On the Robustness of Reciprocal Associations Between Personality and Religiosity in a German Sample

Richard E. Lucas[1] & Julia M. Rohrer[2]

[1] Department of Psychology, Michigan State University

[2] Wilhelm Wundt Institute for Psychology, Leipzig University

Author Note

Correspondence concerning this article should be addressed to Richard E. Lucas, 316 Physics Rd., Michigan State University, East Lansing, MI 48823. E-mail: lucasri@msu.edu

Abstract

Objective: Entringer, Gebauer, and Kroeger (2023) used longitudinal data from a German panel study to examine reciprocal causal effects between personality and religiosity, along with cultural moderators of these effects. The current paper examines the robustness of the original effects to alternative model specifications.

Method: We reanalyzed the same 4-wave data spanning 12 years (total N = 46,316), first replicating the original cross-lagged panel analyses and then extending these analyses in three ways: Using a random-intercept cross-lagged panel model, using observed rather than latent variables, and modeling each trait individually rather than simultaneously.

Results: Correlations between personality and religiosity were all small in size, even when aggregating over 12 years. Lagged effects were very small, and none was robust across all model specifications. Cultural moderators also depended on model specifications.

Conclusions: The very small size of these reciprocal effects, along with their sensitivity to model specifications, suggest that conclusions about causal effects of personality and religiosity should be drawn very cautiously.

Keywords: personality, religiosity, cross-lagged panel model

On the Robustness of Reciprocal Associations Between Personality and Religiosity in a German Sample

A common goal in personality research is to identify robust associations between personality characteristics and consequential behaviors and outcomes. Identifying these associations allows researchers to develop and test hypotheses that inform personality theories. For instance, researchers may study the links between a trait like conscientiousness and an outcome like job achievement. A greater understanding of the processes underlying this association can inform theories about how conscientiousness affects outcomes in the real world. In addition, this research could provide practical guidance for those seeking to improve achievement levels. Moreover, a consideration of the reverse causal direction—understanding whether achievement experiences impact trait levels—can inform theories of personality development and change. Thus, studies that examine the processes underlying such links have great potential further the understanding of how personality characteristics shape peoples lives, and how life experiences shape personality. Because personality traits are stable over time, however, they can be difficult to manipulate experimentally, which means that testing these processes can be challenging. Personality researchers often rely on longitudinal analyses to further their understanding of the causal processes that might underlie such associations.

Recently, Entringer et al. (2023) conducted such an examination, investigating the links between the Big Five personality traits and religiosity in a very large German sample. This research was motivated both by prior theories meant to explain how personality can shape religiosity and by theories that posit that religiosity can affect personality. Considering the effects of personality on religiosity, for instance, a niche-picking perspective suggests that people who have personality traits that are consistent with the behaviors that are typically exhibited in religious contexts should gravitate towards these religious contexts. In addition, the Sociocultural Norm Perspective (Eck & Gebauer, 2022) posits that

personality traits like agreeableness, conscientiousness, and (low) openness to experience produce normative behaviors; when religiosity is normative in a culture, then these traits should cause greater religiosity. Considering the effects of religiosity on personality, complementary theories suggest that religiosity itself can impact these same traits, as religious contexts also promote or even enforce behaviors and views that are consistent with traits like agreeableness and conscientiousness.

To test these ideas, Entringer et al. (2023) relied on a widely used approach for examining reciprocal causal effects in panel data: the cross-lagged panel model (CLPM, Heise, 1970). In the CLPM, each variable (in this case, personality and religiosity) at each occasion is predicted from the same variables assessed at prior waves. With additional assumptions (e.g., no unobserved confounders), lagged associations from one variable to the other (e.g., from Time 1 personality to Time 2 religiosity) can be interpreted as causal effects. Entringer et al. (2023) found evidence for mutual causal effects of agreeableness on religiosity and religiosity on agreeableness. Moreover, the religiosity of the region in which respondents lived moderated the links between some personality traits and religion, including the effects of openness and conscientious on religiosity and religiosity on openness.

Entringer et al.'s (2023) study had a number of desirable features that make it especially well-suited to examining questions about reciprocal associations between personality and religiosity. First, the authors used a very large panel study with four waves of assessment over a period of twelve years. These features should contribute to the robustness of the results. Moreover, the authors used a sophisticated latent-variable version of the CLPM that accounts for measurement error, which can help reduce the likelihood of spurious lagged associations (Lucas, 2023). In addition, the authors examined these associations separately in different German federal states that varied in overall religiosity, which allowed them to examine theoretically relevant contextual moderators of these associations. Finally, the authors conducted many robustness checks to support their primary findings.

Questions About Robustness

Despite these strengths, however, there are reasons to question the robustness of the support for reciprocal causal effects between personality and religiosity. First, when examining these reciprocal effects, Entringer et al. (2023) relied solely on the traditional CLPM, which only includes lagged associations with variables assessed at the immediately prior wave. The use of the CLPM has been championed on the grounds that it tests "between-person prospective effects" (Orth, Clark, Donnellan, & Robins, 2021), whereas some alternative models do not. However, it is questionable whether these between-person causal effects are clearly defined (see Lucas, 2023, for an argument against this interpretation). Moreover, it is unlikely that the CLPM can appropriately test such effects even if they could be defined. This is because a critical assumption underlying the CLPM is that there are no sources of stability in the measures of interest beyond the two that are included in the model: the autoregressive effects reflecting the stability of each variable and the lagged effects between variables. If stable-trait variance also exists in the measures being modeled, then this assumption is violated and the CLPM results in bias lagged associations (Heise, 1970). This bias would invalidate the interpretation of the lagged paths as estimates of the causal effects (regardless of whether those effects are described as between-person or within-person effects).

Recently, Lucas (2023) used simulations to show that this problem is quite severe; spurious lagged associations can be found as often as 100% of the time in realistic scenarios (e.g., when there is just a moderate amount of stable-trait variance, when the stable-trait variance from the two waves is moderately correlated, when sample sizes are moderate to large, and when multiple waves of assessment are included). The size of the bias found in these simulations is high relative to the size of the estimated effects reported by Entringer et al. (2023). Notably, two strengths of Entringer et al.'s study (the very large sample size and the use of a four-wave design) can actually increase the likelihood of finding spurious lagged

effects (Lucas, 2023).

A second issue is that Entringer et al. (2023) chose to model all five traits simultaneously when predicting changes in religiosity. Although the Big Five traits are hypothesized to be relatively independent, in empirical data they are usually not. Thus, when modeling all traits simultaneously, estimated paths from personality to religiosity reflect associations that persist after controlling for all other personality traits. The decision to control for correlated variables comes with interpretational challenges, however, as the association can only be interpreted as an association between religiosity and the variance that is not shared with other traits (Lynam, Hoyle, & Newman, 2006). The nomological networks surrounding specific constructs are rarely developed using residualized trait scores, and thus, the conceptual connections between this residualized variable and the hypothetical construct it is meant to assess are not always clear. Entringer et al. (2023) justified their decision to simultaneously model all five traits by noting that this decision is consistent with prior research. While such comparability is often desirable, it does not mean that the analytic choice is substantively justified. Because of the interpretational challenges, our preference is to interpret unadjusted associations; but at the very least, robustness across modeling choices is important to consider.

A final issue is in regard to the size of the effects that Entringer et al. (2023) found. The authors foreshadow this issue early in the paper, noting that lagged effects are typically much smaller than cross-sectional effects and that prior cross-sectional correlations should often provide an upper bound on the size of any lagged associations[1]. They argue that because the correlations between personality and religiosity tend to be small (e.g., around .19 in their review), the lagged effects should be even smaller. Indeed, the observed lagged effects in their study were quite small, with maximum standardized regression coefficients of

[1] It is important to note that although this would be true under certain very plausible assumptions, the lagged effects could potentially be larger than the aggregated effect if there were negative autoregressive effects or unobserved confounders with associations with the opposite sign.

.039. The estimated moderator effects they found were similarly small in size.

Although we agree that small effects can sometimes be important, such effects provide challenges for interpretation. First, just because effects that are small in size can be important in some contexts does not mean that they are always important; justification for why a particular small effect is important is needed. One common defense of the importance of small effects is that these effects accumulate over time. However, this would usually mean that aggregated between-person correlations should themselves be reasonably large in size. Entringer et al. (2023) did not report zero-order correlations either within waves or after aggregating personality and religiosity across the 12-year period.

Perhaps more importantly, very small effect sizes may simply reflect biases induced by subtle model misspecification or residual confounding rather than true underlying effects. For instance, although Entringer et al.'s (2023) decision to model latent personality factors when examining reciprocal associations has the desirable feature of removing measurement error, it also comes with a cost in terms of model complexity. Although researchers might hope that the items of their measures load cleanly on the factor to which the item belongs (and not to any other), this is not always the case in practice. In such cases, allowing for secondary loadings may be necessary to improve model fit, though questions can remain about whether such post hoc modifications capitalize on chance. Moreover, if the measurement model that is chosen does not fit well, then this misspecification can bias the structural parameters (e.g., lagged effects, see Rhemtulla, van Bork, & Borsboom, 2020). In these cases, questions about the robustness of estimates (especially estimates that are very small in size) can be raised. In the current study, we compare the latent-variable model used by Entringer et al. (which included three cross-loadings and allowed residuals of identical items to be correlated over time) to a simpler observed-variable model that includes only the observed mean scores for each personality trait measure.

The Present Study

In many ways, Entringer et al.'s (2023) paper represents a model example of an attempt to answer the questions they set out to address. They were clear about the causal effects in which they were interested and they carefully described their approach. They shared all code for reproducing their analyses, and they took many steps to ensure the robustness of their results. Nonetheless, additional questions can be raised about effect sizes, model complexity, the decision to model all five personality traits simultaneously, and—most importantly—the decision to use a traditional CLPM instead of a model that accounts for additional sources of stability. The goal of the current analysis is to test the robustness of these results to alternative specifications.

We first simply examine the correlations between religiosity and each of the Big Five traits when each is aggregated across all waves. This provides a simple index of effect size that helps establish whether any lagged causal effects accumulate over time. Next, we test a series of models that examine the robustness of the results reported in Entringer et al. (2023). Specifically, after first replicating their results, we then test all possible combinations of three model modifications. We compare models where the Big Five traits are modeled as latent-traits (using the same measurement model in the original paper, including secondary loadings) versus when they are modeled as observed variables. We compare models where the Big Five traits are entered simultaneously as predictors to those where separate models are run for each trait individually. Finally, we compare the CLPM to the random-intercept cross-lagged panel model (RI-CLPM, Hamaker, Kuiper, & Grasman, 2015), which includes a random intercept to account for stable-trait variance. Combining each pair of comparisons results in eight separate models with results for each of the Big Five traits. In addition, Entringer et al. (2023) examined the moderating contextual effect of state-level religiosity. We also test these moderating effects for each of the Big Five traits. The results will then be compared for robustness.

Methods

This paper uses data from the German Socio-Economic Panel (SOEP) study, which assessed the Big Five personality traits and religiosity four times, at four-year intervals. The inclusion of four waves of assessment allows for the use of both the CLPM and the more complex RI-CLPM. Entringer et al. (2023) provided detailed code for their analyses, however, they did not provide code to extract and clean data from the raw data files. Thus, as an additional test of the reproducibility of the analyses from the description provided in the text, we developed our own code for extracting and cleaning the data based on the decision rules reported in the text. We also relied on a more recent release of the dataset (Version 37, versus Version 35 used in the original paper), though this should not impact the analyses, as the variables of interest were not included in the additional two waves. As we discuss in more detail below, the exact number of participants included in the final samples and the precise estimates from the original model were not perfectly reproduced. For the full sample, this is due primarily to the inclusion of two additional federal states. Yet even for the overlapping states, final sample sizes differed slightly from the original. It is important to note, however, both the final sample sizes and results are quite close to those from the original paper (see below for details), and all substantive conclusions from the original paper were supported in analyses that use the same model. Thus we proceeded with our robustness checks using the sample we extracted. Our full code for extracting, cleaning, and analyzing these data is available at: https://osf.io/uyb7p/.

Participants

The SOEP is a long-running panel study of households in Germany. Households are contacted yearly, and all adult members of sampled households are asked to participate. There were 37 waves of data available for our analyses, and this resulted in a total sample size of 46,316 (versus 44,485 in the original paper). Sample sizes for each of the 16 federal states are reported in Table 1. Additional details about the sample are reported in the

Table 1

Within-state correlations between each personality trait and religiosity.

	Correlation with Religiosity						
State	Agr	Con	Ext	Neu	Opn	Religiosity	N
Schleswig-Holstein	0.10	0.13	0.01	-0.01	0.05	1.59	1605
Hamburg	0.08	-0.01	-0.04	-0.02	0.02	1.52	677
Lower Saxony	0.11	0.08	0.00	0.00	0.04	1.73	4459
Bremen	0.07	-0.01	0.05	0.01	0.04	1.65	336
North Rhine-Westphalia	0.13	0.09	-0.01	0.00	0.02	1.78	9968
Hesse	0.14	0.10	0.01	-0.01	0.03	1.80	3300
Rhineland-Palatinate	0.08	0.06	-0.03	0.00	0.01	1.83	2081
Baden-Wuerttemberg	0.10	0.05	-0.02	0.01	-0.01	1.87	5672
Bavaria	0.07	0.06	-0.05	0.00	-0.04	1.96	7248
Saarland	0.09	0.14	0.04	0.02	0.07	1.77	428
Berlin	0.10	0.08	0.06	-0.05	0.08	1.45	1601
Brandenburg	0.08	-0.01	0.01	-0.02	0.11	1.34	1738
Mecklenburg-Western Pomerania	0.09	0.04	-0.03	-0.02	0.06	1.28	967
Saxony	0.04	0.00	-0.01	0.01	0.10	1.40	2848
Saxony-Anhalt	0.11	0.07	0.00	0.00	0.14	1.33	1637
Thuringia	0.07	0.04	-0.02	0.03	0.11	1.46	1751
Pooled Within	0.10	0.07	-0.01	0.00	0.03		46316
Raw Correlation	0.10	0.06	-0.01	-0.01	0.03		46316

Note. Agr = Agreableness; Con = Conscientiousness; Ext = Extraversion; Neu = Neuroticism; Opn = Openness. Mean religiosity and sample sizes are presented in the rightmost columns.

original paper. Consistent with the original report, we only included participants who lived in the same federal state throughout the entire study.

Measures

Consistent with the original report, we focus on two measures: A single-item measure of religiosity and a 15-item measure of the Big Five personality traits. The religiosity item is: "How often do you attend church, religious events?" Responses are recorded on the following scale: 1 = at least once a week, 2 = at least once a month, 3 = less often, and 4 = never.

This item was reverse scored so higher scores indicate higher levels of religiosity. An index of state-specific cultural religiosity was calculated by averaging religiosity scores of all participants within each federal state. Average scores for each federal state are presented in Table 1.

To measure personality, the SOEP includes the SOEP-BFI (Gerlitz & Schupp, 2005), which is a 15-item short-form version of the original Big Five Inventory (John, Donahue, & Kentle, 2005; John & Srivastava, 1999). Each of the 5 traits is assessed using 3 items using 7-point scales that range from strongly disagree to strongly agree. For latent-variable models, we relied on the same measurement model that Entringer et al. (2023) used (which included three cross-loadings and correlated residuals between the same item in different waves). For observed-variable models, the three items for each trait were averaged to create a single score for each trait in each year.

Open Science Disclosure

Although we cannot share the data used in these analyses, the entire dataset is freely available for scientific use upon request: https://www.diw.de/en/diw_01.c.601584.en/data_access.html. All study materials (including question wording and frequencies) are available here: https://paneldata.org. All code used to extract data from the raw data files and to analyze the data are available in the repository on the Open Science Framework: https://osf.io/uyb7p/. Note that there are two versions of the SOEP data that are made available. The full sample is only provided to researchers from the European Union, whereas some identifying variables (including state of residence) are excluded from the international dataset made available to researchers outside the European Union. As a European-Union user, the second author of this paper conducted all analyses on the full dataset.

Results

We first examine the size of the between-person correlations between each Big Five personality trait and religiosity, with the idea that if there are small (reinforcing, reciprocal) "within-person" effects between the two on a finer time scale, these could accumulate into stronger between-person associations over time (Funder & Ozer, 2019). We computed each person's average score for each personality trait across all available waves and the average religiosity score across all waves. We then examined the correlations among these variables (a) in each state, (b) by pooling the within-state associations, and (c) by examining the raw correlation among all participants, ignoring the state-level structure of these data. These correlations are reported in Table 1.

As can be seen in this table, even after aggregating over a 12-year period, all correlations are quite small. The largest correlation for the full sample is between religiosity and agreeableness, which is just 0.10. The largest correlation among the 80 tested (including those from the full sample and pooled within-state associations) was just 0.14. Thus, correlations between personality and religiosity are quite small. We next turn to models that examine the reciprocal associations between personality and religiosity.

Full Sample Analyses

In their original paper, Entringer et al. (2023) focused on results from each federal state individually and on the meta-analytic averages of these state-level associations. This allowed them to examine whether results varied across states and to test whether state-level religion moderated associations between religiosity and personality. We also use this approach, but we additionally present model comparisons in the full sample, ignoring state of residence. We focus on this latter approach for three reasons. First, as we will show, the estimates from the full sample almost perfectly replicate the meta-analytic averages across the 14 analyzed states, and focusing on one sample instead of 14 (or 16) allows for clearer model comparisons. Second, in the original analyses, two states were dropped because of small sample sizes that

resulted in problems with model convergence. When testing additional models, the specific states in which estimation problems emerge differ across models, which means that any overall differences that result could be due to differences in the models or differences in which states provide estimates. Finally, by ignoring the state-level structure, no participants needed to be excluded due to residence in states for which sample sizes were small. Again, we note that we eventually do discuss results in each state individually in the next section.

The models we tested compare estimates across three dichotomous modeling decisions: (1) The original CLPM versus the potentially more appropriate RI-CLPM, (2) Models that include latent personality traits (with a complex measurement model) versus models that include only a single observed mean score for each trait, and (c) models that include all five traits simultaneously (and hence control for correlations among traits when predicting religiosity) versus models that include just one trait per model. The combination of these factors results in eight sets of estimates for each personality trait. For models that included latent traits for all five traits simultaneously, we used the same measurement model specification included in the original paper.

Model Fit. Fit indexes for all models are presented in Table 2. Consistent with Entringer et al. (2023), we report chi-square with degrees of freedom and p-value, along with robust CFI, robust RMSEA, and SRMR. Evaluating the fit of these models is challenging, as the chi-square is sensitive to sample size and subtle misfit can lead to significant values with large sample sizes, and there are no unambiguous cutoffs for well-fitting models using the other indexes. Recommended cutoffs for the CFI are typically either .90 or .95, and the recommend cutoff for the RMSEA and SRMR is typically .05. Entringer et al. (2023) considered models with CFI values close to (and frequently below) .90 as acceptable, but we generally prefer the .95 cutoff. It is important to note that the CLPM is nested within the RI-CLPM, so the difference in fit between these pairs of models can be tested explicitly, though the large sample sizes make statistical significance testing overly sensitive. Other pairs of models are not nested and cannot be directly compared using the chi-square

Table 2

Fit indexes for full sample models.

Model	Type	Variables	Trait	ChiSq	df	p-value	CFI	RMSEA	SRMR
CLPM	Latent	All		42,804.44	1,835	0.000	0.90	0.02	0.05
RICLPM	Latent	All		33,685.49	1,814	0.000	0.92	0.02	0.04
CLPM	Observed	All		10,972.99	210	0.000	0.92	0.03	0.07
RICLPM	Observed	All		821.96	189	0.000	1.00	0.01	0.02
CLPM	Latent	Single	agr	2,738.34	90	0.000	0.96	0.03	0.04
			cns	3,059.65	90	0.000	0.97	0.03	0.04
			ext	2,745.00	90	0.000	0.98	0.03	0.04
			neu	2,317.45	90	0.000	0.97	0.02	0.03
			opn	2,948.55	90	0.000	0.97	0.03	0.04
RICLPM	Latent	Single	agr	718.73	87	0.000	0.99	0.01	0.02
			cns	915.66	87	0.000	0.99	0.01	0.03
			ext	621.07	87	0.000	1.00	0.01	0.02
			neu	301.01	87	0.000	1.00	0.01	0.01
			opn	901.73	87	0.000	0.99	0.01	0.03
CLPM	Observed	Single	agr	3,326.92	22	0.000	0.93	0.06	0.08
			cns	3,271.77	22	0.000	0.93	0.06	0.08
			ext	3,731.58	22	0.000	0.93	0.06	0.08
			neu	3,445.16	22	0.000	0.93	0.06	0.08
			opn	3,445.27	22	0.000	0.93	0.06	0.08
RICLPM	Observed	Single	agr	73.71	19	0.000	1.00	0.01	0.01
			cns	146.29	19	0.000	1.00	0.01	0.02
			ext	122.16	19	0.000	1.00	0.01	0.01
			neu	128.48	19	0.000	1.00	0.01	0.02
			opn	96.53	19	0.000	1.00	0.01	0.01

Note. Agr = Agreableness; Con = Conscientiousness; Ext = Extraversion; Neu = Neuroticism; Opn = Openness.

difference.

Table 2 shows that these modeling decisions affect fit. Although RMSEA and SRMR values for the original model are acceptable, the CFI for the CLPM with latent variables and all traits included just clears the .90 criterion and does not reach the more stringent .95 threshold. These results are consistent with those from the original paper. Indeed, in the original paper, the CFI for the primary model was even below the cutoff of .90 in 13 of 14 states. These fit indexes suggest some caution is necessary when interpreting the estimates from the models.

Notably, the RI-CLPM with latent variables and all traits included (shown in the second row of Table 2 also has a CFI value below .95 (even though the Chi-Square value is considerably—and significantly—lower for this model than for the CLPM). In fact, even the CFI for a simple measurement model does not exceed the more stringent .95 cutoff, with a value of 0.93. This is also consistent with the results in the original paper, where the CFI for a measurement model specifying metric invariance across states had a CFI value of just .92. The fact that the fit of this unconstrained measurement model is not especially strong provides further justification to examine simpler models with less complicated measurement models. Misspecification in the measurement model could affect estimates of the structural parts, including the lagged effects of interest (Rhemtulla et al., 2020).

Moving to the comparison of the CLPM and the RI-CLPM for the model with observed trait measures, the difference in fit indexes is even clearer. In this case, the CFI for the CLPM is still a borderline acceptable 0.92 whereas for the RI-CLPM it is 1. Moreover, the difference in Chi-square values for the two models is not just significant, but dramatic, dropping from 10,972.99 to 821.96 for models that differ by just 21 degrees of freedom.

It is easy to understand the source of misfit in the CLPM just by considering the implied stability coefficients from such a model and then comparing them to the actual

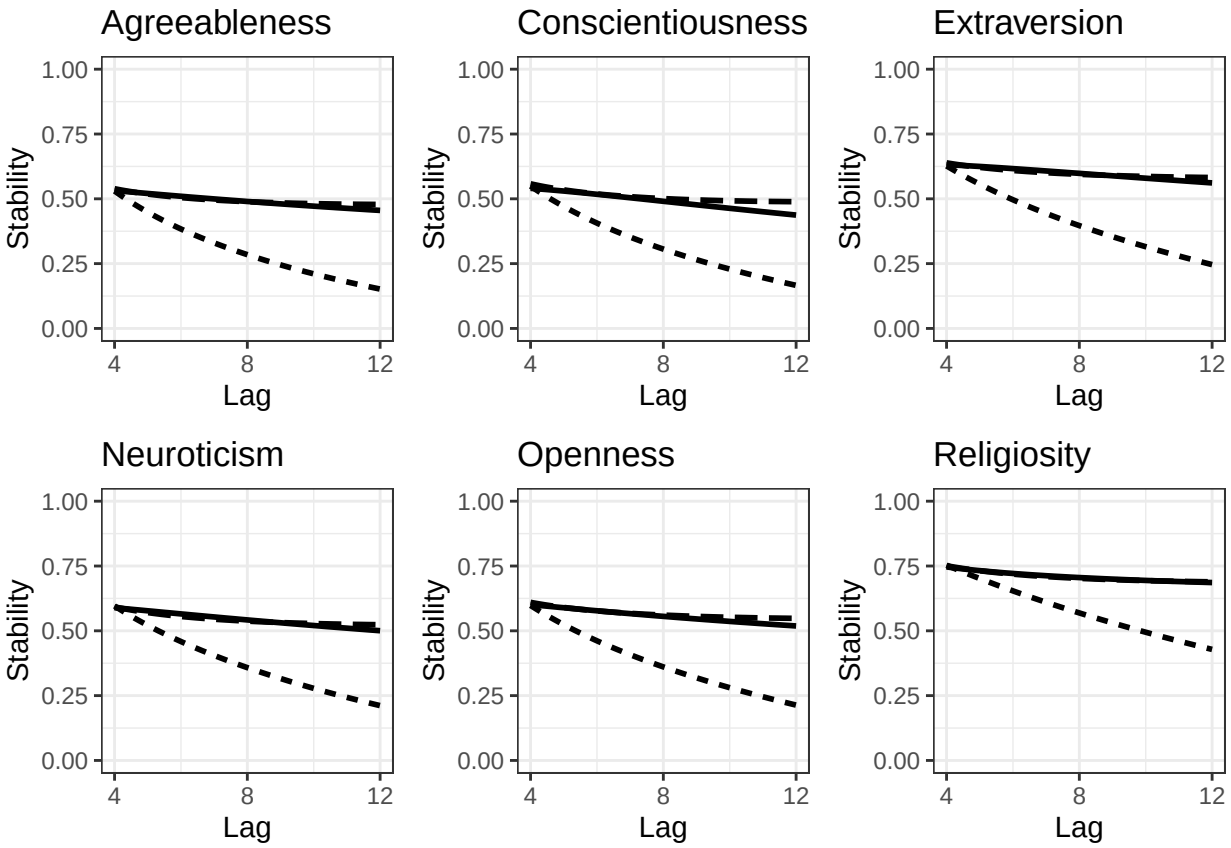

Figure 1. Actual stability coefficients (solid line) and implied stability coefficients from the CLPM (short-dashed line) and RI-CLPM (long-dashed line).

correlations from the data. Lucas (2023) noted that a major problem with the CLPM is that the model implies that stability coefficients should decline quickly with increasing lags (especially when cross-lagged paths are small), yet actual stability coefficients for most variables are quite stable over increasingly long lags. This is also true in these data, as can be seen in Figure 1. This figure plots actual stability coefficients for each variable across 4-, 8-, and 12-year intervals in the solid lines. Stability coefficients start out moderate for 4-year intervals, but decline only slightly across 8- and 12-year intervals. The implied stability coefficients from the CLPM (shown in the lines with short dashes), however, decline much faster than the actual stability coefficients, leading to a predicted stability of approximately half the actual stability for the 12-year intervals. In contrast, the RI-CLPM (shown in the

lines with long dashes) reproduces the patterns of stability coefficients almost perfectly. The CLPM is poorly suited to describing these patterns of stability, which means that the model is mis-specified. This mis-specification will then bias other parameter estimates in the model.

This same pattern of differences in fit indexes emerges in the single-trait models, with the RI-CLPM consistently outperforming the CLPM, especially in the observed-trait models (and less so in the latent-variable models, which include correlations between item-specific residuals that can capture some of the otherwise unmodeled stability in the CLPM[2]). In short, the fit indexes suggest that the measurement model for the latent-trait models may not fit the data well. In addition, the RI-CLPM consistently fits better than the CLPM, which can be explained by the fact that the CLPM cannot account for the slow decline in stability over increasingly long lags, whereas the RI-CLPM can. The final decision—whether to model all traits together versus separately—cannot be informed by fit statistics.

Estimates of Lagged Associations. Consistent with the original paper, our focus is on the cross-lagged paths from each trait to religiosity and from religiosity to each trait. Figure 2 presents the estimated cross-lagged paths from each personality trait to religiosity for each of the eight tested models. Results from the CLPM are presented in grey lines, whereas results for the RI-CLPM are presented as black lines. Results from the latent-variable models are included as solid lines whereas results for the observed-variable models are presented as dashed lines. Finally results for models that include all five traits simultaneously are labeled with triangles, whereas those that model each trait individually

[2] This highlights an interesting (but not at all uncommon) modeling decision by the authors. Although the structural part of the model assumes that the autoregressive process and lagged associations include only lagged effects from the immediately prior wave, the item-specific residuals are allowed to correlate with item-specific residuals *from all other waves*. In combination with the CLPM, this posits that while the underlying personality trait *does not* have a trait-like structure, whatever part of the items that cannot be attributed to the underlying personality trait *does* have a trait-like structure. It is not clear how one would justify such an asymmetry substantively. In any case, using the CLPM with correlated item-specific residuals forces any surplus stability into the residual correlations, thus potentially improving the empirical fit without removing the structural source of the misspecification. In contrast, using the RI-CLPM with correlated item-specific residuals allows the model to allocate any stability to both the trait and the items, depending on whatever provides the best fit to the empirical data.

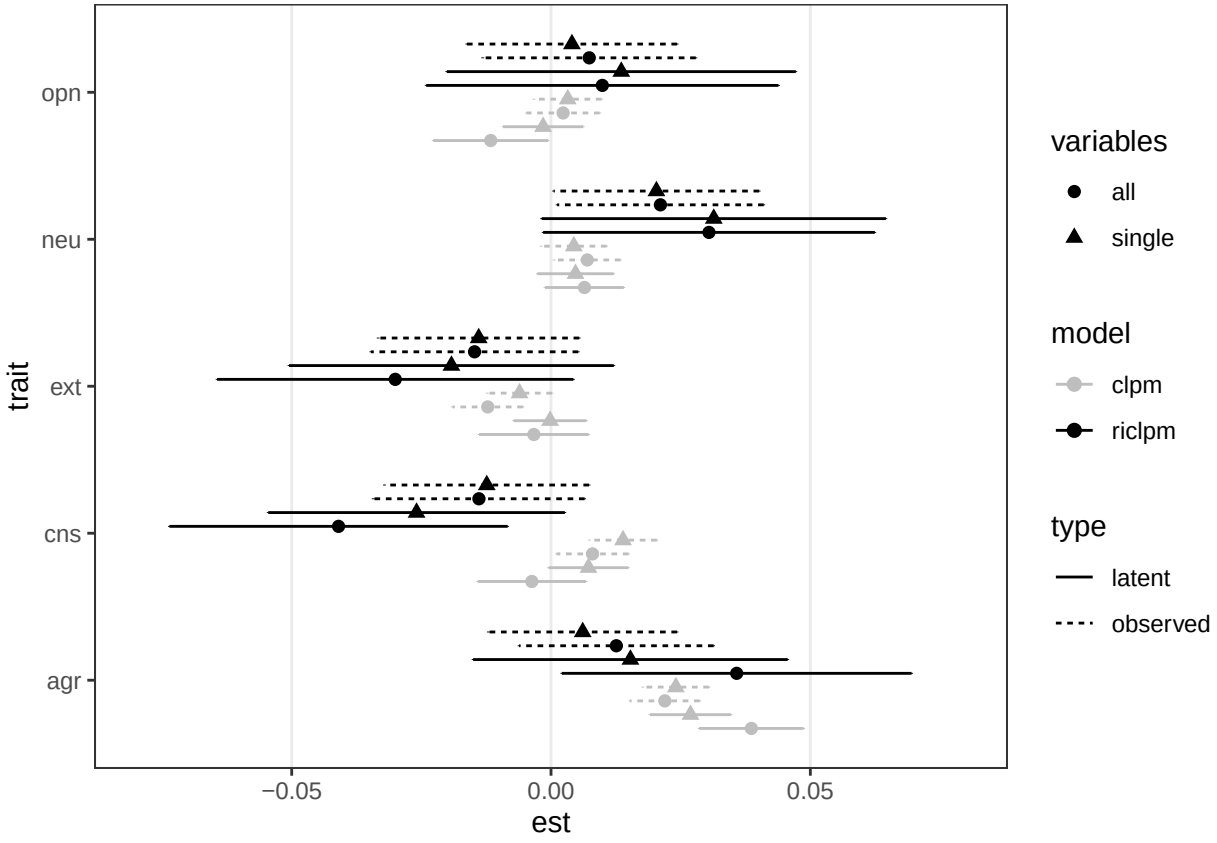

Figure 2. Estimated Lagged Effects of Personality on Religiosity

are labeled with circles. Bars represent 95% confidence intervals for the standardized parameter estimates. Because the standardized estimates vary slightly across waves, we follow Entringer et al. (2023) and present the average of these estimates.

Before describing the results, it is important to acknowledge that the confidence intervals for estimates from the RI-CLPM models are generally considerably larger than those for the CLPM. This pattern is not unique to these data or these model specifications. The RI-CLPM is equivalent to a multilevel model that separates between-person associations from within-person associations, and estimates of the within-person parts of these models often have less bias at the cost of reduced efficiency (see Allison, 2009, for an explanation). Because of this, we highlight both the significance of the effect and the parameter estimates when comparing results across models.

Figure 2 shows that conclusions about the lagged effects of personality on religiosity would differ depending on which model was used. Indeed, there is no effect that emerges consistently across all model specifications. For instance, the largest effect from Entringer et al. (2023) was the lagged association between agreeableness and religiosity, which had an average standardized effect of 0.039. The comparable model from Figure 2 is the CLPM with latent traits and all traits modeled simultaneously, which resulted in an identical average standardized effect of 0.039. This estimate was very similar when the RI-CLPM was used, when latent variables were modeled, and when all traits were included simultaneously. However, estimated effects were smaller (frequently about half the size) and sometimes nonsignificant in the other model specifications. This association between agreeableness and religiosity was the most robust of the effects we examined, and even it varied in size and significance across model specifications.

Importantly, Figure 2 shows that our alternative models do not always result in reduced effect sizes relative to what was reported in Entringer et al. (2023). For instance, in the original study, the lagged effect of neuroticism on religiosity was a nonsignificant .011. In our reanalysis, we found similar average estimate of 0.006 with the same model specification in the full sample. However, the size and significance of the effect varied considerably across specifications. Most notably, the RI-CLPM resulted in estimated lagged effects that were close in size (though not always significant) to the lagged effect of agreeableness, which was the largest effect found in the original study and a primary finding that was highlighted in the discussion. It is sometimes claimed that using the RI-CLPM necessarily results in smaller estimates of lagged effects than the CLPM (e.g., Asendorpf, 2021), but this is not correct. Lucas (2023) showed that if stable-trait variance exists in the measures, results from the CLPM can underestimate the true lagged effects. Although neuroticism received little attention in Entringer et al.'s (2023) original report, the lagged path from neuroticism to religiosity is one of the largest effects found when the arguably more appropriate RI-CLPM is used.

When relying on statistical significance as a criterion for evaluating replication across the original analysis and our full-sample analysis, a single discrepancy arose: the path from openness to religiosity was nonsignificant in the original analyses but significant in our reanalysis. However, the estimate itself was virtually identical. Specifically, the estimate for the path from openness to religiosity was -0.013 in the original analysis versus -0.012 in ours. This does not seem like a consequential difference.

Finally, Figure 2 shows that effects can even reverse direction depending on the chosen model specification. Most notably, although there was no significant effect of conscientiousness in the original study (a finding replicated in the full sample using the same model specification), this effect became significantly negative in one version of the RI-CLPM and significantly positive in two of the other three specifications of the CLPM. These examples show that conclusions about the lagged associations between personality traits and religiosity depend on the precise model specification. Indeed, conclusions about all five traits depend on which model specification is chosen (if the significance of the effect is used to guide interpretation), with the sign of the effect even changing from significantly negative to significantly positive for one of the five traits.

Turning to the lagged paths from religiosity to personality, Entringer et al. (2023) found only one significant effect: the path from religiosity to agreeableness, which was estimated to be a small but significant 0.011. The size and significance of this effect was again replicated in our full-sample analysis (average standardized estimate = 0.017). However, as Figure 3 shows, the size and significance of this lagged effect again varied considerably; two specifications of the CLPM resulted in considerably larger effects for the path from religiosity to agreeableness (with effect sizes comparable to the largest effects of personality on religion), whereas nonsignificantly negative associations emerged in all specifications of the RI-CLPM.

In addition, estimates for three of the four other effects differed depending on which

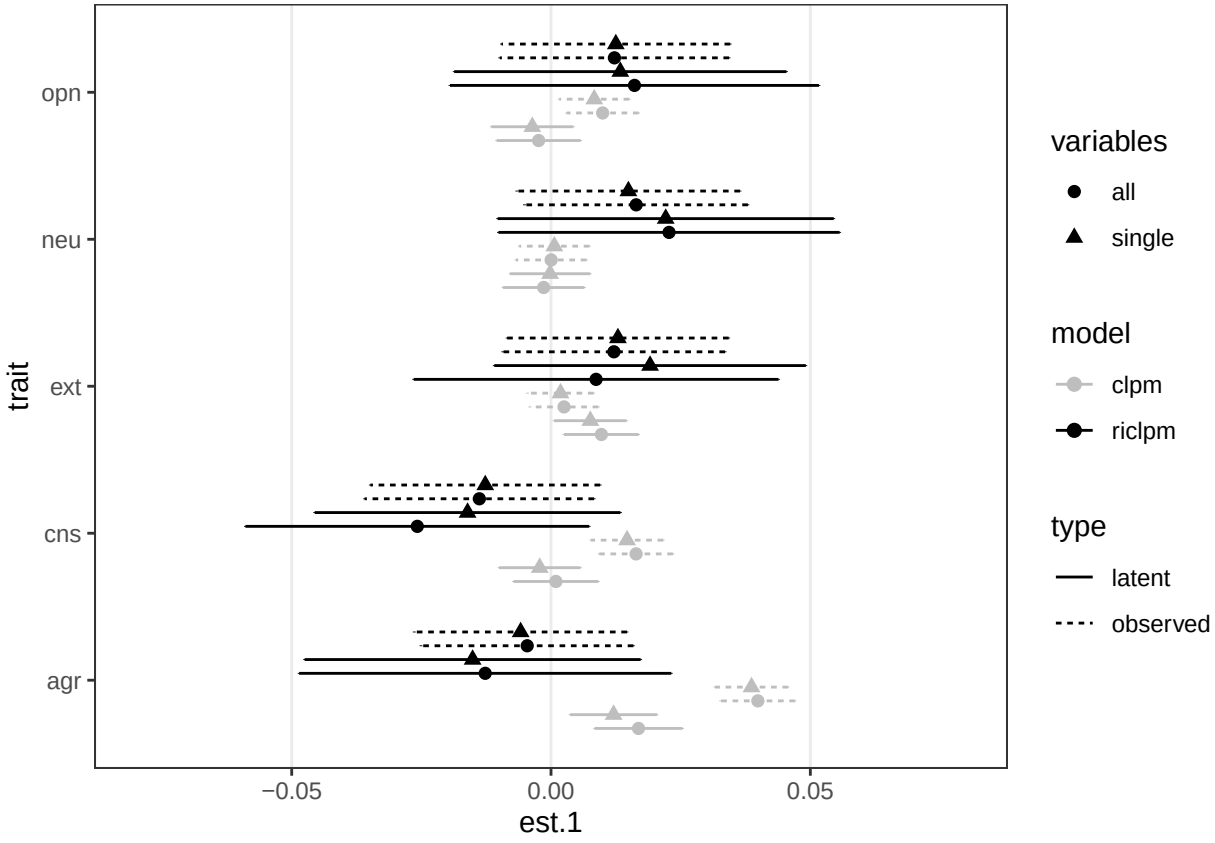

Figure 3. Estimated Lagged Effects of Religiosity on Personality

model specification was used. For instance, the lagged effects of religiosity on openness and conscientiousness—effects that were not significant in the original paper—were significant in both the observed-variable CLPM models. There was also one effect that was significant in the full sample that was not significant when using the same model in the original study: the lagged effect of religiosity on extraversion. The effect was a non-significant .004 in the original paper, whereas it was a significant (though only slightly larger) 0.01 in the current analyses. Again, as was true with the paths from personality to religiosity, conclusions about the paths from religiosity to personality vary depending on which model specification one chooses.

Meta-analysis Across Federal States

As noted previously, the results from the full-sample analyses mostly replicate the average effects from the meta-analysis of results across the 14 federal states analyzed in Entringer et al. (2023). By focusing on the full sample, comparisons across models were simplified, and the effect of any state-specific model estimation problems could be eliminated. We also conducted state-level analyses to examine how modeling choices affect state-level religiosity as a moderator of the lagged effects.

Assessing how model choice affects results across states is challenging, however, as estimation and convergence problems occur at different rates across different states and different models. Even in the original paper, responses from two states (Bremen and Saarland) were dropped from the analyses due to estimation problems in the CLPM. Our own analyses replicated these problems in these two states, as well as in one additional state not identified as problematic in the original analyses. Specifically, although estimates could be obtained for Hamburg, some variances were estimated to be negative. Additional problems (including negative variances, non-positive-definite matrices, and lack of convergence) were encountered with other model specifications, including the CLPM with latent variables and each trait modeled separately and the RI-CLPM with latent variables traits modeled separately or together. Notably, no estimation or convergence problems were encountered with any model (even those in the two states that were omitted from the original paper) that included observed variables instead of latent variables for the Big Five. This is further evidence that the complex measurement model causes problems and that the simpler observed-variable models should be preferred. A full list of parameter estimates for the cross-lagged paths along with a list of estimation and convergence problems are included in Tables 1 through 16 of the supplement.

Because of these estimation problems, we rely only on the four observed-variable models when examining the effects of state-level religiosity on the lagged paths. Specifically,

we first tested each model in each state and then used meta-analytic procedures to test whether state-level religiosity moderated these paths in each model. Parameter estimates and 95% confidence intervals are presented in Figure 4 for the paths from personality to religiosity and Figure 5 for the paths from religiosity to personality.

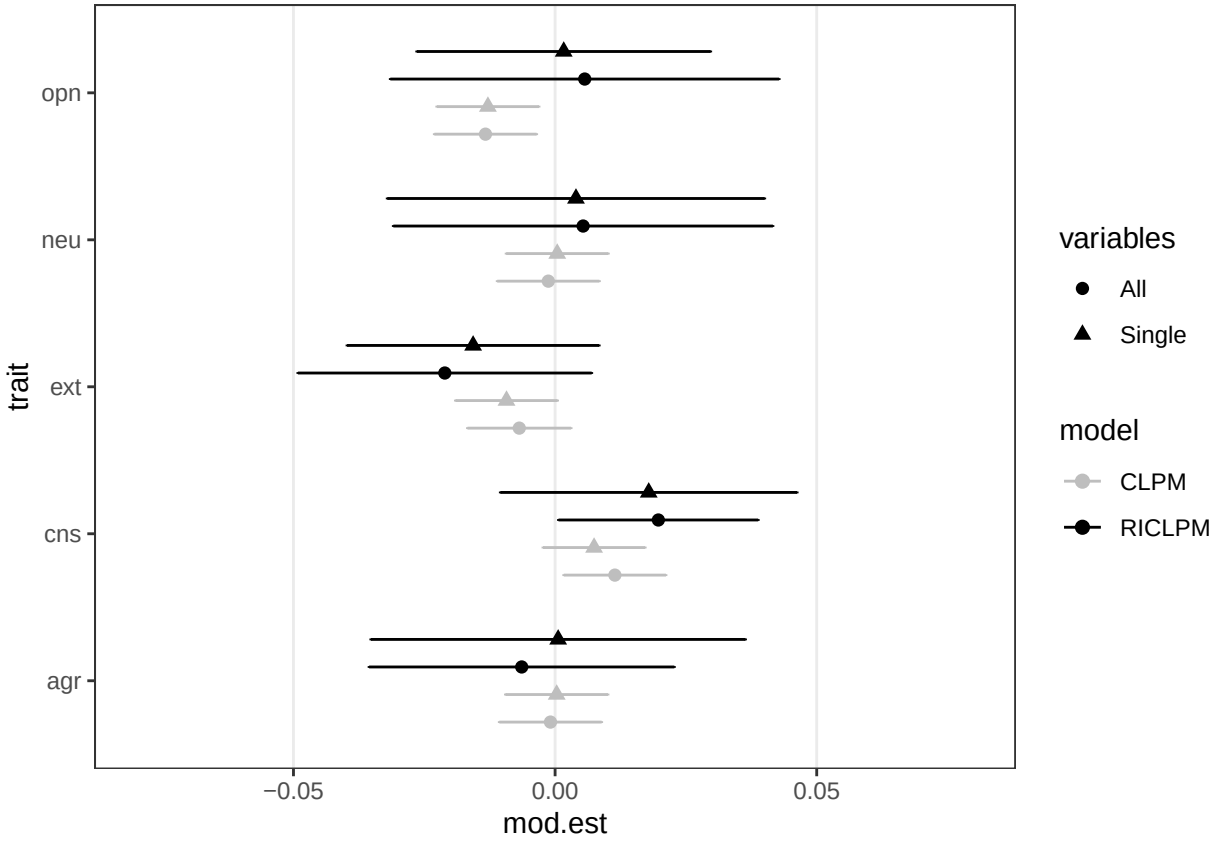

Figure 4. Meta-analytic results for state-level religiosity as a moderator of the lagged effect of traits on religiosity

In the original paper, Entringer et al. (2023) found moderating effects of state-level religiosity for the paths from openness and conscientiousness to religiosity. Figure 4 shows that both of these effects were replicated when the CLPM was used and all traits were modeled simultaneously. This was true even though we focus on the model that uses observed variables for the Big Five traits, whereas Entringer et al. modeled these as latent variables. However, as was true for the simple lagged effects, these results varied depending

on precisely which model specification was used. In the case of openness, both version of the CLPM resulted in significant moderating effects, but neither version of the RI-CLPM effect did. For the conscientiousness effect, when all five traits were modeled simultaneously, both the CLPM and RI-CLPM resulted in significant moderating effects. When conscientiousness was examined on its own, however, these effects were not significant. Moderating effects for the other three traits were consistently small and nonsignificant across the four model specifications.

Entringer et al. (2023) also found moderating effects of state-level religiosity for the paths from religiosity to openness, but the paths to the other traits were not significant. Figure 5 shows that this one significant path also emerged in our version of this model. Again, however, when the RI-CLPM was used, the meta-analytic effect became nonsignificant (and indeed, the sign of this nonsignificant effect reversed). Thus, consistent with the full-sample analyses, we found no moderating effects of state-level religiosity that consistently emerged across model specifications.

Discussion

Psychologists and other social scientists are often interested in the ways that personality affects real-world outcomes along with the reciprocal links between real-world experiences and personality change. Because personality is relatively stable over time and difficult to manipulate, experimental evidence that can inform theories about these processes is often difficult, if not impossible, to obtain. In these situations, longitudinal data analyzed with appropriate quantitative methods can often be the best choice for describing and understanding change and reciprocal effects (though see Rohrer & Murayama, 2021).

Entringer et al. (2023) used this approach to examine the reciprocal links between the Big Five personality traits and religiosity in a large German sample. They noted that there are strong theoretical reasons to expect reciprocal effects, and indeed, in their analyses, they

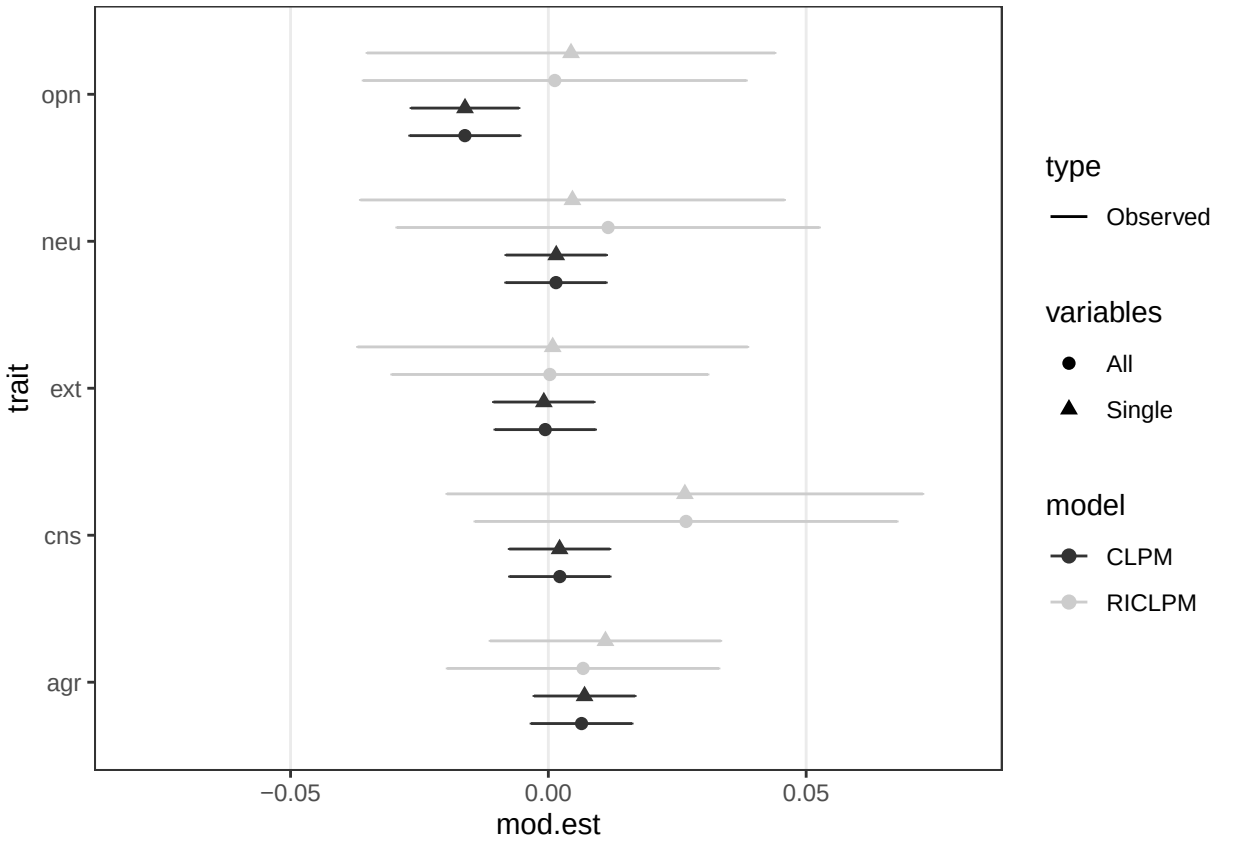

Figure 5. Meta-analytic results for state-level religiosity as a moderator of the lagged effect of religiosity on traits

found some. Specifically, agreeableness prospectively predicted change in religiosity over time across all states, whereas both openness and conscientiousness predicted it differently depending on the religiosity of the region. In addition, religiosity prospectively predicted agreeableness across all states, and there was a moderating effect of regional religiosity on the effect of religiosity on openness. Entringer et al. (2023) concluded that personality and religion do have reciprocal effects on one another over time, and that some of these effects depend on the cultural context.

Although Entringer et al.'s (2023) study has a number of important strengths (including a series of robustness tests), our own reanalyses challenge the robustness of these effects to theoretically justifiable alternative model specifications. In short, when a series of

reasonable alternative models was used to examine these reciprocal effects *no single lagged effect (either culturally-consistent or culture-moderated effects) emerged consistently across model specifications.* Thus, we urge caution in conclusions about reciprocal causal effects between personality and religiosity.

Measurement Models and Model Complexity

The first issue we raised was in regard the complexity of the model that was used in the original study. This original model included items as indicators of latent Big Five personality traits, and it modeled all traits simultaneously. To be sure, this approach is defensible, and it has some advantages over alternatives. Most importantly, in models like the standard CLPM, the existence of measurement error can lead to spurious lagged effects (Lucas, 2023). This is because correctly estimating the lagged effect of a predictor on an outcome after controlling for the prior wave of the outcome requires precise measurement of that outcome at the prior wave (Westfall & Yarkoni, 2016). If outcomes are measured with error, lagged effects can emerge solely due to incomplete control of scores at the prior wave. Using latent variables helps address this concern. Indeed, in our analyses, there were certain lagged effects that emerged only when the observed-variable version of the CLPM was used, but not when the latent-variable version was used (e.g., the lagged effects from conscientiousness and extraversion on religiosity and the lagged effects of religiosity on conscientiousness; see Figures 2 and 3). These effects should be interpreted cautiously, as they could be spurious. Notably, although the existence of measurement error can also lead to spurious lagged effects in the RI-CLPM (Lucas, 2023), no effects emerged in the observed-variable version of the RI-CLPM that did not also emerge in the latent-variable version.

Although modeling latent traits has some advantages, it also comes at the cost of model complexity and possible increases in model misspecification. Especially when all five traits are modeled simultaneously, it is important for items to relate strongly to the latent-trait that they are designed to measure and not to other traits. If there are exceptions,

clear and robust secondary loadings need to be identified and modeled. The authors of the original paper carefully considered the measurement model and tested for measurement invariance across federal states. It is still possible, however, that some important secondary loadings were omitted or even that some of those that were included capitalize on chance in the current sample and do not reflect the true underlying associations among items and constructs. If the measurement model is incorrectly specified, then the structural parts of the model can be affected (Rhemtulla et al., 2020).

There are at least two reasons to be cautious when interpreting results that rely on this complex measurement model. First, the models in the original study did not fit especially well, with CFI values typically falling below the standard .90 threshold. All models in our reanalysis that included latent traits—including an unconstrained measurement model—resulted in CFI values that were below the more stringent standard of .95. Entringer et al. (2023) defended the relatively low CFI values in their study by noting that the CFI for correctly specified models declines as the number of observed variables in the model increases (e.g., Kenny & McCoach, 2003). However, Kenny and McCoach showed that this effect becomes less pronounced as sample sizes increase, and the largest sample sizes in their paper (and those they reviewed) were around 1,000, compared to over 40,000 in the current study. Moreover, Kenny and McCoach also showed that the RMSEA decreases (implying better fit) for *incorrectly* specified models as the number of observed variables increases (they did not report simulation results for SRMR). Thus, the discrepancy between the CFI and the RMSEA (and SRMR) is difficult to interpret. It is impossible to tell whether the relatively low CFIs result from the large number of observed variables or a misspecified model. This is further reason to use robustness across alternative specifications as a guide when interpreting parameter estimates.

A second issue is that the latent-variable models often led to estimation and convergence problems even for the relatively simple CLPM, but especially for the RI-CLPM.

These problems meant that participants from two states needed to be excluded from the analyses in the original study, and many more had to be excluded from our reanalysis. In contrast, when observed variables were modeled, no estimation or convergence problems emerged for any of the tested models in any of the 16 federal states. Although neither the fit nor convergence issues is definitive proof that the measurement model used in these analyses is misspecified, together they suggest that some caution is warranted when interpreting results that rely on this measurement model. In such cases, robustness across alternative specifications is desirable before strong conclusions are drawn, and that robustness did not emerge across the alternative specifications we tested. Indeed, fundamentally different conclusions about the links between personality and religiosity would result from the different model specifications.

Controlling for Other Traits

A second issue has to do with the fact that Entringer et al. (2023) chose to model all traits simultaneously when examining lagged effects. This means that each lagged effect reflects the association between the unique variance in that trait and religiosity, after controlling for all other traits. This decision is not without consequences. Most notably, it comes with important challenges regarding interpretation.

As Lynam et al. (2006) noted, the nomological network surrounding a trait is typically developed using unadjusted associations between trait measures and other constructs. Theories about constructs are typically based on these nomological networks, and both predictions for and interpretations of associations like those investigated in this paper are necessarily based on these existing theories. However, when controlling for other traits, it becomes necessary to explain why only this residual variance—variance that reflects a construct that is not well-understood—is linked with the outcome.

For instance, in this study, results for conscientiousness appeared to be most strongly

affected by the decision to model all traits simultaneously. We ran a simple regression analysis predicting a latent conscientiousness variable in 2005 from latent variables representing each of the other four traits assessed in that wave; these four variables accounted for 35% of the reliable variance in conscientiousness. Any theoretical explanation for the lagged association between conscientiousness and religiosity must explain why it is only the remaining unique variance that predicts this outcome. Moreover, to correctly interpret these effects, it would be necessary to clarify that those who score low on conscientiousness are no more likely than those who score high on conscientiousness to increase in religiosity; it is only those who score low on conscientiousness *relative to their standing on all five other traits* who are likely to do so. Finally, it seems likely that the size of these between-trait correlations will likely be sample- and measure-specific, which makes it even more difficult to draw broad and generalizable conclusions about the processes underlying these associations.

To be sure, modeling all traits simultaneously is not a clearly wrong decision, but it is one that requires justification and careful interpretation. If this strong justification does not exist, then one would hope that this specific decision would not substantively affect results, but in this case they do. A number of specific effects were dependent on whether the trait in question was modeled on its own or simultaneously with the other Big Five traits.

Controlling for Stable Traits

A final issue concerns the analyses that were used in the original paper. Entringer et al. (2023) relied on a standard lag-1 CLPM, where personality and religiosity were predicted only from the same variables measured in the immediately preceding wave. This model implies that there are no other sources of stability that contribute to these measures than what is captured by the stability coefficients and a single lagged effect of the other variable in the model. If other factors contribute to the stability of the measures over time, then the CLPM will be misspecified. This misspecification will bias the parameter estimates, including the lagged effects. Simulation studies show that the size of this bias exceeds the effect sizes

reported by Entringer et al. (2023) under realistic data generating models (Lucas, 2023). One way (though not the only way) to address this misspecification is to include a random intercept for each construct, which can account for some common and theoretically plausible forms of stability, including the existence of stable trait variance. Our analyses show (a) that the actual stability coefficients in these data do not match the implied coefficients from the CLPM, (b) that these actual coefficients do closely match those implied by the RI-CLPM, and (c) that the RI-CLPM fits the data considerably better than the CLPM. Importantly, the size and significance of the lagged effects—both from personality to religiosity and from religiosity to personality—often depend on whether the CLPM or RI-CLPM is used.

What can be made of the fact that the predicted cultural moderating effects also varied depending on whether the CLPM or RI-CLPM was used? One possibility is that the original moderating effects were not due to state-level differences in either the effect of personality on religiosity or religiosity on personality, but to differences in the size of the stable-trait-level associations between personality and religiosity across federal states. Once these associations were accounted for, then the potentially spurious lagged effects would not emerge in states with especially high trait-level associations. Regardless, the effects of these modeling choices again show that conclusions about the reciprocal effects—including moderating effects of culture—are not robust across different reasonable model specifications.

Effect Sizes

It is important to consider these issues in the context of the very small effects identified in the original study and this reanalysis. The largest effect in the original study was a standardized regression coefficient around .04, and all estimated effects in our reanalysis also fell below .05, regardless of the model specification that we used. As noted above, these small effects are an issue not only because they may lack practical significance, but also because even very slight model misspecification or residual confounding can lead to these small effects.

It is worth highlighting that of the 10 primary lagged effects that were the focus of this investigation (personality to religiosity and religiosity to personality for each of the five traits), there was one effect that was slightly more robust than the others. The path from agreeableness to religiosity was significantly greater than zero for five of eight models and it was only nonsignificant in the RI-CLPM models, which had wider confidence intervals (though the absolute size of this estimate did vary considerably across models). Indeed, this is the one trait for which lagged effects were also found in a more recent study that used the RI-CLPM to investigate the same question in a separate dataset. Lenhausen, Schwaba, Gebauer, Entringer, and Bleidorn (2023) used eleven waves of data from a Dutch sample to examine the reciprocal association between personality and three measures of religiosity: belief in God, attendance at religious services, and prayer. In their study, the effects from agreeableness to religiosity were significant, but only when belief in God was used as an outcome. The standardized regression coefficient was a similarly small .027. The estimate for the item that is closest to that used by Entringer et al. (2023) was not significant. Extraversion also significantly predicted changes in belief in God, which is an effect that was not significant in Entringer et al.'s (2023) study and was sometimes significant in the opposite direction in the current reanalysis.

These small effects must be interpreted cautiously because even when using the RI-CLPM, lagged effects can be biased if other sources of variance contribute to the pattern of associations over time. For instance, The Stable-Trait Autoregressive Trait State (STARTS) model (Kenny & Zautra, 2001) includes an additional state component that reflects variance at a particular occasion that is completely unrelated to variance at any other wave. Lucas (2023) used simulations to show that the failure to include such a component when it exists can bias estimates of lagged effects in both the CLPM and RI-CLPM, and an analysis of real data showed that state components are often needed to accurately reproduce the underlying correlation matrices in longitudinal data. Thus, the very small lagged associations that emerged in these studies may not be robust when these more

complex models are used.

Importantly, both the current reanalysis and Lenhausen et al. (2023) reported zero-order correlations between personality and religiosity. In the current study, no correlation exceeded .10 in the full sample, and in Lenhausen et al.'s study no correlation exceeded .12. A common argument for small effects in the context of cross-sectional or short-term longitudinal studies is that these small effects can accumulate over time (Funder & Ozer, 2019). These simple analyses show that happens for personality and religiosity, it fails to induce a substantial correlation between the two, even when both variables are aggregated over time. Without evidence that such accumulation occurs, another justification for the importance of small effect sizes is needed.

Conclusion

Longitudinal data can be extremely helpful when examining the processes that underlie psychological phenomena that are difficult to manipulate experimentally. In the case of personality and religiosity, it would be challenging to develop manipulations powerful enough to substantially impact either of these variables, and therefore, studying their reciprocal associations can likely only be accomplished through the use of longitudinal data and analyses. However, these analyses come with challenges, as there are often many ways to model the data, each of which may come with different underlying assumptions. If there is no clear reason to accept one set of assumptions over the others, then the best outcome is if results are robust to different specifications. In this reanalysis, we showed that none of the effects identified by Entringer et al. (2023) is robust to a set of plausible alternative specifications and all are very small. Thus, we urge caution when interpreting the reciprocal associations between personality and religiosity.

Disclosures

Author Contributions

Richard E. Lucas conceptualized the study and wrote the initial analysis code and the first draft of the paper.

Julia Rohrer wrote additional code and ran all analyses that required access to the raw data, contributed additional ideas for analyses, and contributed to writing and editing the text.

Conflicts of Interest

The author declares that there were no conflicts of interest with respect to the authorship or the publication of this article.

Prior Versions

A preprint of this paper was posted on the PsyArXiv preprint server: .

References

Allison, P. D. (2009). *Fixed effects regression models*. Sage.

Asendorpf, J. B. (2021). Modeling developmental processes. In *The Handbook of Personality Dynamics and Processes* (pp. 815–835). Elsevier.

Entringer, T. M., Gebauer, J. E., & Kroeger, H. (2023). Big Five Personality and Religiosity: Bidirectional Cross-Lagged Effects and their Moderation by Culture. *Journal of Personality*, *n/a*(n/a).

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes*, *4*.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116.

Heise, D. R. (1970). Causal Inference from Panel Data. *Sociological Methodology*, *2*, 3–27. Retrieved from https://www.jstor.org/stable/270780

John, O. P., Donahue, E. M., & Kentle, R. L. (2005). *The "Big Five" Inventory – Versions 4a and 54*. Berkeley, CA.

John, O. P., & Srivastava, S. (1999). The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (pp. pp. 102–138). New York, NY: Guilford Press.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 333–351.

Kenny, D. A., & Zautra, A. (2001). The trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New Methods for the Analysis of Change* (pp. 243–263). Washington, DC: American Psychological Association.

Lenhausen, M., Schwaba, T., Gebauer, J., Entringer, T., & Bleidorn, W. (2023). Transactional Effects Between Personality and Religiosity. *Journal of Personality and Social Psychology*.

Lucas, R. E. (2023). Why the Cross-Lagged Panel Model Is Almost Never the Right Choice. *Advances in Methods and Practices in Psychological Science*, *6*(1).

Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The Perils of Partialling: Cautionary Tales from Aggression and Psychopathy. *Assessment*, *13*(3), 328–341.

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*(4), 1013–1034.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45.

Rohrer, J. M., & Murayama, K. (2021). These are not the effects you are looking for: Causality and the within-/between-person distinction in longitudinal data analysis. *Advances in Methods and Practices in Psychological Science*, *6*(1).

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLOS ONE*, *11*(3), e0152719.