Author Responses

Dear Editor and Reviewers,

Thank you very much for taking the time to read our manuscript and to give thorough feedback to improve the quality of our manuscript. You can find our responses to the suggestions and concerns raised by the reviewers below:

Reviewer #1

1. *This proposal describes a laudable attempt to create a well-thought through measure of perceptions of the science-religion relationship. I agree with the authors that we currently do not have such a tool available and I hope this work can change that. I appreciate the thorough methodological approach that is proposed - as well as the proposed comparison to the Farias et al's well known (but arguable limited) scale, as well as that by Leicht et al.*
   I just want to highlight the following: The five-model approach makes sense and offers more precise way to measure perceptions. I did notice though that I found it a bit hard to clearly distinguish context-switch from compartment. For example; Context-switch Item 7 (In some situations, religious...") would in my understanding fit quite well in compartment. So I guess what I am wondering is what the conceptual difference between "situations" or "circumstances" (Model 2) and "aspects of life", "elements of human experience", or "domains of application" (Model 3) are.

Thank you so much for your positive feedback. Context-Switch and Compartment are indeed quite similar, and to emphasize the differences between the two, we have modified the following text in the introduction section, along with an example to give the reader a clearer picture (pp. 5-6):

"*…Nevertheless, it is possible that one principally believes that science and religion are at odds, which is why they think that they cannot believe in science and religion at the same time. While Barbour does not include this "ambivalent situation" in his taxonomy, qualitative studies have provided support for its existence (Shipman et al., 2002; Taber et al., 2011; Yasri & Mancy, 2012). Since these individuals tend to avoid engaging in both science and religion at the same time, they switch flexibly between "wearing a thinking cap" and "taking a leap of faith," depending on the situations and challenges they encounter. For example, an individual might rely on their scientific reasoning when taking a science exam but switch to their religious beliefs when attending a church service. We call this a context-switch model (Yasri et al., 2013; Zein et al., 2024).*

The compartment model suggests a belief that science and religion are two independent domains and thus, do not necessarily interfere with each other. However, unlike the context-switch view, people endorsing a compartment view may be able to utilize scientific and religious explanations at the same time to approach the same problem from different perspectives or explain separate elements of a phenomenon. For example, when attending a funeral, people endorsing a compartment view may use science to explain the biological cause of death ("this person died from multiple organ failure..."), but they use religious beliefs to make sense of the afterlife ("...and now they are at peace with God"). These two elements (biological and afterlife) are construed as independent, so to them, there is no need to connect the two. Consequently, according to people who adopt the compartment model, science and religion are neither in conflict nor compatible (Gould, 1999; Zein et al., 2024)."

To help participants provide appropriate responses to the context-switch items, we added an instruction to our scale for participants to read before giving their responses:

> "*Please note that some items have "Situation" in them, and it can refer to **time** (e.g., night, morning, afternoon, etc.), **place** (e.g., school, church, home, museum, workplace, etc.), or **events** (e.g., Christmas, exams, pandemic, solar eclipse, funeral, wedding, etc.).*"

2. *Furthermore, I was happy to see references to the work on explanatory co-existence (Legare et al.). Indeed, it would be worth considering the role of culture/country-level differences in ascribed importance to the various models: what I mean is that the conflict and compartment models (Models 1 and 3) might be particularly Western ways of viewing the relationship.*

Thank you! It is true that previous research has pointed to the between-countries differences in viewing the relationship between science and religion, and we have discussed this in several parts of the introduction, such as (pp. 4-5):

"*Empirical evidence has also supported the assertion that the propagation of the "conflict" view (i.e., the belief that science and religion are fundamentally opposed) may only be prevalent in culturally secular countries (Johnson et al., 2020). For example, while religiosity is an important predictor of less favorable attitudes toward science in the U.S., this pattern does not generalize to other countries (Cologna et al., 2024; McPhetres et al., 2020). In some countries (e.g., Nigeria, Qatar, Kuwait, Egypt, India, and the Philippines in McPhetres et al., 2020, and Türkiye, Bangladesh, and Malaysia in Cologna et al., 2024) the association is reversed, with higher levels of religiosity corresponding to more favorable attitudes toward science.*

Indeed, several global surveys have captured a wide variation across countries in perceptions of the relationship between science and religion. (Kostyukov, 2019; Wellcome Trust, 2021). When participants in the 2020 Wellcome Trust Global Monitor (Wellcome Trust, 2021) were asked "When scientific evidence contradicts religious teachings, do you believe in science or religion?", more than 60 percent of participants in European countries (e.g., Germany, the United Kingdom, Belgium, Italy, etc.) were proponents of science, while more than half of the participants in some other countries (e.g., Brazil, Jordan, Türkiye, and Indonesia) sided with religion. Furthermore, the proportions of participants who said they believed in science and those who sided with religion are about the same in the U.S., India, and Uzbekistan. Interestingly, the number of participants who answered, "science and religion do not disagree" and "it depends" are particularly sizeable in some countries, with more than 20 percent of participants in Israel, Thailand, Croatia, Jordan, and Cambodia (Wellcome Trust, 2021)."

It is important to note, however, that our measurement study is based primarily on the assumption that perceptions of the relationship between science and religion differ among *individuals*. We base the assumption of individual differences (in views of the relationship between science and religion) primarily on evidence from qualitative studies examining Barbour's taxonomy. In our view, individual differences in views of the relationship between science and religion and differences across countries are two independent but related issues.

As an example, a study by McPhetres et al. (2020) found the relationship between religiosity and attitude towards science is positive in their Egyptian samples, which may imply that Egyptians, in average, are less likely to perceive a conflict between science and religion. Meanwhile, a qualitative examination of 25 science teachers in Egypt shows that participants

vary in their beliefs about the science-religion relationship, ranging from an extreme conflict (…"*Western scientists don't care about ethics*") to an absolute compatibility ("...*God creates everything including science,*" see Mansour, 2011).

In a similar vein, in secular countries where the relationship between religiosity and trust in science is often found to be negative, such as Israel (Cologna et al., 2024), public perceptions of the relationship between science and religion can also vary among individuals. A qualitative study of Israeli science teachers' and scientists' beliefs about the relationship between science and religion also shows wide individual variation, with no single mental model/view receiving majority of support from the participants (Dodick et al., 2010).

In this study, we are not particularly focused on examining differences between countries, but rather on making sure that our scale works similarly across different populations (i.e., measurement invariance). To that end, we hypothesize that a US American and a German with ***the same level*** of endorsement to conflict view, for example, would have ***similar probabilities*** of responding positively to items reflecting the conflict view. More precisely, these individuals (with the same level of endorsement to conflict view) would be equally likely to choose "strongly agree" with the item "*Science and religion are ultimately at odds, with no possibility of reconciliation,*" regardless of whether they are Germans or U.S. Americans.

It is important to note that we chose Germany and the United States as our reference and target group because the scale was developed in Germany. We believe that it makes sense to choose the United States as the target group because much previous research on this topic has been conducted in the United States (probably the issue is of particular societal importance in this country). After making sure that the scale works similarly in different countries (i.e., Germany and the U.S.), our scale may be useful for measuring perceptions of the relationship between science and religion in a future large-scale survey.

3. *Additionally, I would point the authors to a potentially useful chapter by Rutjens and Preston (2020) on the science-religion relationship, as well as the Negative Perceptions of Science Scale by Morgan et al., 2018 (which includes various items pitting science against religion, not unlike the BISS).*

Thank you for pointing us to a book chapter by Rutjens and Preston! We have included an argument we cite from the book chapter in the first paragraph of our introduction section (p. 3):

"*…Psychological functions served by both religious and scientific explanations, such as the need for explanation, control, and existential meaning, may contribute to a "competition" between scientific and religious beliefs, which in turn, leads to perceptions of conflict between science and religion (Rutjens & Preston, 2020).*"

To test discriminant validity, we chose beliefs in science (BISS) because it is theoretically related to, but different from, the construct we are measuring with our newly developed scale. In this case, we do not want our scale to be too closely related to belief in science, which we have described in the "the present study" section (p. 17):

"*…That said, belief in science focuses on individuals' perceptions of the high value of science and scientific institutions rather than on how individuals conceive of the relationship between science and religion. In other words, individuals may value science and scientific institutions exceptionally highly but still value religion to a similar degree because science and religion deal with different aspects of reality (i.e., the "compartment" view), or because they think that science and religion complement each other (i.e., the "complementary" view). Therefore, while individuals with strong beliefs in science may also hold the "conflict" view*

with a preference for the superiority of science, we hypothesize that belief in science and individuals' perceptions of the science-religion relationship are theoretically distinct."

Therefore, BISS fits this context better than NPSS, so we decide to plan to test the discriminant validity by correlating person parameters estimated from our scale data with person parameters derived from modeling BISS data.

Reviewer #3

1. I read this manuscript with great interest. I think it's very worthwhile the attempt to create a more sophisticated scale to assess how individuals vary in their attitudes and beliefs towards science & religion, and to move away from a Dawkins-like simplistic dualism of inexorable conflict. The authors build on Barbour's well-known model and summarise previous attempts to use this model to develop scales (e.g. Marin & Lindeman, 2021), which they consider limited for only using 4 items to assess Barbour's 4 models of science-religion relationship. They summarise their criticisms of such attempt as such: "The drawbacks of this approach are, first, that it limits participants' ability to fully express the nuance of their beliefs about the science-religion relationship, and, second, that with only one item representing each mental model, measurement precision cannot be warranted." (p.7). Instead, the authors suggest a scale with 45 items which addresses 5 aspects (conflict, context-switch, compartment, complementary, consonance).

   At the theoretical level, there are merits and problems with their suggested measure. It would indeed be worthwhile having an instrument which assesses more fully individuals' variety of beliefs; however, I am not convinced that a scale with 45 items is necessarily better than one with 4 or 5 items. I am not convinced for two reasons: first, I suspect that Marin & Lindeman did not start out with 4 items but had a longer scale which they piloted and cut down to 4, because those few items worked better than having more items; second, the authors state that they derived their 45 items from previous qualitative studies and that this "results in 45 items" (p. 20), but most of these items are simply paraphrases and simple reiterations of other items (e.g for the 'conflict' model we have 'Religion and science always offer contradictory explanations of the world', 'Scientific explanations always conflict with religious beliefs', 'Science and religion are fundamentally opposed', and 'Religious scriptures are in fundamental conflict with scientific evidence' — the same repetition happens for the other dimensions of the scale).

Thank you for your extensive feedback! We fully agree that a self-report measure like ours should be as brief as possible, but – at the same time – as informative (in terms of content validity) as possible. Yet, increasing content validity often means sacrificing brevity. With that in mind, we re-examined our items carefully and scrutinized the extent to which the resulting scale is indeed a good compromise between content validity and brevity. This re-examination led to three major improvements:

First, we ran Pilot Study 1 to estimate the underlying structure of our scale data and, more generally, to test whether the items worked to elicit unfolding responses from participants. The results indicated that our scale data pointed to a unidimensional, bipolar, unfolding construct, and based on this data, we condensed the scale to 27 items.

Second, we rewrote some of these 27 items to better reflect the varying degrees of wording strength, so that some items are very strongly worded but others are mildly worded (i.e.,

"intermediate" items), indicating a transition between different mental models. In rewriting the item content, we followed recommendations from Cao et al. (2015), which are specifying *conditions* (so that participants' endorsement is conditional on a certain condition, e.g., "***When*** *searching for answers to fundamental questions, science and religion sometimes come to different conclusions*"), using *transitions* (so that the description of "extreme" stance of conflict or compatibility can be avoided, e.g., "*Science can answer certain questions that religion cannot, and vice versa,* ***but the combination of the two*** *makes them equally useful*"), and specifying *lower frequencies* (so that participants who **always** or **never** consider the idea suggested by the item will answer "strongly disagree" to these items, e.g., "*It seems to me that science and religion can* ***sometimes*** *go in the same direction*").

Third, we asked our lab members to rate the wording strength of each item (Pilot Study 2), which allowed us to gauge whether the items reflected different levels of wording strength. We also asked the raters to provide feedback on the content of the items, and based on the results of this pilot study, we further finetuned the items.

We have added a new section, "Pilot Data" (pp. 27-31), to present the results of Pilot Study 1 and 2 as follows:

"*We ran three pilot studies to refine the items and performed an initial test of the measurement assumptions. We ran Pilot Study 1 (n = 614, Female = 61.23%, Male = 36.15%, Others = 2.6%, M_{age} = 39.66, SD_{age} = 16.55) by circulating a study invitation to our participant pool from December 2023 to February 2024 while waiting for the first round of review. We performed an unrotated PCA (Tay & Drasgow, 2012) to conduct a dimensionality test, with all 45 items, by imposing 2 components structure to our data since a bipolar unidimensional unfolding construct typically results in two linear principal components (Nandakumar et al., 2002; Roberts et al., 2000; van Schuur & Kiers, 1994). The emergence of "the extra factor" (Maraun & Rossi, 2001; van Schuur & Kiers, 1994) or a "spurious dimension" (Tay & Drasgow, 2012) shows the existence of a unidimensional, bipolar construct. This is particularly evident when the relationship between the first and second principal components is close to zero or negative (Tay & Drasgow, 2012). When loadings from the first component are plotted against the loadings of the second component, unfolding data show a "simplex-like" pattern (Davison, 1977), where the endpoint of the "simplex" curve is folded inward, demonstrating a semicircular pattern (Davison, 1977; Roberts et al., 2000; Tay & Drasgow, 2012).*

We found that this semicircular pattern exists in our data, and that the loadings of the principal components are clustered within each subscale and ordered along the x-axis, suggesting an ordered sequence and transitions between the mental models. It is important to note, however, that we presented the scale to participants in five blocks, each consisting of only items from one subscale, and then randomized the order of items within the block. Thus, in one block, participants saw only items from the conflict subscale and may have been able to adjust their responses accordingly. In this case, although the items appeared to be neatly clustered according to their respective subscale/mental model (see Figure 1), the scale presentation may have confounded the results.

Moreover, interestingly, the plot shows that the loadings of the compartment items are grouped on the left side, closer to the conflict, while items of the context-switch subscale are grouped closer to the complementary, which slightly differs from our initial hypothesis. The correlation between the first and second principal components is negative (r(43) = -.31, 95% CI [-.55, -.01], p = .042), which supports our assumption that the construct is unidimensional and bipolar (Tay & Drasgow, 2012). We present the PCA plot of the first and the second principal components in Figure 1.

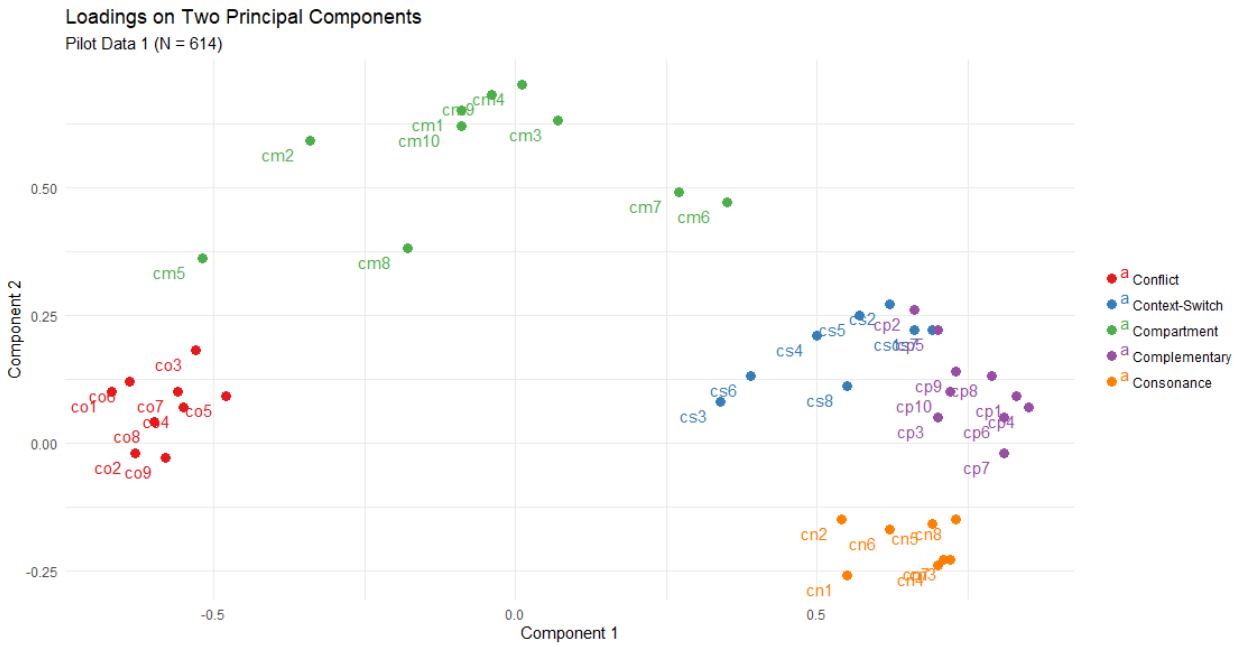

Figure 1. PCA Plot, Pilot Data 1 (n = 614)

We fitted a one-dimensional GGUM and GPCM model to Pilot Data 1, but the models initially did not converge. We suspected that the convergence issues were caused by collinearity between items. Therefore, we computed Yen's Q3 statistics (Yen, 1984) to detect these items and identified three problematic items (i.e., two "conflict" items and one "compartment" item). We subsequently removed these items, re-specified our models, and re-ran our analyses.

The GPCM and the GGUM models with the remaining 42 items were successfully converged. The GGUM model had a good fit to the data ($M_2(693) = 1{,}400.78$, $p < .001$, RMSEA = .040, 90% CI [.037, .043], SRMSR = .069, CFI = .988, TLI = .986), while the GPCM did not ($M_2(735) = 8{,}444.155$, $p < .001$, RMSEA = .120, 90% CI [.117, .112], SRMSR = .165, CFI = .858, TLI = .851). Model comparison (Table 1) shows that, generally, the GGUM model fits the data better.

Table 1. Model Comparison Between Unfolding (GGUM) and Dominance (GPCM) Model, Pilot Study 1

Model	Number of Parameters	LL	AIC	BIC	SABIC	HQ
GGUM	210	-26,669.52	53,911.72	54,687.24	54,020.53	54,120
GPCM	168	-29,784.67	59,923.33	60,705.67	60,143.73	60,227.57

Based on the Pilot Study 1 and feedback from an anonymous reviewer, we condensed the scale items from 45 to 27 items, with only 5-6 items representing each mental model. We further calibrated these items to better reflect an unfolding pattern by rewriting them to have varying degrees of wording strength. Each subscale contains very strongly worded items (e.g., "To me, it is completely inconceivable that science and religion are in conflict since they clearly convey the same fundamental truths") and the more "intermediate" or neutral items (Cao et al., 2015, e.g., "It seems to me that science and religion can sometimes go in the same direction").

To empirically test whether the scale items indeed reflect different levels of wording strength, we asked 8 of our lab members (Pilot Study 2) to rate the wording strength of our revised scale (27 items) from 0 (very mildly worded) to 10 (very strongly worded). To assist the raters in their evaluations, we provided them with specific instructions that included the description

and a concrete example of each mental model/subscale. Intraclass correlation (ICC) analysis suggests that the average scores from all raters are consistent (ICC2k = 0.87, 95% CI [0.78, 0.93]), and the rating variability between items is significantly greater than between raters (F(26, 182) = 8.6, p < .001), denoting sufficient interrater reliability. We also asked raters to provide general feedback on the items, and based on the feedback, we refined these items again to ensure wording strength varies meaningfully, and, most importantly, represent the prototypicality of each mental model.

Additionally, we also asked raters to place all 27 items on a continuum ranging from 0 (completely conflict) to 10 (completely compatible). In doing so, we essentially asked raters to "predict" item locations on the continuum of conflict – compatibility. Average ratings from all raters are highly consistent (ICC2k = 0.94, 95% CI [0.90, 0.97]), and the rating variability between items is significantly greater than between raters (F(26, 182) = 18, p < .001), suggesting good interrater reliability. Interestingly, we saw the similar pattern we found in Pilot Data 1, that the raters placed items reflecting context-switch view as closer to compatibility while items reflecting compartment view were generally closer to conflict."

The detailed results of Pilot Study 2 can be found at the end of this letter. All pilot data are also available in an OSF repository.

2. One other reason why in this particular case 'less may be more', and work better, is because there are huge cultural variations in how science, religion, and their relationship are understood. I could give plenty of anecdotes I have personally encountered working across cultures, but let me give a simple one: if you talk to any anthropologist who works with indigenous peoples in the Amazon, and ask them about science-religion relationship for those peoples, they will all tell you that there is no such distinction between religious and scientific knowledge. A psychologist would naturally counter-argue that an indigenous worldview is a minority, or even an exception; but an historian would remind us that such distinction is a product of relatively recent European-centred history, which does not generalise to all cultures. Recent cross-cultural work on what atheists and agnostics believe in testifies to these differences. In response to a single item question 'The scientific methods is the only reliable path to knowledge' (which implicitly mirrors the 'competition' model of Barbour's taxonomy by excluding religion and other forms of knowledge), both unbelievers and the general population of six countries (European, Asian, and American) answered in surprisingly different ways (https://www.stmarys.ac.uk/research/centres/benedict-xvi/docs/benedict-centre-understanding-unbelief-report.pdf).
   In summary, I don't understand the rationale for having those 45 items, a great part of which are redundant and simply paraphrasing the same statements.

We fully agree that much of the evidence suggests that people in different cultures (or countries) view the relationship between science and religion differently. Specifically, people living in Christian-majority European and North American countries are more likely to hold a conflict belief. We have acknowledged this in our manuscript as follows (pp. 4-5):

"*Empirical evidence has also supported the assertion that the propagation of the "conflict" view (i.e., the belief that science and religion are fundamentally opposed) may only be prevalent in culturally secular countries (Johnson et al., 2020). For example, while religiosity is an important predictor of less favorable attitudes toward science in the U.S., this pattern does not generalize to other countries (Cologna et al., 2024; McPhetres et al., 2020). In*

some countries (e.g., Nigeria, Qatar, Kuwait, Egypt, India, and the Philippines in McPhetres et al., 2020, and Türkiye, Bangladesh, and Malaysia in Cologna et al., 2024) the association is reversed, with higher levels of religiosity corresponding to more favorable attitudes toward science.

Indeed, several global surveys have captured a wide variation across countries in perceptions of the relationship between science and religion. (Kostyukov, 2019; Wellcome Trust, 2021). When participants in the 2020 Wellcome Trust Global Monitor (Wellcome Trust, 2021) were asked "When scientific evidence contradicts religious teachings, do you believe in science or religion?", more than 60 percent of participants in European countries (e.g., Germany, the United Kingdom, Belgium, Italy, etc.) were proponents of science, while more than half of the participants in some other countries (e.g., Brazil, Jordan, Türkiye, and Indonesia) sided with religion. Furthermore, the proportions of participants who said they believed in science and those who sided with religion are about the same in the U.S., India, and Uzbekistan. Interestingly, the number of participants who answered, "science and religion do not disagree" and "it depends" are particularly sizeable in some countries, with more than 20 percent of participants in Israel, Thailand, Croatia, Jordan, and Cambodia (Wellcome Trust, 2021)."

And indeed, most large-scale surveys or studies have focused solely on measuring "perception of conflict" and have ignored other possible mental models. Against this background, we believe it is very important to include items that reflect a wide range of viewpoints that people may hold when thinking about the relationship, and (the extension of) Barbour's taxonomy serves this purpose well. Believing that "*no such distinction between religious and scientific knowledge*" is a prototypical *consonance* model that we have already included in the scale.

The primary goal of this study is to offer items that reflect the content of different mental models so that participants will hopefully find that some items match the nuances in their beliefs (i.e., the *types* of mental model). With our proposed design, the scale will also be able to locate participants' belief in a *continuum* of conflict–compatibility.

3. Another theoretical concern can be phrased as a question: what exactly are the authors trying to achieve with the construction of this scale? Measuring beliefs of this kind has some interest per se, I would agree, but the study and the scale would be tremendously more meaningful if the authors were testing the significance of these beliefs in regard to how they might be related to a set of psychological characteristics or how they might help predict others types of beliefs or behaviours. For example, the authors mention the use of other scales in order to test for construct validity, particularly the Belief in Science scale. I went back and re-read that paper because of the authors' statement that "Farias and colleagues (2013) assert that individuals with strong beliefs in science are skeptical of, or even reject, scientifically unsubstantiated beliefs, including religious explanations". This strikes me as a secondary interest of that work; the merit of that particular paper is its evaluation of a simple hypothesis: that believing in the superiority of science can alleviate stress and anxiety in non-religious individuals. The items are clearly tapping into a strong form of science belief which one might call 'scientism'. What makes the scale interesting is its ability to predict a psychological function (alleviation of stress and anxiety), something which in the literature has usually been associated with religious beliefs.

   The use of that scale in the current study, by contrast, made me think: why go through the trouble of developing a new scale of belief in the relationship between

science and belief and not test its potential relevance? You don't need to have an experiment for this; one can simply think of hypotheses related to proximal beliefs or behaviors that can be assessed either via scales or vignettes. The authors make the case that they don't want to test for cognitive differences, as they assume that the scale makes no assumptions about 'cognitive mastery or intellectual abilities'. That is fine, but what other aspects might be relevant to test when developing such a scale?

Overall, I feel that the authors have not given too much thought to these theoretical aspects because of their over-emphasis on testing the 'unfolding' versus 'dominance' models. The section on 'The Present Study' which should clearly describe the main hypotheses of the study is long and nebulous; the methodological aspects of scale testing they are suggesting are, admittedly, complex but they should NOT detract from spelling out clearly what this new scale is expected to achieve/explain/predict. The authors themselves acknowledge (p19) that understanding how "individuals conceive the relationship between science and religion can inform the process of developing science (or religious) education curricula that emphasize fostering critical thinking and constructive dialogue about the collaborative role of science and religion in the public domain." I would encourage the authors to think now about how to make this study more meaningful, and not at a later point.

We are grateful for this comment, which gave us pause. We agree that justifying the purpose of our entire endeavor is important and something we should have done more carefully in the original version of the paper. We have now added a new section titled "***Future Applications of the Scale***" on pp. 19-20 in order to provide such a justification, and we hope that our new section convinces the reviewer. The section reads as follows:

"*While Barbour's typology plays an influential role in guiding interested scholars in various fields, such as philosophy of science (e.g., Damper, 2022a, 2022b), science education (e.g., Woolley et al., 2023; Yasri & Mancy, 2012), psychology of religion and cognitive science of religion (e.g., Legare & Visala, 2011; Marin & Lindeman, 2021), and sociology of religion (e.g., Baker, 2012; Evans, 2011), the taxonomy has yet to be empirically tested in a quantitative framework. In this sense, designing a scale that reflects Barbour's taxonomical work can serve a twofold function. First, it allows us to provide a "severe test" of Barbour's theory, particularly with respect to the claim that people differ in their views of the relationship between science and religion, and that this difference comes in five types. Second, the scale can be a useful tool for future research projects that may unpack many interesting questions, as the following examples show.*

Scholars argue that individuals dynamically engage with scientific and religious explanations throughout their lives (Davoodi et al., 2019; Ecklund & Park, 2009). At school age, individuals begin to refine their scientific reasoning skills. During this time, they are expected to gradually separate their scientific beliefs from their religious beliefs, with the former eventually replacing the latter (i.e., "conceptual change," cf. Posner et al., 1982). Consequently, how individuals perceive the relationship between science and religion may also change over time. Similarly, major life events, such as living abroad and interacting with people from different cultural and religious backgrounds, may also contribute to changes in individuals' perceptions of the relationship between science and religion (Rios & Aveyard, 2019; Rios & Roth, 2020). In this sense, our scale will serve as a useful tool for investigating these questions much more systematically and rigorously than previous research did.

Another interesting and related question is how individuals relate to scientific and religious explanations in times of major societal crises, such as pandemics (Ayub et al., 2023; Jackson et al., 2020), anthropogenic climate crises (Arbuckle, 2017; Jenkins et al., 2018), or

armed conflicts (Shai, 2022; Tarusarira, 2014). In these times of crisis, individuals are confronted with existential threats, and both science and religion offer to meet these existential needs (Rutjens & Preston, 2020). Therefore, it is possible that individuals who do not perceive a strong conflict between science and religion can flexibly use scientific and religious explanations during an existential crisis.

Relatedly, discerning how individuals relate to scientific and religious explanations is critical for optimizing science communication strategies (Elsdon-Baker & Lightman, 2020) when discussing controversial topics related to the role of science and religion in the public sphere, such as evolution, genetic engineering, space exploration, vaccines, and many others (Drummond & Fischhoff, 2019; Lobato & Zimmerman, 2019). Again, our scale will be useful in this context. It is important to note, however, that our current study focuses on rigorously testing the core tenet of Barbour's theory, which we believe is a critical first step before answering any of the substantive questions sketched here."

As noted above, creating the scale and testing Barbour's theory is a major focus of this study. We feel that exploring any of these questions in this present study is a digression from the main story of this manuscript, while these questions deserve a thorough (but separate) research program. To maximize future usability of the scale, we are committed to writing and making publicly available an annotated R script so that future researchers interested in using our scale to investigate these questions can derive the estimated person parameters (θ) from a GGUM model of the scale data. This allows researchers to use the estimated θ for their analysis, rather than the average/sum scores, which typically do not work with unfolding data.

In addition to this, we will still conduct discriminant (i.e., belief in science), convergent (i.e., perceptions of science-religion relationship by Leicht et al., 2021), and criterion-related validity (i.e., centrality of religiosity) so that at least we still have some "predictions" to test.

4. My main concerns are with the theoretical nexus of the study. But I should add that, at the methodological level, I am concerned by three aspects: first, why don't they run a pilot study before moving into the major data collection, simply to finetune the items (and get rid of some). That is the common good practice. I don't want to preach about the importance of running good pilots but, given a recent major flop in clinical psychology which involved many millions of dollars testing over 8,000 school children in England to assess the effects of mindfulness only to find out that 80% of the children were not motivated enough to care about doing the intervention, I feel one should be reminded that good pilots are more likely to make good studies.

Thanks for your critical feedback. We agree that pilot studies significantly improve the quality of the scale before proceeding to a major data collection. We have described how we collected Pilot Data 1 and 2 in response to point #1, and in addition to these pilot studies, we conducted a Pilot Study 3 to ensure that the refined version of the scale worked as we expected. We presented the results of Pilot Study 3 as follows (pp. 31-33):

"*To test whether our refined scale works well as intended, we collected Pilot Data 3 by inviting PsyWeb subscribers (n = 1,111, Female = 73.17%, Male = 25.11%, Others = 1.65%, Mage = 46.12, SDage = 16.20), and ran PCA with all 27 items, imposing two principal components structure. This time, we presented 27 items in three blocks, with items from all subscales intermixed. In the first and second blocks, we included two items from each subscale, for a total of ten items per block. The third block consisted of the remaining seven items. Additionally, we randomized the order of the items, so the scale items were presented differently for each participant. Nonetheless, items were still clustered within the same*

subscale (Figure 2), but the items were less neatly clustered compared to Pilot Data 1 (Figure 1). The correlation between the first and second principal components is moderately negative (r(25) = -.51, 95% CI [-.75, -.16], p = .007), further supporting the existence of a unidimensional, bipolar construct (Tay & Drasgow, 2012).

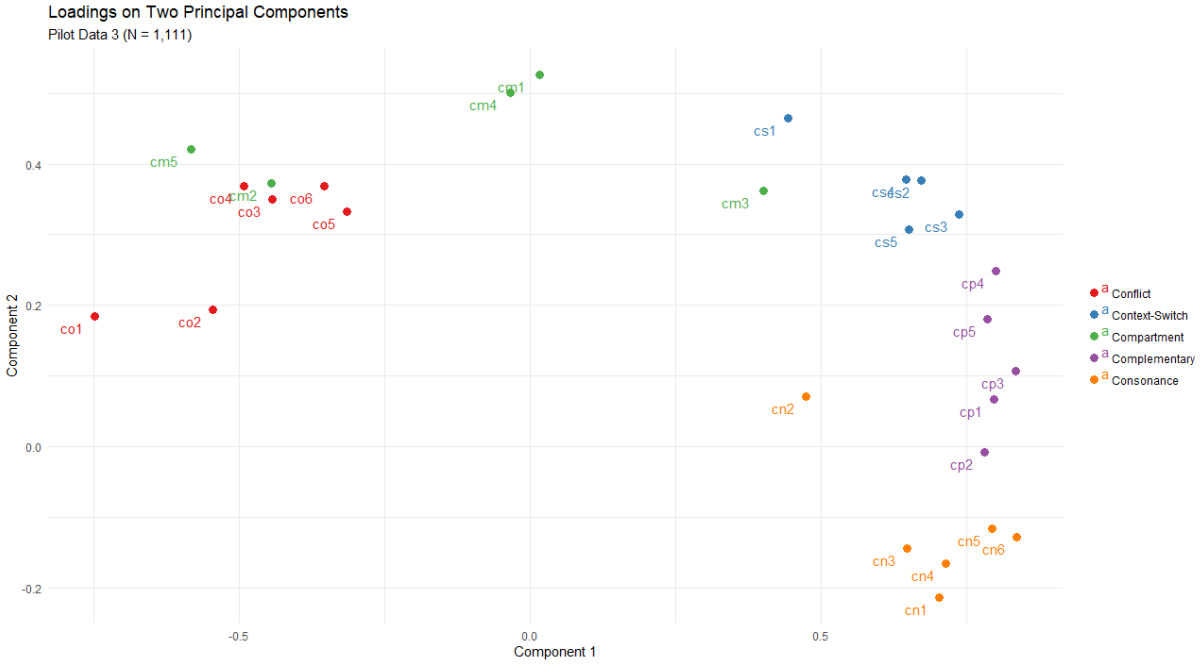

Figure 2. PCA Plot, Pilot Data 3 (n = 1,111)

Again, we fitted a one-dimensional GGUM and GPCM model to Pilot Data 3, and both models were converged properly. The GGUM model fit the data well (M2(243) = 1,056.22, p < .001, RMSEA = .054, 90% CI [.051, .058], SRMSR = .075, CFI = .984, TLI = .981), while GPCM did not (M2(270) = 3,217.62, p < .001, RMSEA = .099, 90% CI [.096, .102], SRMSR = .080, CFI = .886, TLI = .875). Replicating our findings from Pilot Study 1, the GGUM fits our data better (Table 2)."

Table 2. Model Comparison Between Unfolding (GGUM) and Dominance (GPCM) Model, Pilot Study 3

Model	Number of Parameters	LL	AIC	BIC	SABIC	HQ
GGUM	135	-30,620.44	61,510.89	62,187.65	61,758.85	61,766.79
GPCM	108	-30,945.31	62,106.62	62,648.02	62,304.99	62,311.33

5. Second, I found the rationale to undertake the scale validation in the USA and Germany utterly unconvincing. If these countries were not chosen simply because of convenience (co-authors being from each country), then I would strongly recommend the authors to choose truly culturally diverse countries (e.g. USA and India).

The scale was originally developed in Germany, so we have tested it (in our pilot studies) and will test it again (in this current study) in German samples. Then, we choose the USA as a reference group because most previous studies on the same topic have been conducted with US samples (probably the issue is of particular societal importance in this country). We have added the following text in the last part of "Present Study" section (p. 18-19) as follows:

"*…We chose the United States as the target group because much previous research on this topic has been conducted with U.S. samples – probably because the issue is of particular*

societal importance in this country (J. O. Baker, 2012; Ecklund & Park, 2009; Scheitle, 2011)."

...and deleted the arguments that delineate the differences between Germany and the United States, which we also agree are unnecessary and unconvincing.

6. Third, the rationale to include different groups (religious vs atheists vs agnostics) is also missing. What is the benefit — theoretical or methodological — of including these groups? And why these groups in particular and not a variety of religious ones, such as Buddhists and Muslims?

Thanks for pointing this out! We agree that estimating measurement invariance between religious vs. atheists vs. agnostics is less justified. While testing equivalence between different religious groups is more justified and interesting, we decided not to do this because we need to recruit a large number of participants (~700-800) for each religious group, which we could not do in this present research. This is because a GGUM model requires a larger sample size than the other IRT models due to its parameterization complexity.

7. A side note on this passage of the manuscript: "Since culture plays an important role in shaping how individuals interpret the relationship between science and religion (Johnson et al., 2020), we expect the items of our scale to work similarly (i.e., similar item discrimination and threshold parameters) across groups (i.e., countries and group membership)."

   I am confused: if culture plays an important role in shaping individuals' interpretations, why should the items work similarly across cultures and group membership?

We apologize for the lack of clarity in the previous version of our manuscript. We have deleted the sentence "Since culture plays…" to improve clarity and added a few sentences to clarify what we are trying to accomplish with the planned DIF/measurement invariance analysis as follows (p. 18):

"*Finally, it is important to make sure that the items yield similar response patterns from two individuals with similar levels of perceptions, regardless of where these individuals come from. Therefore, we expect the items of our scale to work similarly (i.e., similar item discrimination and threshold parameters) across countries. Since our scale is developed in Germany, we will compare data from a German sample (as a reference group) with data from a U.S. sample to scrutinize its measurement invariance. More precisely,* ***we expect that participants with the same level of perceptions (e.g., similar levels of conflict belief)****, regardless of their country of residence (Germany vs. the United States)* ***have the same probability of responding to the response category*** *(e.g., equally likely to choose "strongly agree" with the item "Science and religion are ultimately at odds, with no possibility of harmony").*"

The sentence "Since culture plays…" confuses the reader between culture as a factor explaining why two samples *averagely* differ in their perceptions of the relationship between science and religion and culture as a source of measurement bias. Therefore, we removed this.

8. The title: what is the adverb 'mentally' adding? Do you need it? The section entitled 'Measuring How Individuals Relating Science to Religion' should say 'relate'.

Corrected as suggested, thanks!

Reviewer #4

1. This is an interesting proposal to develop a questionnaire measuring how people perceive the relation between science and religion. I think that theoretically, the proposal is strong and the writing of the proposal is well. However, the writing of the data analysis plan is sloppy and contains errors (see below for some examples but there are many more). Most importantly, the analysis proposed is ad-hoc, suboptimal, and will potentially cause the results to be uninterpretable. The authors use all kinds of "ad hoc tricks" to conduct the analysis, while a statistically sound analysis based on explicit models is straightforward to conduct using the mirt package that the authors use. See below for some detailed comments (in chronological order).

Thank you so much for your critical feedback! With your feedback, we were able to improve the quality, accuracy, and transparency of our analysis plan.

2. "Since the "dominance" and "unfolding" models make different assumptions about the relationship between response probability and latent trait, misfitting a "dominance" model to "unfolding" data can lead to several issues and inaccuracies. This misapplication can result in poor model fit and incorrect estimations of item parameters."

   Please give a reference (as violating an assumption will not necessarily result in parameter bias)

Thank you for your correction! You are right that misfitting dominance model to unfolding data can only affect model and item fit, but not necessarily parameter bias. Therefore, we corrected this as follows (p. 14):

"*Since the "dominance" and "unfolding" models make different assumptions about the relationship between response probability and latent trait, misfitting a "dominance" model to "unfolding" data can lead to several issues and inaccuracies, such as* ***poor model and item fit*** *(Roberts et al., 2000; Stark et al., 2006)."*

3. Page 14
   "…or, more precisely, the points on the trait continuum at which participants move from one response category to the next (i.e., the step functions)."

   The authors discuss IRT models as if these are deterministic while they are probabilistic. That is, and IRT model does not use a step function but a probabilistic function (socalled 'category response functions') which model the probability of responding in a certain category. As such, there are no 'points on the trait continuum at which participants move from one response category to the next'. The threshold parameters determine the location of the probabilistic function on the trait continuum.

   -- thresholds (b, i.e., trait levels at which participants move from one response category to another) See above, thresholds are not interpreted like that

Thank you for this correction. We changed the text as follows (p. 15):

"*Since our proposed scale uses polytomous responses, it is important to model how trait levels are related to the probability of endorsing a specific response category (e.g., "strongly disagree") – or, more precisely, the points on the trait continuum* ***at which participants are***

equally likely *to endorse one response category and its next adjacent category (i.e., the "step functions" or item thresholds or b)."*

…and deleted the latter as it was an unnecessary repetition.

4. --"graded partial credit model - GCPM"

   This model does not exist. The authors probably mean the generalized partial credit model. In addition the abbreviation should be GPCM

Thank you for this correction! We have corrected this error and double-checked that it does not appear anywhere in the manuscript.

5. *Page 15
   --incorporates both item location (δi) and discrimination parameters Here the authors suddenly use a subscript i, without explaining where i stands for. In addition, i should then also be used for the discrimination and threshold parameter from the dominance model.
   --Importantly, both GCPM and GGUM rest on the assumption that the conflict-compatibility continuum (Figure 1) is unidimensional and bipolar.

   This is only true for the unidimensional GPCM and GGUM. For this study, multidimensional models seem the most appropriate choice.

We removed the *i* subscript to avoid the confusion. Thank you for pointing this out!

6. *Page 17
   --we expect that participants with the same level of compatibility perceptions, regardless of their country of residence (Germany vs. the U.S.) or religiosity (religious vs. agnostics and atheists) have the same probability of responding to the items.

   I think the authors mean 'the probability of a certain response' (probability of responding relates to missing data which has nothing to do with the current analysis)

We changed the text as follows (p. 18):

"*More precisely, we expect that participants with the same level of perceptions (e.g., similar levels of conflict belief), regardless of their country of residence (Germany vs. the United States) have* ***the same probability of responding to the response category*** *(e.g., equally likely to choose "strongly agree" with the item "Science and religion are ultimately at odds, with no possibility of harmony").*"

7. *Page 19
   --"Since…"
   this sentence ends abruptly

Corrected, thank you.

8. *Page 20
   --"This results in 45 items…"
   no need for a new paragraph here (previous paragraph is only one sentence)

Corrected, thank you.

9. --Participants will be asked to indicate their agreement level to these items in four possible responses; strongly disagree (1), disagree (2), agree (3), and strongly agree (4).
   I would strongly advice to use a midpoint, forcing subjects to have an opinion will affect the validity of the scale

Thank you for your suggestions! However, we decided not to use the midpoint option as this is explicitly suggested by Dalal et al. (2014), as midpoints does not consistent with the theory of unfolding response. Individuals who stand in the middle of a continuum (i.e., "neutral" stance) will endorse "intermediate" or neutral items instead, and thus, midpoints are not needed.

In practice, incorporating midpoint into the scale can jeopardize model fit, as shown in Dalal et al.'s (2014) study. Many previous studies that used unfolding scales also did not use midpoints (Cao et al., 2018; Freund & Lohbeck, 2021; Roberts et al., 2000; Roberts & Laughlin, 1996; Sun et al., 2021).

10. *Page 22
    "Before running the IRT models, we will test the dimensionality of each mental model. This step is crucial because an IRT model requires an assumption that the scale is unidimensional and locally independent (Bock & Gibbons, 2021)."
    you should use multidimensional IRT models for these data

Thank you for this suggestion. Yet, it is important to note that the theory of conflict-compatibility continuum that we would like to test in this study suggests that the construct is unidimensional and bipolar even though we also assume that the *types* of mental models are qualitatively distinct.

We are inspired by a previous study by Freund & Lohbeck (2021) which describes this idea and tests it against the motivational types in the Self-Determination Theory (SDT) framework. In this study, Freund & Lohbeck (2021) showed that while the motivation types are qualitatively different (e.g., *introjected* implies that motivation arises from internalized pressures and obligations, while *intrinsic* suggests that motivation comes from personal satisfaction and enjoyment of activities), they also reflect certain *levels* of autonomous behavior regulation (e.g., *introjection* is less autonomous than *intrinsic*). Freund & Lohbeck (2021) also show that motivation (in the SDT framework) is unidimensional and bipolar, with motivation types arranged on a continuum from low to high autonomy, and that the unfolding model fits their data better than the dominance model.

We also propose the same idea that mental model *types* are indeed qualitatively different, but they involve certain *levels of* conflict/compatibility. For example, individuals who hold the *compartment* model tend to view science and religion as having different methods that are valid to answer completely independent sets of questions, while those who hold the consonance model see no distinction between science and religion. These mental model *types* are substantially different, but they posit a certain *level of conflict/compatibility*, with compartment being around in the middle of the continuum, and consonance being closer to the "compatibility" side.

We also found the evidence for unidimensionality in our pilot study 1 and 3, and we present it to answer your point #11 below.

11. *Page 22
    --"To that end, we will perform a parallel analysis with a polychoric correlation matrix as input to determine the number of optimal factors.

    Next, we will also perform principal component analysis (PCA) to identify grouping patterns among the items (De Ayala & Hertzog, 1991). Items loading on the first two principal components and yielding item-level communalities greater than 0.3 will be considered optimally unidimensional for further analyses, according to evidence from simulation studies (Roberts, 2018; Roberts et al., 2000; Roberts & Laughlin, 1996)."

    Parallel analysis and PCA implicitly assume the dominance model (in fact, parallel analysis on the polychoric correlation matrix is equivalent to the exploratory graded response model, which is a dominance model highly related to the the GPCM used by the authors). So I would not use those here. You should look at the fit of the multidimensional IRT model and see if there is any misfit in the factor structure. If there is no notable misfit, there is some evidence supporting the theoretical structure

Thank you for pointing this out! You are correct that irt.fa() function that we initially planned to use essentially performs an exploratory IRT, and thus, we removed this part and discarded it from our analysis plan.

However, we will still plan to run PCA, because it has been shown computationally that PCA, despite its underlying "dominance" model assumption, can capture unfolding patterns in scale data (Davison, 1977; Ross & Cliff, 1964), which we also showed in our Pilot Study 1 (pp. 27-29).

We ran three pilot studies to refine the items and performed an initial test of the measurement assumptions. We ran Pilot Study 1 (n = 614, Female = 61.23%, Male = 36.15%, Others = 2.6%, M_{age} = 39.66, SD_{age} = 16.55) by circulating a study invitation to our participant pool from December 2023 to February 2024 while waiting for the first round of review. We performed an unrotated PCA (Tay & Drasgow, 2012) to conduct a dimensionality test, with all 45 items, by imposing 2 components structure to our data since a bipolar unidimensional unfolding construct typically results in two linear principal components (Nandakumar et al., 2002; Roberts et al., 2000; van Schuur & Kiers, 1994). The emergence of "the extra factor" (Maraun & Rossi, 2001; van Schuur & Kiers, 1994) or a "spurious dimension" (Tay & Drasgow, 2012) shows the existence of a unidimensional, bipolar construct. This is particularly evident when the relationship between the first and second principal components is close to zero or negative (Tay & Drasgow, 2012). When loadings from the first component are plotted against the loadings of the second component, unfolding data show a "simplex-like" pattern (Davison, 1977), where the endpoint of the "simplex" curve is folded inward, demonstrating a semicircular pattern (Davison, 1977; Roberts et al., 2000; Tay & Drasgow, 2012).

We found that this semicircular pattern exists in our data, and that the loadings of the principal components are clustered within each subscale and ordered along the x-axis, suggesting an ordered sequence and transitions between the mental models. It is important to note, however, that we presented the scale to participants in five blocks, each consisting of only items from one subscale, and then randomized the order of items within the block. Thus, in one block, participants saw only items from the conflict subscale and may have been able to adjust their responses accordingly. In this case, although the items appeared to be neatly clustered according to their respective subscale/mental model (see Figure 1), the scale presentation may have confounded the results.

Moreover, interestingly, the plot shows that the loadings of the compartment items are grouped on the left side, closer to the conflict, while items of the context-switch subscale are grouped closer to the complementary, which slightly differs from our initial hypothesis. The correlation between the first and second principal components is negative (r(43) = -.31, 95% CI [-.55, -.01], p = .042), which supports our assumption that the construct is unidimensional and bipolar (Tay & Drasgow, 2012). We present the PCA plot of the first and the second principal components in Figure 1.

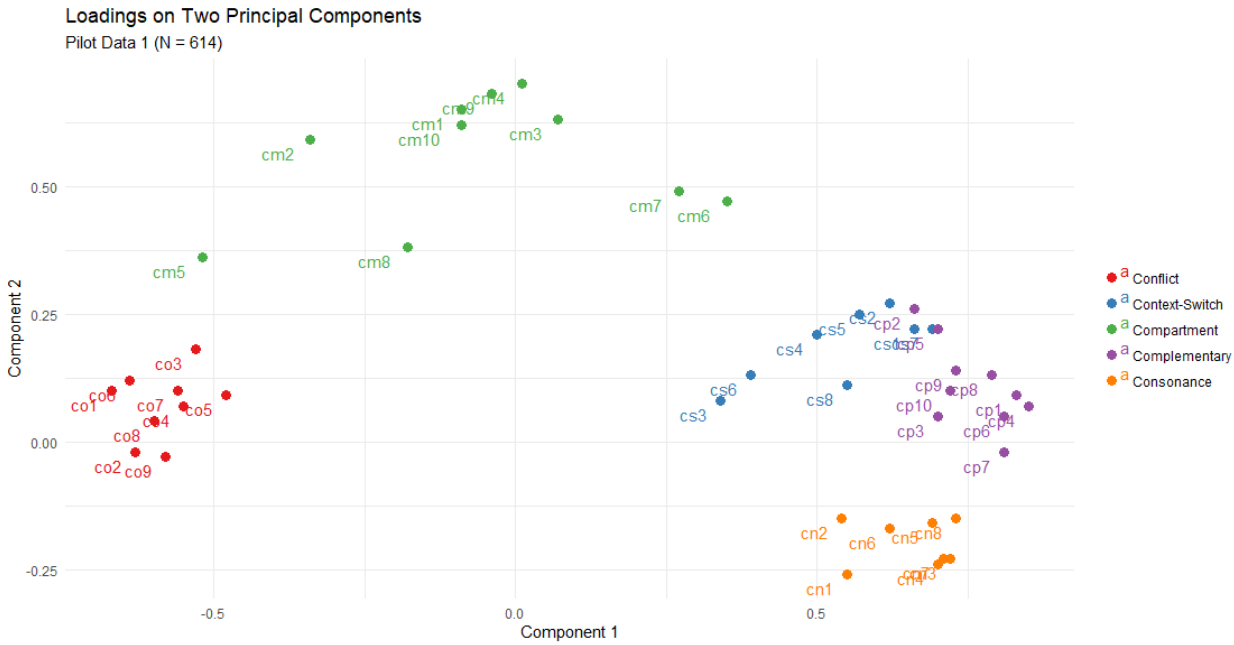

Figure 1. PCA Plot, Pilot Data 1 (n = 614)

We fitted a one-dimensional GGUM and GPCM model to Pilot Data 1, but the models initially did not converge. We suspected that the convergence issues were caused by collinearity between items. Therefore, we computed Yen's Q3 statistics (Yen, 1984) to detect these items and identified three problematic items (i.e., two "conflict" items and one "compartment" item). We subsequently removed these items, re-specified our models, and re-ran our analyses.

The GPCM and the GGUM models with the remaining 42 items were successfully converged. The GGUM model had a good fit to the data ($M_2(693) = 1{,}400.78$, $p < .001$, RMSEA = .040, 90% CI [.037, .043], SRMSR = .069, CFI = .988, TLI = .986), while the GPCM did not ($M_2(735) = 8{,}444.155$, $p < .001$, RMSEA = .120, 90% CI [.117, .112], SRMSR = .165, CFI = .858, TLI = .851). Model comparison (Table 1) shows that, generally, the GGUM model fits the data better."

Table 1. Model Comparison Between Unfolding (GGUM) and Dominance (GPCM) Model, Pilot Study 1

Model	Number of Parameters	LL	AIC	BIC	SABIC	HQ
GGUM	210	-26,669.52	53,911.72	54,687.24	54,020.53	54,120
GPCM	168	-29,784.67	59,923.33	60,705.67	60,143.73	60,227.57

…and Pilot Study 3 (pp. 31-33) as follows:

"*To test whether our refined scale works well as intended, we collected Pilot Data 3 by inviting PsyWeb subscribers (n = 1,111, Female = 73.17%, Male = 25.11%, Others = 1.65%, Mage = 46.12, SDage = 16.20), and ran PCA with all 27 items, imposing two principal*

components structure. This time, we presented 27 items in three blocks, with items from all subscales intermixed. In the first and second blocks, we included two items from each subscale, for a total of ten items per block. The third block consisted of the remaining seven items. Additionally, we randomized the order of the items, so the scale items were presented differently for each participant. Nonetheless, items were still clustered within the same subscale (Figure 2), but the items were less neatly clustered compared to Pilot Data 1 (Figure 1). The correlation between the first and second principal components is moderately negative (r(25) = -.51, 95% CI [-.75, -.16], p = .007), further supporting the existence of a unidimensional, bipolar construct (Tay & Drasgow, 2012).

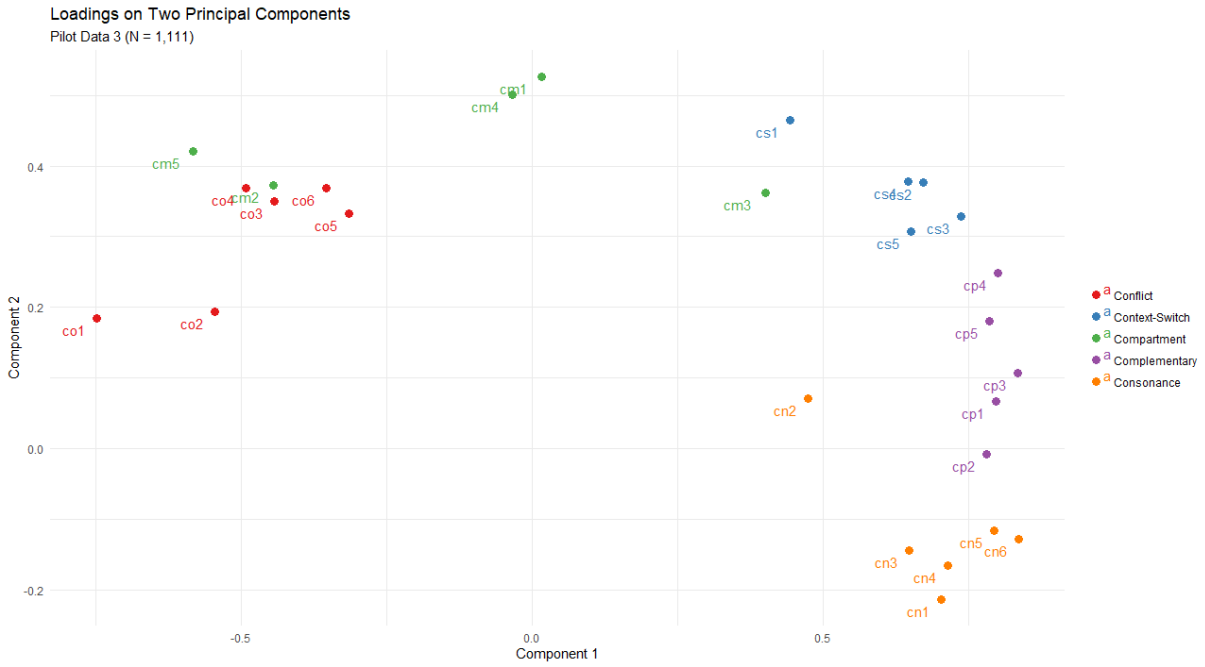

Figure 2. PCA Plot, Pilot Data 3 (n = 1,111)

Again, we fitted a one-dimensional GGUM and GPCM model to Pilot Data 3, and both models were converged properly. The GGUM model fit the data well (M2(243) = 1,056.22, p < .001, RMSEA = .054, 90% CI [.051, .058], SRMSR = .075, CFI = .984, TLI = .981), while GPCM did not (M2(270) = 3,217.62, p < .001, RMSEA = .099, 90% CI [.096, .102], SRMSR = .080, CFI = .886, TLI = .875). Replicating our findings from Pilot Study 1, the GGUM fits our data better (Table 2).”

Table 2. Model Comparison Between Unfolding (GGUM) and Dominance (GPCM) Model, Pilot Study 3

Model	Number of Parameters	LL	AIC	BIC	SABIC	HQ
GGUM	135	-30,620.44	61,510.89	62,187.65	61,758.85	61,766.79
GPCM	108	-30,945.31	62,106.62	62,648.02	62,304.99	62,311.33

12. *Page 23
    --"Lower values of LL, AIC, BIC, AICc, and SABIC indicate a better fit to the data (Kang et al., 2009)."

    Higher values of LL indicate better fit to the data. In addition, are these fit measures suitable in comparing these two models? (have they been used before? Have they been shown to identify the correct model?)

Thank you for this! We corrected the text as follows (p. 25):

"*Higher values of LL and lower values of AIC, BIC, AICc, and SABIC indicate a better fit to the data (Kang et al., 2009).*"

Yes, a study by Freund & Lohbeck (2021) we mentioned to address your point #11 also used the same criteria to test GPCM against GGUM.

13. --"To interpret
    model parameters (a and b), we will use a guide from F. B. Baker and Seock-Ho (2017), with parameters exceeding 0.00, 0.35, 0.65, 1.35, and 1.70, will be interpreted as very low, low, moderate, high, and very high,…"

    First, these are highly arbitrary cut-off as they depend on the variance of θ in the data. You can better consult the standardized estimates. Second, these cut-off only apply to the discrimination parameters (not to the thresholds as these range from -Inf to Inf)

Thank you for your helpful suggestion! We totally agree that the cut-off is arbitrary and indeed, we will calculate the standardized estimates as well, so that we modified the text as follows (p. 25):

"*To interpret item discrimination (a), we build upon F. B. Baker and Seock-Ho (2017), who suggested that parameters exceeding 0.00, 0.35, 0.65, 1.35, and 1.70, should be interpreted as very low, low, moderate, high, and very high, respectively. It is important to note, however, that these cut-offs are somewhat arbitrary and can vary depending on the variance of the latent trait (θ) in the data. Therefore, we will also calculate standardized discrimination parameters to account for the specific distribution of θ in our sample.*"

14. Page 24

    the procedures to test for validity as discussed on page 24 is highly uncommon and ad-hoc. First, estimating θ and then correlating it to other variables (or other θ parameters) neglects the standard errors of θ attenuating correlations and power. Correlations between θ and other variables should be conducted within the IRT model of choice (at the latent level) which is straightforward in the mirt package used by the authors. In this way, measurement error is accounted for and the ad-hoc CFA procedure described is not needed.

Thank you so much for your critical feedback. We realize that correlating factor scores (from a CFA model) and estimated person parameters (from an IRT model) is not an optimal approach. We just found out that it is indeed possible to estimate a multidimensional IRT model (with mirt) with mixed item types. Therefore, even if our scale data follow the "unfolding" model, while the criterion we use to test discriminant (i.e., belief in science), convergent (i.e., perception of compatibility by Leicht et al., 2021), and criterion-related validity (i.e., centrality of religiosity) follows the "dominance" model, we can still estimate them simultaneously in a multidimensional IRT model.

Therefore, we decided to modify our plan as follows (pp. 25-26):

"*We will also estimate a multidimensional IRT model. In this process, we will simultaneously estimate our scale model and BISS model, and then correlate the two constructs at the latent level to test for discriminant validity. In specifying this model, we will set items reflecting perceptions of the relationship between science and religion as "unfolding" (i.e., GGUM), while belief in science items as "dominance" (e.g., GPCM).*

Currently, there is no consensus on the cut-off for a correlation between two constructs that indicates problems with discriminant validity. To fill this gap, Rönkkö and Cho (2022) propose a classification system, rather than a specific cut-off, in which the upper limit of the 95% confidence interval (CI) of correlations below 0.8 may indicate the absence of discriminant validity problems. Therefore, we expect that the upper bound of the 95% CI of the latent correlation between perceptions of the relationship between science and religion, and belief in science to be less than 0.8 (Rönkkö & Cho, 2022).

Next, we will compare the newly developed scale with the scale proposed by Leicht et al. (2012). To do so, we will follow the same procedure as in the discriminant validity test above. That is, we will specify a multidimensional IRT model in which items reflecting the latent dimensions (i.e., explanations and human-world interactions) of the Leicht et al. (2012) scale are specified as "dominance" (i.e., GPCM) items, while our scale items are specified as "unfolding" (i.e., GGUM), and then, correlate them with each other at the latent level."

Strength of item wording – Results of Pilot Study 2

We asked a number of raters (n = 8) to rate the strength of wording of each subscale, by rating each item from 0 (very mildly worded) to 10 (very strongly worded). To help raters evaluate, we provided them with specific instructions, with the description of each model/subscale and an example of it.

For example:

> "*Complementary view refers to the idea that scientific and religious explanations are mutually supportive and thus can be combined, with science explaining the factual aspect while religion focuses on normative content. However, this combination of explanations often tends to be vaguely stated.*
>
> *For example, someone may think that their failure on an exam is due to the fact that they are underprepared (scientific explanation), but their religious beliefs may tell them that the failure is God's punishment, teaching them a valuable lesson about the importance of preparation. However, how these two explanations work together (e.g., who actually causes the failure? The person or God?) is often very vague and is less important to a person with a complementary view.*
>
> *Below are the items that reflect* ***Complementary*** *subscale. Please click, drag, and drop each item to the position on the continuum where its attitude/tone strength best fits. When evaluating the attitude strength of these items, please focus* ***only on this subscale*** *and ignore any other subscales you have seen previously.*"

Intraclass correlation (ICC) analysis suggest *average scores* from all raters are consistent (ICC2k = 0.87, 95% CI [0.78, 0.93]), and the rating variability ***between*** items are significantly greater than ***within*** items (F(26, 182) = 8.6, p < .001), implying sufficient interrater reliability.

Items in these tables below are sorted from highest (strongly worded) to lowest (mildly worded) based on their mean ratings.

Conflict

Item No.	Original Items	Revised Items	Why?	*M*
2	Entscheidungen auf der Basis von Wissenschaft zu treffen, erfordert die vollständige Aufgabe religiöser Überzeugungen. *Making decisions on the basis of science requires the complete abandonment of religious beliefs.*	Um Entscheidungen auf der Basis von Wissenschaft zu treffen, muss man seine religiösen Überzeugungen vollständig aufgeben. *In order to make decisions on the basis of science you need to completely abandon your religious beliefs.*	Fixing its structure.	9.41

Item No.	Original Items	Revised Items	Why?	*M*
1	Wissenschaft und Religion sind letztlich unvereinbar und können nicht miteinander versöhnt werden. *Science and religion are ultimately at odds, with no possibility of reconciliation.*	Wissenschaft und Religion sind letztlich unvereinbar und können nicht in ***Einklang gebracht*** werden. *Science and religion are ultimately at odds, with no possibility of* ***harmony****.*	A rater suggested that *Einklang gebracht* fits better in this context, and even, more straightforward than "Versöhnung".	8.69
3	Religion und Wissenschaft bieten immer widersprüchliche Erklärungen für jedes Phänomen in der Welt. *Religion and science always offer contradictory explanations for every phenomenon in the world.*	Religion und Wissenschaft bieten widersprüchliche Erklärungen für ***so gut*** wie jedes Phänomen in der Welt. *Religion and science offer contradictory explanations* ***for almost*** *every phenomenon in the world.*	Making this item less strongly worded so that it is placed in around the middle. A rater suggested this as well.	7.97
4	Meistens scheinen Wissenschaft und Religion gegensätzliche Standpunkte zu vertreten. *More often than not, science and religion seem to offer opposing viewpoints.*	No change.		4.32
5	Es gibt einige Fälle, in denen religiöse Schriften und wissenschaftliche Erkenntnisse nicht übereinstimmen. *There are some instances where religious scriptures and scientific evidence don't align.*	No change.		2.81
6	Bei der Suche nach Antworten auf grundlegende Fragen kommen Wissenschaft und Religion manchmal zu unterschiedlichen Ergebnissen. *When searching for answers to fundamental questions, science and religion sometimes come to different conclusions.*	No change.		1.91

Context-Switch

Item No.	Original Items	Revised Items	Why?	*M*
1	Ich kann mich je nach Situation entweder voll auf die Wissenschaft	Ich kann je nach Situation ***komplett umschalten, ob ich mich auf*** die	A rater suggested to simplify the sentence structure.	7.89

Item No.	Original Items	Revised Items	Why?	*M*
	verlassen oder mich meinem Glauben hingeben. *I can fully switch between relying on science and embracing my faith, depending on the situation.*	Wissenschaft oder auf meinen Glauben ***verlasse***. *Depending on the situation, I can fully switch between relying on science* ***or my faith****.*		
2	In bestimmten Situationen bevorzuge ich stark die Wissenschaft, in anderen verlasse ich mich ganz auf meine religiösen Überzeugungen. *I strongly favor science in certain situations, but embrace my religious beliefs entirely in others.*	In bestimmten Situationen bevorzuge ich stark die Wissenschaft, in anderen verlasse ich mich ganz auf meine religiösen Überzeugungen. *In certain situations, I strongly favor science, but in others, I entirely embrace my religious beliefs.*	Fixing its structure so that English and German version is equivalent.	7.65
4	Es ist nicht ungewöhnlich, dass ich flexibel von wissenschaftlichen Erkenntnissen zu meinem Glauben wechsle oder umgekehrt. *It's not unusual for me to flexibly switch from relying on scientific evidence to embracing my faith, or vice versa.*	Ich kann flexibel wechseln zwischen meinem Vertrauen in wissenschaftliche Erkenntnisse und meinem Glauben. *I can flexibly switch between my trust in scientific evidence and my faith.*	Fixing and simplifying its structure.	3.8
3	In bestimmten Situationen vertrete ich wissenschaftliche Standpunkte, in anderen hingegen bekenne ich mich eher zu religiösen Überzeugungen. *In specific situations, I endorse scientific viewpoints, but in others, I rather embrace religious beliefs.*	In bestimmten Situationen **kann** ich wissenschaftliche Standpunkte vertreten, in anderen hingegen bekenne ich mich eher zu religiösen Überzeugungen. *In certain situations, I* ***can*** *endorse scientific viewpoints, but in others, I rather embrace religious beliefs.*	Fixing its structure.	3.62
5	Die Bedeutung, die ich der Wissenschaft im Vergleich zur Religion beimesse, kann je nach der Situation variieren. *The importance I place on science relative to religion can vary, depending on the situations.*	Die Bedeutung, die ich der Wissenschaft im Vergleich zur Religion beimesse, kann je nach Situation variieren. *The importance I place on science relative to religion can vary, depending on the situations.*	Removing "der" before "Situation".	2.5

Compartment

Item No.	Original Items	Revised Items	Why?	*M*
2	Die Ziele von Wissenschaft und Religion sind so unterschiedlich, dass die eine nicht die Funktion der anderen übernehmen kann. *The purposes of science and religion are so unique that one cannot serve the function of the other.*	Die Ziele von Wissenschaft und Religion sind so unterschiedlich, dass die eine nicht die Funktion der anderen übernehmen kann. *The purposes of science and religion are so **different** that one cannot serve the function of the other.*	Fixing the English translation	8.1
4	Wissenschaft und Religion nutzen unterschiedliche Methoden, um die Welt zu verstehen, und jede ist nur innerhalb ihres eigenen Bereichs gültig. *Science and religion use distinctive methods to make sense of the world, each valid only within its own domain.*	Wissenschaft und Religion nutzen **sehr verschiedene** Methoden, um die Welt zu verstehen, und ***jede Methode ist nur in ihrem jeweiligen*** Bereich gültig. *Science and religion use very distinctive methods to make sense of the world, and **each method** is valid only within its own domain.*	Making it clear that the validity statement refers to the methods and simplifying the sentence structure.	6.16
1	Wissenschaft und Religion bieten Erklärungen für völlig unterschiedliche Fragestellungen. *Science and religion offer explanations for entirely separate sets of questions.*	No change.		6.07
5	Wissenschaftliche und religiöse Erklärungen mögen in ihren jeweiligen Bereichen funktionieren, bleiben aber voneinander getrennt. *Scientific and religious explanations may work in their respective domains, but they remain separate.*	No change.		5.8
3	Meiner Ansicht nach befassen sich Wissenschaft und Religion mit verschiedenen Aspekten unserer Existenz, jede auf ihre eigene Art und Weise. *In my view, science and religion deal with different aspects of our existence, each in its own unique way.*	Meiner Ansicht nach befassen sich Wissenschaft und Religion mit verschiedenen Aspekten ***eines Phänomens***, jeder auf ihre eigene Art und Weise. *In my view, science and religion deal with different aspects of **a phenomenon**, each in its own unique way.*	Changing “existence” to “a phenomenon” to make it more general, not so specific to the “existential” theme.	2.52

Complementary

Item No.	Original Items	Revised Items	Why?	*M*
1	Nur wenn wir die Wissenschaft durch die Religion ergänzen oder umgekehrt, können wir ein umfassendes Verständnis von der Welt erlangen. *Only by complementing science with religion, or vice versa, can we gain a comprehensive understanding of the world.*	Nur wenn **sich** Wissenschaft und Religion ergänzen, können wir ein umfassendes Verständnis von der Welt erlangen. *Only when science and religion complement each other, can we gain a comprehensive understanding of the world.*	Removing “die” and “wir”	8.38
3	Wissenschaft und Religion können die Lücken des jeweils anderen ausfüllen, um ein vollständiges Bild unserer Welt zu ergeben. *Science and religion can fill in the gaps of the other to form a complete picture of our world.*	Wissenschaft und Religion können ihre **jeweiligen** Lücken **gegenseitig ausfüllen**, um ein vollständiges Bild unserer Welt zu ergeben. *Science and religion can mutually fill in each other’s gaps to form a complete picture of our world.*	Fixing grammar	5.29
2	Was die Wissenschaft offenbart und was wir von der Religion lernen, sollte kombiniert werden, um zu verstehen, warum wir existieren. *What science reveals and what we learn from religion should be combined in order to understand why we exist.*	Was Wissenschaft ***zeigt*** und was wir von Religion lernen ***können***, sollte ***zusammengeführt*** werden, ***um die Welt zu verstehen***. *What science shows and what we can learn from religion **should be brought together** in order to **understand the world.***	A rater mentioned that „Offenbart” is usually used in religious/spiritual context, so changed it into “zeigt.” Two raters mentioned that “question of existence” is rather specific, so I decided to make the item more general.	5.15
5	Gemeinsam tragen Wissenschaft und Religion zu einem umfassenderen Verständnis der Welt bei, jede aus ihrem eigenen Blickwinkel. *Together, science and religion are useful for a broader understanding of the world, each from its own angle.*	Gemeinsam ***können*** Wissenschaft und Religion unser Verständnis der Welt ***erweitern***, jeweils aus ihrem eigenen Blickwinkel. *Together, science and **religion can expand** our understanding of the world, each from their own angle.*	Making the item milder to place it further away from item 4. I decided to make this one milder (not item 4, despite its lower mean), because *SD* of this item is higher than item 4, indicating less consensus in the rating score (than item 4 below).	4.62
4	Die Wissenschaft kann bestimmte Fragen beantworten, die die Religion nicht beantworten kann, und umgekehrt, aber beide sind gleichermaßen nützlich, um ein vollständiges Bild unserer Welt zu zeichnen.	Wissenschaft kann bestimmte Fragen beantworten, die Religion nicht beantworten kann, und umgekehrt, aber ***die Kombination aus beiden*** macht sie gleichermaßen nützlich. *Science can answer certain questions that religion cannot, and vice versa, but*	A rater mentioned that this item a bit too similar to the items of “compartment,” so I added “*die Kombination der beiden*” to better imply a complementary view.	4.37

Item No.	Original Items	Revised Items	Why?	*M*
	Science can answer certain questions that religion cannot, and vice versa, but both are equally beneficial in painting a complete picture of our world.	***the combination of the two*** *makes them equally useful.*		

Consonance

Item No.	Original Items	Revised Items	Why?	*M*
1	Die Vorstellung, dass Wissenschaft und Religion miteinander im Konflikt stehen, ergibt für mich keinen Sinn, da sie im Wesentlichen von denselben Wahrheiten sprechen. *The idea that science and religion are in conflict doesn't make sense to me because they are essentially talking about the same truths.*	***Für mich ist es völlig unvorstellbar***, dass Wissenschaft und Religion in Konflikt miteinander stehen, da sie ***eindeutig*** dieselben ***grundlegenden*** Wahrheiten ***vermitteln***. To me, it is ***completely inconceivable*** that science and religion are in conflict since they clearly convey ***the same fundamental truths***.	Making this item stronger to place it further away from the revised version of item 3.	7.13
5	Wenn ich die Zusammenhänge zwischen Wissenschaft und Religion sehe, fühlt es sich an, als wären sie Teil derselben Einheit. *When I see the connections between science and religion, it seems like they are part of the same unity.*	Wenn ich über die Zusammenhänge zwischen Wissenschaft und Religion **nachdenke**, **scheint es**, als **seien** sie Teil derselben Einheit. *When I* ***think about*** *the connections between science and religion, it seems like they are part of the same unity.*		6.11
2	Über Wissenschaft nachzudenken ist für mich auch ein Aspekt meines spirituellen Lebens. *To me, thinking about science is also an aspect of my spiritual life.*	No change.		5.66
3	Ich denke, dass Wissenschaft und Religion aus der gleichen Quelle stammen, auch wenn sie oberflächlich betrachtet unterschiedlich erscheinen mögen. *I think that science and religion come from the same source, even though*	***Ich bin fest davon überzeugt***, dass Wissenschaft und Religion dieselbe **Wurzel** haben, auch wenn sie oberflächlich betrachtet unterschiedlich erscheinen mögen. ***I firmly believe*** that science and religion have the same **root**, even though they may appear different on the surface.	This item has the same mean ratings as item 2, so it should be made stronger than item 2, but milder than item 1.	5.66

Item No.	Original Items	Revised Items	Why?	*M*
	they may appear different on the surface.			
6	Die Vorstellung, dass Wissenschaft und Religion im Einklang stehen, ergibt Sinn für mich. *The idea that science and religion are in harmony makes sense to me.*	No change.		4.6
4	Ich glaube, dass Wissenschaft und Religion im Grunde in dieselbe Richtung weisen. *I believe that, essentially, science and religion point in the same direction.*	***Mir scheint***, dass Wissenschaft und Religion ***manchmal*** in die gleiche Richtung gehen ***können***. ***It seems to me*** that science and religion can ***sometimes*** go in the same direction.	I made this item milder to place it further away from item 6 above.	4

Predicted Item Locations – Results of Pilot Study 2

Additionally, we also asked raters (*n* = 8) to place all items on a continuum ranging from 0 = completely conflict to 10 = completely compatible. Interestingly, the raters viewed items reflecting context-switch view as closer to compatibility while items reflecting compartment view were generally closer to conflict - but this still makes sense theoretically.

ICC shows that average scores from all raters are consistent (ICC2*k* = 0.94, 95% CI [0.90, 0.97]), and the rating variability between items are significantly greater than within items ($F(26, 182) = 18$, $p < .001$), implying good interrater reliability.

Items' ratings on the continuum of conflict – compatibility are presented in the table below, and items are sorted from lowest (most conflict) to highest (most compatible) based on their mean ratings.

Item No.	Subscale	Item	*M*	SD	*Mdn*	*SE*
1	Conflict	Science and religion are ultimately at odds, with no possibility of reconciliation.	0.21	0.21	0.2	0.07
2	Conflict	Making decisions on the basis of science requires the complete abandonment of religious beliefs.	0.32	0.61	0.14	0.23
4	Conflict	More often than not, science and religion seem to offer opposing viewpoints.	0.84	0.68	0.59	0.26
3	Conflict	Religion and science always offer contradictory explanations for every phenomenon in the world.	1.47	3.16	0.13	1.2
2	Compartment	The purposes of science and religion are so unique that one cannot serve the function of the other.	2.4	1.78	2.45	0.67
5	Compartment	Scientific and religious explanations may work in their respective domains, but they remain separate.	2.62	1.65	2.58	0.58
4	Compartment	Science and religion use distinctive methods to make sense of the world, each valid only within its own domain.	2.75	2.21	3.59	0.84
1	Compartment	Science and religion offer explanations for entirely separate sets of questions.	2.98	2.21	3.04	0.78

Item No.	Subscale	Item	*M*	SD	*Mdn*	*SE*
5	Conflict	There are some instances where religious scriptures and scientific evidence don't align.	3.13	1.54	3.24	0.58
6	Conflict	When searching for answers to fundamental questions, science and religion sometimes come to different conclusions.	3.83	1.67	3.79	0.59
5	Context-switch	The importance I place on science relative to religion can vary, depending on the situations.	4.03	2.25	4.18	0.85
3	Compartment	In my view, science and religion deal with different aspects of our existence, each in its own unique way.	4.64	1.26	4.98	0.48
3	Context-switch	In specific situations, I endorse scientific viewpoints, but in others, I rather embrace religious beliefs.	5.28	2.46	5.23	0.93
4	Complementary	Science can answer certain questions that religion cannot, and vice versa, but both are equally beneficial in painting a complete picture of our world	5.52	2.05	5.61	0.72
1	Context-switch	I can fully switch between relying on science and embracing my faith, depending on the situation.	5.58	2.2	5.7	0.78
2	Context-switch	I strongly favor science in certain situations, but embrace my religious beliefs entirely in others.	5.6	2.26	5.76	0.85
4	Context-switch	It's not unusual for me to flexibly switch from relying on scientific evidence to embracing my faith, or vice versa.	5.84	2.2	5.94	0.78
5	Complementary	Together, science and religion are useful for a broader understanding of the world, each from its own angle.	5.99	1.39	6.16	0.53

Item No.	Subscale	Item	*M*	SD	*Mdn*	*SE*
3	Complementary	Science and religion can fill in the gaps of the other to form a complete picture of our world.	7.1	1.19	7.45	0.45
2	Consonance	To me, thinking about science is also an aspect of my spiritual life.	7.27	1.43	7.83	0.54
4	Consonance	I believe that, essentially, science and religion point in the same direction	7.82	1.6	7.39	0.61
3	Consonance	I think that science and religion come from the same source, even though they may appear different on the surface.	8.37	1.72	9.11	0.65
2	Complementary	What science reveals and what we learn from religion should be combined in order to understand why we exist.	8.5	1.17	8.77	0.44
1	Complementary	Only by complementing science with religion, or vice versa, can we gain a comprehensive understanding of the world.	8.51	1.42	8.86	0.5
1	Consonance	The idea that science and religion are in conflict doesn't make sense to me because they are essentially talking about the same truths.	8.59	3.14	9.68	1.11
6	Consonance	The idea that science and religion are in harmony makes sense to me.	8.67	1.53	9.4	0.54
5	Consonance	When I see the connections between science and religion, it seems like they are part of the same unity	8.89	1.68	9.51	0.64