Many Analysts Religion Project: Stage 1

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

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

1. Introduction

This write-up and accompanying `R` code serve as a pre-registration for the Many Analysts Religion Project (for further details, see `http://bit.ly/MARPinfo`). The overall research questions it asks are: 1) Do religious people report higher well-being? 2) Does the relation between religiosity and well-being depend on how important people consider religion to be in their country (i.e., perceived cultural norms of religion)? In this document, we detail our analysis plan and to convey a sense of what we expect our results to resemble, we also provide some details of a simple simulation we performed based on the provided Data Documentation.

2. Analysis Plan

Informally following a Bayesian workflow (e.g., Gelman et al., 2020), we will perform the following steps in an iterative manner: 1) initial model formulation and specification; 2) prior predictive checks; 3) modify priors; 4) provisionally accept priors; 5) diagnostic checks; 6) posterior predictive checks; 7) model comparison; and 8) plotting and inference for focal coefficients. There are, however, a few assumptions that must be checked before we proceed with this workflow after obtaining the data. These have to do with the relationship between the target predictors about religiosity and the cultural norm variables. We therefore propose to first assess the internal reliability of these items. Specifically, we need to ensure that a) the questions that comprise our target variables are, in fact, intercorrelated, and b) the target predictors are not tokens of the same underlying construct.

2.1. Test Item Reliability

Since we will be using multiple-item constructs, the first order of business is to run test-item reliability analyses on the focal data. This will assess and ensure appropriate internal construct validity. To do this, we will conduct Bayesian reliability analyses using the `Bayesrel` package (Pfadt et al., 2020) for `R` (R Core Team, 2020) to obtain a McDonald's omega with 95% credible intervals. For the purposes of our plan, we operationalize the focal constructs as follows (cf., the Data Documentation):

- *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 will add up 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 from 10 (min.) to 50 (max.).

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

- *Religiosity*: Our primary predictor variable will be the summation of all the Religiosity items, except variables **rel_3** and **rel_4**, as these two variables fall outside the scale format of the remaining items. This summation is then transformed onto a 0-1 scale.
- *Cultural Norms*: We include the summation of both items, **cnorm_1** and **cnorm_2**. This summation is then transformed onto a 0-1 scale.

2.2. General Model Details

As we are taking the summation of the well-being items, all specifications model our outcome as distributed normally with mean μ_i and standard deviation σ. All varying effect coefficient parameters' priors are distributed with multi-variate normal distributions defined by their respective means, μ_p and the variance-covariance matrix $\mathbf{SRS}$, where $\mathbf{S}$ is a diagonal matrix of each parameters standard deviations, σ_p, and $\mathbf{R}$ is a correlation matrix with a defined prior distribution in the LKJcorr family (Lewandowski et al., 2009) where $\eta = 2$. The remaining priors are defined as follows: $\mu_p \sim \text{Normal}(0, 1)$, $\sigma_p \sim \text{Exponential}(1)$, and the main intercept, $\alpha \sim \text{Normal}(30, 2)$. We define the prior of α in this manner because it effectively serves as a midpoint for the possible values on the scale. All specifications treat country as varying intercepts with various models treating religiosity and/or cultural norms as varying effects.

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 interpreting these covariates as (causal) effects; see also Westreich and Greenland, 2013), we fit models both with and without demographic variables. The *Religiosity* and (in Scenario 1 discussed below) *Cultural Norms* items are transformed onto a 0-1 scale for ease of interpretation (i.e., holding these variables at zero is meaningful).

In principle, one could have specified a *fully* varying effects model, where all covariates, including the demographics and attention check, interacted with varying intercepts and slopes across countries. Arguably, this might be a more realistic approximation of the actual data-generating processes. However, for ease of interpretation and as we have no clear theoretical justifications to do so, we refrain from this here. Our full model in Scenario 1 (Model 1 below) therefore only includes varying slopes and intercepts for the focal terms, namely religiosity and perceived cultural norms of religion and their interaction. We will not attempt interpretation of the demographic covariates, other than possibly for exploratory purposes and validity checks.

Note also that all our models assume the outcome (*Well-Being*) to be approximately normally distributed. While we find this a reasonable preliminary assumption, we recognize that the actual data may call for different outcome distributions, such as a Student's t or beta. In that case, we will then fit such models in addition to the pre-registered models.

2.3. Two Scenarios

While self-reported religiosity and perceived cultural norms of religiosity are presented as two separate constructs in the project documentation, they may, in fact, be largely overlapping (see, for example, the similar wording of items **rel_8** and **cnorm_1** and **rel_9** and **cnorm_2**). Because we do not yet know if the *Religiosity* questions and the *Cultural Norms* questions are in fact measuring the same underlying construct (e.g., a perceived general importance of religion or propensity toward conformity), we plan to perform reliability analyses on the combined *Religiosity* and *Cultural Norms* items using the same procedure in the previous subsection. If they are not substantially overlapping (denoted *Religiosity* $\neq$ *Cultural Norms* below), we follow Scenario 1 (see below), whereas if they are substantially overlapping (denoted *Religiosity* $\approx$ *Cultural Norms* below), we follow Scenario 2 (see below). By "substantially overlapping" we mean that the *Religiosity* and *Cultural Norms* items come out as nearly-indistinguishable in reliability and drop-item analyses.

Below, we detail our focal models in `brms` syntax. Here, the outcome variable is on the left side of the tilde. 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)`.

2.3.1. Scenario 1: Religiosity ≠ Cultural Norms

Model 1[1]. 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.c`), perceived socioeconomic status (`SES.c`), education (`Edu.c`), and attention check (`AC.r`, reverse-scored so that "passed" is the reference group).

```
m1 <- brm(WB ~ Rel*Norm + (Rel*Norm | Country)
        + Gender + Age.c + SES.c + Edu.c + AC.r,
        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.r, ...)
```

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.c + SES.c + Edu.c + AC.r, ...)
```

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

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

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

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

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

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

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

2.3.2. *Scenario 2: Religiosity ≈ Cultural Norms*

If the *Religiosity* and *Cultural Norms* items substantially overlap (see above), we will not be able to straightforwardly assess whether the relationship between religiosity and well-being is moderated by perceived local importance of religion. Therefore, in Scenario 2, we include the *Cultural Norms* items in the *Religiosity* scale, and then group-mean center *Religiosity*. By doing this, we interpret the analytical output as *Religiosity's* effect on individual well-being as a deviation from group-mean religiosity. This formulation indirectly addresses the "person-culture fit" hypothesis presented in the project's Theoretical Background document, with the group-mean centered religiosity serving as a proxy for locally perceived cultural importance of religion.

Model 7[2]. This is the full model that allows varying effects of group-mean centered religiosity (`Rel.c`) across countries and simple effects for demographic variables and the attention check (see Model 1 description).

```
m7 <- brm(WB ~ Rel.c + (Rel.c | Country)
        + Gender + Age.c + SES.c + Edu.c + AC.r,
        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.c + (Rel.c | Country) + AC.r, ...)
```

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

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

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

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

2.4. *Model Comparison*

Our modelling approach includes fitting a series of models ranging in different levels of complexity. Such an approach calls for some kind of metrics whereby models can be compared for their respective fit and predictive performance. However, to our knowledge, there exists no universally applicable and accepted set of model comparison metrics (e.g., McElreath, 2020, ch. 7).

One recently proposed technique is approximate leave-one-out cross-validation (LOO) (Vehtari et al., 2017). Simply put, LOO involves withholding data points and checking how the fitted model performs in predicting the held-out data. LOO is gaining popularity, but has also recently come under criticism, for instance for performing less well in "idealized" data sets with simple variable

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

structures and samples sizes nearing infinity (e.g., Gronau and Wagenmakers, 2019a). Other researchers argue for alternative strategies, such as Bayes Factor (e.g., Gronau and Wagenmakers, 2019b). A "Bayes factor B comparing an alternative hypothesis to the null hypothesis means that the data are B times more likely under the alternative than under the null" (Dienes, 2014, p. 4). However, Bayes Factor itself is criticized on a number of grounds, for instance the assumption that one of the models under consideration represents the "true" data-generating process (e.g., Vehtari et al., 2019).

For the sake of this pre-registration, we commit to approximate leave-one-out cross-validation with corresponding information criteria and Akaike model weights, using the `loo` (Vehtari et al., 2017) and `brms` packages (see accompanying `R` code for simple implementation). These algorithms will return: expected log posterior density, ELPD (a higher value is better), a measure of overall model fit and out-of-sample predictive accuracy, also known simply as the log score or log-likelihood (e.g., Gelman et al., 2014); the difference in ELPD values with the best model as the reference; the standard error (SE) of the ELPD differences; 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 based on the LOOIC score (a higher value is better) (cf., e.g., Bürkner and Charpentier, 2020). We do not, however, exclude the possibility of applying additional model comparison techniques as well.

2.5. Missing Values

There are few missing values (NA henceforth) in the target variables. For those we include in our models, only age (27 NAs) and perceived socioeconomic status (5 NAs) have missing values (cf., the Data Documentation). As our modelling approach relies on comparison of several models varying in complexity and number of variables, and since such model comparison may be sensititive to the exact sample size, we will conduct complete-cases analyses, unless the missing values are clustered (e.g., in certain countries). If they are clustered, however, we will perform imputation of the missing data points, using the imputation function in the `brms` package (Bürkner, 2017, 2018) building on the `mice` package (Buuren and Groothuis-Oudshoorn, 2011) for `R`.

3. Parameter recovery and model comparison on simulated data

For the sake of this pre-registration, we illustrate and perform prior predictive checks and model validation on simulated data that takes a simplified form of the actual data set (as deduced from the Data Documentation). We will analyze the data using the `brms` package, an `R` interface to `Stan` (Stan Development Team, 2020), using weakly informative priors.

3.1. Parameter Recovery for Simulated Data

Figure 1 plots coefficients returned for the full model in Scenario 1, Model 1, fitted to the simulated data. In short, all analyses across Models 1-10 return reasonable diagnostics and estimates that are sensible considering the arbitrary simulated values (see `R` code below and `.R` file for further details)[3].

[3]The sign of attention check (AC.r) is reversed as an artifact of reverse-scoring the variable. See `.R` file for details.

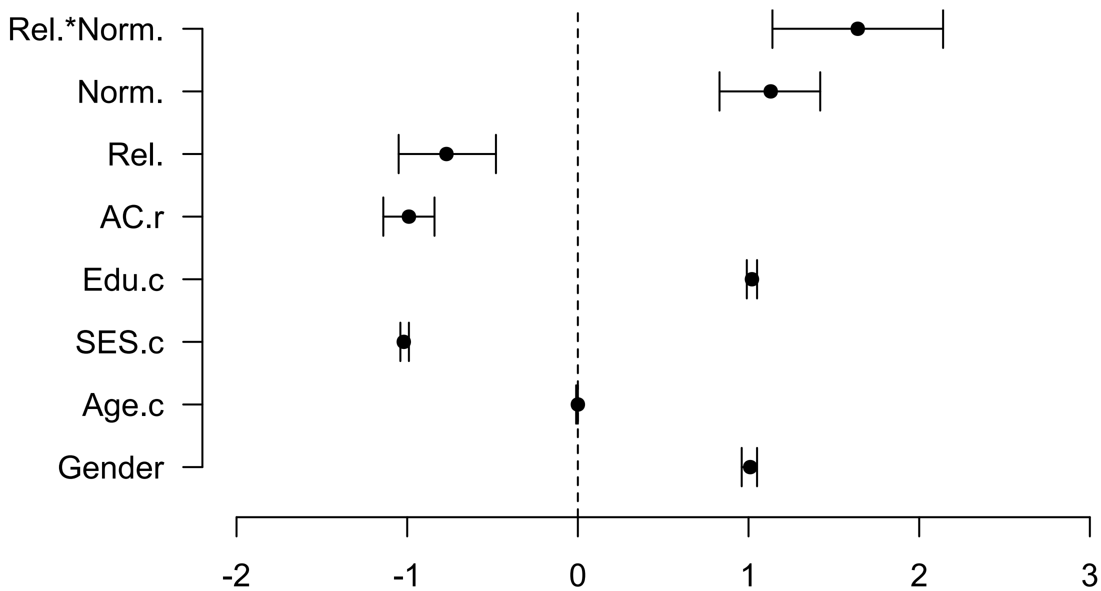

Figure 1: **Parameter recovery from Model 1.** We do not include the intercept here since the correctly recovered estimate (≈30) is out of bounds for the plot.

We set the simulated effects as follows:

```
alpha <- rnorm(N, mean = 30, sd = 2) # Intercept
beta1 <- -1 # Religiosity
beta2 <- 1 # Cultural Norms
beta3 <- 2 # Interaction btw. Rel and Norm
beta4 <- 1 # Attention Check
beta5 <- 1 # Gender
beta6 <- 0 # Age
beta7 <- 1 # Education
beta8 <- -1 # SES

WB <- alpha + beta1*Rel + beta2*Norm + beta3*Rel*Norm + beta4*AC
        + beta5*Gender + beta6*Age + beta7*Edu + beta8*SES
```

Of course, our simulated dataset is idealized (e.g., no multicollinearity among the predictors) and crude (e.g., we did not specify varying effects across countries), and successful parameter recovery is therefore no guarantee that our models will perform well in a more complex real-world dataset.

3.2. Model Comparison on Simulated Data

In Table 1, we present model comparison results from Scenario 1 models, Models 1-6. Encouragingly, LOO prefers the model that best resembles the “true” simulated data (Model 3).

	ELPD difference	SE of EPLD difference	P-LOO	LOOIC	Akaike-Weight
Model 3	–	–	27.5	40510.1	0.602
Model 1	-0.4	1.18	37.7	40511.0	0.398
Model 5	-21.4	6.2	26.7	40553.0	0.000
Model 4	-4608.5	76.1	15.5	49727.2	0.000
Model 2	-4609.3	76.1	22.2	49728.8	0.000
Model 6	-4615.9	76.2	14.7	49742.0	0.000

Table 1: **Model comparison of Models 1-6 fitted to simulated data.** 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).

However, Model 3 performs near-identically to the most complex model, Model 1 (but note that Model 3 out-competes Model 1 on P-LOO, the effective number of parameters). It is well-known that model comparison metrics often favor the more complex models (e.g., McElreath, 2020, ch. 7). Our synthetic results indicate that allowing countries to have varying slopes for the interaction term does not increase predictive performance, which is the "right" answer in this context. Model 5, which does not include an interaction term, performs noticeably worse, and Models 2, 4 and 6 perform worse still, likely because they do not include the demographic variables and, for Model 6, the interaction term.

Acknowledgments

Bendixen and Purzycki thank the Aarhus University Research Foundation for support.

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., 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.

Dienes, Z., 2014. Using Bayes to get the most out of non-significant results. Frontiers in Psychology 5. doi:10.3389/fpsyg.2014.00781.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing 24, 997–1016. doi:10.1007/s11222-013-9416-2.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.C., Modrák, M., 2020. Bayesian Workflow. arXiv:2011.01808 [stat] URL: http://arxiv.org/abs/2011.01808.

Gronau, Q.F., Wagenmakers, E.J., 2019a. Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. Computational Brain & Behavior 2, 1–11. doi:10.1007/s42113-018-0011-7.

Gronau, Q.F., Wagenmakers, E.J., 2019b. Rejoinder: More Limitations of Bayesian Leave-One-Out Cross-Validation. Computational Brain & Behavior 2, 35–47. doi:10.1007/s42113-018-0022-4.

Lewandowski, D., Kurowicka, D., Joe, H., 2009. Generating random correlation matrices based on vines and extended onion method. Journal of multivariate analysis 100, 1989–2001.

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

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

R Core Team, 2020. 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.

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.

Vehtari, A., Simpson, D.P., Yao, Y., Gelman, A., 2019. Limitations of "Limitations of Bayesian Leave-one-out Cross-Validation for Model Selection". Computational Brain & Behavior 2, 22–27. doi:10.1007/s42113-018-0020-6.

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.