought not to cohere with the other items, and therefore, we are better able to identify and interpret the construct being measured. This derives from and, in turn, informs the theory behind a particular set of items. In addition to validating a factor structure derived from EFA, Velicer and Jackson (1990) contended that "when two (or more) legitimate competing theories exist, a confirmatory strategy can be employed to determine which provides a better fit to the data" (p. 21).

Research using a confirmatory, rather than an exploratory, strategy has been sparse when it comes to analyzing intrinsic and extrinsic religiosity. This is an important limitation of the existing literature because factor structures derived through EFA are not always supported by CFA in new samples (van Prooijen & van der Kloot, 2001). Both substantive and methodological explanations exist for these discrepancies. Substantive explanations usually center on differences in the sample characteristics, such as when EFA is performed on a sample of undergraduate students and CFA is performed on a more diverse sample from the community or a sample collected by another group of researchers. Differences in age, socioeconomic status, geography, culture, and a number of other characteristics may explain why the factor structure was not validated in the new sample. However, methodological explanations may be just as compelling. For example, Fabrigar, Wegener, MacCallum, and Strahan (1999) demonstrated that researchers sometimes make "questionable decisions" (p. 272) when performing EFA (e.g., criteria for selecting the optimal number of factors, choice of rotation method) that can greatly impact the derived factor structure (see Fabrigar & Wegener, 2012 for recommendations when performing EFA). Furthermore, CFA imposes more restrictions on the items and thus creates more opportunities for misfit. Typically, low factor loadings from the EFA model (e.g., those less than |.3|) are constrained to zero in the subsequent CFA model. However, because an item's cross-loadings account for some of the item's variance when performing EFA, constraining an item's cross-loadings to zero may lead to the CFA model providing poorer fit to the data than the EFA model. To rule out substantive explanations, van Prooijen and van der Kloot (2001) argued that "the ultimate test of possible methodological explanations for the differences in EFA and CFA results should necessarily involve the *same* samples" (p. 780). Substantively, we would expect EFA and CFA to lead to the same conclusions when applied to the same sample. Thus, only methodological explanations could account for any discrepancies. Later in this article, we demonstrate that EFA and CFA can lead to different conclusions regarding the measurement of intrinsic and extrinsic religiosity.

Finally, the limited use of a confirmatory strategy in existing research has implications for assessing measurement invariance. Psychologists interested in religion have long acknowledged that religious motivations might vary across groups, and theoretical and empirical attention to this issue has increased (Saroglou & Cohen, 2011, 2013). In an attempt to validate measures of intrinsic and extrinsic religiosity across cultures, researchers have performed PCA, EFA, or (more rarely) CFA in non-U.S., non-Christian samples. However, this approach may provide misleading answers about the suitability of the ideas and measures across groups with potentially different cultural norms about what it means to be appropriately religious (A. B. Cohen et al., 2005). Consistently finding support for a certain number of factors does not establish the level of measurement invariance needed to use these measures in a culturally diverse sample (e.g., when estimating regression or path models). Even still, when analyses of intrinsic and extrinsic religiosity are conducted outside the United States (e.g., among Polish Catholics; Brewczynski & MacDonald, 2006) or in non-Christian samples (e.g., among Muslims; Flere & Lavric, 2008; Ghorbani, Watson, Ghramaleki, Morris, & Hood, 2002; Ji & Ibrahim, 2007; Khan et al., 2015), the number of factors, factor loadings, and relevance of some items can be quite different from U.S. samples (cf. Maltby, 1999, who focused on an Irish population).

To our knowledge, two other articles have used a confirmatory strategy to assess intrinsic and extrinsic religiosity, but neither formally investigated measurement invariance.[1] Brewczynski and MacDonald (2006) estimated a series of CFA models using a Polish Catholic sample and concluded that three factors provided the best fit. However, six of the global fit indices fell below conventional thresholds (p. 71), and several modifications were needed to improve fit (p. 72). Flere and Lavric (2008) performed PCA using Bosnian Muslim, Serbian Orthodox, Slovenic Catholic, and U.S. Protestant samples. They then performed CFA to validate their PCA results rather than to test more mainstream theories about the factor structure of intrinsic and extrinsic religiosity or to establish measurement invariance. In this article, we formally test measurement invariance across culturally and religiously diverse samples.

## Overview of Current Goals

Precisely determining the factor structure of the intrinsic and extrinsic religious orientation scale is important for theoretical reasons—without a clear understanding of the measurement properties of the scale, it is not possible to articulate the theoretical construct ostensibly being operationalized (Millsap, 2011). Therefore, we first aimed to produce more definitive answers regarding how to score a very commonly used measure of intrinsic and extrinsic religiosity. We focus on Gorsuch and McPherson's (1989) measure, which has 1,032 Google Scholar citations as of July 2017. This measure is a somewhat updated set of items from Allport and Ross's (1967) highly cited original set of items. Clearer answers about how to score the scales will enable a better theoretical interpretation of the constructs of intrinsic and extrinsic religiosity. This issue needs to be more rigorously investigated because the various scales used to measure these constructs have virtually become synonymous over the last 50 years (since Allport and Ross) with the theoretical definitions of the constructs.

Second, we wished to investigate whether the scales, when scored appropriately, would show measurement invariance across culturally and religiously diverse samples. Using a series of CFA models to establish measurement invariance provides evidence that a set of items is measuring the same latent constructs (say, intrinsic religiosity) with the same structure across different groups. In Study 2, we chose to focus on U.S. Protestants, Turkish Muslims, and Irish Catholics because this combination of groups affords a strong theoretical test of the structure of religious orientation. Of primary importance, we determined that each group should be studied in a country in which it is the majority religion, rather than confounding majority/minority status. This is important because nationality and religion often interact (A. B. Cohen, Gorvine, & Gorvine, 2013; Johnson & Cohen, 2013; Saroglou & Cohen, 2013).

Of course, there are many countries in which either Catholicism or Islam is the dominant religion. We selected Ireland and Turkey because these are both relatively Westernized countries which are as similar as possible in factors such as the role of religion in society. Furthermore, people in the United States, Ireland, and Turkey enjoy relative religious freedom (Warner, Kilinc, Hale, Cohen, & Johnson, 2015). Nevertheless, we acknowledge that these three groups differ in more than just religion (e.g., collectivism, geography, currency, demographics), and deciding which characteristics define a culture and which serve as confounds is one of the most important methodological issues in cross-cultural research (D. Cohen, 2007). To disentangle religion and nationality, we assessed measurement invariance of intrinsic religiosity across U.S. Protestants, Catholics, and Muslims as well as investigated whether intrinsic religiosity, as we are suggesting scoring it, predicts warmth toward atheists and gay men/lesbians (Study 3).

## Study 1: EFA and CFA in a U.S. Protestant Sample

Consistent with most existing research, we first adopted an exploratory strategy to examine the underlying structure of Gorsuch and McPherson's (1989) measure in a sample of U.S. Protestants, arguably the group for which this measure is most appropriate for capturing intrinsic and extrinsic religious motivations (A. B. Cohen et al., 2005). We then performed CFA on the same sample of participants. Although we view EFA and CFA as complementary, CFA is arguably most appropriate in this case because four competing theories (outlined below) exist for how to score this measure. Nevertheless, contrasting the results from EFA and CFA illustrates why existing research on intrinsic and extrinsic religiosity, which has primarily relied on an exploratory strategy, may have led to an incomplete understanding of these constructs.

### Method

*Participants and procedure.* Participants were 310 U.S. Protestants (140 males, 1 missing) who completed an online survey of the Gorsuch and McPherson's "Age-Universal" Religious Orientation Scale (appendix) using a 1 (*strongly disagree*) to 7 (*strongly agree*) scale. The 18-24 years age range response was the median for this sample. The survey was part of a larger study of volunteerism among U.S. Protestants reported elsewhere (Johnson, Cohen, & Okun, 2016).

*Data analytic strategy.* Based on the 14 items listed in appendix, we performed EFA while extracting one to five factors. We specified a geomin rotation which allows the factors to be correlated. To determine the optimal number of factors, we examined the eigenvalues in a scree plot and conducted a parallel analysis (Horn, 1965) using the PARALLEL option in Mplus 7.4. Briefly, parallel analysis generates a random sample with the same number of cases (here, $N = 310$) and items (14) as the original sample, computes eigenvalues from the random sample's correlation matrix, and compares the eigenvalues from the original sample with the eigenvalues from the random sample. By specifying PARALLEL = 100, we generated 100 random samples and averaged across the 100 sets of eigenvalues to yield a single set of eigenvalues. The optimal number of factors is the number of original sample eigenvalues that are greater than the random sample eigenvalues. We assessed the global fit of each solution using the Satorra–Bentler scaled $\chi^2$ test of exact fit, root mean square error of approximation (RMSEA; Steiger, 1989), comparative fit index (CFI; Bentler, 1990), Tucker–Lewis index (TLI; Tucker & Lewis, 1973), and standardized root mean square residual (SRMR; Bollen, 1989). Generally, an RMSEA of .05 or lower, CFI of .95 or higher, TLI of .95 or higher, and SRMR of .06 or lower indicates close fit, though West, Taylor, and Wu (2012) cautioned against the reification of these cutoffs. To investigate local fit, we examined the factor loadings, residuals, and modification indices.

Next, we estimated CFA models that tested four commonly used ways of scoring this measure. Although Gorsuch and McPherson have ER scored as including two separate subscales (ER-Social and ER-Personal), it is also common to create a single ER Scale (e.g., A. B. Cohen & Hill, 2007). We also tested two alternatives for three items such as "It doesn't matter much what I believe so long as I am good" (Items IR6 to IR8 in appendix). There has been historical debate about whether these items are negatively scored IR items or whether they belong with ER. Gorsuch and McPherson, for example, recommended an IR Scale (including IR-reversed items), an Extrinsic-Social Scale, and an Extrinsic-Personal Scale. Allport and Ross instead had the IR-reversed items on their Extrinsic Scale as measures of utilitarian religious orientation. The model descriptors in appendix indicate which items are being assigned to what factors, with how researchers would probably theoretically interpret the items and factors if they were scored that way.

First, we reverse-scored Items IR6 to IR8 and tested a two-factor model, with Items IR1 to IR8 on the first factor and Items ER1 to ER6 on the second factor (Model 1 in appendix). Including the negatively keyed items on the first factor is consistent with Gorsuch and McPherson's recommendation. Furthermore, computing a single ER Scale is a common practice, sometimes because researchers do not have distinct hypotheses for the social and personal components of ER, even if they do for IR and ER in general. Then, we tested a two-factor model with Items IR1 to IR5 on the first factor and Items ER1 to ER6 and Items IR6 to IR8 (three items originally intended to measure a utilitarian component of extrinsic religiosity on the Allport and Ross Scale) on the second factor (Model 2). Conceptually, the two factors are traditionally understood to correspond to intrinsic religiosity (IR) and extrinsic religiosity (ER), respectively.

Next, we tested a three-factor model with Items IR1 to IR8 on the first factor; Items ER1, ER2, and ER6 on the second factor; and Items ER3 to ER5 on the third factor (Model 3). Conceptually, the three factors correspond to IR (including the three reverse-scored "utilitarian" items), a social component of ER (ERS), and a personal component of ER (ERP). Finally, we tested a four-factor model (Model 4) with Items IR1 to IR5 on the first factor; Items ER1, ER2, and ER6 on the second factor; Items ER3 to ER5 on the third factor; and Items IR6 to IR8 on the fourth factor. Conceptually, the four factors correspond to IR, ERS, ERP, and the difficult to appropriately characterize "utilitarian" component of ER (ERU). Henceforth, we refer to these four models as the (a) IR(ERU)-ER, (b) IR-ER(ERU), (c) IR(ERU)-ERS-ERP, and (d) IR-ERS-ERP-ERU models. We assessed the global fit of the four models using the Satorra–Bentler scaled $\chi^2$ test of exact fit, RMSEA, CFI, TLI, and SRMR and the local fit using the factor loadings, residuals, and modification indices. We tested the relative fit of nested models using Satorra–Bentler scaled $\chi^2$ difference tests (Satorra, 2000), which we denote by $\Delta\chi^2$ to avoid confusion with the $\chi^2$ test of exact fit. To examine the relative fit of nonnested models, we used the Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and sample size–adjusted BIC (SABIC; Sclove, 1987). The AIC, BIC, and SABIC consider fit and parsimony, with lower values being favored. All of the EFA and CFA models were estimated using full information maximum likelihood (FIML) estimation. Unlike listwise or pairwise deletion, FIML estimation does not exclude participants with missing scores (eight of the 310 participants had missing scores on one or two items) but instead uses all the available data to estimate parameters. Using the MLR option in Mplus 7.4, we requested the Satorra–Bentler scaled $\chi^2$ test of exact fit and used a sandwich estimator for the standard error computations to adjust for nonnormality (discrete items are nonnormal by definition).

## Results

Table S1 in the online supplement presents the correlations among the items. Figure S1 is a scree plot of the eigenvalues from the original sample's correlation matrix (connected by the solid line) and the random sample eigenvalues from Horn's parallel analysis (connected by the dashed line). The solid line's "elbow" and the number of factors before the solid and dashed lines intersect (i.e., the number of original sample eigenvalues that are greater than the random sample eigenvalues) suggest that three factors should be extracted. Furthermore, the fourth eigenvalue is less than 1, meaning that extracting a fourth factor would explain less than one item's variance (each of the 14 items has a variance of 1, and the sum of the eigenvalues equals the total variance of 14).[2] With three factors, the $\chi^2$ test of exact fit was significant, $\chi^2(52) = 143.305$, $p < .001$, but the global fit indices suggested adequate fit to the data, RMSEA = .075, CFI = .939, SRMR = .033. Table S2 in the online supplement reports the factor loadings, with those greater than |0.500| in bold. Consistent with Gorsuch and McPherson's recommendation, Items IR1 to IR8 comprised the first factor (intrinsic religiosity); Items ER1, ER2, and ER6 comprised the second factor (social component of extrinsic religiosity); and Items ER3 to ER5 comprised the third factor (personal component of extrinsic religiosity). However, Items IR6, IR7, and IR8 had the highest cross-loadings (−0.222, −0.322, and −0.243, respectively, on the third factor). When extracting an additional factor, Items IR6, IR7, and IR8 formed the fourth factor. These cross-loadings become important when moving from EFA to CFA, as we describe next.

Tables S3 to S6 in the online supplement report the standardized factor loadings and factor correlations for the four models tested using CFA. Table 1 summarizes the relative and global fit indices.

*Model 1.* Model 1 had IR and ER as two separate scales, with items such as "Although I am religious . . ." as reverse-scored IR items. The global fit indices suggested poor fit, $\chi^2(76) = 368.007$, $p < .001$, RMSEA = .111, CFI = .806, TLI = .768, SRMR = .100. As shown in Table S3, the standardized factor loadings for Items ER1, ER2, and ER6 were high (0.704-0.840), whereas those for Items ER3 to ER5 were low (0.161-0.211). This pattern of standardized factor loadings is not surprising given the item correlations. The correlations between Items ER1, ER2, and ER6 and Items ER3 to ER5 ranged from .090 to .187, which suggests that these six items do not form one factor.

*Model 2.* Model 2 also had IR and ER as two separate scales but with items such as "Although I am religious . . ." as positively-scored ER items. Items ER1, ER2, and ER6 were again grouped with Items ER3 to ER5 (along with Items IR6 to IR8), which we already described as inappropriate based on the item correlations. In Table S4, we again see low standardized factor loadings for Items ER3 to ER5 (0.166-0.230). The standardized factor loadings for Items IR6 to IR8 were low (0.185-0.244), which is not surprising given that the correlations between Items IR6 to IR8 and Items ER1 to ER6 ranged from .009 to .187. Again, this factor structure seemed

**Table 1.** Relative and Global Fit Indices for Confirmatory Factor Analysis Models.

| | IR(ERU)-ER model | IR-ER(ERU) model | IR(ERU)-ERS-ERP model | IR-ERS-ERP-ERU model |
|---|---|---|---|---|
| $\chi^2$ test of exact fit (*df*) | 368.007 (76) | 598.905 (76) | 261.581 (74) | 122.808 (71) |
| Scaling correction factor | 1.086 | 1.101 | 1.068 | 1.075 |
| AIC | 15,480.527 | 15,740.400 | 15,364.088 | 15,222.747 |
| BIC | 15,641.200 | 15,901.072 | 15,532.233 | 15,402.102 |
| SABIC | 15,504.820 | 15,764.693 | 15,389.510 | 15,249.864 |
| RMSEA (90% CI) | .111 [.100, .123] | .149 [.138, .160] | .090 [.079, .102] | .049 [.034, .063] |
| CFI | .806 | .653 | .875 | .966 |
| TLI | .768 | .584 | .847 | .956 |
| SRMR | .100 | .182 | .081 | .045 |

*Note.* IR = intrinsic religiosity; ERU = "utilitarian" component of ER; ER = extrinsic religiosity; ERS = social component of ER; ERP = personal component of ER; AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample size–adjusted BIC; RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual.

highly implausible based on the item correlations, and the global fit indices suggested poor fit for this model, $\chi^2(76) = 598.905$, $p < .001$, RMSEA = .149, CFI = .653, TLI = .584, SRMR = .182.

*Model 3.* In Model 3, Items IR1 to IR8 formed the IR factor; however, extrinsic religiosity was split into social and personal factors: ERS and ERP. This model fits significantly better than Model 1, $\Delta\chi^2(2) = 68.018$, $p < .001$, which again suggests that Items ER1 to ER6 do not form one factor, as posited by Model 1. The relative and global fit indices reported in Table 1 also indicate that Model 3 more closely fits the data than Model 1. Nonetheless, the global fit indices did not indicate adequate or close fit for Model 3, $\chi^2(74) = 261.581$, RMSEA = .090, CFI = .875, TLI = .847, SRMR = .081. The ERS and ERP factors were positively correlated, $r = .242$, $p = .002$, but none of the other factor correlations were significant, as seen in Table S5. Although the standardized factor loadings were high (0.528-0.864) across all three factors, Items IR6 to IR8 had lower standardized factor loadings (0.528-0.621) than the other items on the IR factor (0.673-0.864).

*Model 4.* Model 4 differed from the previous model by splitting Items IR1 to IR8 into two factors: IR and ERU. Items IR6 to IR8 were also not reverse-scored in this model. The IR and ERU factors were strongly correlated, $r = -.708$, but splitting Items IR1 to IR8 into two factors significantly improved model fit relative to Model 3, $\Delta\chi^2(3) = 162.482$, $p < .001$. The relative and global fit indices reported in Table 1 also indicate that Model 4 more closely fits the data than Model 3. As shown in Table S6, the standardized factor loadings were high (0.593-0.863) across all four factors. As expected, the ERS, ERP, and ERU factors were positively correlated (see Table S6). The $\chi^2$ test of exact fit was significant, $\chi^2(71) = 122.808$, $p < .001$, but the global fit indices suggested that Model 4 adequately fits the data, RMSEA = .049, CFI = .966, TLI = .956, SRMR = .045, and all the normalized residuals were less than 3 in absolute value.

## Discussion

Given the results from Study 1, the question becomes what to do with Items IR6 to IR8. EFA, which was developed to identify major factors underlying a correlation matrix, supported three factors with these items functioning as reverse-keyed IR items. However, this factor structure was clearly rejected by CFA (Model 3), and the adequate fit of Model 4 suggests that these items form a distinct factor. While it was for a longtime *de rigeur* to include negatively keyed items in scales to guard against indiscriminate responding (and one will recall Allport and Ross's labeling of people who scored high on both intrinsic and extrinsic religiosity as "indiscriminately pro-religious" and even as "muddle-headed"), we now know that negatively keyed items are cognitively processed differently from positively keyed items and, therefore, often form method factors that are essentially statistical artifacts (DiStefano & Motl, 2006)—and this problem can be magnified in certain cultures (Lindwall et al., 2012). The double-barreled wording of these items further complicates their role in the various models.

To resolve this issue, we must turn to theory. We believe that the conceptual interpretation of this factor is unclear. We could refer to these three items as extrinsic items, as Allport did—but these items on the basis of face validity do not tap an instrumental or utilitarian use of religion. A utilitarian, or instrumental, use of religion is theoretically meant to indicate the use of religion to achieve some other nonreligious goal, like making business contacts, deriving personal comfort, or achieving social benefits. Items measuring the instrumental use of religion (like the ERS or ERP items) would be better constructed as, "Religion is useful to me to achieve [thus and such a goal]." Rather, from a face validity perspective, the items appear to tap the idea that religion is not very important to the participant. Given the lack of theoretical support, problematic pattern of correlations, and poor fit to the data, we recommend dropping these three items and scoring the remaining items as three subscales: IR, ERS, and ERP. In the current sample, a CFA model with these three factors provided close fit to the data, $\chi^2(41) = 78.865$, $p < .001$, RMSEA = .055, CFI = .966, TLI = .954, SRMR = .045.

## Study 2: Measurement Invariance in a Sample of U.S. Protestants, Irish Catholics, and Turkish Muslims

As a result of Study 1, we think the best way to proceed is with a reduced set of items measuring a three-factor structure based on IR without the reverse-scored items (five items), ERS (three items), and ERP (three items). To avoid model modifications that capitalize on chance, we apply this model to a different sample of U.S. Protestants in Study 2, and we examine measurement invariance by comparing these U.S. Protestants with Irish Catholics and Turkish Muslims.

### *Method*

*Participants and procedure.* Participants were 308 U.S. Protestants (139 males), 329 Irish Catholics (111 males), and 339 Turkish Muslims (169 males). Participants reported being 18 to 24 (12%, 38%, and 43% for U.S. Protestant, Irish Catholic, and Turkish Muslim participants, respectively), 25 to 36 (29%, 5%, and 35%), 37 to 55 (32%, 23%, and 21%), 56 to 70 (22%, 26%, and 1%), or 71 or over (5%, 8%, and 0%), suggesting that the Turkish Muslim participants were generally younger than the U.S. Protestant or Irish Catholic participants.

Participants in the United States were workers on Amazon.com's Mechanical Turk website who completed Gorsuch and McPherson's "Age-Universal" Religious Orientation Scale (appendix) as part of a larger online study of volunteerism. All the items were administered, but we did not analyze the items we recommended dropping. Participants in (Dublin) Ireland and (Istanbul) Turkey were university students and community members who completed a paper and pencil version of the Gorsuch and McPherson's "Age-Universal" Religious Orientation Scale (in English or Turkish, respectively) as part of a larger study of charitable giving in their local university, church hall, or mosque. The results of the larger study are reported elsewhere (Warner et al., 2015). Participants in Ireland and Turkey were paid 25 Euros or 25 Turkish Lira, respectively, for their participation.

*Data analytic strategy.* Using CFA, we tested measurement invariance across U.S. Protestant, Irish Catholic, and Turkish Muslim participants for a three-factor model with Items IR1 to IR5 on the first factor (IR); Items ER1, ER2, and ER6 on the second factor (ERS); and Items ER3 to ER5 on the third factor (ERP). Table S7 in the online supplement presents the correlations among the items for U.S. Protestant, Irish Catholic, and Turkish Muslim participants. We sequentially fit the configural, metric (i.e., weak), scalar (i.e., strong), and strict invariance models. The configural invariance model posits that the same pattern of fixed and freely estimated factor loadings holds across groups. Said differently, the number of factors and the items that load on each factor do not vary across groups. To define the metric, we selected one item on each factor and constrained its factor loading to unity and its intercept to zero. The metric, scalar, and strict invariance models are increasingly more restrictive. The metric invariance model constrains the factor loadings to be equal across groups; the scalar invariance model constrains the factor loadings and intercepts to be equal across groups; and the strict invariance model constrains the factor loadings, intercepts, and unique variances to be equal across groups. As in Study 1, we used FIML estimation to handle missing data (18 U.S. Protestant, 15 Irish Catholic, and 27 Turkish Muslim participants had missing scores on up to four items) along with the MLR option in Mplus 7.4. We assessed the global fit using the Satorra–Bentler scaled $\chi^2$ test of exact fit, RMSEA, CFI, TLI, and SRMR (see Study 1 for details) and the local fit using the factor loadings, residuals, and modification indices. Because the configural, metric, scalar, and strict invariance models are nested, we investigated relative fit via Satorra–Bentler scaled $\chi^2$ difference tests (Satorra, 2000).

Due to violation of metric invariance for the three-factor model (discussed below), we also tested measurement invariance across U.S. Protestant, Irish Catholic, and Turkish Muslim participants for a one-factor model of Items IR1 to IR5 using the same procedure. We focused on these items because we did not recommend changing the scoring of the two factors measuring social and personal components of extrinsic religiosity in Study 1.

### *Results and Discussion*

Table 2 summarizes the global fit indices from the configural, metric, scalar, and strict invariance models comparing U.S. Protestant, Irish Catholic, and Turkish Muslim participants. The global fit indices suggested adequate or poor fit of the configural invariance model, $\chi^2(123) = 452.235$, $p < .001$, RMSEA = .091, CFI = .903, TLI = .870, SRMR = .086, and the metric invariance model significantly deteriorated fit relative to the configural invariance model, $\Delta\chi^2(16) = 127.958$, $p < .001$. Because the global fit indices suggested poor fit of the metric invariance model, $\chi^2(139) = 584.278$, $p < .001$, RMSEA = .099, CFI = .869, TLI = .845, SRMR = .123, we do not consider the scalar and strict invariance models.

To identify the sources of misfit, we examined the group-specific models, as the overall $\chi^2$ test of exact fit equals the sum of the group-specific $\chi^2$ values. For Turkish Muslim participants, the correlation pattern implies that the three-factor structure is implausible, as the correlations between Item IR2

**Table 2.** Global Fit Indices for Measurement Invariance of Three-Factor Model Across U.S. Protestants, Irish Catholics, and Turkish Muslims.

| | Configural | Metric | Scalar | Strict |
|---|---|---|---|---|
| $\chi^2$ test of exact fit (*df*) | 452.235 (123) | 584.278 (139) | 1,127.377 (155) | 1,322.822 (177) |
| Scaling correction factor | 1.071 | 1.078 | 1.069 | 1.145 |
| RMSEA (90% CI) | .091 [.082, .100] | .099 [.091, .108] | .139 [.131, .147] | .141 [.134, .148] |
| CFI | .903 | .869 | .714 | .663 |
| TLI | .870 | .845 | .696 | .686 |
| SRMR | .086 | .123 | .179 | .222 |

*Note.* The metric invariance model significantly deteriorated fit relative to the configural invariance model, $\Delta\chi^2(16) = 127.958$, $p < .001$, and the scalar invariance model significantly deteriorated fit relative to the metric invariance model, $\Delta\chi^2(16) = 588.417$, $p < .001$. RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual.

and the remaining four items on the IR factor were much lower than would be expected (range = .132-.207), as were the correlations between Item ER1 and the remaining two items on the ERS factor (.133 and .119). A three-factor CFA model supported this conclusion, RMSEA = .099, CFI = .849, TLI = .797, SRMR = .105. We have no theoretical explanation for this pattern of correlations and the resulting poor model fit; it may be a feature of the translation, Turkish cultural context, Muslim religion, some combination of these factors, or even other factors. We are not aware of previous literature that would offer any insight. In terms of the misfit of the ER1 item, we can speculate that, in the Turkish culture, one may enjoy praying in the company of friends and associates.

When using CFA to test the proposed three-factor structure for Irish Catholic participants, the $\chi^2$ test of exact fit was significant, $\chi^2(41) = 163.062$, $p < .001$, and the global fit indices suggested adequate or poor fit, RMSEA = .095, CFI = .902, TLI = .869, SRMR = .088. The largest modification index suggested allowing Item ER1 to load on the IR factor. Referring to Table S7, Item ER1 correlated with the IR items (range = .396-.494) about as highly as it correlated with Items ER2 (.491) and ER6 (.403). In our opinion, however, making such a modification would not be justified by any current theory of intrinsic and extrinsic religiosity, as extrinsic and intrinsic items are not intended to correlate with each other. The second largest modification index proposed correlating the uniquenesses for Items ER2 ("I attend religious services mainly because I enjoy seeing people I know there") and ER6 ("I attend religious services mostly to spend time with friends"), which are very similar items because they are both about deriving social benefits from religion.

Finally, when using CFA to test the proposed three-factor structure for U.S. Protestant participants, the $\chi^2$ test of exact fit was significant, $\chi^2(41) = 113.594$, $p < .001$, but the global fit indices suggested adequate fit to the data, RMSEA = .076, CFI = .941, TLI = .921, SRMR = .056, and all the normalized residuals were less than 3 in absolute value. The largest modification index proposed correlating the uniquenesses for Items IR4 ("I try hard to live all my life according to my religious beliefs") and IR5 ("My whole approach to life is based on my religion"), which are very similar items. In fact, inasmuch as intrinsic religiosity is usually theoretically characterized as having religion be the master motive of one's life, these two items seem closest in content to tapping that religious orientation. Although we did not impose post hoc modifications, the modification index for this theoretically justifiable residual correlation would reduce the model's $\chi^2$ value by roughly 31.432, producing a rather substantial improvement in fit.

For completeness, we tested measurement invariance of the one-factor model of Items IR1 to IR5. Overall, the global fit indices suggested adequate fit of the configural invariance model, $\chi^2(15) = 67.096$, $p < .001$, RMSEA = .103, CFI = .966, TLI = .932, SRMR = .027. However, the metric invariance model significantly deteriorated fit relative to the configural invariance model, $\Delta\chi^2(8) = 114.967$, $p < .001$, and the global fit indices suggested poor fit, $\chi^2(23) = 169.037$, $p < .001$, RMSEA = .140, CFI = .905, TLI = .876, SRMR = .165. This misfit is not surprising when constraining the factor loadings to be equal across groups given the low correlations between Item IR2 and the remaining four items among Turkish Muslim participants, as described above.

In sum, configural invariance did not hold across U.S. Protestant, Irish Catholic, and Turkish Muslim participants. The proposed three-factor structure was problematic in the Irish Catholic sample and highly implausible in the Turkish Muslim sample but provided adequate fit in the U.S. Protestant sample. As such, we do not think that Studies 1 and 2 support the use of IR (without Items IR6 to IR8), ERS, and ERP except for U.S. Protestants. The results from Study 2 also do not support the use of the one-factor model of Items IR1 to IR5 across the three groups.

## Study 3: Measurement Invariance in a Sample of U.S. Protestants, Catholics, and Muslims

As described in Study 2, measurement invariance did not hold across U.S. Protestants, Irish Catholics, and Turkish

Muslims for the one-factor model of Items IR1 to IR5. However, we cannot say whether these results occurred due to religious differences or differences on other characteristics such as nationality or culture. Thus, the purpose of Study 3 was to assess measurement invariance across U.S. Protestants, Catholics, and Muslims.

To ensure that our revised, five-item intrinsic religiosity scale still predicts what intrinsic religiosity is supposed to predict, in this study, we also investigated criterion validity of the five-item version of intrinsic religiosity based on feelings of warmth toward atheists and gay men/lesbians. In addition to many articles on the factor structure of IR and ER that have proliferated in the decades since Allport and Ross (1967), loads of articles have been published on what kinds of prejudice are related to religiosity (Hall, Matz, & Wood, 2010). This literature is quite complex, with these relations depending on whether the prejudices are seen as proscribed by the religion (e.g., ethnic prejudice) or as being tolerated or even encouraged by the religion, such as prejudice against atheists (Gervais & Norenzayan, 2013) or against gay men and lesbians. Whitley (2009) showed in a meta-analysis that intrinsic religiosity is indeed related to more negative attitudes toward gay men and lesbians, with a Cohen's $d$ of 0.482, which equates to an $r$ of .234.

In the data set for this study, warmth toward a variety of outgroups was assessed: Catholics, Protestants, Jews, Muslims, Mormons, scientists, atheists, and gay men/lesbians. We focused on predicting warmth toward atheists and gay men/lesbians, and not the other groups, because prejudices against atheists and gay men/lesbians are the most tolerated by the religious groups represented in our study.

## Method

*Participants and procedure.* Participants ($N$ = 616) were selected from a larger study investigating whether and how social identity threat affects individuals from diverse religious backgrounds (Pasek & Cook, under review). Data were collected in partnership with Qualtrics and panel recruiting companies with the express aim of recruiting a geographically and demographically diverse nonprobability national sample of 1,000 participants—equal numbers of Protestants, Catholics, Jews, and Muslims—residing in the United States. Eligible adults were paid approximately US$2 to complete the online survey. As reported in Pasek and Cook (under review), complete data from eligible participants were collected from 975 participants (note that Pasek and Cook further removed five participants due to missing data not relevant for our analyses). In the analyses presented here, we excluded Jewish participants ($n$ = 247) because Studies 1 and 2 did not include Jewish participants. Furthermore, Judaism is rather unique when considering the role of faith versus practice and the overlap between ethnicity and religious identity (A. B. Cohen et al., 2013; A. B. Cohen et al., 2005; A. B. Cohen & Hill, 2007; A. B. Cohen, Siegel, & Rozin, 2003; Silverman, Johnson, & Cohen, 2016). Also, because Study 3 aimed to disentangle religion and nationality, we excluded participants born outside of the United States ($n$ = 112). Of the 616 participants in the final sample, 238 were Protestant (100 males), 229 were Catholic (112 males), and 149 were Muslim (65 males, 1 other, 1 missing). The Muslim participants were generally younger (median = 28, range = 18-74) than the Protestant (median = 60, range = 19-88) and Catholic (median = 55, range = 19-85) participants.

Participants completed Items IR1 to IR8 of Gorsuch and McPherson's (1989) "Age-Universal" Religious Orientation Scale (appendix) using a 1 (*not at all true*) to 5 (*very true*) scale. Participants also responded to the items "How warmly do you feel toward atheists?" and "How warmly do you feel toward gay men and lesbians?" on a 0 to 10 feeling thermometer, with higher scores indicating more warmth.

*Data analytic strategy.* Using CFA, we tested measurement invariance across U.S. Protestant, Catholic, and Muslim participants for a one-factor model of Items IR1 to IR5. We sequentially fit the configural, metric, scalar, and strict invariance models (see Study 2 for details). To define the metric, we constrained one item's factor loading to unity and its intercept to zero. As in Study 1, we used FIML estimation to handle missing data (two Protestant, three Catholic, and five Muslim participants had missing scores on one or two items) along with the MLR option in Mplus 7.4. We assessed the global fit using the Satorra–Bentler scaled $\chi^2$ test of exact fit, RMSEA, CFI, TLI, and SRMR (see Study 1 for details) and the local fit using the factor loadings, residuals, and modification indices. Because the configural, metric, scalar, and strict invariance models are nested, we investigated relative fit via Satorra–Bentler scaled $\chi^2$ difference tests (Satorra, 2000).

We assessed criterion validity based on whether the five-item version of intrinsic religiosity predicted feelings of warmth toward atheists and gay men/lesbians. Separately for U.S. Protestant, Catholic, and Muslim participants, we regressed feelings of warmth toward atheists on the mean of Items IR1 to IR5. We then regressed feelings of warmth toward gay men/lesbians on the mean of Items IR1 to IR5, with and without excluding the nine Protestant, 13 Catholic, and 11 Muslim participants who reported a sexual orientation other than heterosexual. Because excluding these participants did not appreciably change the results, we report the results based on the full sample.

## Results and Discussion

Table S8 in the online supplement presents the correlations among Items IR1 to IR8 for U.S. Protestant, Catholic, and Muslim participants. Items IR6 to IR8 did not correlate highly with the remaining five items (e.g., range = .056-.184 for U.S. Muslim participants) and are thus not considered here. Table 3 summarizes the global fit indices for the series of CFA models

**Table 3.** Global Fit Indices for Measurement Invariance of One-Factor Model Across U.S. Protestants, Catholics, and Muslims.

| | Configural | Metric | Scalar | Partial scalar |
|---|---|---|---|---|
| $\chi^2$ test of exact fit (*df*) | 75.201 (15) | 79.708 (23) | 172.504 (31) | 96.163 (29) |
| Scaling correction factor | 1.096 | 1.072 | 1.039 | 1.050 |
| RMSEA (90% CI) | .140 [.109, .172] | .110 [.084, .136] | .149 [.128, .171] | .106 [.083, .130] |
| CFI | .954 | .956 | .891 | .948 |
| TLI | .907 | .943 | .895 | .947 |
| SRMR | .029 | .037 | .093 | .043 |

*Note.* The metric invariance model did not significantly deteriorate fit relative to the configural invariance model, $\Delta\chi^2(8) = 2.967$, $p = .936$. The scalar invariance model significantly deteriorated fit relative to the metric invariance model, $\Delta\chi^2(8) = 99.381$, $p < .001$. The partial scalar invariance model, which allowed Item IR1's intercept to vary across the three groups, also significantly deteriorated fit relative to the metric invariance model, $\Delta\chi^2(6) = 16.080$, $p = .013$. RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean square residual.

estimated to assess measurement invariance of the five-item version of intrinsic religiosity. Other than the RMSEA, the global fit indices suggested adequate fit of the configural invariance model, $\chi^2(15) = 75.201$, $p < .001$, RMSEA = .140, CFI = .954, TLI = .907, SRMR = .029, and the metric invariance model did not significantly deteriorate fit relative to the configural invariance model, $\Delta\chi^2(8) = 2.967$, $p = .936$. The global fit indices also suggested adequate fit of the metric invariance model, $\chi^2(23) = 79.708$, $p < .001$, RMSEA = .110, CFI = .956, TLI = .943, SRMR = .037. Finally, we examined scalar invariance (i.e., the systematic tendency to endorse an item at a higher or lower level apart from one's standing on the construct) by imposing equality constraints on the intercepts. These constraints produced a dramatic increase in the $\chi^2$ value, $\Delta\chi^2(8) = 99.381$, $p < .001$, and the global fit indices suggested poor fit, RMSEA = .149, CFI = .891, TLI = .895, SRMR = .093. Much of this reduction in fit owes to intercept differences on Item IR1 ("I enjoy reading about my religion"), the magnitudes of which were commensurate with large standardized mean differences by A. B. Cohen's (1988) standards. We tested partial scalar invariance by allowing this intercept to vary across the three groups. Doing so improved fit, RMSEA = .106, CFI = .948, TLI = .947, SRMR = .043, though the partial scalar invariance model still significantly deteriorated fit relative to the metric invariance model, $\Delta\chi^2(6) = 16.080$, $p = .013$. Intercept differences of this magnitude probably preclude the use of this scale to evaluate cross-cultural mean differences.

When assessing criterion validity, greater intrinsic religiosity predicted significantly less warmth toward atheists, $b = -0.952$, $z = -5.926$, $p < .001$, and gay men/lesbians, $b = -0.989$, $z = -6.346$, $p < .001$, for U.S. Protestant participants. For U.S. Catholic participants, greater intrinsic religiosity predicted marginally less warmth toward atheists, $b = -0.313$, $z = -1.626$, $p = .104$, but not gay men/lesbians, $b = 0.137$, $z = 0.675$, $p = .500$. For U.S. Muslim participants, greater intrinsic religiosity predicted significantly less warmth toward atheists, $b = -1.265$, $z = -4.666$, $p < .001$, and gay men/lesbians, $b = -1.513$, $z = -5.924$, $p < .001$. The correlations between intrinsic religiosity and feelings of warmth toward gay men/lesbians for U.S. Protestant and Muslim participants ($r = -.353$ and $-.394$, respectively) were somewhat greater in magnitude than the correlation reported in Whitley's meta-analysis ($r = -.234$), though Whitley dichotomized intrinsic religiosity and relied on articles that did not score intrinsic religiosity as we are proposing here. Table S9 in the online supplement reports the correlations between each item (including Items IR6 to IR8) and prejudice toward atheists and gay men/lesbians.

## General Discussion

The many personality psychologists, health psychologists, and psychologists of religion working in the wake of Allport have written hundreds of articles about religious identity and motivation. Our first goal in this research was to examine the factor structure of intrinsic and extrinsic religiosity among U.S. Protestants and then to see whether any widely used and theoretically defensible factor structure would also be valid in other cultural groups. We hoped to be able to validate an appropriate way of scoring these scales, with which the field could more precisely assess the relations of religiosity to constructs of interest to social and personality psychologists, such as the Big Five or values (Saroglou, Delpierre, & Dernelle, 2004; Saroglou & Muñoz-García, 2008), subjective well-being (A. B. Cohen, 2002), or prejudice (Hunsberger & Jackson, 2005; Rowatt, Johnson Shen, LaBouff, & Gonzalez, 2013).

In Study 1, we estimated a series of CFA models comparing four competing theories for how to score Gorsuch and McPherson's commonly used measure of intrinsic and extrinsic religiosity. Our results suggested that a four-factor model was most appropriate, with three reverse-keyed items usually included as IR items (but sometimes identified as problematic) forming their own factor. We ultimately recommended dropping these items and scoring the remaining items as three subscales: IR (with five positively keyed intrinsic religious orientation items), ERS (with three positively keyed extrinsic religious orientation items assessing the importance of the social aspects of religion), and ERP (with three positively keyed extrinsic religious orientation items assessing the importance of religion in coping). Our shortened, five-item measure of IR still predicted less warmth toward atheists and gay men/lesbians, in a pattern that was fairly consistent across U.S. Protestants, Catholics, and Muslims, in Study 3.

Some researchers might still prefer to retain the three problematic, negatively keyed items, as many previous articles have done so. The correlations of these three items with our prejudice items (Table S9) did not seem wildly different than those of the other items, and this might seem to bolster the viewpoint that these items should be retained. We would caution though that retaining these items has both statistical and theoretical drawbacks. From a statistical perspective, a factor defined by all eight items was clearly rejected by CFA, and the inclusion of the three problematic items may distort relations between intrinsic religiosity (as defined by the five positively keyed items) and other variables. From a theoretical perspective, it is unclear what these three items are measuring, partly because they are double-barreled and assume that the respondent is religious.

Using the recommended revised factor structure (IR, ERS, and ERP) in Study 2, we investigated whether measurement invariance held across three groups: U.S. Protestants, Irish Catholics, and Turkish Muslims. Given how often Gorsuch and McPherson's measure is used in cross-cultural research, we were hoping to find evidence that these items were measuring the same latent constructs with the same structure across these three groups. Of course, even if we did find strong evidence for invariance across these three groups, there are still a lot of countries and religions that we did not investigate and so this would not provide a test of the universality of the measure. However, this would have been a first step toward rigorously establishing that these scales are useful across groups. We selected U.S. Protestants, Irish Catholics, and Turkish Muslims so that invariance across different pairs of groups would be informative about when the scales would or would not be appropriate. For example, if the scales were invariant across U.S. Protestants and Irish Catholics but not Turkish Muslims, it would point to the scales either being appropriate in English or for Christians but not in the Turkish language or for Muslims. If the scales were invariant across Turkish Muslims and Irish Catholics but not U.S. Protestants, then it might hint that the scales are reliable in Europe or perhaps mainly in collectivistic religions. Invariance across all three groups might hint at the generalizability of the theoretical structure and measurement of religious motivations, which is to some extent what the psychology of religion field seems to currently assume, and suggest that the invariance of the IR/ER measure should next be evaluated in other non-Western religious contexts. However, Study 2 did not support the invariance of the scales across any pair. We did, however, find in Study 3 that the five-item IR Scale achieved metric invariance across U.S. Protestants, Catholics, and Muslims, meaning this scale seems to be appropriate for use in correlations and regressions though not mean comparisons (Steenkamp & Baumgartner, 1998).

There are any number of cultural factors that could shape the structure of religious motivations. One that has been given some theoretical attention is ways that religious groups can be more individualistic or collectivistic in their motivations (A. B. Cohen et al., 2005), but there are doubtless many others. We hope researchers will be mindful of both measurement limitations and theoretical work showing that how we conceptualize and measure religiousness ought to reflect group norms and religious theology about what it means to be a religious person.

## Appendix

Gorsuch and McPherson's (1989) "Age-Universal" Religious Orientation Scale.

| Item | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| IR1. I enjoy reading about my religion. | IR | IR | IR | IR |
| IR2. It is important to me to spend time in private thought and prayer. | IR | IR | IR | IR |
| IR3. I have often had a strong sense of God's presence. | IR | IR | IR | IR |
| IR4. I try hard to live all my life according to my religious beliefs. | IR | IR | IR | IR |
| IR5. My whole approach to life is based on my religion. | IR | IR | IR | IR |
| IR6. Although I believe in my religion, many other things are more important in life. | IR | ER | IR | ERU |
| IR7. Although I am religious, I don't let it affect my daily life. | IR | ER | IR | ERU |
| IR8. It doesn't matter much what I believe so long as I am good. | IR | ER | IR | ERU |
| ER1. I attend religious services because it helps me to make friends. | ER | ER | ERS | ERS |
| ER2. I attend religious services mainly because I enjoy seeing people I know there. | ER | ER | ERS | ERS |
| ER3. I pray mainly to gain relief or protection. | ER | ER | ERP | ERP |
| ER4. What religion offers me most is comfort in times of trouble or sorrow. | ER | ER | ERP | ERP |
| ER5. Prayer is for peace and happiness. | ER | ER | ERP | ERP |
| ER6. I attend religious services mostly to spend time with friends. | ER | ER | ERS | ERS |

*Note.* The "IR" and "ER" labels in the Item column reflect where these items would be scored if there were one IR and one ER Scale, which is one commonly used way of scoring the scales. The model numbers are from Study 1 only. IR = intrinsic religiosity; ER = extrinsic religiosity; ERU = "utilitarian" component of ER; ERS = social component of ER; ERP = personal component of ER.

## Authors' Note

A.B.C and G.L.M. contributed equally to this work.

## Acknowledgments

This article is dedicated to the memory of our colleague, Prof. Roger Millsap. The authors are also very grateful for the assistance and cooperation of the religious officials who allowed access to their religious communities.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors thank the University of Notre Dame Science of Generosity Program and the Institute for Social Science Research at Arizona State University (ASU) for their financial support of this research and the ASU Center for the Study of Religion and Conflict for logistical support.

## Supplemental Material

Supplementary material is available online with this article.

## Notes

1. In their abstract, Gorsuch and McPherson (1989) stated that "Factor analyses of traditional and age-universal measures of intrinsic and extrinsic religion have identified two subcategories of extrinsic religiousness, suggesting the original scales need revision. In this study, confirmatory multiple group factor analysis confirmed this suspicion" (p. 348). However, later the authors clarified, "multiple group factors were extracted using the communalities from an exploratory factor analysis" (p. 349).
2. Consistent with Fabrigar, Wegener, MacCallum, and Strahan (1999), we do not recommend using the Kaiser criterion for extracting factors based on the number of eigenvalues greater than 1. We report the fourth eigenvalue here because it indicates that extracting a fourth factor would explain very little of the total variance.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

Allport, G. W., & Ross, J. M. (1967). Personal religious orientation and prejudice. *Journal of Personality and Social Psychology*, *5*, 432-443.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley.

Brewczynski, J., & MacDonald, D. A. (2006). Confirmatory factor analysis of the Allport and Ross Religious Orientation Scale with a Polish sample. *The International Journal for the Psychology of Religion*, *16*, 63-76.

Cohen, A. B. (2002). The importance of spirituality in well-being for Jews and Christians. *Journal of Happiness Studies*, *3*, 287-310.

Cohen, A. B., Gorvine, B. J., & Gorvine, H. (2013). The religion, spirituality, and psychology of Jews. In K. I. Pargament (Editor-in-Chief), J. J. Exline & J. W. Jones (Eds.), *APA Handbook of Psychology, Religion, and Spirituality: Vol. 1. Context, theory, and research* (pp. 665-679). Washington, DC: American Psychological Association.

Cohen, A. B., Hall, D. E., Koenig, H. G., & Meador, K. G. (2005). Social versus individual motivation: Implications for normative definitions of religious orientation. *Personality and Social Psychology Review*, *9*, 48-61.

Cohen, A. B., & Hill, P. C. (2007). Religion as culture: Religious individualism and collectivism among American Catholics, Jews, and Protestants. *Journal of Personality*, *75*, 709-742.

Cohen, A. B., Siegel, J. I., & Rozin, P. (2003). Faith versus practice: Different bases for religiosity judgments by Jews and Protestants. *European Journal of Social Psychology*, *33*, 287-295.

Cohen, D. (2007). Methods in cultural psychology. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 196-236). New York, NY: Guilford Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, *13*, 440-464.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York, NY: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Flere, S., & Lavric, M. (2008). Is intrinsic religious orientation a culturally specific American Protestant concept? The fusion of intrinsic and extrinsic religious orientation among non-Protestants. *European Journal of Social Psychology*, *38*, 521-530.

Gervais, W. M., & Norenzayan, A. (2013). Religion and the origins of anti-atheist prejudice. In S. Clarke, R. Powell, & J. Savulescu (Eds.), *Religion, intolerance, and conflict: A scientific and conceptual investigation* (pp. 126-145). Oxford, UK: Oxford University Press.

Ghorbani, N., Watson, P. J., Ghramaleki, A. F., Morris, R. J., & Hood, R. W., Jr. (2002). Muslim-Christian Religious Orientation Scales: Distinctions, correlations, and cross-cultural analysis in Iran and the United States. *The International Journal for the Psychology of Religion*, *12*, 69-91.

Gorsuch, R. L., & McPherson, S. E. (1989). Intrinsic/extrinsic measurement: I/E-Revised and single-item scales. *Journal for the Scientific Study of Religion*, *28*, 348-354.

Gorsuch, R. L., & Venable, G. D. (1983). Development of an "age-universal" I-E scale. *Journal for the Scientific Study of Religion*, *22*, 181-187.

Hall, D. L., Matz, D. C., & Wood, W. (2010). Why don't we practice what we preach? A meta-analytic review of religious racism. *Personality and Social Psychology Review*, *14*, 126-139.

Hill, P. C., & Hood, R. W., Jr. (Eds.). (1999). *Measures of religiosity*. Birmingham, AL: Religious Education Press.

Hill, P. C., & Pargament, K. I. (2003). Advances in the conceptualization and measurement of religion and spirituality: Implications for physical and mental health. *American Psychologist*, *58*, 64-74.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Hunsberger, B., & Jackson, L. M. (2005). Religion, meaning, and prejudice. *Journal of Social Issues*, *61*, 807-826.

Ji, C.-H. C., & Ibrahim, Y. (2007). Islamic doctrinal orthodoxy and religious orientations: Scale development and validation. *The International Journal for the Psychology of Religion*, *17*, 189-208.

Johnson, K. A., & Cohen, A. B. (2013). The relation between religious and national cultures. In V. Saroglou (Ed.), *Religion, personality, and social behavior* (pp. 338-360). New York, NY: Psychology Press.

Johnson, K. A., Cohen, A. B., & Okun, M. A. (2016). God is watching . . . but also watching over you: The influence of benevolent God representations on secular volunteerism among Christians. *Psychology of Religion and Spirituality*, *8*, 363-374.

Khan, Z. H., Watson, P. J., & Chen, Z. (2015). Meanings of animal sacrifice during Eid-ul-Adha: Relationships with religious orientations and Muslim experiential religiousness in Pakistan. *Archive for the Psychology of Religion*, *37*, 37-53.

Kirkpatrick, L. A. (1989). A psychometric analysis of the Allport-Ross and Feagin measures of intrinsic-extrinsic religious orientation. In M. L. Lynn & D. O. Moberg (Eds.), *Research in the social scientific study of religion* (Vol. 1, pp. 1-31). Greenwich, CT: JAI Press.

Kirkpatrick, L. A., & Hood, R. W., Jr. (1990). Intrinsic-extrinsic religious orientation: The boon or bane of contemporary psychology of religion? *Journal for the Scientific Study of Religion*, *29*, 442-462.

Koenig, H. G., & Büssing, A. (2010). The Duke Religion Index (DUREL): A five-item measure for use in epidemological studies. *Religions*, *1*, 78-85.

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, *94*, 196-204.

Maltby, J. (1999). The internal structure of a derived, revised, and amended measure of the Religious Orientation Scale: The "Age Universal" I-E Scale-12. *Social Behavior and Personality*, *27*, 407-412.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. London, England: Routledge.

Momtaz, Y. A., Hamid, T. A., Yahaya, N., & Ibrahim, R. (2010). Religiosity among older Muslim Malaysians: Gender perspective. *Journal of Muslim Mental Health*, *5*, 210-220.

Pargament, K. I. (1992). Of means and ends: Religion and the search for significance. *The International Journal for the Psychology of Religion*, *2*, 201-229.

Pasek, M. H., & Cook, J. E. (under review). Religion from the target's perspective: A portrait of religious threat and its consequences in the United States.

Rowatt, W. C., Johnson Shen, M., LaBouff, J. P., & Gonzalez, A. (2013). Religious fundamentalism, right-wing authoritarianism, and prejudice: Insights from meta-analyses, implicit social cognition, and social neuroscience. In R. F. Paloutzian & C. L. Park (Eds.), *Handbook of the psychology of religion and spirituality* (2nd ed., pp. 457-475). New York, NY: Guilford Press.

Saroglou, V., & Cohen, A. B. (2011). Psychology of culture and religion: Introduction to the JCCP special issue. *Journal of Cross-Cultural Psychology*, *42*, 1309-1319.

Saroglou, V., & Cohen, A. B. (2013). Cultural and cross-cultural psychology of religion. In R. F. Paloutzian & C. F. Park (Eds.), *Handbook of the psychology and religion and spirituality* (2nd ed., pp. 330-354). New York, NY: Guilford Press.

Saroglou, V., Delpierre, V., & Dernelle, R. (2004). Values and religiosity: A meta-analysis of studies using Schwartz's model. *Personality and Individual Differences*, *37*, 721-734.

Saroglou, V., & Muñoz-García, A. (2008). Individual differences in religion and spirituality: An issue of personality traits and/or values. *Journal for the Scientific Study of Religion*, *47*, 83-101.

Satorra, A. (2000). Scaled and adjusted restricted tests in multisample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis* (pp. 233-247). London, England: Kluwer Academic.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333-343.

Silverman, G., Johnson, K. A., & Cohen, A. B. (2016). To believe or not to believe, that is not the question: The complexity of Jewish beliefs about God. *Psychology of Religion and Spirituality*, *8*, 119-130.

Smither, J. W., & Walker, A. G. (2015). The relationship between core self-evaluations, views of God, and intrinsic/extrinsic religious motivation. *Psychological Reports: Sociocultural Issues in Psychology*, *116*, 647-662.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78-90.

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.

Trimble, D. E. (1997). The Religious Orientation Scale: Review and meta-analysis of social desirability effects. *Educational and Psychological Measurement*, *57*, 970-986.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10.

van Prooijen, J., & van der Kloot, W. A. (2001). Confirmatory analysis of exploratively obtained factor structures. *Educational and Psychological Measurement*, *61*, 777-792.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*, 1-28.

Warner, C. M., Kilinc, R., Hale, C. W., Cohen, A. B., & Johnson, K. A. (2015). Religion and public goods provision: Experimental and interview evidence from Catholicism and Islam in Europe. *Comparative Politics*, *47*, 189-209.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). New York, NY: Guilford Press.

Whitley, B. E. (2009). Religiosity and attitudes toward lesbians and gay men: A meta-analysis. *The International Journal for the Psychology of Religion*, *19*, 21-38.

Worthington, E. L., Jr., Wade, N. G., Hight, T. L., Ripley, J. S., McCullough, M. E., Berry, J. W., . . . Bursley, K. H. (2003). The Religious Commitment Inventory-10: Development, refinement, and validation of a brief scale for research and counseling. *Journal of Counseling Psychology*, *50*, 84-96.