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

The entire dataset is freely available for scientific use from the German Institute for Economic Research (DIW): https://www.diw.de/en/diw_01.c.601584.en/data_access.html

Ethical permission for the study was granted by the Scientific Advisory Board of the DIW Berlin, Germany.

The authors declare that they have no conflicts of interest related to this research.

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 to further the understanding of how personality characteristics shape people's 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 (Bleidorn et al., 2022; Stieger et al., 2021). 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, as well as causal effects of openness and religiosity on religiosity and religiosity on openness. 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 conscientiousness 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. 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 biased lagged associations (Heise, 1970). This bias would invalidate the interpretation of the lagged paths as estimates of the causal effects. Importantly, stable-trait variance is often a plausible default assumption for psychological constructs, for example due to stable genetic influences (e.g., on personality) or due to stable environmental influences (e.g., of family background on religiosity).

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. Indeed, the authors of the original paper used a similar approach in a conceptual follow-up (Lenhausen, Schwaba, Gebauer, Entringer, & Bleidorn, 2023). At the very least, robustness across modeling choices is important to consider.

The decision to include all trait simultaneously—and to do so using latent traits—also adds to model complexity, and if that complexity does not accurately reflect the underlying data generating processes, then this can affect parameter estimates. In this particular case, concerns can be raised about the complex measurement model that is required when all traits are included and modeled as latent traits. 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 (as was true in the original analyses), 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.[1]

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[2]. They argued 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 .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 (Anvari et al., 2021). One common defense of the importance of small effects is that these effects accumulate over time.

---

[1] It is also important to note that the increased model complexity leads to greater potential for errors in the specification of the model. Entringer et al. (2023) noted in their published paper that they intended to include all cross-sectional correlations among the Big Five traits in their model. However, these paths were not explicitly specified in the model code. Lavaan, the program used for this analyses, does have default settings that add correlations between latent variables, even when not explicitly specified. However, it appears that these default settings only add such paths for purely exogenous variables (i.e., those that predict but are not themselves predicted by any other variables) and purely endogenous variables (i.e. those that are predicted but do not themselves predict other variables), but not for those that are both predictors and outcomes. Thus, it appears that in the original analyses, cross-sectional correlations were only included for the first and last waves. Because the Big Five are often correlated with one another, excluding these paths may affect both model fit and other estimated parameters in the model.

[2] 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.

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. We address this issue by examining these aggregated correlations. Perhaps more importantly, very small effect sizes may simply reflect biases induced by subtle model misspecification or residual confounding rather than true underlying effects. Thus, each of the model specification issues that we have highlighted—even if they reflect just minor misspecification—could account for the very small observed effects. Again, in such cases, robustness to alternative specification provides an important check on the influence of these factors.

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 (at least in a simple way). 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.

Although debates about the merits of the CLPM have (at least in the psychological literature) typically focused on the comparison between the CLPM and RI-CLPM (Hamaker et al., 2015; Lucas, 2023; Lüdtke & Robitzsch, 2022), alternative models do exist. For instance, versions of the dynamic panel model (DPM, Dishop & DeShon, 2021) address the possible existence of stable-trait variance (and other time-invariant confounders) in a slightly different way than the RI-CLPM (see Murayama & Gfrörer, 2022). Indeed, one objection that has been raised about the RI-CLPM is that when the random intercept is included, the dynamic processes are modeled using residualized scores, which means that the stable trait variance for one variable is not allowed to affect change in the other (Orth, Clark, Donnellan, Robins, & Robins, 2021). The DPM, in contrast, does not residualize wave-specific variance when examining the dynamic processes, which means that links between stable-trait variance for one variable and change in another can be examined. Because debates about the CLPM have primarily emphasized the RI-CLPM as an alternative, we focus on these comparisons in this paper. We do, however, report results of the DPM as supplemental analyses.

Methods

This paper uses data from the German Socio-Economic Panel (SOEP) study (doi:10.5684/soep-core.v37o, Goebel et al., 2019), 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 we rely on variables included in both versions. 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

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.

states are reported in Table 1. Additional details about the sample are reported in the 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.

Analytic Strategy

For the primary analyses, we tested all combinations of models that vary on three dimensions: latent versus observed, single-trait versus simultaneous inclusion of all traits, and CLPM versus RI-CLPM. For the latent-variable models, we relied on the same measurement model that Entringer et al. (2023) used (which included correlated residuals between the same item in different waves). Consistent with the original analyses, for latent variable models that included all traits simultaneously, three cross-loadings were also added (details of the measurement-model specification are included in the original paper and in our code on our OSF site). For observed-variable models, the three items for each trait were averaged to create a single score for each trait in each year.

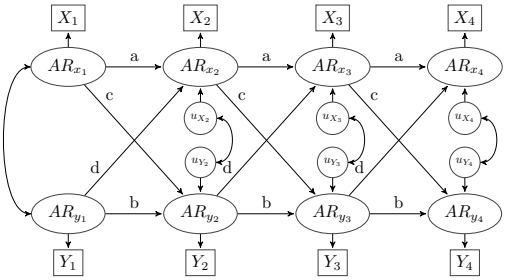
To specify a CLPM, each trait at each wave was predicted from the same trait and religiosity at the prior wave, and religiosity at each wave was predicted from religiosity and the trait scores at the prior wave. Correlations between each trait and religiosity at the first wave were freed, and residuals for traits and religiosity were allowed to correlate at later waves. For models that include all traits simultaneously, all Big Five traits were allowed to correlate within each wave, but lagged paths between traits were not included.

The RI-CLPM is similar to the CLPM, but with an added random intercept for each

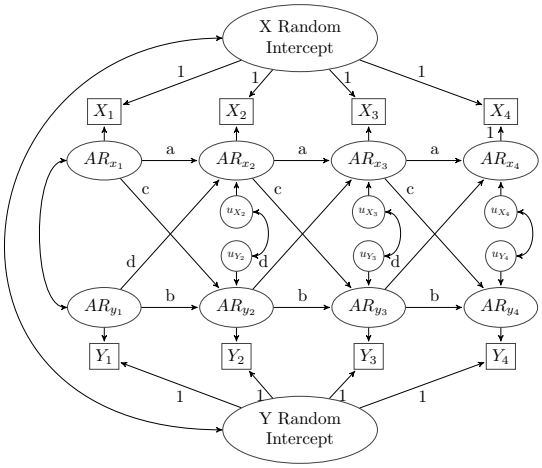
trait and for religiosity (Hamaker et al., 2015). Specifically, scores for each variable (i.e., a specific trait or religiosity) were specified to load on a corresponding latent trait representing the random intercept (with loadings fixed to 1), and a residual was estimated for each variable at each wave. The autoregressive and cross-lagged paths were then modeled using these residualized scores. In both the CLPM and RI-CLPM, stability coefficients and cross-lagged paths were constrained to be equal across waves.

As a supplemental analysis, we also tested the DPM (Dishop & DeShon, 2021), which models unobserved heterogeneity in a slightly different way than the RI-CLPM (see Murayama & Gfrörer, 2022 for a discussion of the relation between the DPM and RI-CLPM). In this model, scores at each wave after the first are specified to load on a latent variable (with loadings fixed to 1). The correlations between each of the two latent traits and the first wave of assessment for the trait and for religiosity are freely estimated. Finally, unlike the RI-CLPM, the dynamic processes (including stability and cross-lagged paths) are not modeled using residualized scores and are instead modeled using the observed indicators (or the latent traits that capture reliable variance in a wave for those models that use latent traits). Both latent-variable and observed-variable models were tested, though models that included all five traits simultaneously did not converge. These models were only tested using the full sample.

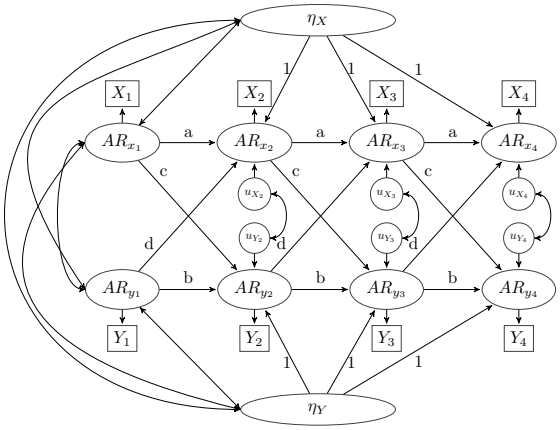
Figure 1 provides details for the univariate, manifest-variable versions of each of the three models for comparison. Details of all model specifications are included in the project repository on the corresponding OSF site. Note that in this figure, the CLPM and DPM include an observed variable and a corresponding latent variable for each measure at each wave. Because there is no error term associated with the observed variable, the latent variable is equivalent to the observed variable, which means that these models are equivalent to models that only included observed variables. The model is presented in this way to highlight the relationships among the various models.



Panel A: Cross-Lagged Panel Model



Panel B: Random-Intercept Cross-Lagged Panel Model



Panel C: Dynamic Panel Model

Figure 1. Diagram of the univariate, manifest-variable versions of the cross-lagged panel model, random-intercept cross-lagged panel model, and dynamic panel model. Paths with the same subscript are constrained to be equal (within a model). Some additional constraints needed to identify the model are not shown. For latent variable models, the observed variables are replaced with latent variables, each of which has three items as indicators.

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 emerged differed across models, which means that any overall differences that resulted 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. As noted earlier, however, within-wave correlations between pairs of Big Five traits were inadvertently excluded from the second and third waves in the original paper, and thus, after comparing the original model with and without these paths, we then include them in all subsequent analyses[3].

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 difference.

First, Table 2 shows that the unconstrained measurement model fits reasonably well. The second and third lines compare the model from the original paper, which only included within-wave correlations between religiosity and each of the Big Five traits for the first and last waves, and a modified model that allowed for correlations among the traits in all four waves. As can be seen in this figure, the change in chi-square is quite dramatic for a difference of just 20 degrees of freedom. An examination of the estimated correlations shows why freeing these paths is necessary: Of the 40 estimated correlations, 38 are significant, and

[3] In an earlier draft of this paper, submitted as a preprint, we made the same error as the original authors in all models that included latent traits. Models that used manifest variables for the traits did include the appropriate correlations.

Table 2

Fit indexes for full sample models.

Model	Type	Variables	Trait	ChiSq	df	p-value	CFI	RMSEA	SRMR
MEASUREMENT	Latent	All		26,219.54	1,445	0.000	0.95	0.03	0.04
CLPM*	Latent	All		42,804.44	1,835	0.000	0.90	0.02	0.05
CLPM	Latent	All		31,793.22	1,835	0.000	0.93	0.03	0.04
RI-CLPM	Latent	All		28,037.77	1,814	0.000	0.95	0.03	0.04
CLPM	Observed	All		11,062.31	230	0.000	0.88	0.08	0.07
RI-CLPM	Observed	All		914.57	209	0.000	0.99	0.02	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
RI-CLPM	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
RI-CLPM	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. The * for the first CLPM indicates that not all within-wave correlations among the Big Five traits were included, as specified in the original paper. The other multi-trait models include these correlations.

many are in the range of .40 to .60, with maximum correlations among traits or their wave-specific residuals of .67. Because omitting these correlations is a form of model misspecification that can affect the rest of the parameter estimates, we include them in all multi-trait models, though we do compare the lagged effects from the original model to this more appropriate one.

Table 2 shows that the primary modeling decisions do 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 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 (though it appears that these models did not allow for within-wave correlations among the Big Five traits, which, as noted above, detrimentally affects fit). These fit indexes suggest some caution is necessary when interpreting the estimates from the models. In contrast, the RI-CLPM with latent variables and all traits included (shown in the third row of Table 2) exceeds even the more stringent recommended standards for all fit indices, and the Chi-Square value is considerably—and significantly—lower for this model than for the CLPM.

Moving to the comparison of the CLPM and the RI-CLPM for the models 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 0.88 whereas for the RI-CLPM it is 0.99. Moreover, the difference in Chi-square values for the two models is not just significant, but dramatic, dropping from 11,062.31 to 914.57 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 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

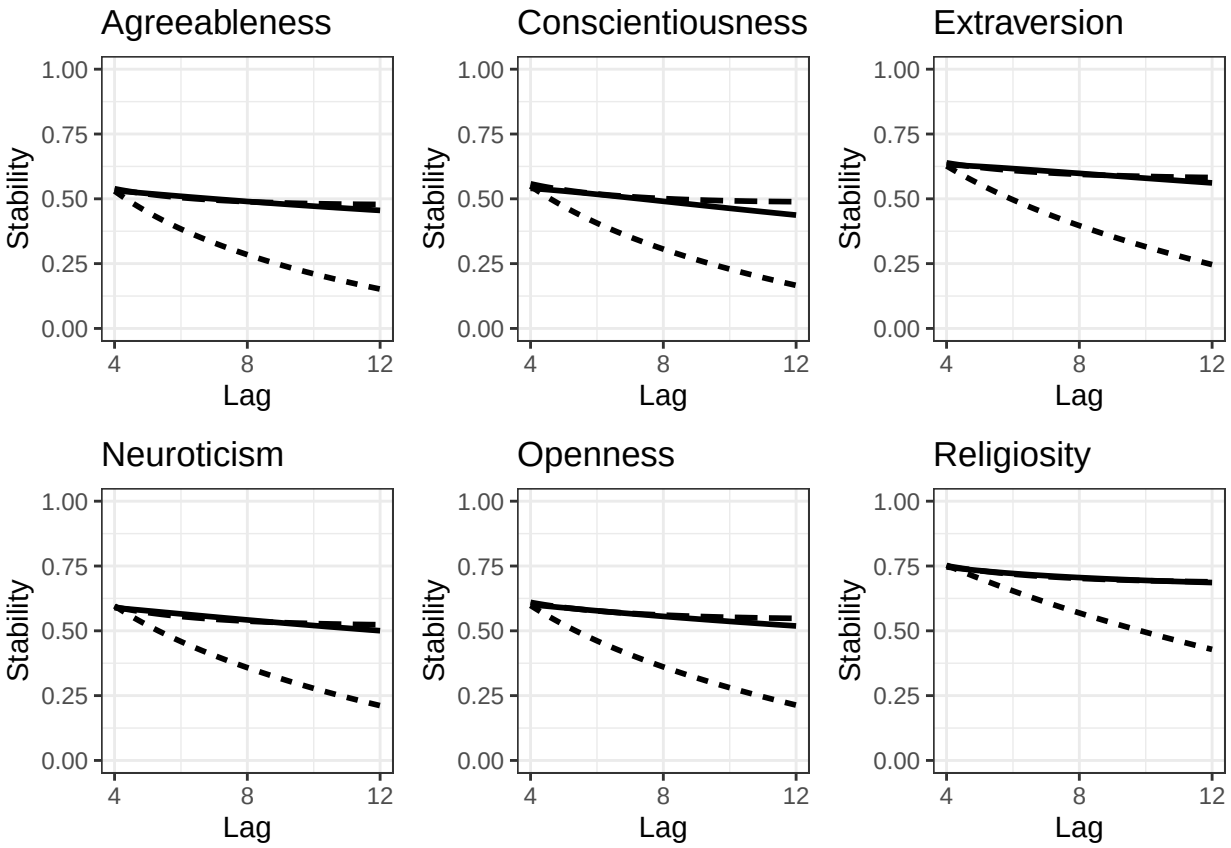


Figure 2. Actual (smoothed) stability coefficients (solid line) and implied stability coefficients from the CLPM (short-dashed line) and RI-CLPM (long-dashed line).

(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 2. 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 misspecified. This misspecification 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[4]). 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 3 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 are labeled with circles. Bars represent 95% confidence intervals for the standardized parameter estimates. Because the standardized estimates vary slightly across waves, we

[4] 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.

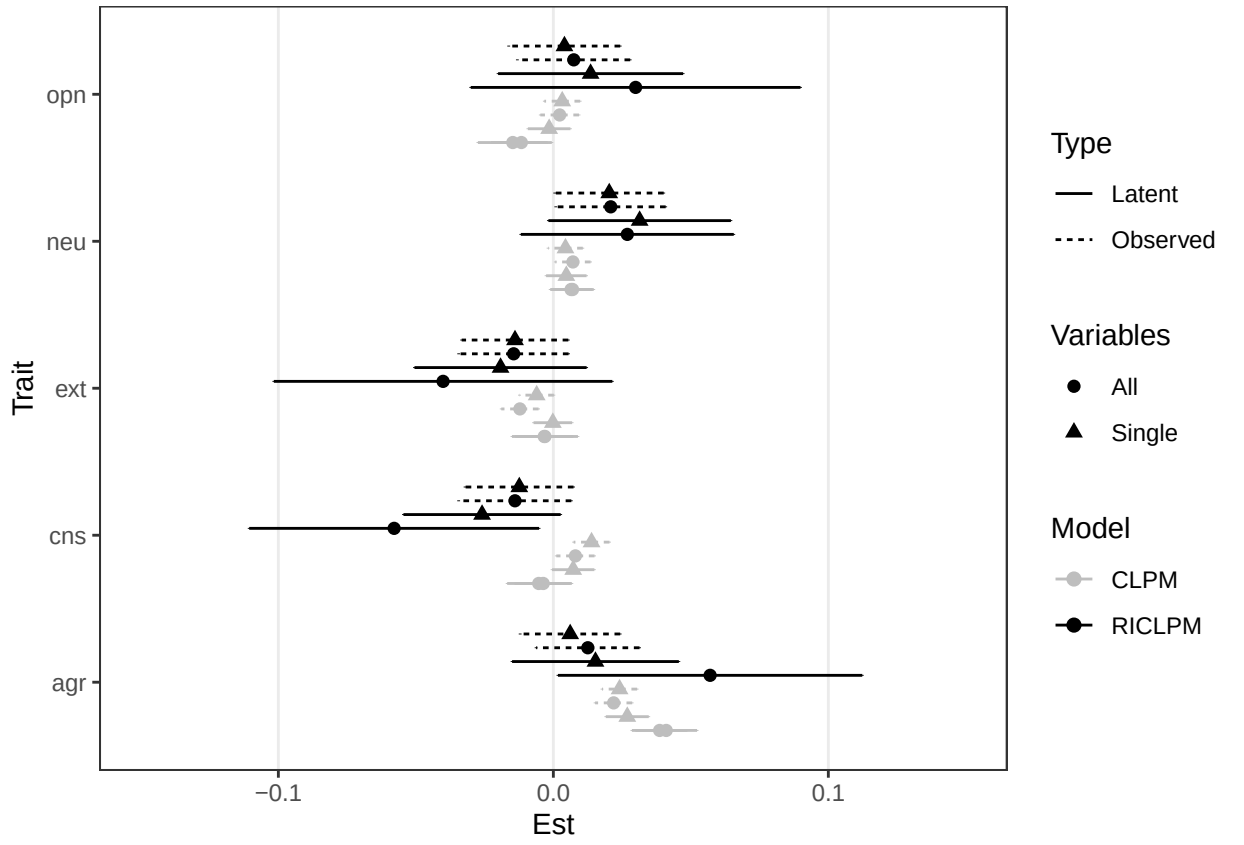


Figure 3. Estimated Lagged Effects of Personality on Religiosity

follow Entringer et al. (2023) and present the average of these estimates. Note that there are two points for the CLPM with latent variables and all traits modeled simultaneously (the original model), one as fit in the original paper and one that allows for within-wave correlations between each pair of Big Five traits. Although allowing for these correlations results in considerably better model fit, the similarity in estimates that results shows that this modeling decision has little effect on the parameter estimates (at least for the CLPM).

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 3 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: The effect of each trait is significant in at least one model, and none is significant across all 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 3 is the CLPM with latent traits and all traits modeled simultaneously, which resulted in a very similar average standardized effect of 0.041[5]. This estimate was similar in size, significant, and even slightly larger when the RI-CLPM was used, but only when latent variables were modeled and when all traits were included simultaneously. 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 3 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 0.007 with an almost identical model specification in the full sample. However, the size and significance of the effect varied 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

[5] Note that the estimate was actually identical to that from the original model when we tested a model that only included the within-wave correlations between each pair of Big Five traits for the first and last wave, as in the original paper.

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 -0.015 in ours. This does not seem like a consequential difference.

Finally, Figure 3 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

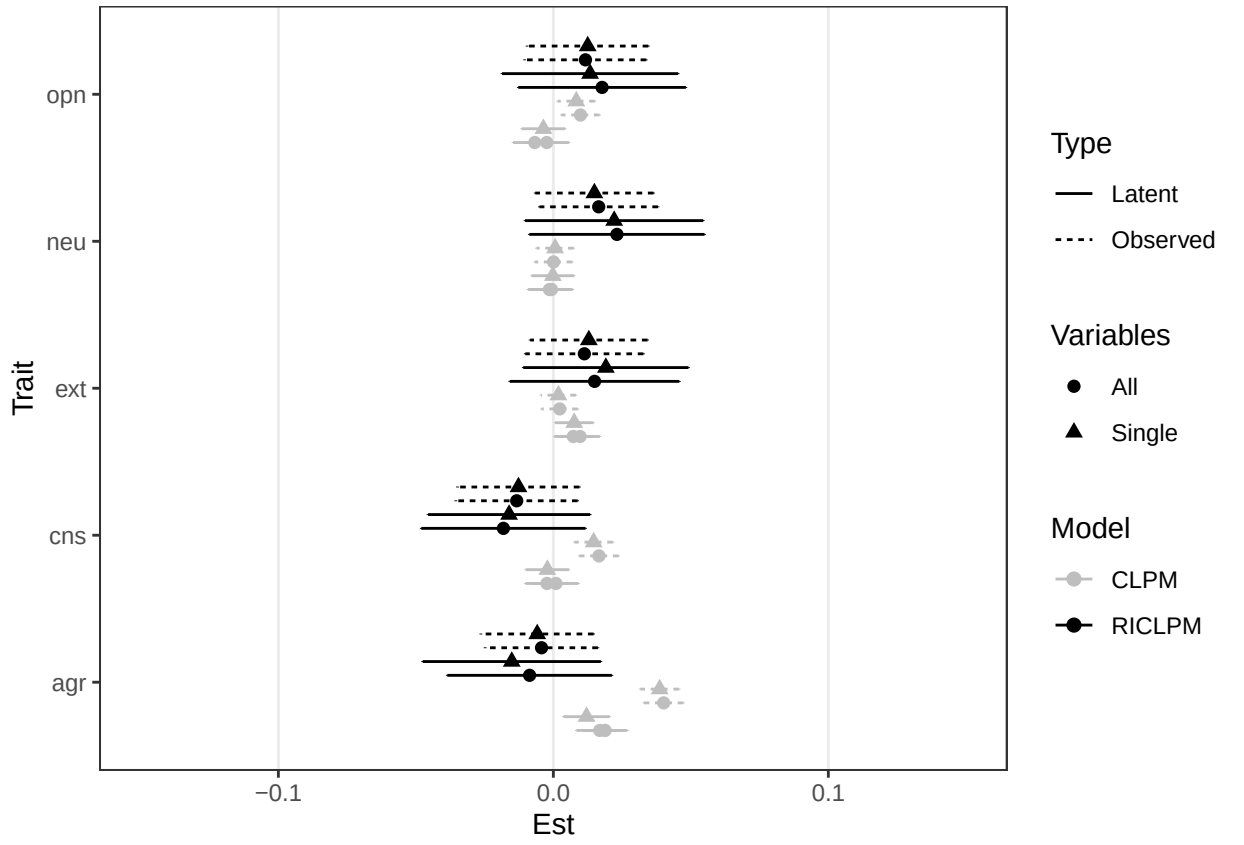


Figure 4. Estimated Lagged Effects of Religiosity on Personality

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 = 0.019). However, Figure 4 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 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) 0.007 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 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 5 (left panel for the paths from personality to religiosity and right panel for the paths from religiosity to personality).

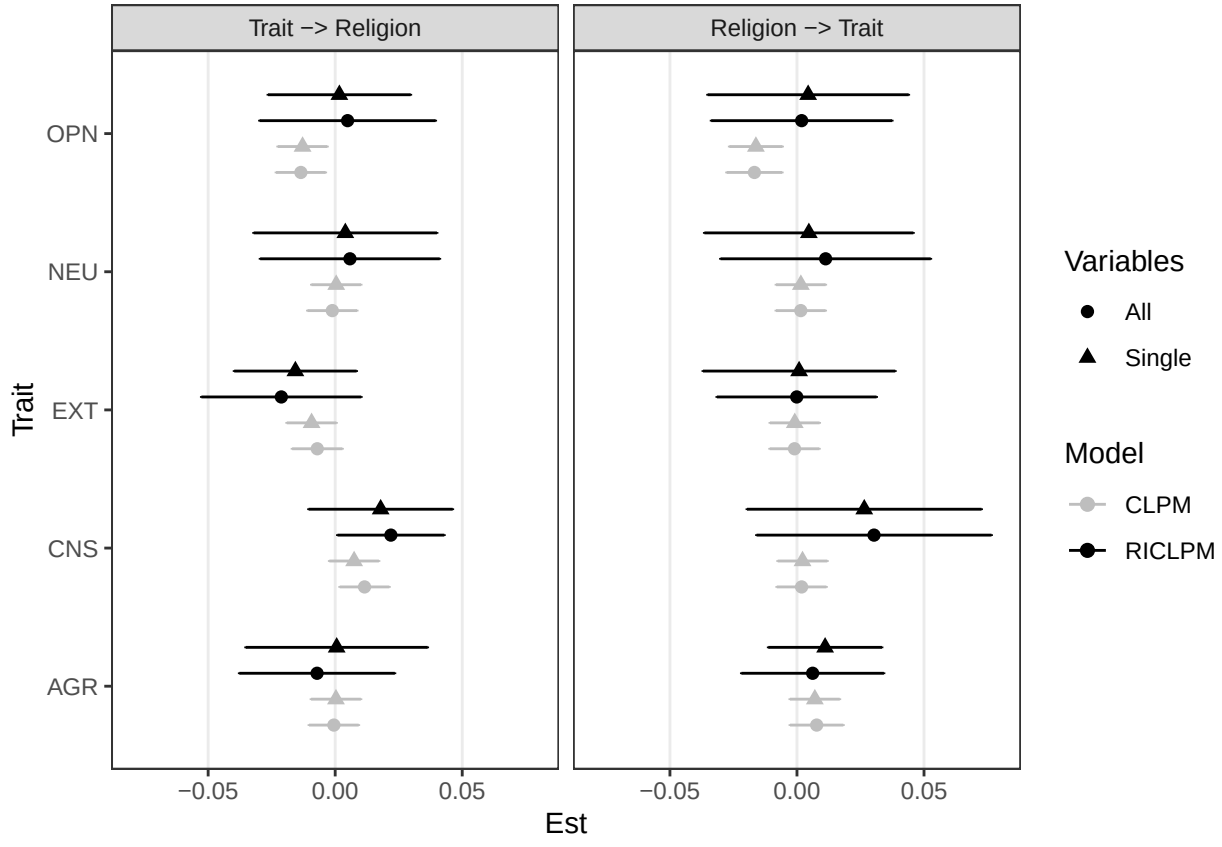


Figure 5. Meta-analytic results for state-level religiosity as a moderator of the lagged effect of traits on religiosity

In the original paper, Entringer et al. (2023) found moderating effects of state-level religiosity for the paths from openness and conscientiousness to religiosity. The left panel of Figure 5 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 versions 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 (though similar in size to models that included all five traits). 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. The right panel of 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.

Supplemental Analyses: The Dynamic Panel Model

Debates about the utility of the standard CLPM have often focused on the RI-CLPM as a preferable alternative (e.g., Hamaker, 2023; Hamaker et al., 2015; Lucas, 2023; Lüdtke & Robitzsch, 2022). Concerns have been raised, however, about the fact that the RI-CLPM residualizes wave-specific variance when examining dynamic processes (e.g., Orth et al.,

2021). Critics of the RI-CLPM claim that this feature changes the question being asked. These critics have not, however, provided a coherent justification for this critique from a causal inference perspective. Indeed, as we explain below, we believe that the critique reflects a form of conceptual confusion about how specific analyses map on to theoretical questions. Moreover, alternative models, like the Dynamic Panel Model (DPM), can address the concerns about the CLPM without relying on residualization.

Considering the issue of conceptual confusion, we disagree with Orth et al. (2021) that residualization within the RI-CLPM fundamentally changes the nature of the question that is being asked. To clarify the matter, it is helpful to distinguish between the theoretical estimand—the target quantity we are interested in to address our substantive research question—and the empirical estimand, which can be derived from observable data (Lundberg, Johnson, & Stewart, 2021). Whether or not the empirical estimand corresponds to the theoretical estimand will depend on whether so-called identification assumptions are met. Both the CLPM and the RI-CLPM are typically employed to address the same theoretical estimand. They are both meant to answer the causal question: Would a change in personality *cause* a change in religiosity? However, each model targets a different empirical estimand, and thus, each recovers the theoretical estimand of interest under different assumptions. A primary argument against the CLPM is that the assumptions required to map its empirical estimand to the theoretical estimand almost never hold. In contrast, the RI-CLPM requires somewhat less restrictive (and thus more realistic) assumptions (see Murayama & Gfrörer, 2022).

Considering a possible alternative model that can account for time-invariant confounding, the DPM does not residualize occasion-specific variance when examining dynamic processes over time. Instead, in the DPM, any effects of time-invariant predictors are specified to flow through each occasion when predicting change in the outcome variable, while still accounting for the trait-like associations between non-adjacent waves (see the

online supplement for the code we used to run it and detailed results). We initially tried running a model with all five traits included simultaneously (with observed variables), but the model did not converge[6]. Therefore, in these supplemental analyses, we focused on selecting one set of model specifications across which the three model types could be compared. Specifically, we compared the results of the latent-variable model with each trait examined as a separate model across the CLPM, RI-CLPM, and DPM. Details about model fit for these models, along with an observed-variable version of the model are presented in the online supplement.

Figure 6 shows the results of these analyses, with the paths from personality to religion in the left panel and the paths from religion to personality in the right. Consistent with the earlier model comparisons, the results from the original paper are not robust across these alternative specifications. None of the four significant average effects reported in the original paper (effects from agreeableness and openness to religiosity and effects from religiosity to agreeableness and extraversion) were significant when the DPM was used to model these associations. Indeed, the estimates from the DPM are generally (though not always) more similar to estimates from the RI-CLPM than the CLPM. Thus, despite concerns that the RI-CLPM fundamentally changes the question that is being asked (as compared to the simpler CLPM), results from the DPM (a model that is not subject to these critiques) results in estimates that are generally closer to those from the RI-CLPM than the CLPM.

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

[6] Convergence problems often emerge with more complex versions of the RI-CLPM such as the Stable Trait, Autoregressive Trait, State model, but less has been published about potential issues with the DPM. Given that all of these models need at least three waves of data for identification, the fact that convergence problems emerged in models with just four waves of data is not surprising.

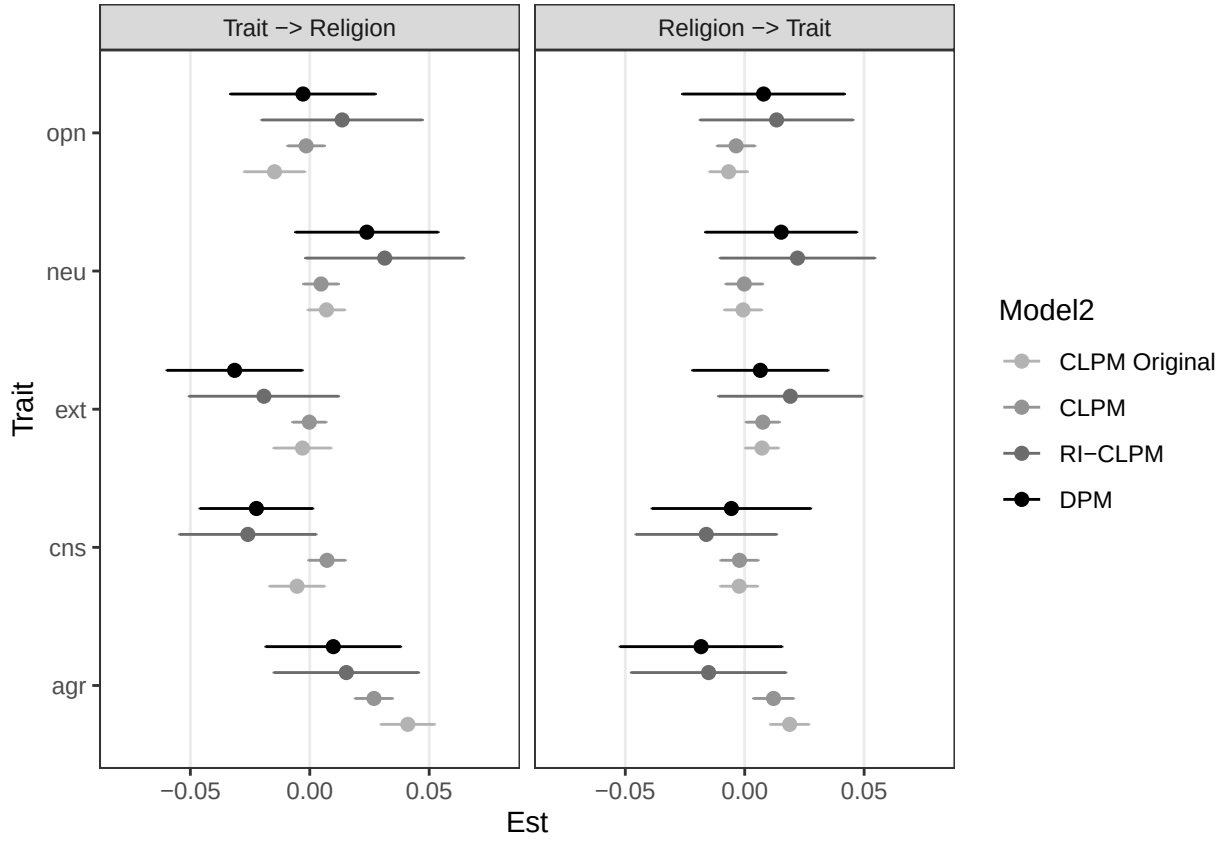


Figure 6. Comparing results of the observed-variable and latent-variable versions of the Dynamic Panel Model to the original results. All models other than the Original CLPM use separate models for each trait and traits are modeled as latent variables.

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 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 to 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 Figure 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, 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 some CFI values falling below the standard .90 threshold. 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.

Finally, it is important to note that concerns about measurement error and the spurious effects that might result can be addressed in the single-variable latent-trait models, models that are subject to fewer concerns about model complexity. Even when focusing just on the comparison between these simpler models and the original model, results were not robust.

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. The differences across models are limited: the effect of this specific

modeling choice was most evident in models that relied on the latent-variable RI-CLPM. The existence of these differences, however, is cause for concern about the robustness of the effects.

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 often 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.

Interpreting Results from the RI-CLPM

In response to concerns about the biased estimates that result from the use of the standard CLPM, many researchers advocate incorporating something like a random intercept or stable-trait factor to account for time-invariant confounders and other forms of unobserved heterogeneity (e.g., Bailey, Hübner, Zitzmann, Hecht, & Murayama, 2023; Hamaker et al., 2015; Lucas, 2023). Orth et al. (2021), however, argued that because the RI-CLPM and related models focus on residualized scores when examining dynamic processes, the question of interest is fundamentally changed when the RI-CLPM is used instead of the CLPM.

One important response to Orth et al.'s critique is to note that even if that were true and the RI-CLPM is not able to answer the question of interest, this does not mean that one should rely on the results of the simpler CLPM. It would be inappropriate to use this model in cases where clear violations of critical assumptions exist. And if no alternative model currently exists that can simultaneously answer the question and address these violations, researchers might need to just admit that the question of interest cannot be addressed with the data in hand, at least until methods advance further.

Fortunately, for this particular issue, an alternative model does exist—one that is not subject to the concerns that Orth et al. (2021) and others have raised (e.g., Asendorpf, 2021). As noted earlier, the DPM incorporates additional latent traits to deal with unobserved heterogeneity in a manner similar to what the RI-CLPM does, but without residualizing variables when modeling dynamic processes. Murayama and Gfrörer (2022)

noted that in many cases, one might expect the results from the DPM to be very similar to those from the RI-CLPM, and indeed, that seemed to be the case in this study. The results from the DPM-based analyses did not consistently replicate those from the original study, and estimates from these models were, in general, more similar to those from the RI-CLPM than those from the CLPM. This comparison of these three models adds to debates about the appropriate response to the critiques that have been raised of the CLPM.

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 no estimated effects across all models in our reanalysis exceeded .06. 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 et al. (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 effect from agreeableness to religiosity was 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.

Limitations

This study is not without limitations, though many are shared with the original paper to which ours responds. For instance, although the use of a four-wave study spanning twelve years improves upon many studies that examine reciprocal processes, the complexity of the models involved mean that even these four waves may not be enough. These models require at least three waves of data, and may still be empirically under-identified with four. Furthermore, even more sophisticated models that require even more waves of data (such as the STARTS; Kenny & Zautra, 2001) might be necessary to accurately test these lagged effects. Additional sources of misspecification that are due to the necessary model simplification may have prevented us from finding lagged effects that did exist. And finally, both our study and the original on which it was based use very brief measures of personality and a single-item measure of religiosity. Psychometric limitations may also have an effect on the estimates and the conclusions that we drew from these results.

It is important to note, however, that one study that improved on the current one by including many more waves and longer measures of personality and relgiosity did not replicate the important culture-independent results from Entringer et al. (2023), at least with the religiosity measure that was closest to what was used in the original (Lenhausen et al., 2023). Thus, despite these limitations, our primary conclusion that the results reported in the original study are not robust is justified.

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. While this does not imply that the effects are non-existent, it means that we must exercise 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: https://osf.io/preprints/psyarxiv/zt9mg.

References

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

Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., ... Orben, A. (2021). *Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters.* PsyArXiv. https://doi.org/10.31234/osf.io/g3vtr

Asendorpf, J. B. (2021). Modeling developmental processes. In *The Handbook of Personality Dynamics and Processes* (pp. 815–835). Elsevier. https://doi.org/10.1016/B978-0-12-813995-0.00031-5

Bailey, D., Hübner, N., Zitzmann, S., Hecht, M., & Murayama, K. (2023). *Illusory traits: Wrong but sometimes useful*. PsyArXiv. https://doi.org/10.31234/osf.io/rz8h4

Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *148*(7-8), 588–619. https://doi.org/10.1037/bul0000365

Dishop, C. R., & DeShon, R. P. (2021). A tutorial on Bollen and Brand's approach to modeling dynamics while attending to dynamic panel bias. *Psychological Methods*. https://doi.org/10.1037/met0000333

Eck, J., & Gebauer, J. E. (2022). A sociocultural norm perspective on Big Five prediction. *Journal of Personality and Social Psychology*, *122*(3), 554–575. https://doi.org/10.1037/pspp0000387

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). https://doi.org/10.1111/jopy.12770

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. https://doi.org/10.1177/2515245919847202

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten

Persönlichkeitsmerkmale im SOEP. *DIW Research Notes*, *4*.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie Und Statistik*, *239*(2), 345–360. https://doi.org/gfxztr

Hamaker, E. L. (2023). The Within-Between Dispute in Cross-Lagged Panel Research and How to Move Forward. *Psychological Methods*.

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. https://doi.org/f67cvh

Heise, D. R. (1970). Causal Inference from Panel Data. *Sociological Methodology*, *2*, 3–27. https://doi.org/10.2307/270780

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. https://doi.org/10.1207/S15328007SEM1003_1

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*. https://doi.org/10.1037/pspp0000466

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). https://doi.org/10.1177/25152459231158378

Lüdtke, O., & Robitzsch, A. (2022). *A comparison of different approaches for estimating cross-lagged effects from a causal inference perspective*. PsyArXiv. https://doi.org/10.31234/osf.io/gcvb4

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, *86*(3), 532–565. https://doi.org/10.1177/00031224211004187

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. https://doi.org/10.1177/1073191106290562

Murayama, K., & Gfrörer, T. (2022). *Thinking clearly about time-invariant confounders in cross-lagged panel models: A guide for model choice from causal inference perspective*. PsyArXiv. https://doi.org/10.31234/osf.io/bt9xr

Orth, U., Clark, D. A., Donnellan, M. B., Robins, R. W., & 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. https://doi.org/gg7zfw

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. https://doi.org/10.1037/met0000220

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). https://doi.org/10.1177/25152459221140842

Stieger, M., Flückiger, C., Rüegger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2021). Changing personality traits with the help of a digital personality change intervention. *Proceedings of the National Academy of Sciences*, *118*(8), e2017548118. https://doi.org/10.1073/pnas.2017548118

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719