Does Religiosity Predict Well-Being Across 24 Countries?
Many Analysts Religion Project: Stage 2
(Team 008)

Theiss Bendixen[a,*], Anne Lundahl Mauritsen[a], Uffe Schjoedt[a], Benjamin Grant Purzycki[a]

[a] *Department of the Study of Religion, Aarhus University, DK*

Introduction

This technical report presents analyses from our pre-registered plan for the Many Analysts Religion Project (MARP; for further details of the project, see http://bit.ly/MARPinfo). The document is structured as follows. Section 1 highlights the general summary of our results and follows the template provided by the MARP organizers. Section 2 spells out our focal analyses in more detail while Section 3 spells out related secondary analyses (see below for discussion). In the Appendix, we report further results from both pre-registered predictions (Appendix A), and examine a novel question about social desirability (Appendix B). In our pre-registration, we simulated data to build and check our statistical models as well as spelled out the general analysis plan and data transformations prior to seeing the real data. The pre-registration manuscript and code (Stage 1) can be found here: https://osf.io/kujg3/.

The overall research questions this project asks are: Q1) Do religious people report higher well-being? Q2) Does the relation between religiosity and self-reported well-being depend on perceived cultural norms of religion? Q2 is motivated by the "person-culture fit" hypothesis, the idea that it is not religiosity in itself that is associated with well-being but rather whether an individual's own level of religiosity fits that of the local culture. A highly religious person will express greater well-being in a highly religious culture (or in a culture where religion is considered important for the average person), whereas a non-religious person will express higher well-being in a non-religious culture, all else being equal. A mismatch between an individual's and their culture's level of religiosity will likewise be associated with lower well-being, again all else being equal (for details, see MARP document "Theoretical Background"). In the main models, we ultimately find support for Q1 but fail to find support for the "person-culture fit" hypothesis (Q2) (but see Appendix A).

1. General Summary

1.1. Analytic approach

We conducted a series of Bayesian multilevel regressions modeling the outcome likelihood (*Well-Being*) as Gaussian distributed. We fit models of varying complexity and employed approximate leave-one-out cross-validation (Vehtari et al., 2017) to compare the models' out-of-sample predictive performance.

Q2 calls for the assessment of the possible existence of an interaction effect between *Religiosity* and perceived importance of religiosity (*Cultural Norms*), so that the association between well-being and religiosity is moderated by perceived importance of religiosity in one's country. However,

[*] Corresponding author (Team 008): tb@cas.au.dk

because we suspected that the two focal predictors – religiosity and cultural norms – possibly were partially overlapping measures of the same underlying construct, we pre-registered two analytic scenarios.

If the two focal predictors were found to be sufficiently distinct, Scenario 1 would model them as separate predictors and their interaction. If the two focal predictors were found to be nearly indistinguishable, Scenario 2 would collapse the items and instead address Q2 more indirectly, by modeling the association between well-being and the deviance from country-mean religiosity, where country-mean religiosity serves as a proxy for the importance of religion for an "average person" (cf., MARP document "Data Documentation") in each country. For our main analysis, we carried out Scenario 2. In the Appendix, we report results of Scenario 1 (Models 1-6).

1.2. Operationalization of independent variables

While our primary predictor variables across models include questions about *Religiosity*, we make a few critical distinctions across classes of our models. For Scenario 2 (Models 7-10), we use *Religious Fit*, **RF** as our primary predictor. We calculate **RF**–a ratio of individuals' religious fit between their group and the global sample–as follows:

$$\mathbf{RF} = \frac{\sum \text{religiosity}_i - \mu_j}{\max_g(\sum \text{religiosity}_i - \mu_j)}$$

RF, then, is calculated by summing each individual's religiosity score, $\sum \text{religiosity}_i$, then country-mean ($\mu_j$) centering that value. We divide this centered value by the global maximum deviance score. That is, we divide the country-mean centered religiosity scores by the greatest country-mean centered religiosity score in the entire data set, hence $\max_g$.

Out of the entire global data set, the maximum an individual deviated from their country mean was ∼6.34. If, for example, an individual deviated from their own sample by -3 (i.e., they were 3 points below their country subsample mean), their **RF** score would be -0.47. If, however, an individual was 6 points *above* their group mean, their **RF** would be 0.95. In this way, the higher the value of **RF**, the more one's deviance from their group corresponds to the most *anyone* in the data could deviate from their respective group.

Importantly, while this transformation has the advantage of turning the predictor's maximum value into 1, it would *not* necessarily do the same for negative scores. In other words, it does not necessarily scale between -1 and 1. In our case, the lowest group-mean centered score for an individual was -6.27. This conveniently rendered the lowest **RF** score as -0.99. Someone could have, however, reported very low religiosity in a context of very high religiosity and therefore had an $\mathbf{RF} < -1$. While an **RF** value of 1 is useful for our purposes, it is neither ideal nor necessarily generalizable.

Across models 11-14 (Section 3), we use the same procedure, but instead take the *absolute* value of **RF**. This allows us to examine the role of general religious fit so that being "too" religious or not religious enough by the same amount in a context is treated as similarly deviant. Transformations of *Religiosity* data for Models 1-6 are found in the Appendix.

1.3. Operationalization of dependent variables

As our focal outcome variable for both Q1 and Q2, we focused on the psychological (`wb_psych_1` – `wb_psych_6`) and social aspects (`wb_soc_1` and `wb_soc_2`) of well-being as well as the general satisfaction with life and health (`wb_gen_1` and `wb_gen_2`). We summed the scores of these 10 items, each of which were self-reported on scales from 1 to 5. This gives individual well-being scores ranging between 10 (min.) and 50 (max.).

1.4. Covariates

Demographic variables: Our demographic variables include age, education, gender, and perceived socioeconomic status, since these are the individual-level variables where, in our view, some theoretical justification exists for inclusion. We country-mean center all demographic variables, except gender, where male is the reference. Intercepts then represent estimates for country-average males.

Attention check: Response options of the attention check item ranged from 1-7, with 1 being the "right" answer (indicating attention check passed). For ease of interpretation, we transform this into a binary variable, where 0 = passed and 1 = failed. This ensures that intercepts represent participants who passed the attention check.

1.5. Effect size units

We report regression coefficients (posterior means) for the intercepts and our focal predictor (*Religious Fit*), including 95% credible intervals, from the full main model (Model 7).

1.6. Effect sizes for Q1 and Q2

The global intercept was estimated at 35.49 [34.62, 36.36] and represents the estimated *Well-Being* score (on the raw 10-50 scale; see above) for a male that passed the attention check and is country-average in religiosity, age, perceived socioeconomic status and education. Across countries, intercepts ranged from 28.57 [28.01, 29.13] to 38.93 [38.44, 39.45] (see Figure 2).

The global effect size for *Religious Fit* is 2.70 [2.08, 3.25], which is interpreted as the increase (on the raw 10-50 scale; see above) in well-being relative to the intercept for a maximally religious male with country-average demographic characteristics who passed the attention check. Across countries, intercepts ranged from 1.55 [0.13, 2.79] to 4.21 [2.49, 6.35] (see Figures 2 and 3).

These estimates constitute effect sizes for both Q1 and Q2, since they simultaneously get at the association between *Well-Being* and *Religiosity* (Q1) as well as the association with *Well-Being* and deviance from *Religiosity* (proxy for Q2).

1.7. P-value or Bayes Factor for Q1 and Q2

Not relevant. Our analytic approach neither relied upon nor applied these metrics.

1.8. Additional analyses

We re-ran a slightly modified set of analyses to assess whether the relationship between *Well-Being* and deviance from country-mean *Religiosity* was non-linear. Here, we took the *absolute* deviance of *Religious Fit* (e.g., a *Religious Fit* value of -0.5 is simply converted into an absolute deviance of 0.5) as the focal predictor. These analyses indicated that absolute deviance from country-mean *Religiosity* was also robustly associated with increased *Well-Being* across countries (see Figure 4), meaning that deviating in both directions (positively as well as negatively) from country-mean religiosity is both associated with increased well-being. This contradicts the "person-culture fit" hypothesis (which predicts highest well-being at country-mean level religiosity), to the extent that country-mean religiosity in these sub-samples is a valid proxy for the actual cultural importance of religion for the average person in each country. The global and cross-country intercepts were qualitatively similar to the main model (Model 7) reported above (see Figure 4). The global effect size for *Religious Deviance* in the full model (Model 11) was estimated at 2.13 [1.42, 2.80], which represents the estimated increase (on the raw 10-50 scale; see above) in well-being compared to the intercept for a male with country-average demographic characteristics who passed the attention check and who deviates maximally from country-mean religiosity (compared to being country-mean level religious). Across countries, estimates ranged from 1.53 [-0.61, 2.71] to 2.73 [1.49, 5.14] (see Figure 4).

For full transparency, we also conducted our pre-registered Scenario 1, which modeled *Religiosity* and *Cultural Norms* as separate predictors on 0 (global minimum) to 1 (global maximum) scales and included their interaction. Here, we report from the full model (Model 1): Globally, we found evidence for an interaction effect (2.20 [1.31, 2.75]), although the size varied across countries (from 1.16 [-2.72, 3.47] to 3.72 [1.55, 6.89], interpreted as the outcome on the raw 10-50 well-being score; see Appendix and Figure A.6). We are, however, somewhat skeptical of the validity of this finding, since test item reliability analyses indicate that the *Religiosity* and *Cultural Norms* items are partially overlapping and might be measures of the same underlying construct (e.g., a generally perceived importance of religion). See Sections 2.2–2.4 and Appendix A for details.

Finally, we attempted to explore the influence of social desirability in these data. Previous research has suggested that survey items asking about subjective, intra-psychological matters might be more susceptible to social desirability effects and that items about measurable, behavioral aspects are less so (Jones and Elliott, 2017; Møller et al., 2020). To explore this, we picked the *Religiosity* item that was the most "behavioral" (i.e., service attendance, `rel_1`) as predictor for the set of *Well-Being* items asking about physical well-being (`wb_phys_1` – `wb_phys_7`), which we summed and transformed to a 10-50 scale so that it is comparable to the main models' outcome. The model is otherwise essentially the same as Model 7, i.e., including varying intercepts and slopes across countries for the new *Religiosity* variable, simple effects for demographic variables and attention check. However, since the single *Religiosity* item is an ordered categorical variable, we now model this variable as a monotonic function (in the MARP data, all religiosity items were transformed from their original categorical scale to a continuous 0-1 variable so as to be comparable across differing scale ranges, but here, we back-transformed the selected variable to recover the original categorical form in order to model it monotonically). Despite substantial between and within-country variation – as in the main models – the association between *Religiosity* and *Well-Being* is notably reduced compared to the main results. The global effect size for *Religiosity* in this Social Desirability Model was estimated to 0.12 [0.00, 0.24], which represents the estimated increase in well-being compared to the global intercept (38.11 [37.32, 38.84]) for a male with country-average demographic characteristics who passed the attention check and who is maximally religious (compared to minimally religious). Across countries, estimates ranged from -0.07 [-0.43, 0.21] to 0.30 [0.09, 0.52] (see Figure B.7). In several countries, then, there were substantial probability mass around and below zero. Although one should be careful in drawing conclusions from this simple analysis, taken together the results could be interpreted as evidence for the hypothesis that behavioral aspects are less susceptible to social desirability. See Appendix B for details.

2. Main Analyses and Results (Models 7-10)

2.1. Outline

In this section, we report our main pre-registered analyses in more detail (see our pre-registration (Stage 1) document here: https://osf.io/kujg3/). The outline is as follows. First, we present our selected focal variables (*Well-Being*, *Religious Fit*, and *Cultural Norms*) and their constituent items. Second, we report test item reliability of these focal variables. Third, we discuss the overall model specification and data transformations, including demographic variables and cases of missing data. Subsequently, we detail our main models in both formal notation and working code. These models were build, checked and pre-registered using simulated data prior to seeing the real data. Following the pre-registered Bayesian workflow, we then report model diagnostics and perform model comparison and posterior predictive checks. We then summarize results, which support Q1, followed by a brief conclusion and discussion of limitations of our analytic approach.

The discussion of limitations feeds into another round of analyses that take the same general form as just outlined, but differs in one crucial way, in that the focal predictor (*Religious Fit*) is transformed so that it now represents the absolute deviance from country-mean religiosity (again, *Religious Deviance*). This procedure allows us to address a key prediction from the "person-culture fit" hypothesis, namely that any deviance (whether positive or negative) from country-mean level of religiosity should be associated with decreased levels of well-being. We fail to find support for this prediction (Q2), as absolute deviance from country-mean level appears to be associated with increased self-reported well-being.

2.2. Variable Selection

We operationalize our focal constructs as follows (cf., the MARP document "Data Documentation" for item details):

- *Well-Being*: As our focal outcome variable, we focus on the psychological (**wb_psych_1** – **wb_psych_6**) and social aspects (**wb_soc_1** and **wb_soc_2**) of well-being as well as the general satisfaction with life and health (**wb_gen_1** and **wb_gen_2**). We add up the scores of these 10 items, each of which were self-reported on scales from 1 to 5. This yields individual well-being scores ranging between 10 (min.) and 50 (max.).

- *Religiosity*: Our primary predictor variable is the summation of all the *Religiosity* items, except variables **rel_3** and **rel_4**, as these two variables pertain to self-identification and denomination, respectively, and thus arguably fall outside of the psychometric scale format of the remaining items.

- *Cultural Norms*: We include the summation of both items, **cnorm_1** and **cnorm_2**.

2.3. Test Item Reliability

To ensure appropriate internal construct validity of the focal variables, we conducted Bayesian reliability analyses using the Bayesrel package (Pfadt et al., 2020) for R (R Core Team, 2021). Here, we report McDonald's omega with 95% "uncertainty intervals".

- *Well-Being:* 0.88 [0.87, 0.88]. Dropping any one of the items does not improve internal consistency.

- *Religiosity:* 0.94 [0.93, 0.94]. Dropping any one of the items does not improve internal consistency.
- *Cultural Norms:* 0.85 [0.85, 0.86]. This construct consists of only two items, so drop-item analysis is not possible.
- *Religiosity + Cultural Norms:* 0.93 [0.92, 0.93]. Dropping any one of the items does not markedly improve internal consistency.

2.4. General Model Details – Scenario 2

We deem the reliability results for *Well-Being*, *Religiosity* and *Cultural Norms* to be acceptable. However, as the *Religiosity* and *Cultural Norms* items appear to measure the same underlying construct, it might be problematic to include both items as separate predictors and model their interaction (Scenario 1 in the pre-registration). For our main analyses, we therefore pursue Scenario 2 (Models 7-10), as specified in the Stage 1 document. However, we did conduct Scenario 1 as well and we briefly report on it in the Appendix for full transparency. All models were fitted using the `brms` package (Bürkner, 2017, 2018), an interface to `Stan` (Stan Development Team, 2020) in `R` (R Core Team, 2021).

2.5. Data Transformations

As we use the summation of the selected *Well-Being* items, all specifications model our outcome likelihood as Gaussian distributed with mean μ_i and standard deviation σ. Below, we outline the structure of all predictors.

2.5.1. Religious Deviance

For Scenario 2, we take the summation of *Religiosity* and *Cultural Norms* items, mean-center by country and rescale the sum onto a range from -1 to 1 (see above). By doing so, we can interpret the output as the effect on *Well-Being* of deviating maximally from group-mean religiosity[1]. This indirectly addresses the "person-culture fit" hypothesis underlying the project's research question Q2.

2.5.2. Demographic variables

Our demographic variables include gender, age, education and perceived socioeconomic status, since these are the individual-level variables where some theoretical justification exists for inclusion, in our view. However, as we want to avoid the pitfalls of "garbage-can regressions" (Achen, 2005) (e.g., including a long list of "control" covariates and perhaps, more problematic still, interpreting the estimates of these covariates as (causal) effects; see also Westreich and Greenland, 2013), we fit models both with and without demographic variables, and we do not report or interpret their estimates[2]. Demographic variables in all analyses are country-mean centered, but for gender where male is the reference, so that intercepts represent demographically country-average males.

[1] In the pre-registration, we stated that we would transform *Religiosity* data onto a 0 – 1 scale. However, this will not work for the mean-centering of Scenario 2, where we want the score of zero to represent being perfectly group-average religious – and not, say, "non-religious".

[2] In order to be comparable for interpretation, the outcome, the demographic variables as well as the focal predictor from the *Religiosity* data could be standardized. Further, since socioeconomic status and education are kinds of ordered categories, they could be modeled as monotonic functions (Bürkner and Charpentier, 2020). However, as we are – for current purposes – not interested in the separate impact of socioeconomic status and education on well-being

2.5.3. Attention Check

Response options of the attention check item ranged from 1-7, with 1 being the “right” answer (indicating attention check passed). For ease of interpretation, we transform this into a binary variable, where 0 = passed and 1 = failed. This ensures that all intercepts represent participants who passed the attention check.[3]

2.6. Missing Data and Nonsensical Values

The data set generally contains very few missing data points relevant for our analytic purposes. In the pre-registration, we committed to perform complete-cases analysis depending on the clustering of the missingness. The five NAs (i.e., missing values) in perceived socioeconomic status (`ses`) do not appear clustered in any specific country, whereas (`age`) is clustered to an extent. Out of 27 NAs in (`age`), 11 appear in the Morocco sub-sample. Given the size of the total as well as the country-level sample sizes, we suspect this poses a very minor problem. However, on principle we perform imputation of missing values using the `mice` package[4] (Buuren and Groothuis-Oudshoorn, 2011) for the missing values in `age` and fit the imputed data set to the full main model (Model 7). The results of the imputed model is consistent with those of the main models. In addition to missingness, the `age` variable contains a few nonsensical values, specifically age in years below 18, including sub-teen ages. In total, 19 participants report ages below 18, ranging from 0 to 17 years. As these values do not appear clustered, we exclude these from all analyses. In total, this yields a complete data set with N = 10,484.

2.7. Formal Main Models

We formally define our focal models below. Across all models, we treat well-being scores, y_i as distributed normally with mean μ_i and standard deviation σ (Line 1). Our linear models vary in a few ways, all of which are summarized in Line 2 and defined in Lines 4 through 7. All models include **A**, **K**, and **R**: **A** includes the main intercept, α, and group-specific offsets, α_j (Line 3), **K** includes the attention-check indicator variable (0 = passed) (Line 4); and **R** includes the *Religious Fit* predictor (Line 5). Some models include **D**, the demographic variables for: an indicator variable

(see our Stage 1 document), but instead simply include these covariates to hold them constant at a meaningful value (i.e., country-mean), we refrained from modeling these variables monotonically. This is also our justification for not modeling the demographic variables with fully varying effects, as we also outline in our pre-registration.

[3] We did not explicitly pre-register this transformation (although our simulated data took this form; see `R` code for Stage 1), as we assumed (wrongly) that the variable had already been transformed this way (cf., the document “Data Documentation”).

[4] This package imputes m data sets which are then stored in a single object. The default method is so-called “predictive mean matching”, a method that is considered robust particularly in contexts with large sample sizes and relatively few missing values (Van Buuren, 2018, sect 3.4). In an ideal scenario, we would then feed these m data sets to the `brms` function, `brm_multiple`, in order to propagate the uncertainty in the imputation to the model fitting. However, in order to perform the relevant data transformations (i.e., centering, re-scaling, etc.), we use the `mice::complete` function. This yields a specific imputed value for each missing value. Another option would be to use “one-step” imputation, which performs imputation during model fitting (Bürkner, 2020). However, the data transformations are then performed on the incomplete (un-imputed) data before fitting, which is also potentially problematic. While these procedures will very likely not result in differing outcomes in this case due to the substantial sample size and relatively few missing values, the principled best-practice approach would perhaps be to program `Stan` directly to perform all imputation and data transformations during model fitting. Exploring and evaluating consequences of differing imputation techniques goes beyond this write-up, however. In the `R` code, we illustrate with some convenience functions how to check the sensibility and diagnostics of the imputed data.

for participant sex (1 = male; 2 = other; 3 = female), country-mean centered age, socioeconomic status, and education (Line 6). Finally, some models include a by-group varying effect for *Religious Fit*, **G** (Line 7). See our code below.

$$\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \mathbf{A} + \mathbf{K} + \mathbf{R} + \mathbf{D} + \mathbf{G} && (2)\\
\mathbf{A} &= \alpha + \alpha_j && (3)\\
\mathbf{K} &= \beta^{\text{check}} k_i && (4)\\
\mathbf{R} &= \beta^{\text{religious fit}} r_i && (5)\\
\mathbf{D} &= \beta^{\text{sex}} s_i + \beta^{\text{age}} a_i + \beta^{\text{econ}} c_i + \beta^{\text{education}} d_i && (6)\\
\mathbf{G} &= \beta^{R}_{\text{COUNTRY}[i]} r_i && (7)
\end{aligned}$$

We define priors for the intercepts and predictors as follows:

$$\alpha, \alpha_j \sim \text{Normal}(30, 2)$$
$$\beta_p \sim \text{Normal}(0, 1)$$

We assigned our intercepts with a prior normal distribution around a value of 30 and a standard deviation of 2. We did this because 30 effectively serves as a midpoint for the possible values on the scale (10-50) while the value of 2 was a relatively conservative guess at how much variation there would be in the answers. Since two standard deviations roughly cover 95% probability of a normal distribution, this prior mostly expects variation to be within approximately ±4 outcome scores from the midpoint. Similarly, the prior for the beta coefficients mostly expects variation to be within approximately ±2 outcome scores. These sets of priors are then fairly regularizing as they are skeptical toward strong deviations from the midpoint of the outcome scale and toward strong relationships between the outcome and the predictors. However, since this data set contains thousands of observations, these priors are not so strong that they will substantially influence inference.

The remaining multi-level structure is formally defined as follows. The varying effects across countries bound in a variance-covariance matrix with multivariate priors of a Gaussian variety (with means of 0 as these are already included in the linear model above:

$$\begin{bmatrix} \alpha_j \\ \beta^{R}_{\text{COUNTRY}} \end{bmatrix} \sim \text{Multivariate Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{SRS}\right)$$

Here, **S** is a diagonal matrix of intercept and predictors' standard deviations, σ_p,

$$\mathbf{S} = \begin{bmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_{\beta^R} \end{bmatrix}$$
$$\sigma_p \sim \text{Exponential}(1)$$

while **R** is their correlation matrix which has a prior defined distribution of

$$\mathbf{R} \sim \text{LKJCorr}(2)$$

which is generally considered a weakly regularizing prior (McElreath, 2020, p. 442-443).

2.8. Main Model Specifications in R Code

Following this formal model, we now turn to the `R` code for our main model specifications in `brms` syntax. For those unfamiliar with the syntax, the outcome variable is on the left side of the ∼. Varying slopes are on the left side of the bar, while varying intercepts are on the right side. So, for example, varying religiosity across countries would look something like this: `(Rel | Country)`, whereas a model with only varying intercepts for country resembles `(1 | Country)`.

Priors. See formal description for details of prior choice.

```
aprior     <- set_prior("normal(30,2)", class = "Intercept")
bprior     <- set_prior("normal(0,1)", class = "b")
lkjprior   <- set_prior("lkj_corr_cholesky(2)", class = "cor")
sdprior    <- set_prior("exponential(1)", class = "sd")
sigmaprior <- set_prior("exponential(1)", class = "sigma")
```

Model 7[5]. This is the full model that allows varying effects of group-mean centered religiosity (`Rel`) across countries and simple effects for demographic variables and the attention check.

```
m7 <- brm(WB ~ Rel + (Rel | Country) +
 Gender + Age + SES.c + Edu + AC,
 data = data,
 family = gaussian(),
 prior = aprior + bprior + lkjprior + sdprior + sigmaprior,
 sample_prior = TRUE,
 control = list(adapt_delta = 0.99),
 chains = 4, iter = 2000)
```

Model 8. This is the same as Model 7 but it excludes demographics.

```
m8 <- brm(WB ~ Rel + (Rel | Country) + AC, ...)
```

Model 9. This model includes varying intercepts for countries with simple effects for *Religious Fit*, demographic variables, and the attention check.

```
m9 <- brm(WB ~ Rel + (1 | Country) +
 Gender + Age + SES + Edu + AC, ...)
```

Model 10. This is the same as Model 9 but it excludes demographics.

```
m10 <- brm(WB ~ Rel + (1 | Country) + AC, ...)
```

[5] For Model 8 through 10 we truncate the model code for concision. See accompanying `R` code for complete model specifications.

2.9. Results

2.9.1. Model Diagnostics and Model Comparison

All models returned reasonable diagnostic values for all parameters, with only the intercept occasionally showing some traces of convergence problems (“effective samples” < 1000; $\hat{R} \approx$1.01-1.02) for some models. For model comparison (Table 1), we employ approximate leave-one-out cross-validation (LOO), which emphasizes out-of-sample predictive accuracy. See above for model details.

	ELPD difference	SE of EPLD difference	P-LOO	LOOIC	Akaike-Weight
Model 7	–	–	45.8	66498.7	0.9998
Model 9	-9.2	4.3	32.3	66517.0	0.0001
Model 8	-860.0	44.4	45.4	68218.8	0.0000
Model 10	-898.7	45.0	26.6	68296.1	0.0000

Table 1: **Model comparison of Models 7-10.** Models are ranked for their predictive performance according to LOO. The Table reports the following: the difference in ELPD values with the best model as the reference; the standard error (SE) of the ELPD difference; the effective number of parameters, P-LOO (a lower value is better); leave-one-out information criterion, LOOIC (a lower value is better); and Akaike model weights (a higher value is better).

LOO strongly favors the full model (Model 7), which, to reiterate, includes demographic variables as well as varying slopes of *Religious Fit* across countries. Model 7 is also the most complex model (indexed by P-LOO), which tends to be favored by model comparison metrics. The next-best model is Model 9, which includes demographic variables but only varying intercepts of *Religious Fit* across countries. Model 9 outperforms Model 7 in “effective number of parameters” (P-LOO), while still retaining some predictive accuracy. The two models without demographic variables (Model 8 and 10) perform notably worse. Note, however, that LOO is designed for predictive and not (causally) inferential purposes, and these values should be interpreted in this light.

2.9.2. Posterior Predictive Checks

Using the `bayesplot` package (Gabry and Mahr, 2021; Gabry et al., 2019), we perform simple posterior predictive checks of Model 7 for each country, illustrated in Figure 1. In essence, these plots are visualizations of how the model, conditional on the model and the data, predicts new samples from each country.

The posterior predictive check reveals reasonable predictive accuracy overall, although the distribution of observed outcomes in some countries (the *Well-Being* scores; the dark-blue line) take on shapes that are slightly more skewed and/or concentrated than what the modeled normal distribution (the light-blue lines) allows for. Also, since we did not use a truncated outcome distribution, the model predicts values slightly outside the possible range (i.e., *Well-Being* scores below 10 and above 50). This can be seen in Figure 1 in that the light-blue predictive lines “over-shoot” the observed outcome distribution in some cases. The dotted line marks the estimated global average of *Well-Being* and helps illustrate degree of “shrinkage”: the amount that the model adaptively regularizes its estimate in any specific country toward the global mean, in order to avoid over-fitting and improve overall predictive performance. In other words, the dotted line represents the estimate – the grand mean – from “complete-pooling” (i.e., all countries are analyzed together), and the light blues lines are estimates from “partial-pooling” (i.e., countries are analyzed separately but each country’s estimate is informed by and shrunk towards the grand mean).

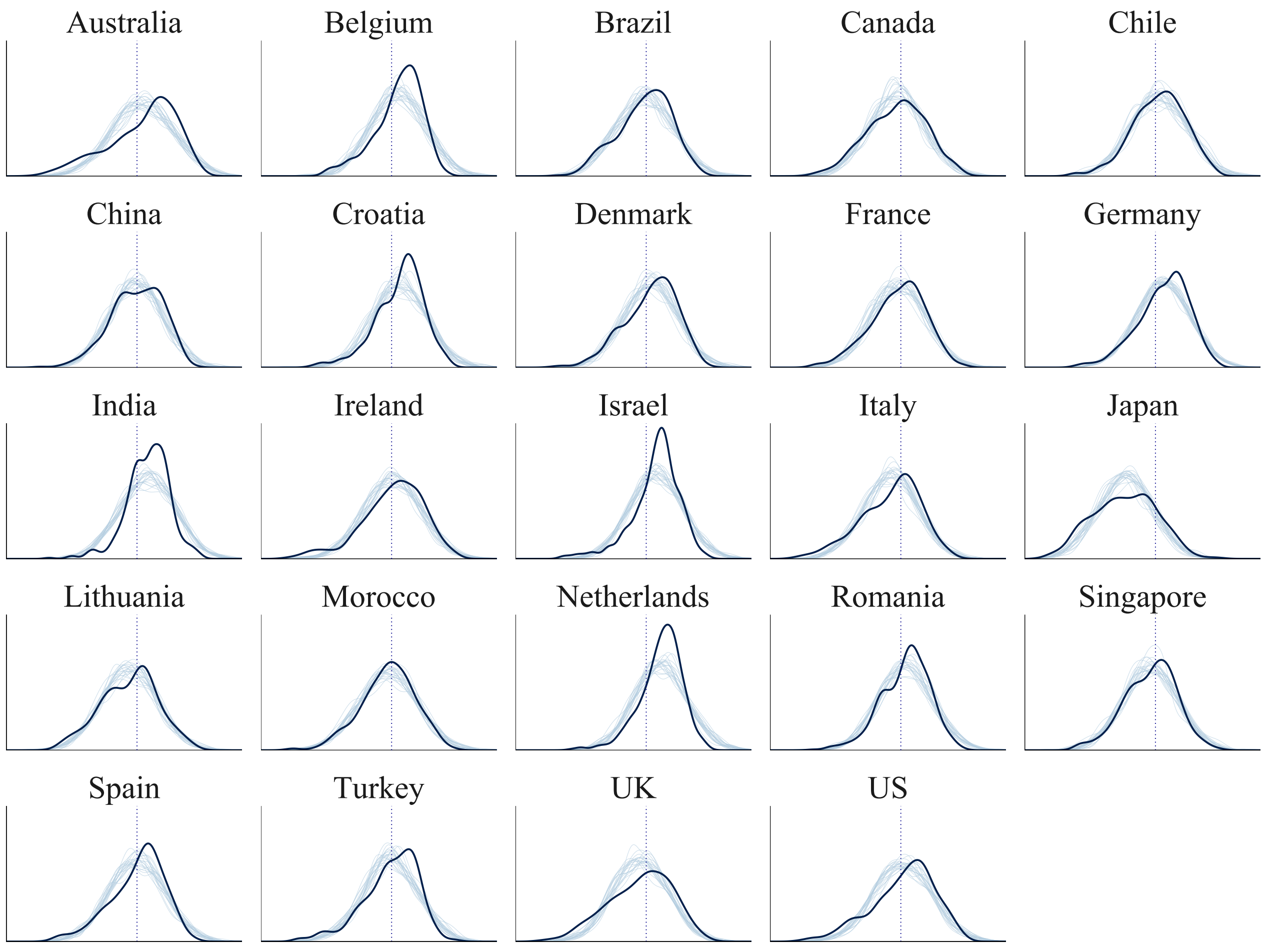

Figure 1: **Posterior predictive checks.** The dark-blue line is the distribution of observed outcomes (the *Well-Being* scores) for each country. The light-blue lines are 20 predicted samples drawn from the posterior distribution of Model 7 for each country. To illustrate shrinkage, the dotted line represents the estimated global average of *Well-Being*, 35.5 [34.6, 36.4] (Figure 2).

2.9.3. Model 7 Results

In Figure 2, produced using the `brmstools` package[6] (Vuorre, 2018), we report results from Model 7. Results for Model 8-10 are qualitatively similar (see Table 2). Figure 2 reports cross-country posterior distributions and posterior means with 95% “credible intervals”. Recall that all demographic variables are country-mean centered, that *Religiosity* is country-mean centered on a -1 to 1 scale (see above for formula) and that for the *Attention Check*, $0 =$ “passed”. Therefore, the intercept represents the estimated *Well-Being* score for a male that passed the attention check and is country-average in religiosity, age, perceived socioeconomic status and education. Likewise, the posterior means for *Religious Fit* reflects the estimated *Well-Being* score for a globally maximally religious male with country-average demographic characteristics who passed the attention check.

Figure 3, produced using the `tidybayes` package (Kay, 2020), illustrates the robust positive

[6] This package is deprecated, but it serves our specific purpose here.

cross-national association between *Religious Fit* and *Well-Being*. Blue lines are posterior predictive draws from Model 7 (100 draws for each country) overlain on the raw data points (slightly jittered for visualization).

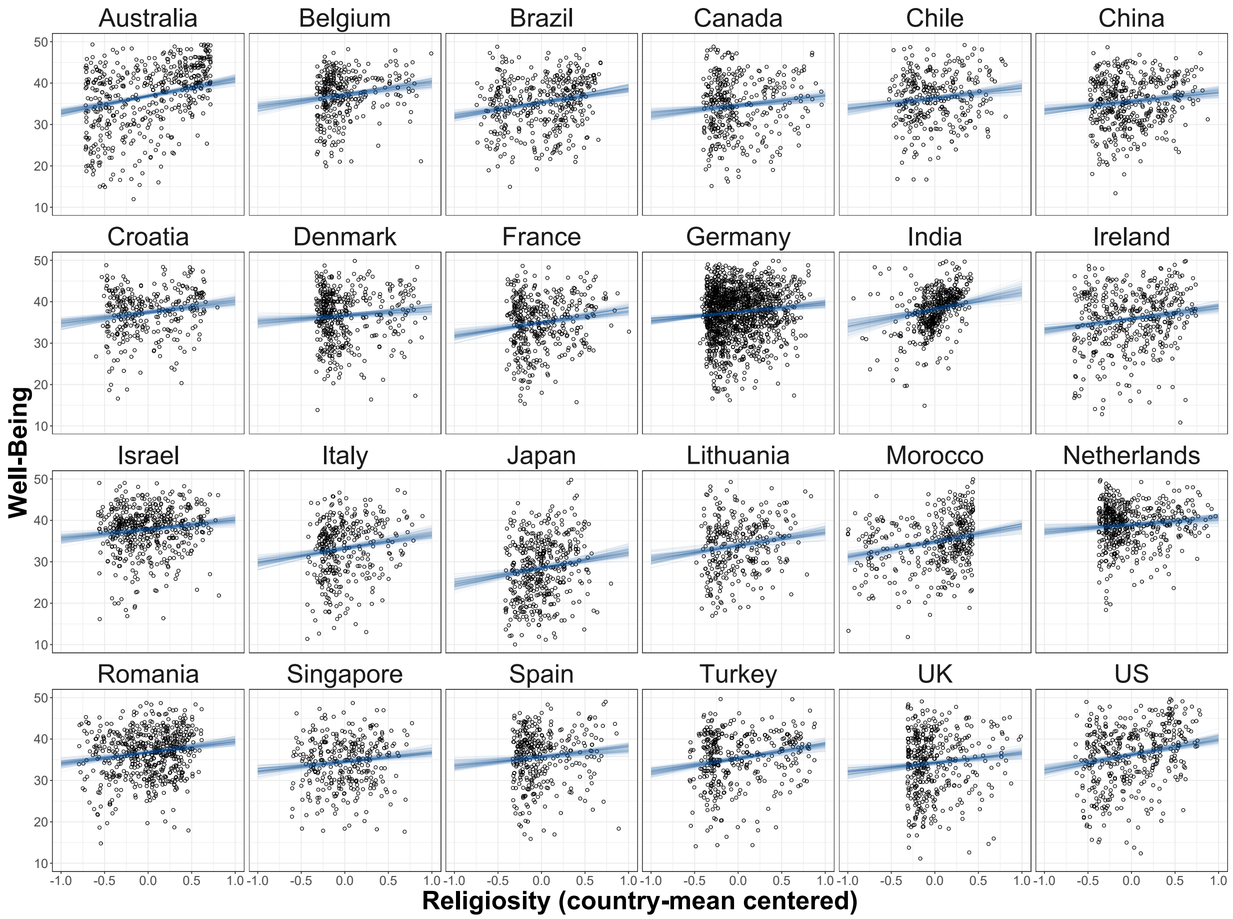

Figure 3: **Posterior predictive draws for Model 7.** 100 draws (blue lines) for each country overlain on raw data points (slightly jittered for visualization).

2.10. Discussion

While there is notable within- and between-country variation, the association between religiosity and well-being is robustly positive across countries, lending affirmative support for Q1. Further, there is evidence that deviance from country-mean level of religiosity is positively associated with increasing well-being, which contradicts the "person-culture fit" hypothesis (which predicts most well-being at country-mean level of religiosity) and hence lends no support for Q2.

As with all self-report data, the MARP data are vulnerable to various demand-characteristics and related biases, including social desirability. We attempt to address some of this concern in the Appendix. We should also be careful with concluding that the relationship between religiosity and well-being is causal. While the results are seemingly robust when holding some possible confounders constant (Model 7 and 9), this is not in itself sufficient to conclude a causal relationship.

Another important caveat to make explicit here is that the analytical approach employed in Scenario 2 assumes that the group-mean level of religiosity in our sample, to which we centered

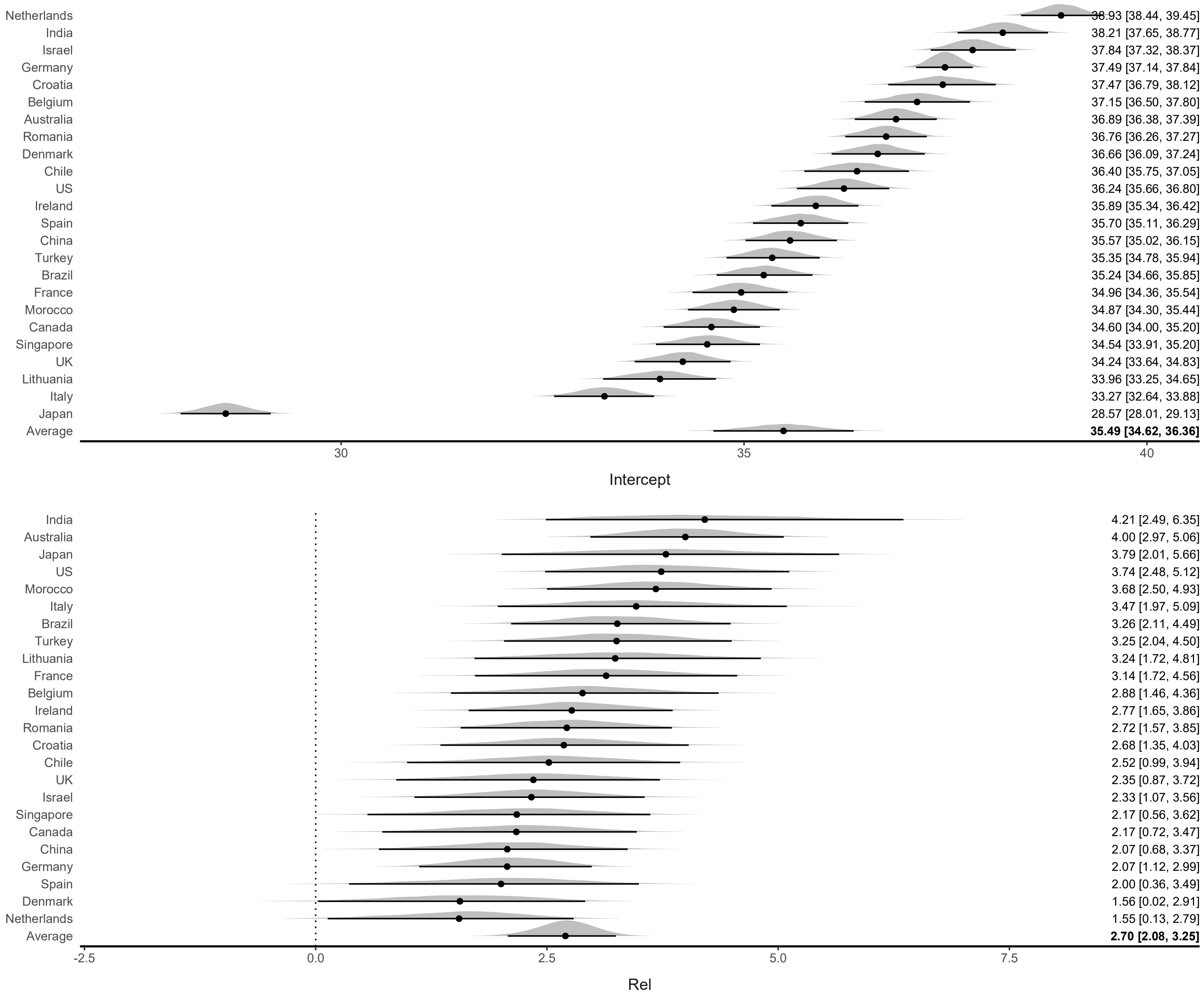

Figure 2: **Country-level results for Model 7.** Top plot: Intercept parameters. Bottom plot: Religiosity parameters. Posterior means (dots) and posterior distributions (grey). Lower and upper bounds (black lines and brackets) represent 95% *credible intervals*. Y-axes are estimated outcomes on raw *Well-Being* score (10-50). The Intercept represents the estimated *Well-Being* score for a male that passed the attention check and is country-average in religiosity, age, perceived socioeconomic status and education. Likewise, the posterior means for *Religiosity* ("Rel") reflects the estimated *Well-Being* score for a maximally religious male with country-average demographic characteristics who passed the attention check. Countries are estimate-ranked.

participants, is somewhat representative of the *actual* country-level religiosity. However, generalizations and inference of any sort from this kind of data set arguably always relies on some assumption of representativeness.

Another important note about this approach is that it assumes that the relationship between religiosity and well-being is approximated by a linear model. However, the focal relationship might plausibly be non-linear. For instance, the "person–culture fit" (see MARP document "Theoretical Background") predicts that people who deviate from the group-mean – by being more or less religious than their surroundings – should report lower well-being, calling perhaps for a quadratic relationship. Taking this prediction seriously will entail transforming the *Religiosity* score, so that each participant's score represents the *absolute* deviance from the group-mean and making below-mean and above-mean deviations equivalent in terms of sign (e.g., a country-mean centered score of -0.5 is simply an absolute deviance of 0.5 from the country-mean). This can be achieved with the simple `abs` function in `R`. Although we did not formally pre-register this variant of Scenario 2, it is in the spirit of the verbal description (see Stage 1 document). We therefore conducted it and report results below.

3. Scenario 2 Analyses and Results (Models 11-14)

This subsection repeats the main analyses with the one difference that we use the absolute value of country-mean centered religion data (*Religious Deviance*) to assess the association between deviance from country-mean religiosity on well-being, motivated by the "person–culture fit" hypothesis. As these models take the exact same statistical form as Model 7-10, we jump straight to results.

3.1. Results

In Figure 4, we report results from Model 11, which mirrors the full main model (Model 7). Results for Model 12-14 (which mirror the remaining main Models 8 through 11) are qualitatively similar (see Table 2). Again, the intercept represents the estimated *Well-Being* score for a male that passed the attention check and is country-average in religiosity, age, perceived socioeconomic status and education. The posterior means for *Religious Deviance* now reflects the estimated *Well-Being* score for a male with country-average demographic characteristics who passed the attention check and who deviates *maximally* from country-mean religiosity (compared to being country-mean level religious).

3.2. Discussion

To reiterate, the "person-culture fit" hypothesis (Q2) predicts highest well-being at country-mean level of religiosity. However, maximally deviating from country-mean religiosity also has a robustly positive association with well-being, despite substantial within- and between-country variation. We therefore fail to find support for the "person-culture fit" hypothesis (Q2), in as much as the country-mean levels of religiosity in our sample is a valid proxy for the actual perceived importance of religion for the average person in each country.

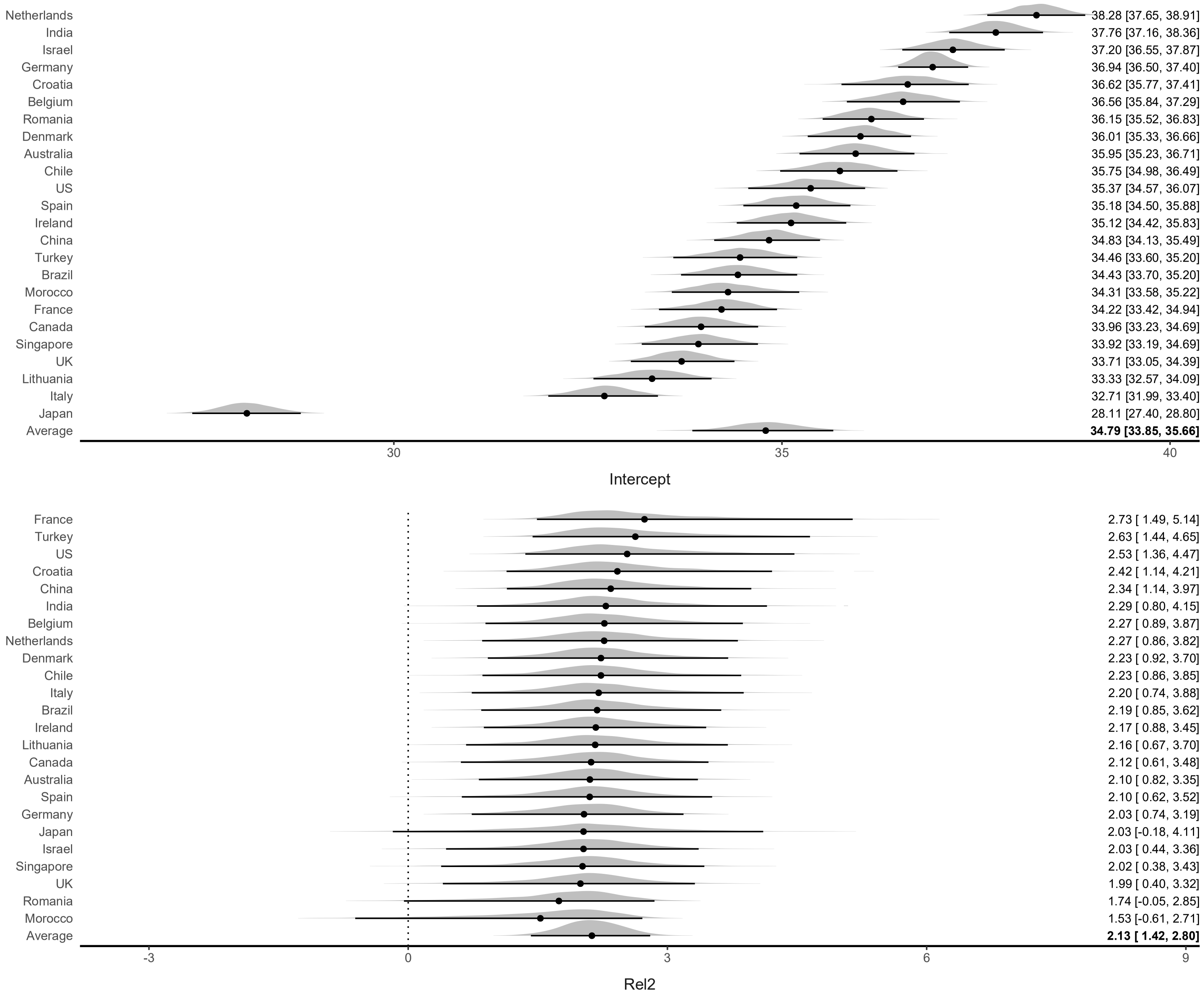

Figure 4: **Country-level results for Model 11.** Top plot: Intercept parameters. Bottom plot: *Religiosity* parameters. Posterior means (dots) and posterior distributions (grey). Lower and upper bounds (black lines and brackets) represent 95% *credible intervals*. Y-axes are estimated outcomes on raw *Well-Being* score (10-50). Estimates for "Rel2" now represents the association with well-being for *absolute* deviance from country-mean religiosity (i.e., *Religious Deviance*). Countries are estimate-ranked.

Covariates	M7	M8	M9	M10	M11	M12	M13	M14[a]
Intercept	35.49	35.37	35.46	35.39	34.79	34.66	34.79	34.65
	(34.62, 36.36)	(34.34, 36.29)	(34.61, 36.33)	(34.53, 36.21)	(33.85, 35.66)	(33.80, 35.47)	(33.85, 35.60)	(33.78, 35.51)
Religiosity[b]	2.70	3.00	2.82	3.49	2.13	2.41	2.15	2.46
	(2.08, 3.25)	(2.08, 3.78)	(2.48, 3.15)	(3.13, 3.85)	(1.42, 2.80)	(1.67, 3.13)	(1.53, 2.75)	(1.82, 3.09)
Attention Check	-0.94	-0.65	-0.94	-0.60	-0.56	-0.15	-0.56	-0.15
	(-1.55, -0.32)	(-1.32, 0.03)	(-1.57, -0.32)	(-1.26, 0.06)	(-1.15,0.06)	(-0.82, 0.50)	(-1.17, 0.06)	(-0.83, 0.52)
Gender–Other	-2.26	–	-2.30	–	-2.44	–	-2.45	–
	(-3.46, -1.09)	–	(-3.53, -1.10)	–	(-3.58, -1.29)	–	(-3.64, -1.29)	–
Gender–Woman	-0.20	–	-0.21	–	-0.08	–	-0.08	–
	(-0.44, 0.03)	–	(-0.45, 0.02)	–	(-0.33, 0.17)	–	(-0.31, 0.16)	–
Age	0.04	–	0.04	–	0.04	–	0.04	–
	(0.03, 0.05)	–	(0.03, 0.05)	–	(0.03, 0.05)	–	(0.03, 0.05)	–
Socioeconomic Status	1.39	–	1.40	–	1.44	–	1.44	–
	(1.32, 1.46)	–	(1.32, 1.47)	–	(1.37, 1.51)	–	(1.37, 1.51)	–
Education	0.31	–	0.33	–	0.33	–	0.33	–
	(0.22, 0.41)	–	(0.23, 0.43)	–	(0.23, 0.43)	–	(0.24, 0.43)	–

Table 2: **Global estimates across main models.** Posterior means with 95% *credible intervals* in parentheses. Note that demographic covariates should not be interpreted as actual (causal) effects on *Well-Being*, as we did not build our models for this purpose (Westreich and Greenland, 2013). We report them simply for transparency (see Stage 1 document). Note also that in Models 7 through 10 *Religiosity* represents maximally religious compared to country-mean religiosity, whereas in Models 11 through 14, *Religiosity* represents absolute deviance from country-mean level religiosity. [a]Model 14 sampled the intercept inefficiently, with $\hat{R} \approx 1.03$ and effective sample size < 400. Therefore, for this particular model, we increased the number of iterations to 4000. [b]Recall that these variables are transformed differently between M7-M10 and M11-M14. See Section 1 for model definitions.

Author Contributions

All authors contributed to the design of the models. TB and BGP wrote the manuscript, with critical inputs from ALM and US. TB wrote the bulk of the `R` code, with critical inputs from BGP and double-checked key parts of the code with ALM. All authors approved of the final draft.

Acknowledgments

We thank the organizers of MARP. Bendixen and Purzycki thank the Aarhus University Research Foundation for support. Purzycki and Schjoedt acknowledge support from the Culture of Schooling Project funded by the John Templeton Trust and the Issachar Fund.

Appendix

In the Appendix, we: 1) model our pre-registered Scenario 1 (Appendix A), and 2) conduct an alternative model, which indirectly aims to assess degree of social desirability in the main models (Appendix B).

Appendix A. Scenario 1

For Scenario 1, the *Religiosity* and *Cultural Norms* items are transformed onto a 0 (scale minimum) to 1 (scale maximum) scale for ease of interpretation (i.e., holding these variables at zero is meaningful). In other words, we simply divide each individuals religiosity score out of a total possible of 7 for *Religiosity* and 2 for *Cultural Norms*. See the Stage 1 document and R code for further model description and specifications.

Appendix A.1. Model Specification in R Code

Here, we display the model specifications for Scenario 1, first in working R code and then in formal notation.

Model 1[7]. Full model predicting individual well-being (WB) with varying intercepts and slopes across countries for religiosity (Rel) and perceived cultural norms of religion (Norm) and their interaction, and simple effects for Gender and group-mean centered demographic variables age (Age), perceived socioeconomic status (SES), education (Edu), and attention check (AC).

```
m1 <- brm(WB ~ Rel*Norm + (Rel*Norm | Country) +
 Gender + Age + SES + Edu + AC,
 data = data,
 family = gaussian(),
 prior = aprior + bprior + lkjprior + sdprior + sigmaprior,
 sample_prior = TRUE,
 control = list(adapt_delta = 0.99),
 chains = 4, iter = 2000)
```

Model 2. This is the same as Model 1 but it excludes demographics.

```
m2 <- brm(WB ~ Rel*Norm + (Rel*Norm | Country) + AC, ...)
```

Model 3. This is an interaction model with varying intercepts across countries with religiosity, cultural norms, demographic variables, and the attention check as simple effects.

```
m3 <- brm(WB ~ Rel * Norm + (1 | Country) +
 Gender + Age + SES + Edu + AC, ...)
```

Model 4. This model is the same as Model 3 but it excludes demographics.

```
m4 <- brm(WB ~ Rel*Norm + (1 | Country) + AC, ...)
```

Model 5. This model has varying intercepts for countries and simple effects for religiosity, cultural norms, demographic variables, and the attention check.

[7] For Model 2 through 6 we truncate the model code for concision. See accompanying R code for complete model specifications.

```
m5 <- brm(WB ~ Rel + Norm + (1 | Country) +
 Gender + Age + SES + Edu + AC, ...)
```

Model 6. This is the same as Model 5 but it excludes demographics.

```
m6 <- brm(WB ~ Rel + Norm + (1 | Country) + AC, ...)
```

Appendix A.2. Formal Model

Below, we formally define the full model (Model 1 from the code above). As in the main models detailed in previous sections, we again treat well-being score, y_i, as distributed normally with mean μ_i and standard deviation σ. This model follows closely with that defined above, with the exception of an additional religious norms consideration, $\mathbf{N}$, and a varying interaction effect of norms and religiosity, $\mathbf{G}$.

$$
\begin{aligned}
&y_i \sim \text{Normal}(\mu_i, \sigma)\\
&\mu_i = \mathbf{A} + \mathbf{K} + \mathbf{N} + \mathbf{D} + \mathbf{G}\\
&\mathbf{A} = \alpha + \alpha_j\\
&\mathbf{K} = \beta^{\text{check}} k_i\\
&\mathbf{N} = \beta^{\text{religiosity}} r_i + \beta^{\text{norms}} n_i + \beta^{\text{relnorms}} r_i n_i\\
&\mathbf{D} = \beta^{\text{sex}} s_i + \beta^{\text{age}} a_i + \beta^{\text{econ}} c_i + \beta^{\text{education}} d_i\\
&\mathbf{G} = \beta^{R}_{\text{COUNTRY}[i]} r_i + \beta^{N}_{\text{COUNTRY}[i]} n_i + \beta^{RN}_{\text{COUNTRY}[i]} r_i n_i
\end{aligned}
$$

While our intercept and predictor priors are defined in the same way as above,

$$
\begin{gathered}
\alpha, \alpha_j \sim \text{Normal}(30, 2)\\
\beta_p \sim \text{Normal}(0, 1)
\end{gathered}
$$

the remaining multi-level structure differs due to the cross-sample varying interaction effect between religiosity and religious norms. The variables comprising this interaction effect are bound in a variance-covariance matrix with multivariate priors of a Gaussian variety (with means of 0 as these are already included in the linear model in Lines 12-13):

$$
\begin{bmatrix} \alpha_j \\ \beta^{R}_{\text{COUNTRY}} \\ \beta^{N}_{\text{COUNTRY}} \\ \beta^{RN}_{\text{COUNTRY}} \end{bmatrix} \sim \text{Multivariate Normal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{SRS} \right)
$$

Here, $\mathbf{S}$ is a diagonal matrix of intercept and predictors' standard deviations, σ_p,

$$
\begin{gathered}
\mathbf{S} = \begin{bmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_{\beta^R} & 0 & 0 \\ 0 & 0 & \sigma_{\beta^N} & 0 \\ 0 & 0 & 0 & \sigma_{\beta^{RN}} \end{bmatrix}\\
\sigma_p \sim \text{Exponential}(1)
\end{gathered}
$$

while **R** is their correlation matrix which is also assigned a prior distribution of LKJCorr(2).

Appendix A.3. Additional Models

In addition to the pre-registrered models, we fitted three further models. These are variations of the pre-registrered Model 5, and are respectively denoted Model 5.extra, Model 5.extra.rel and Model 5.extra.norm in Table (Table A.3). These were fitted to further explore the impact on predictive accuracy of including an interaction term between Religiosity and Cultural Norms (Models 1-4) compared with an additive model, where Religiosity and Cultural Norms were allowed to have varying intercepts and slopes for each country (Model 5.extra). The extra models also allows us to explore the possible collinearity between *Cultural Norms* and *Religiosity*, by fitting two models that are identical except for the exclusion of one of the variables (Model 5.extra.rel and Model 5.extra.norm).

These essential objectives – that is, exploring the existence of an interaction and the possible problem with including *Cultural Norms* and *Religiosity* as separate variables in the same model – were not directly possible with the pre-registrered models. Following the format of the Stage 1 document, we briefly describe these extra models, including their formula in `brms` syntax ("...” denotes truncation; see code for full specification). We use the same priors found in the main models.

Model 5.extra. This model allows *Cultural Norms* and *Religiosity* to have intercepts across countries and simple effects for demographic variables and the attention check.

```
m5.extra <- brm(WB ~ Rel + Norm + (Rel + Norm | Country) +
 Gender + Age + SES + Edu + AC, ...)
```

Model 5.extra.rel. This model resembles Model 5.extra, but excludes *Cultural Norms.*

```
m5.extra.rel <- brm(WB ~ Rel + (Rel | Country) +
 Gender + Age + SES + Edu + AC, ...)
```

Model 5.extra.norm. This model resembles Model 5.extra, but excludes *Religiosity.*

```
m5.extra.norm <- brm(WB ~ Norm + (Norm | Country) +
 Gender + Age + SES + Edu + AC, ...)
```

Appendix A.4. Model Diagnostics and Model Comparison

All models returned reasonable diagnostic values for all parameters, with only the intercept consistently showing some traces of convergence problems ("effective sample size" < 1000; Rhat $\approx$1.01-1.02) for some models. As in the main models, we employ approximate leave-one-out cross-validation (LOO) (Vehtari et al., 2017) for model comparison (Table A.3).

	ELPD diff.	SE of EPLD diff.	P-LOO	LOOIC	Akaike-Weight
Model 1	–	–	52.0	66484.1	0.987
Model 3	-5.0	3.7	34.1	66494.1	0.007
*Model 5.extra	-5.1	3.4	54.4	66494.4	0.006
*Model 5.extra.rel	-13.3	5.4	45.0	66510.7	0.000
Model 5	-17.8	5.5	33.6	66519.7	0.000
*Model 5.norm	-103.0	14.4	50.2	66690.0	0.000
Model 2	-860.1	44.3	51.2	68204.3	0.000
Model 4	-894.7	44.4	28.1	68273.5	0.000
Model 6	-905.7	44.9	27.8	68295.5	0.000

Table A.3: **Model comparison of Models 1-6.** Models are ranked for their predictive performance according to LOO. The Table reports the following: the difference in ELPD values with the best model as the reference; the standard error (SE) of the ELPD difference; the effective number of parameters, P-LOO (a lower value is better); leave-one-out information criterion, LOOIC (a lower value is better); and Akaike model weights (a higher value is better). Asterisks denote additional models that were not pre-registered.

LOO strongly favors the full model (Model 1), which, to reiterate, includes demographic variables as well as varying intercepts and slopes of the main effects of *Religiosity* and *Cultural Norms* and their interaction across countries. The next-best model is Model 3, which includes demographic variables but only varying intercepts of the interaction between *Religiosity* and *Cultural Norms* and their main effects across countries.

Interestingly, Model 5.extra, which does not include the interaction term, performs equally well as Model 3, and only slightly worse than Model 1, indicating that leaving out the interaction term (which, granted, in Model 3 is only allowed to vary in intercepts across countries) does not hurt predictive accuracy dramatically. Model 3, however, does outperform Model 5.extra on "effective number of parameters" (P-LOO).

Also note that Model 5.extra.rel (model excluding Cultural Norms) and Model 5.extra.norm (model excluding Religiosity) differ quite substantially in their predictive accuracy, suggesting that including both focal predictors, (*Religiosity* and *Cultural Norms*), do contribute to improved predictive performance. Moreover, the models that do not include the demographic variables (Model 2, 4 and 6) perform notably worse than the alternative models.

Appendix A.5. Model 1 Results

Figure A.5, produced using the `tidybayes` package (Kay, 2020), illustrates the interaction effect across countries (Model 1). Black lines are posterior predictive draws (100 draws for each country) for the association between *Cultural Norms* and *Well-Being*, when *Religiosity* is set to 0 (globally minimal religious). Blue lines are posterior predictive draws (100 draws for each country) for the association between *Cultural Norms* and *Well-Being*, when *Religiosity* is set to 1 (globally maximal religious). Draws are overlayed on the raw data points (slightly jittered for visualization). Blue points are for individuals, who have a *Religiosity* score >0.5, whereas black points are for *Religiosity* scores ≤0.5. Taken together, then, the black and blue lines and points illustrate the varying interaction effects. The model predicts that, all else being equal, well-being is higher for an individual that is maximally religious and considers religion to be maximally important than for an individual that is minimally religious and considers religion to be maximally important, though to a varying degree across countries (e.g., the effect appears lower in, say, Denmark and the

Netherlands, where the black and blue lines are closer and partially overlap, than in, say, Australia, India, and the US, where the black and blue lines are further apart).

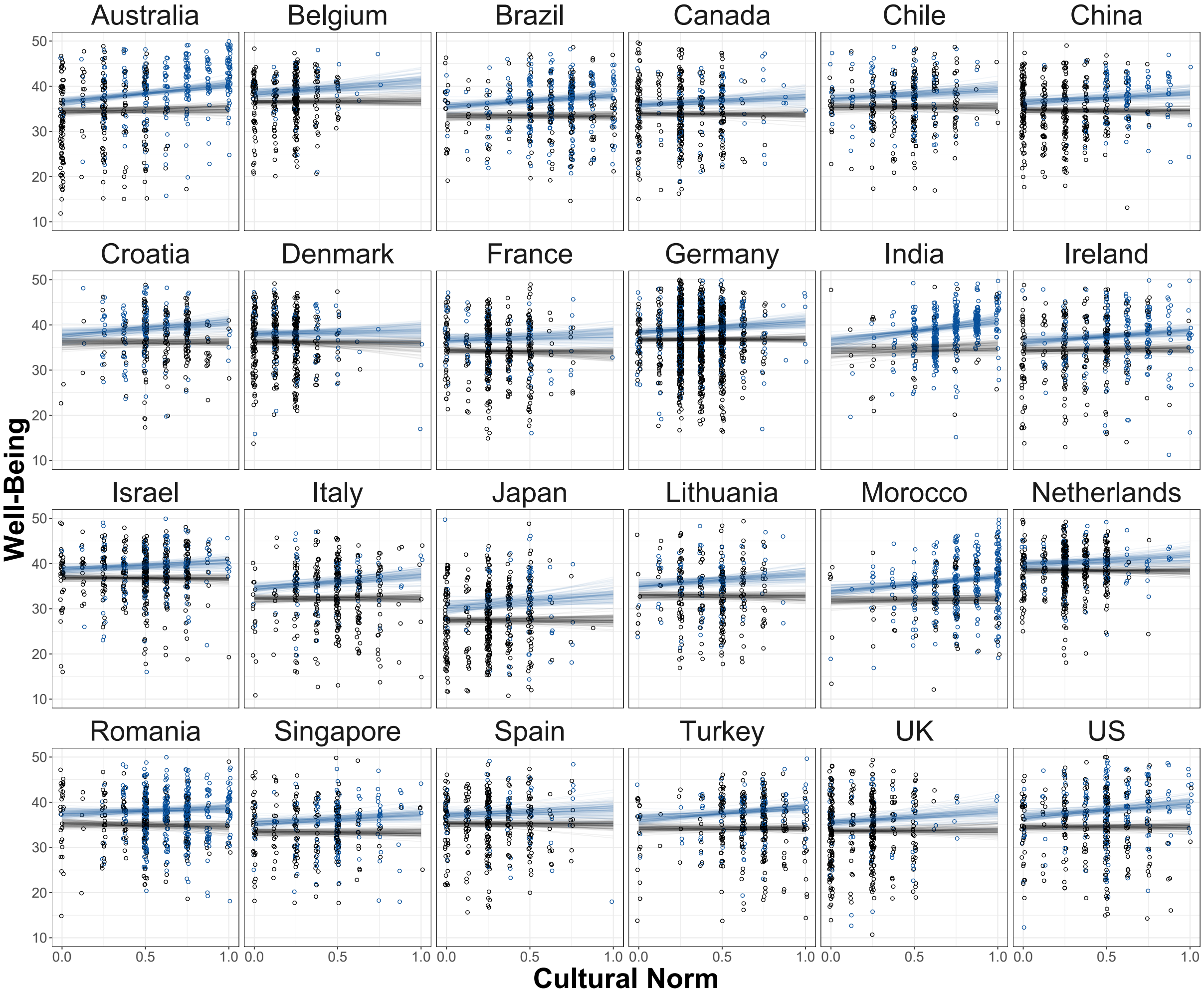

Figure A.5: **Posterior predicted draws for Model 1.** Black lines are posterior predictive draws (100 draws for each country) for the association between *Cultural Norms* and *Well-Being*, when *Religiosity* is set to 0 (globally minimal religious). Blue lines are posterior predictive draws (100 draws for each country) for the association between *Cultural Norms* and *Well-Being*, when *Religiosity* is set to 1 (globally maximal religious). Draws are overlain on the raw data points (slightly jittered for visualization). Blue points are for individuals, who have a *Religiosity* score >0.5, black points for *Religiosity* scores ≤ 0.5. Together, the black and blue lines and points illustrate the interaction effect. The raw data points look almost categorical in nature, since they are the (transformed) summation of only two items (the *Cultural Norms* items).

Figure A.6 reports cross-country posterior distributions and posterior means with 95% *credible intervals* from Model 1. Recall that all demograhic variables are country-mean centered, that *Religiosity* and *Cultural Norms* (perceived importance of religion) range from 0 (minimum) to 1 (maximum) and that for the *Attention Check*, 0 = "passed". Therefore, the intercept represents the estimated *Well-Being* score for a male that passed the attention check, who is minimally religious, perceive religion to be of minimal importance, and is country-average in age, perceived socioeconomic status and education. The posterior means for *Religiosity* reflect the estimated *Well-Being* score for a maximally religious male with minimally perceived importance of religion and country-average demographic characteristics who passed the attention check. Likewise, the posterior means for *Cultural Norms* reflect the estimated *Well-Being* scores for a minimally religious male with country-average demographic characteristics who passed the attention check and who perceives religion to be maximally important. The interaction can be interpreted in two equivalent ways: When *Religiosity* is set to 1 (i.e., maximal religiosity), then add the interaction estimate to the main effect estimate for *Cultural Norms*; or, alternatively, when *Cultural Norms* is set to 1 (i.e., religion is perceived to be maximally important), then add the interaction estimate to the main effect estimate for *Religiosity*.

However, note that Model 2-6 (including the extra variants of Model 5) are not qualitatively similar to Model 1 in estimating the association between *Well-Being* and *Cultural Norms*. In Model 1, this association is essentially zero (conditional on the interaction), while the association between *Well-Being* and *Religiosity* remains robustly positive here, as in the main models. However, in the additive varying effects model (Model 5.extra), the global association between *Well-Being* and *Cultural Norms* is slightly positive (0.73 [-0.01, 1.46]) and it is larger still (1.47 [0.57, 2.30]) in the model that excludes *Religiosity* (Model 5.extra.norm)[8]. The instability of estimates may be caused by some measure of overlap between our focal predictors.

Appendix A.6. Conclusion

Taken at face value, there appears to be a positive interaction effect between *Religiosity* and *Cultural Norms* (perceived importance of religion), according to Model 1 (see Figures A.5 and A.6). However, we are somewhat skeptical of the validity of this finding. First, it is possible that the *Religiosity* and *Cultural Norms* items partially account for the same underlying psychological construct (e.g., a generally perceived importance of a religious lifestyle). Second, the estimates for *Cultural Norms* depend markedly on the inclusion of *Religiosity* in models[9]. Third, it is possible that the main drivers of the apparant interaction effect is driven primarily by the robust positive association between *Well-Being* and *Religiosity* found in the main models, indicating some collinearity or masking effects.

Appendix B. Social Desirability Model

Previous research has suggested that survey items asking about subjective, intra-psychological matters might be more susceptible to social desirability effects and that items tapping into measurable, behavioral aspects are less so (Jones and Elliott, 2017; Møller et al., 2020). To explore this, we picked the *Religiosity* item that is most "behavioral" (i.e., service attendance, `rel_1`) as predictor for the set of *Well-Being* items asking about physical well-being (`wb_phys_1` – `wb_phys_7`), which we summed and transformed to a 10-50 scale so that it is comparable to the main models' outcome. The model is otherwise essentially the same as Model 7, i.e., including varying intercepts and slopes across countries for the new *Religiosity* variable, simple effects for demograhic variables and

[8] Upon inspection of the cross-country estimates, the positive global association between *Well-Being* and *Cultural Norms* reported here appear clearly driven by specific countries.

[9] The same is not the case in the opposite direction, perhaps because the *Cultural Norms* construct, which only includes two item (see e.g., Figure A.5), contains less variation than *Religiosity*, which in the Scenario 1-analyses consist of seven items

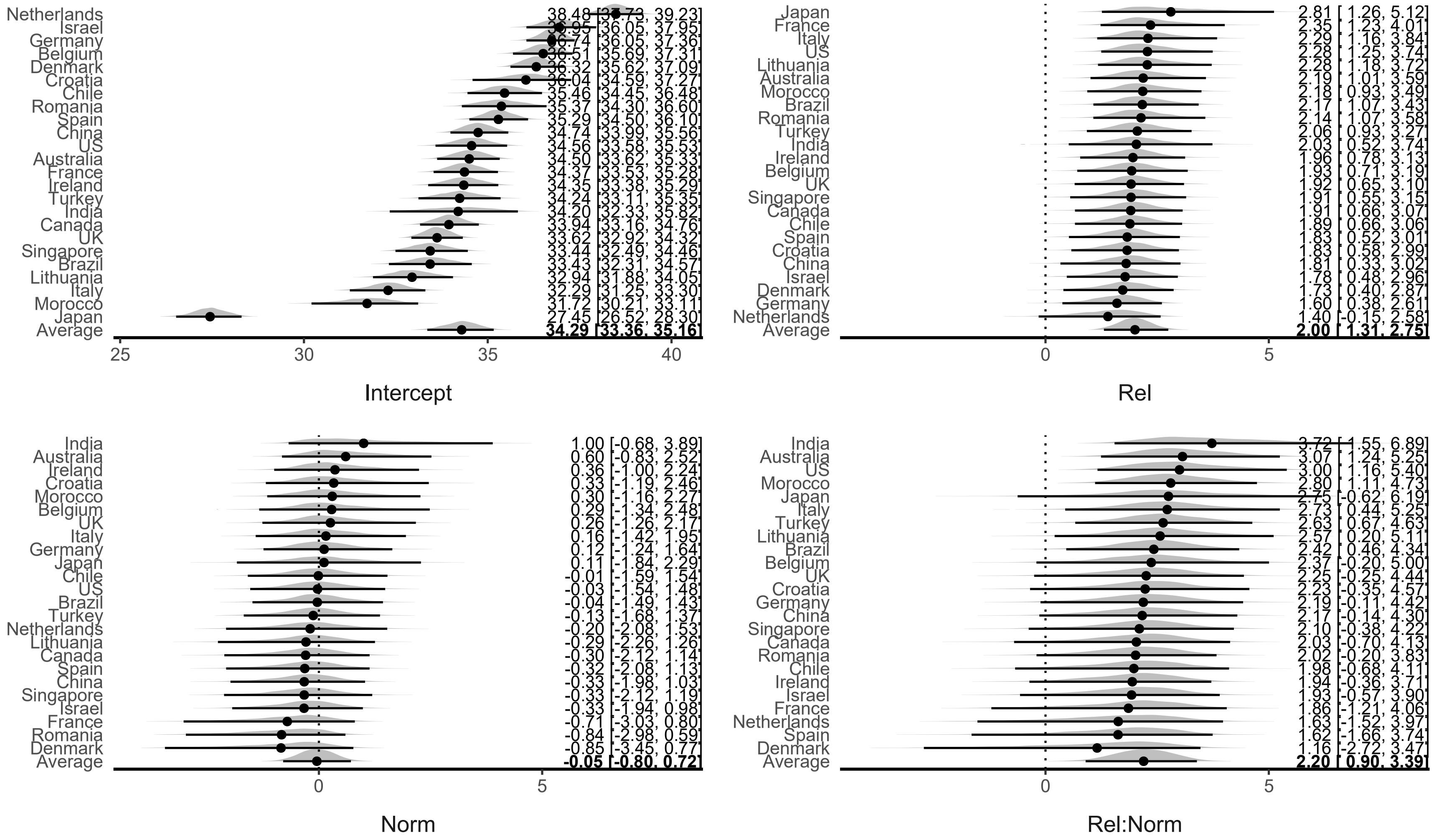

Figure A.6: **Country-level results for Model 1.** Top-left: Intercept parameters. Bottom-left: Main effect parameters for *Cultural Norms*. Top-right: Main effect parameters for *Religiosity*. Bottom-right: Interaction term parameters. Posterior means (dots) and posterior distributions (grey). Lower and upper bounds (black lines and brackets) represent 95% *credible intervals*. Y-axes are estimated outcomes on raw *Well-Being* score (10-50). Countries are estimate-ranked in each sub-plot.

attention check. However, since the single *Religiosity* item is an ordered categorical variable[10], we now model this variable as a monotonic function (Bürkner and Charpentier, 2020), with a weakly informative prior in the form of a Dirichlet distribution with $\alpha = \{2, 2, 2, 2, 2, 2\}$[11], a conventional choice for ordered categorical variables (McElreath, 2020, p. 391-396).

Appendix B.1. Results

In Figure B.7, we report results from the "Social desirability model". The intercepts represent the estimated *Well-Being* scores for a male that passed the attention check, who is (globally) minimally ("behaviorally") religious and is country-average in age, perceived socioeconomic status and education. The posterior means for *Religiosity* reflect the estimated *Well-Being* score for a (globally) maximally ("behaviorally") religious male with country-average demographic characteristics who passed the attention check. Despite substantial between and within-country variation – as in the main models – the association between *Religiosity* and *Well-Being* is notably reduced compared to the main results. In several countries, there is substantial probability mass on both sides of zero, suggesting a weak or no relationship.

Appendix B.2. Conclusion

Although one should be careful with drawing conclusions from this simple analysis, the results could be taken as evidence for the idea that behavioral aspects are less susceptible to demand-characteristics in a survey context, such as social desirability (Jones and Elliott, 2017; Møller et al., 2020). Of course, an alternative interpretation is that *Religiosity* (here operationalized as self-reported service attendance) is simply a more unreliable predictor for physical well-being than for psychological well-being. Indeed, for our main models we deliberately picked measures of psychological and social well-being under the assumption that these are measures where religiosity might have notable influence. As they say: More research is needed.

[10] In the MARP data all religiosity items were transformed from their original categorical scale to a continuous 0-1 variable so as to be comparable across differing scale ranges. Here, we back-transformed the selected variable to recover the original categorical form in order to model it monotonically.

[11] The item is on a scale from 1 to 7, and the Dirichlet prior takes a series of α's equalling the number of response options minus 1 (in this case, $7 - 1 = 6$).

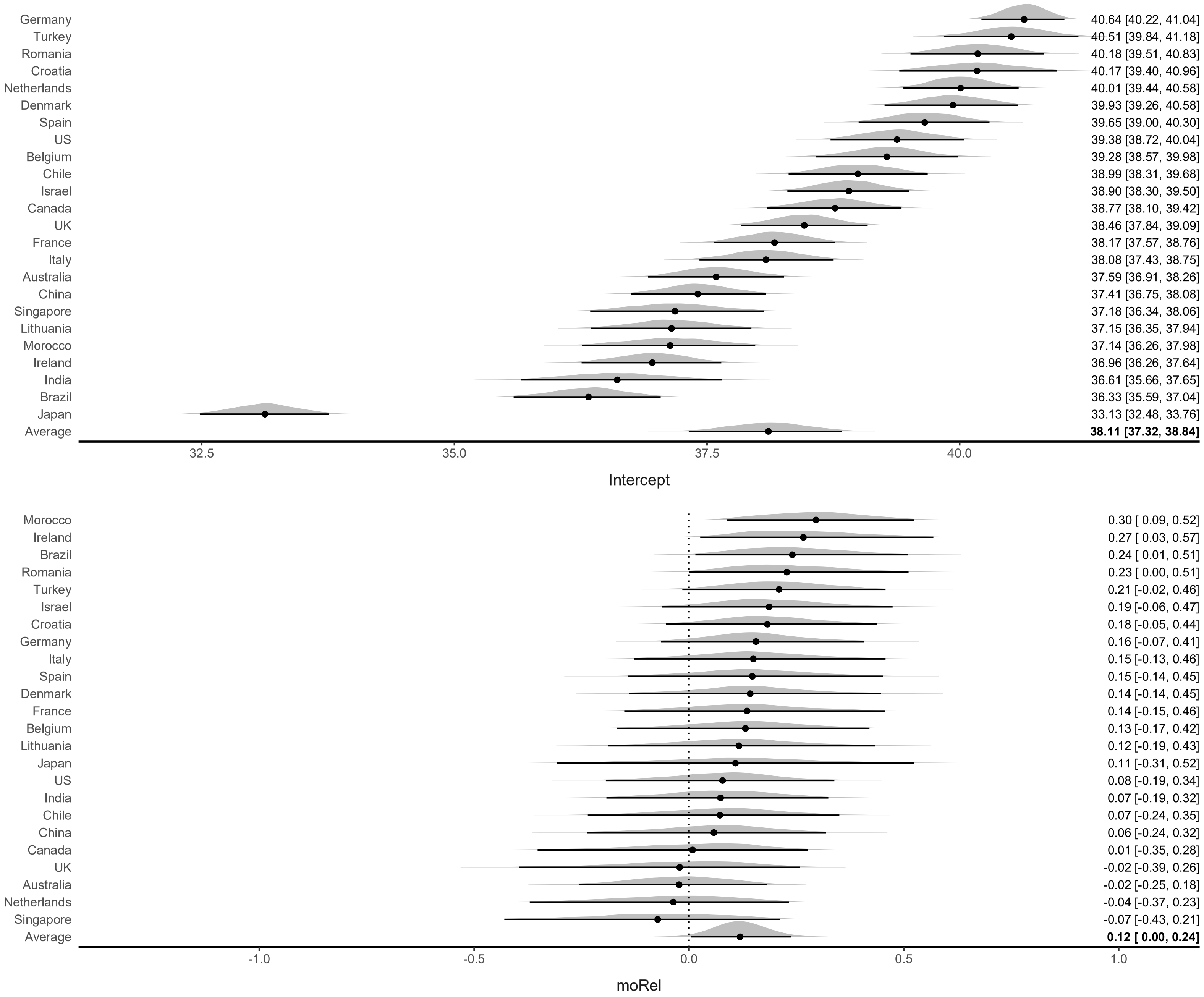

Figure B.7: **Country-level results for "Social desirability model".** Posterior means (dots) and posterior distributions (grey). Lower and upper bounds (black lines and brackets) represent 95% *credible intervals*. Y-axes are estimated outcomes on *Well-Being* score (10-50). Countries are estimate-ranked.

References

Achen, C.H., 2005. Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong. Conflict Management and Peace Science 22, 327–339.

Bürkner, P.C., 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software 80, 1–28. doi:10.18637/jss.v080.i01.

Bürkner, P.C., 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal 10, 395–411. doi:10.32614/RJ-2018-017.

Bürkner, P.C., 2020. Handle Missing Values with brms. URL: https://cran.r-project.org/web/packages/brms/vignettes/brms_missings.html.

Bürkner, P.C., Charpentier, E., 2020. Modelling monotonic effects of ordinal predictors in Bayesian regression models. British Journal of Mathematical and Statistical Psychology 73, 420–451. doi:https://doi.org/10.1111/bmsp.12195.

Buuren, S.v., Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45, 1–67. doi:10.18637/jss.v045.i03.

Gabry, J., Mahr, T., 2021. bayesplot: Plotting for bayesian models. URL: https://mc-stan.org/bayesplot/. r package version 1.8.0.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A., 2019. Visualization in bayesian workflow. J. R. Stat. Soc. A 182, 389–402. doi:10.1111/rssa.12378.

Jones, A.E., Elliott, M., 2017. Examining Social Desirability in Measures of Religion and Spirituality Using the Bogus Pipeline. Review of Religious Research 59, 47–64. doi:10.1007/s13644-016-0261-6.

Kay, M., 2020. tidybayes: Tidy Data and Geoms for Bayesian Models. URL: http://mjskay.github.io/tidybayes/, doi:10.5281/zenodo.1308151. r package version 2.3.1.

McElreath, R., 2020. Statistical Rethinking: A Bayesian course with examples in R and Stan. Second ed., CRC Press.

Møller, A.B., Pedersen, H.F., Ørnbøl, E., Jensen, J.S., Purzycki, B.G., Schjoedt, U., 2020. Beyond the Socially Desirable: Longitudinal Evidence on Individual Prayer-Wellbeing Associations. The International Journal for the Psychology of Religion 30, 275–287. doi:10.1080/10508619.2020.1753330.

Pfadt, J.M., van den Bergh, D., Goosen, J., 2020. Bayesrel: Bayesian Reliability Estimation.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Stan Development Team, 2020. RStan: the R interface to Stan. URL: http://mc-stan.org/. r package version 2.21.2.

Van Buuren, S., 2018. Flexible imputation of missing data. Second ed., CRC press.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27, 1413–1432. doi:10.1007/s11222-016-9696-4.

Vuorre, M., 2018. brmstools. Deprecated.

Westreich, D., Greenland, S., 2013. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. American Journal of Epidemiology 177, 292–298. doi:10.1093/aje/kws412.