Examining Measurement Invariance of The Pios and Pios-R across Christian, Jewish, and Muslim Groups

Johanna Anita Younce
younce94@gmail.com

Follow this and additional works at: https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations

Part of the Clinical Psychology Commons, and the Psychiatric and Mental Health Commons

Recommended Citation

This Dissertation/Thesis is brought to you for free and open access by the Graduate Research & Artistry at Huskie Commons. It has been accepted for inclusion in Graduate Research Theses & Dissertations by an authorized administrator of Huskie Commons. For more information, please contact jschumacher@niu.edu.
ABSTRACT

EXAMINING MEASUREMENT INVARIANCE OF THE PIOS AND PIOS-R ACROSS CHRISTIAN, JEWISH, AND MUSLIM GROUPS

Johanna A. Younce, Ph.D.
Department of Psychology
Northern Illinois University, 2023
Kevin Wu, Director

The Penn Inventory of Scrupulosity (PIOS) was developed in a sample of primarily Christian participants. Since scrupulosity is intertwined with religion, it is important to establish that the PIOS and its revised version (PIOS-R) are invariant across the religions with which the measure is primarily utilized: Christianity, Islam, and Judaism. Additionally, there was inconsistency in the literature as to the factor structure of the PIOS/PIOS-R, so three of the models in the literature each were examined and compared for the PIOS and the PIOS-R. 718 participants were recruited through MTurk and reported affiliation with either Christianity ($n = 274$), Islam ($n = 243$), or Judaism ($n = 201$). CFA results revealed that only a modified version of the two-factor model had satisfactory fit for the full-length PIOS. For the PIOS-R, a two-factor model had best fit. Both versions displayed full invariance as well as equivalent factor variances and covariances across all groups. While inconsistent with my hypothesis, this provides support for the continued use of the PIOS/PIOS-R in each of these populations. Latent mean analyses revealed that the Jewish group was lower on scrupulosity than the other two groups. The PIOS/PIOS-R showed excellent internal reliability and generally good convergent and discriminant validity. The PIOS-R performed the same or better than the PIOS in this study, and the author concluded that there is no empirical basis for the use of the full PIOS moving forward.
Future research should examine invariance of the PIOS-R in less broadly-defined religious groups (e.g., across denominations).

*Keywords:* PIOS, PIOS-R, measurement invariance, scrupulosity, OCD
EXAMINING MEASUREMENT INVARIANCE OF THE PIOS AND PIOS-R
ACROSS CHRISTIAN, JEWISH, AND MUSLIM GROUPS

BY

JOHANNA ANITA YOUNCE
©2023 Johanna Anita Younce

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

Doctoral Director:
Kevin Wu
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Kevin Wu, my advisor, for his invaluable guidance, support, and mentorship throughout my academic journey. Dr. Wu’s unwavering dedication to excellence, his insightful feedback, and his encouragement have been instrumental in shaping my research and personal growth. When I needed it, Dr. Wu was always truly supportive and willing to offer his time and energy. It is rare to find a mentor like him.

I am also immensely grateful to my parents, Hanna and Mark Younce for their unconditional love and generous financial support. My parents’ endless encouragement, belief in my abilities, and words of wisdom have been a constant source of strength throughout my academic journey. I owe them a debt of gratitude for instilling in me a strong work ethic, perseverance, and the values that have shaped me into the person I am today.

Furthermore, I want to extend my appreciation to my partner, Garrett Anderson, for his unwavering support, patience, and understanding during the completion of this dissertation. His presence in my life has been a constant source of motivation and comfort. Garrett’s love and encouragement have helped me maintain balance and perspective during challenging times.

I would also like to thank my friends and colleagues for their support, discussions, and encouragement throughout this journey. Their contributions and the contributions of Dr. Kevin Wu, my parents, and my partner, Garrett Anderson, have played an integral role in my academic and personal growth. Without these individuals, this work would not have been possible.
# TABLE OF CONTENTS

| LIST OF TABLES | ................................................................................................................. | v |
| LIST OF FIGURES | ................................................................................................................. | vi |
| LIST OF APPENDICES | ................................................................................................................. | vii |

Chapter

1. INTRODUCTION ............................................................................................................. 1
   Scrupulosity ............................................................................................................. 2
   Measurement of Scrupulosity ............................................................................... 4
   Religion and Scrupulosity ...................................................................................... 38
   PIOS Invariance ....................................................................................................... 45
   Current Study ........................................................................................................ 47

2. METHOD ......................................................................................................................... 53
   Participants ............................................................................................................. 53
   Measures ............................................................................................................... 58
   Procedure ............................................................................................................... 65

3. DATA ANALYSIS PLAN ............................................................................................... 68
   Confirmatory Factor Analyses .............................................................................. 68
   Measurement Invariance Analysis ....................................................................... 80

4. RESULTS ....................................................................................................................... 82
Chapter Page
Data Screening ................................................................. 82
Confirmatory Factor Analyses ........................................... 83
Measurement Invariance Analysis ..................................... 96
Population Heterogeneity .................................................. 100
Internal Consistency ....................................................... 103
Convergent and Discriminant Validity .............................. 104

5. DISCUSSION ........................................................................... 107
PIOS-R Findings ................................................................. 108
PIOS Findings and Relevance .............................................. 117
Limitations ........................................................................... 119
Broad Implications ............................................................. 122

REFERENCES ............................................................................ 125
APPENDICES ........................................................................... 138
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Descriptive Statistics</td>
<td>84</td>
</tr>
<tr>
<td>2. Goodness-of-Fit Statistics of the Penn Inventory of Scrupulosity in Full Sample ($N = 718$)</td>
<td>86</td>
</tr>
<tr>
<td>3. Goodness-of-Fit Statistics for the Penn Inventory of Scrupulosity</td>
<td>94</td>
</tr>
<tr>
<td>4. Standardized Factor Loadings for the Penn Inventory of Scrupulosity</td>
<td>95</td>
</tr>
<tr>
<td>5. Goodness-of-Fit Statistics for Measurement Invariance Analyses</td>
<td>97</td>
</tr>
<tr>
<td>6. Zero-order Correlation Matrix for Study Variables in Full Sample (Ns range from 713-718)</td>
<td>105</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The One-Factor Model of the PIOS</td>
<td>71</td>
</tr>
<tr>
<td>2. The Two-Factor Model of PIOS</td>
<td>72</td>
</tr>
<tr>
<td>3. The Bifactor Model of the PIOS</td>
<td>73</td>
</tr>
<tr>
<td>4. The One-Factor Model of the PIOS-R</td>
<td>76</td>
</tr>
<tr>
<td>5. The Two-Factor Model of the PIOS-R</td>
<td>77</td>
</tr>
<tr>
<td>6. The Bifactor Model of the PIOS-R</td>
<td>78</td>
</tr>
</tbody>
</table>
### LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. RELIGIOUS AFFILIATION ITEMS</strong></td>
<td>138</td>
</tr>
<tr>
<td><strong>B. PENN INVENTORY OF SCRUPULOSITY ORIGINAL AND REVISED VERSIONS</strong></td>
<td>140</td>
</tr>
<tr>
<td><strong>C. DIMENSIONAL OBSESSIVE-COMPULSIVE SCALE WITH SCRUPULOSITY/RELIGIOUS THOUGHTS SUBSCALE</strong></td>
<td>142</td>
</tr>
<tr>
<td><strong>D. SANTA CLARA STRENGTH OF RELIGIOUS FAITH QUESTIONNAIRE</strong></td>
<td>153</td>
</tr>
<tr>
<td><strong>E. DEPRESSION, ANXIETY AND STRESS SCALE – 21, DEPRESSION AND ANXIETY SUBSCALES</strong></td>
<td>155</td>
</tr>
<tr>
<td><strong>F. POSITIVE AND NEGATIVE AFFECT SCHEDULE, NEGATIVE AFFECT SUBSCALE</strong></td>
<td>157</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

The Penn Inventory of Scrupulosity (PIOS; Abramowitz et al., 2002) and Penn Inventory of Scrupulosity-Revised (PIOS-R; Olatunji et al., 2007) are the most commonly used self-report questionnaires measuring scrupulosity. Scrupulosity, or scrupulous obsessive-compulsive disorder (OCD) is a relatively common symptom dimension of OCD focused on moral- and religious-based obsessions and compulsions (Mataix-Cols et al., 2002). The PIOS/PIOS-R have been used to examine levels of scrupulosity in individuals of several different religious affiliations, including Protestant and Catholic Christians (e.g., Nelson et al., 2006), Orthodox and non-Orthodox Jews (e.g., Pirutinsky & Rosmarin, 2018), Muslims (e.g., Inozu, Clark, & Karanci, 2012), and Mormons (e.g., Judd et al., 2020). Since the PIOS/PIOS-R measure a construct highly relevant to religion, it is important to establish measurement invariance across all of the religious groups in which the scale is used. This is especially true given that Abramowitz et al. (2002) did not report the religious demographics of the sample on which the scale was developed. Testing measurement invariance determines whether the measurement properties of a scale are equivalent in different populations (Brown, 2015). The testing of measurement invariance is relatively new, with the traditional measurement invariance testing procedure developed by Jöreskog in 1971. Measurement invariance testing has become more common since then, and procedures using multi-group confirmatory factor analyses (MGCFAs) have been developed for more complex models (Van de Schoot et al., 2015). However, there still is work to be done in
establishing measurement invariance in many commonly used psychological measures, including the PIOS/PIOS-R.

Researchers need to establish measurement invariance of the PIOS/PIOS-R across religious groups before we can (a) have confidence that the scale has the same meaning and can be interpreted in the same way across the religious groups or (b) draw conclusions about potential differences between religious groups (Kline, 2016). It is concerning, then, that many researchers have used the PIOS/PIOS-R in different religious groups and made comparisons between religious groups (e.g., Buchholz et al., 2019; Inozu, Clark, & Karanci, 2012) despite the lack of measurement invariance studies on the PIOS/PIOS-R. To the author’s knowledge, only one published study has tested measurement invariance of the PIOS (Pirutinsky & Rosmarin, 2018) and no studies have tested the PIOS-R. Further, the aforementioned study examined invariance only between Orthodox and non-Orthodox Jews. The current study sought to address this gap in the literature by examining the measurement invariance of both the PIOS and PIOS-R in the three religious groups most commonly studied within the scrupulosity literature: Christian, Jewish, and Muslim. As a precursor to measurement invariance analyses, multiple measurement models from the literature were tested in a religiously diverse sample. Secondary goals of the study were to examine certain aspects of population heterogeneity, to examine internal consistency of the measures, and to test the convergent and discriminant validity of the measures.

Scrupulosity

Scrupulosity is a relatively understudied symptom dimension of OCD. OCD is a mental disorder characterized by obsessions (persistent intrusive thoughts, urges, or images that cause
anxiety or distress) and compulsions (repetitive behaviors or mental actions that the individuals perform in an attempt to reduce anxiety; American Psychiatric Association [APA], 2013). OCD is a disorder with high rates of severity (Kessler et al., 2005) and suicidality (Angelakis et al., 2015). It is a heterogeneous disorder, with many possible symptom themes, one or more of which an individual with OCD may experience. Typical symptom themes include checking, contamination, ordering and symmetry, scrupulosity, and harm (APA, 2013). It is important to be able to validly and reliably measure each symptom dimension, since those with different symptom dimensions may respond differently to treatment (M. T. Williams et al., 2013) and therefore should be given explicit attention in the literature. Additionally, scrupulous OCD is associated with unique features that may result in poorer treatment outcomes; such features include poorer insight, higher perceptual distortion, and higher magical ideation when compared to those without religious symptoms or with other primary symptom dimensions (Mataix-Cols et al., 2002; Tolin et al., 2001; Wu et al., 2018). These unique challenges highlight the importance of studying scrupulous OCD individually.

Scrupulous OCD itself is heterogeneous in presentation, depending in part on the individual’s religious or moral beliefs (Abramowitz & Jacoby, 2014; Olatunji et al., 2007). An individual who identifies as Christian and has scrupulous OCD may experience obsessive doubt that they do not love God enough or experience compulsions to pray every time a “bad” thought enters their mind. A Muslim individual with scrupulosity may experience fear or doubt that swallowing their saliva during Ramadan broke their fast. A Jewish individual with scrupulous obsessions may experience obsessive thoughts that their food is not 100% kosher and may perform extensive repetitive checking rituals. Notably, many of the behaviors that scrupulous
individuals fear are sinful are not considered wrongful by either religious authorities or other individuals of the same religion (Abramowitz & Jacoby, 2014).

Scrupulosity is the fifth most common obsessional theme in the US and North America (Foa et al., 1995; Mataix-Cols et al., 2002). Approximately 5-6% of individuals with OCD have religious obsessions as their primary symptom dimension (Foa et al., 1995; Tolin et al., 2001), and approximately 22-24% of those with OCD experience religious obsessions (Mataix-Cols et al., 2002). In predominantly Muslim countries, however, the rates of religious themes in individuals with OCD are much higher. In a sample of Muslim individuals with OCD in Saudi Arabia, religious themes were the most common (Mahgoub & Abdel-Hafeiz, 1991). In an Egyptian sample, religious obsessions tied with contamination obsessions as the most common obsessional themes (both 60% prevalence; Okasha et al., 1994). Scrupulosity also appears to be a more common OCD symptom theme within ultra-Orthodox Jewish communities compared to general US and North American samples; 68% of an ultra-Orthodox OCD sample reported primary religious symptoms (Greenberg & Shefler, 2002). This highlights the importance of accurately measuring scrupulosity in Muslim and Jewish samples as well as Christian samples, so that this common symptom theme can be better understood by researchers and clinicians.

Measurement of Scrupulosity

Currently, the PIOS and PIOS-R are the most commonly used, dimensional measures of scrupulosity in the literature. Prior to the development of the PIOS, most research in the area of scrupulous OCD was conducted on samples of OCD patients (e.g., Steketee et al., 1991; Tolin et al., 2001). Often, researchers made use of the Yale-Brown Obsessive-Compulsive Scale
(Y-BOCS; Goodman et al., 1989) Symptom Checklist. However, that instrument is intended to be administered to individuals with an OCD diagnosis to determine which types of symptoms are experienced and to identify primary symptom dimensions. The checklist includes three scrupulosity items (one of which is “other” and provides space to write in scrupulous obsessions) and the opportunity to code specific symptoms as “principal.” This method is entirely appropriate for use with OCD populations, but it does not offer a dimensional measure of scrupulosity, limiting the options for research. For example, checklist measures are not appropriate for examination of scrupulosity in unselected or nonclinical samples, since these measures assume that the respondent has OCD. For this reason and others (i.e., promoting increased research on scrupulosity, examining the prevalence of scrupulous OCD in community samples, assisting religious counselors in detecting scrupulosity so they can provide appropriate referrals, and detecting clinical scrupulosity in those seeking treatment for OCD), Abramowitz et al. (2002) developed the PIOS.¹

¹ Two new measures of scrupulosity have appeared in the year this dissertation was proposed (2021). The Scrupulous Thoughts and Behaviours Questionnaire (Ong et al., 2021) was published in April 2021, and a new DOCS subscale, titled Scrupulosity/Religiosity (DOCS-SR; Wetterneck et al., 2021) was available online in April 2021. Clearly, there was not time for evaluation of either measure’s psychometric properties by other laboratories by the time of writing this introduction. The new Ong and colleagues measure builds upon the PIOS/PIOS-R by including items targeting compulsions as well as obsessions. The DOCS-SR measures scrupulosity in the style of the DOCS, which focuses less on specific manifestations of scrupulosity. Since there was limited information available about these new measures, the focus of the current study remained on the PIOS/PIOS-R.
PIOS Development

Abramowitz et al. (2002) reported two initial studies. Study 1 developed the PIOS and established subscales; Study 2 explored convergent and discriminant validity.

Study 1

The authors used a sample of 215 undergraduate psychology students in Study 1. They reported a mean age of 18.8 years (SD = 1.01) and a majority male (58.6%) sample. No other demographic information was provided for this sample, including no breakdown of the religious affiliations.

Abramowitz et al. (2002) began scale development by generating a pool of 77 items based on experience working with individuals with scrupulous OCD. The items were created to have good face validity and relate specifically to scrupulous OCD themes. Initially, there were separate response options for rating frequency and distress: frequency response options ranged from 0 (never) to 4 (constantly) and distress response options ranged from 0 (not at all distressing) to 4 (extremely). The authors found that responses on the two rating scales were highly correlated, so the response scale that was more normally distributed (frequency) was retained. The authors then removed 34 items due to skewness values greater than 1. Next, 7 items were removed due to redundant item content, as determined by high inter-item correlations and similar wording. The items with the higher item-total correlations were retained. Item-total correlations for the remaining 36 items were recalculated, and items with r values below .30 were excluded. This was done repeatedly until internal consistency could not be improved further, leading to the exclusion of 7 items. Using the remaining 29 items, a principal
components analysis (PCA) with varimax rotation was conducted. There were four components with eigenvalues greater than 1, but the scree plot indicated two meaningful components, so the authors concluded that there were two components. Ten items then were removed due to non-salient loadings on either component or due to even cross-loadings (< .10 difference). This process resulted in the final 19 items of the original PIOS and two subscales: Fear of Sin (FOS; 12 items) and Fear of God (FOG; 7 items). Total scores on the PIOS range from 0 to 76, FOS scores range from 0 to 48, and FOG scores range from 0 to 28. The correlation between subscale scores was .67, \( p < .001 \). Other researchers have found similar correlation coefficients, \( rs = .69-.85, ps < .05 \) (Nelson et al., 2006; Pirutinsky & Rosmarin, 2018). See Reliability section in Psychometric Properties section for commentary about the size of this correlation. Internal consistency reliability was acceptable for the full scale (Cronbach’s alpha of .93) and for both FOS (\( \alpha = .90 \)) and FOG (\( \alpha = .88 \)).

**Study 2**

In Study 2, the authors used a sample of 197 undergraduate psychology students from the same university as in Study 1. More demographic information was provided for this sample. A mean age of 18.8 (\( SD = .93 \)) and a majority male (57.9%) sample was reported. No racial/ethnic data were reported, but religious affiliation was recorded. The sample was 28.4% Catholic, 21.3% Jewish, 20.8% Protestant, 16.8% “other,” and 12.7% atheist/agnostic.

Study 2 provided evidence for convergent and discriminant validity. Specifically, findings showed significant, medium-sized (\( rs = .29-.36 \)), positive correlations between the PIOS (total and subscale scores) and the Maudsley Obsessional Compulsive Inventory (MOCI;
Hodgson & Rachman, 1977), a general measure of obsessive-compulsive (OC) symptomatology. There also were significant, medium-sized correlations between the PIOS (total and subscale scores) and religiosity (measured using a single item created for the study; $rs = .31-.36$). There were no significant relationships between the PIOS scores and a discriminant measure of anger, the Anger Expression Scale ($rs = .08-.13$, $ns$; Spielberger, 1985). Together, the Abramowitz et al. (2002) findings suggested that the PIOS was an acceptably reliable and valid measure of scrupulosity. It has been used in many published studies since its development.

**PIOS-R Development**

Approximately five years after the PIOS was developed, Olatunji et al. (2007) sought to further refine the PIOS. To do so, the researchers asked 352 undergraduate students to complete a survey including the PIOS. The sample had a mean age of 21.34 ($SD = 6.38$) and were majority female (59.4%). Participants primarily identified as White (90.1%). Notably, the authors did not report information about religious affiliation.

Olatunji et al. (2007) first examined the factor structure of the PIOS by comparing one-factor and two-factor models in confirmatory factor analysis. They determined that the fit of the one-factor model was unsatisfactory and that the two-factor model showed better fit, as demonstrated by improvements in all fit indices and an AIC improvement of 260.08 (see Factor Structure section). However, the authors found some areas of localized strain in the two-factor model which resulted in the addition of a dual loading of item 2 on both factors, resulting in improved overall fit. When this was respecified, however, the path to FOG became negative. The authors chose to specify that item 2 load only on FOS, with no significant decrease in model fit.
Although the adjusted two-factor model fit well, Olatunji et al. (2007) sought to examine the items for possible refinements. They began the process by analyzing localized areas of strain by examining modification indices. Two pairs of items showed evidence of correlated residuals: items 2 and 3, and items 5 and 6. These item pairs also displayed high inter-item correlations ($r > .50$). The authors concluded that substantial covariance between the items was likely due to similarity in content. When the error covariances between the item pairs were freely estimated, there was significant improvement in model fit. Modification indices were further examined, and two additional pairs of items were identified as areas of strain: items 15 and 17, and items 10 and 12. These item pairs also were associated with high modification indices and inter-item correlations. They also were recognized as having substantial content overlap. The error covariances between item pairs were freed, resulting in significant improvement in model fit. Finally, items 16 and 18 showed high modification indices and inter-item correlations as well as content overlap. Freeing the error covariance between the items resulted in further improvement in model fit. The final model, with five correlated residuals, was identified as the best-fitting measurement model for the PIOS.

Next, the authors examined the items for redundant content and potential removal. Using the same criteria cited by Abramowitz et al. (2002) for elimination of items, items 2, 6, 10, and 15 were removed. When comparing the fit of the correlated 19-item model to the fit of the 15-item model, based on AIC values (improvement of 114.17), the 15-item model had better fit. Additionally, all factor loadings were significant and salient ($>.66$). Thus, the authors retained the 15-item revised PIOS (PIOS-R).
Since internal consistency reliability is sometimes reduced with the removal of items, internal consistency estimates were calculated for the PIOS-R. Findings revealed good internal consistency for the total score ($\alpha = .94$) and acceptable inter-item correlations. The FOS and FOG subscales also displayed good internal consistency ($\alpha = .92$ and .91, respectively). Additionally, the PIOS-R and its subscales showed near perfect correlation with the 19-item PIOS and its respective subscales ($rs = .99$). These data suggest that the removal of the four items did not result in poorer internal consistency estimates, and the overall meaning of the measure was preserved.

Olatunji et al. (2007) provided information on the convergent and discriminant validity of the PIOS-R. The PIOS-R total and subscale scores showed small to medium positive correlations ($rs = .30-.43$, $ps < .05$) with the Obsessive-Compulsive Inventory-Revised (OCI-R; Foa et al., 2002), a measure of general OC symptomatology; small to medium positive correlations ($rs = .30-.40$, $ps < .05$) with the Trait subscale of the State-Trait Anxiety Inventory (STAI-T; Spielberger et al., 1983), a measure of trait anxiety; and small positive correlations ($rs = .23-.31$, $ps < .05$) with the Negative Affect subscale of the Positive and Negative Affect Schedule (PANAS-NA; Watson et al., 1988). The PIOS-R and its subscales showed small, significant positive correlations ($rs = .22-.27$) with the Disgust Emotion Scale (Kleinknecht et al., 1997); very small to small but significant correlations ($rs = .19-.25$) with the State subscale of the STAI (STAI-S); very small to small but significant correlations ($rs = .13-.23$) with the Multidimensional Blood/Injury Phobia Inventory (Wenzel & Holt, 2003); and very small correlations ($rs = .14-.15$, $ps < .05$) with the Spider Phobia Questionnaire (Klorman et al., 1974). It was uncorrelated ($rs = -.08 – -.02$, $ns$) with the Positive Affect subscale of the PANAS.
(PANAS-PA). These findings all were in the expected direction, and the strengths of the correlations generally reflected the conceptual associations among constructs (e.g., the correlations between the PIOS-R and OC symptomatology were strongest). Additionally, the PIOS-R total score still was significantly, meaningfully correlated with OC symptomatology after controlling for trait anxiety and negative affect (partial $r = .33$). This provided further evidence that the validity of the measure was relatively unchanged by the removal of the four items.

By all metrics examined by Olatunji et al. (2007), the PIOS-R displayed better model fit and otherwise equivalent psychometric properties. Thus, one would expect that it is the preferred version of the PIOS. Many researchers have adopted the PIOS-R (e.g., Fergus & Rowatt, 2014a; Pirutinsky et al., 2015). However, many researchers continue to use the 19-item PIOS (e.g., Judd et al., 2020; Witzig & Pollard, 2013). Thus, the current study will examine the factor structure and measurement invariance properties of both versions of the instrument, in part, to examine whether the PIOS-R is a superior option that should replace the PIOS.

**Factor Structure**

In the field of psychology, researchers often need to measure constructs that are not directly observable (i.e., latent factors), so it is necessary to measure several related observable variables (i.e., indicators). For example, responses to items on a questionnaire are frequently used as indicators of a latent construct. Before making inferences based on these indicators, it is necessary to determine whether the indicators are truly measuring what they are meant to measure. In other words, it must be determined that the test is adequately valid.
A consistent factor structure provides evidence for good construct validity. For example, if the items of the PIOS/PIOS-R are professed to be assessments of different manifestations of scrupulous OCD, then they should all load onto the same factor, which is representative of scrupulosity. However, if items that are purported to measure the same construct do not load on the same factor, then this may be indicative of poor construct validity. In other words, one or more items may be improper operational definitions of the true latent construct. If validity of the test is in question, then research conclusions based on the test may be inaccurate, and clinical use of the test may be inadvisable because scores on the test may not be interpretable.

Assessment of factor structure also assists in determining how a measure should be scored. For example, if there are multiple factors, this would indicate that the test should involve subscales that correspond to the factors. Additionally, factor loadings offer information useful for deciding which items should be included on which subscales.

**PIOS**

The factor structure of the original 19-item PIOS has been examined in a handful of published studies. Abramowitz et al. (2002) was first, using a PCA with varimax rotation, but it was an exploratory analysis conducted as part of the process of developing the questionnaire. The analysis revealed a two-factor model in which all factor loadings were significant and salient (> .46). To this author’s knowledge, only three additional studies have been conducted that involved exploratory or confirmatory factor analyses of the original PIOS (i.e., Huppert & Fradkin, 2016; Olatunji et al., 2007; Pirutinsky & Rosmarin, 2018), and the findings are mixed. One study also examined the factor structure of a Turkish version of the PIOS (Inozu, Keser, &
Across these studies, three models were tested, including one-factor, two-factor, and bifactor models.

**One-Factor Model.** Olatunji et al. (2007), Huppert and Fradkin (2016), and Pirutinsky and Rosmarin (2018) all tested a one-factor model in their respective studies. Olatunji et al. (2007) did so in a sample of 352 undergraduate students (more demographic information listed under the PIOS-R Development section). Findings revealed poor to acceptable fit indices: an RMSEA of .11, CFI of .90, TLI of .88, and SRMR of .07. The factor loadings were all significant and salient (> .60). Overall, the authors concluded that the fit of the one-factor model was poor, especially in comparison to the fit of the two-factor model (see Two-Factor Model section).

Huppert and Fradkin (2016) tested the one-factor model using a WLSMV estimator and a polychoric correlation matrix in a sample of 132 OCD outpatient clients. The sample had an average age of 33.57 (SD = 13.13) and was majority female (50.4%) and White (94.7%). The authors reported religious affiliation of the sample, which was 36.8% Catholic, 21.2% Protestant, 18.8% Jewish, 6.0% affiliated with another religion, and 14.3% not affiliated with a religion. Fit indices were poor to good, with an RMSEA of .136, CFI of .96, TLI of .96, and WRMR of 1.44. In this study, the incremental fit indices showed that the model fit better than the null model, but the absolute fit indices were not in the acceptable range.

Pirutinsky and Rosmarin (2018) also examined the one-factor model using a WLSMV estimator and a polychoric correlation matrix. The sample included 586 community members in the US and Canada, 318 of whom identified as Orthodox Jewish, and 268 of whom identified as non-Orthodox Jewish. The sample had an average age of 37.43 (SD = 15.43) and was majority
female (68%). They only reported change scores for the fit indices for the one-factor model in comparison to the bifactor model (see Bifactor Model section). The one-factor model had poorer fit overall ($\chi^2[20] = 38.15, p = .009$) and the authors concluded that the bifactor model had better fit. However, in large samples such as this one ($N = 586$), chi-square tests can be biased due to inflated Type I error rates. Thus, a significant chi-square difference test alone is not sufficient to conclude that fit is significantly worse for the one-factor model. The changes in the fit indices show worse fit, but the differences were small and some differences may be trivial: $\Delta$RMSEA of .02, resulting in a change from reasonable fit for the bifactor model to mediocre fit for the one-factor model; $\Delta$TLI of .007, which still indicted good fit; $\Delta$WRMR of .46, which indicates worse fit, but the WRMR index for the bifactor model was in the poor range as well (WRMR = 1.6); and $\Delta$CFI of .008, representing no categorial change in fit (the index is still in the “good” range) and, according to Cheung and Rensvold (2002), a change in the CFI of less than .01 may represent a trivial change in fit.

Overall, there is minimal support for the one-factor model of the PIOS. In their exploratory factor analyses, Abramowitz et al. (2002) found that a two-factor solution fit the data best. Olatunji et al. (2007) found overall poor fit for the one-factor model, although they did not use the WLSMV estimator with their CFA, calling some of the results into question. Huppert and Fradkin (2016) also found relatively poor fit, although the comparative fit indices displayed good fit. Findings from Pirutinsky and Rosmarin (2018) suggested that the one-factor model had acceptable fit overall, but the bifactor model had better fit according to the chi-square difference test. Other measures of change in fit (i.e., $\Delta$CFI) indicated that differences in fit between the one-factor and bifactor models were trivial. Generally, however, one would expect less ambiguous
findings from a frequently used measure of a psychological construct. Thus, researchers have examined other models that might offer stronger results.

**Two-Factor Model.** Four studies tested the two-factor model established by Abramowitz et al. (2002), including the Turkish form. Olatunji et al. (2007) used a confirmatory factor analysis (CFA) to test the two-factor model. It showed good to acceptable fit as indicated by the RMSEA (.08), CFI (.95), TLI (.95), and SRMR (.06). Notably, when compared to the one-factor model, AIC improved, indicating better fit. However, the authors observed an area of localized strain, as indicated by a high modification index for item 2. When the model was respecified such that item 2 loaded onto the FOS factor and not the FOG factor, the model showed significant improvement. The improvement was illustrated by improvements in most fit indices: RMSEA (.07), CFI (.96), TLI (.95), SRMR (.057). All item loadings were significant and salient.

Huppert and Fradkin (2016) examined the two-factor model and found goodness of fit indices indicated mixed support. CFI and TLI values were good (.97 and .96, respectively), but the RMSEA and WRMR were poor (.12 and 1.26, respectively). Notably, when a one-factor model (see One-Factor Model section) was nested with the two-factor, the $\chi^2$ difference test revealed that the two-factor model had better fit, $\chi^2(1) = 31.85$, $p < .001$. Pirutinsky and Rosmarin (2018) also tested the two-factor model. They only reported change in fit index values but indicated that the fit was significantly worse for this model compared to a bifactor model (see Bifactor Model section; $\chi^2[20] = 256.20$, $p < .0001$, $\Delta$RMSEA = .50, $\Delta$CFI = .34, $\Delta$TLI = .39, $\Delta$WRMR = 8.94). This suggested that the two-factor model did not have the best fit of the three models tested in this sample.
The Olatunji et al. (2007) study was the only English version study that found a two-factor solution fit best. This could be due to differences in sample make-up (undergraduate students versus OCD outpatient clients and a Jewish community sample) or resulting from differences in the estimation method employed. Olatunji et al. (2007) did not report what estimator was used, but they did report that the latent factors were scaled by fixing the variances to 1.00, suggesting that a WLSMV estimator was not used. Analyses in the other two studies involved a WLSMV estimator (Huppert & Fradkin, 2016; Pirutinsky & Rosmarin, 2018) and, importantly, a WLSMV estimator is recommended for CFAs involving categorical indicators (Brown, 2015; Kline, 2016). It is unknown whether an ordinary ML estimator was used in the Olatunji et al. (2007) study, but it is important to note that the use of this estimator with categorical data can cause several problems, including weakened estimates of relationships among indicators, identification of factors that reflect item difficulty or extremeness rather than true substantive factors, and incorrect test statistics and standard errors (Brown, 2015). Thus, findings from Olatunji et al. (2007) may need to be treated with caution.

Inozu, Keser, and Karanci (2017) tested the two-factor model set forth by Abramowitz et al. (2002) in a sample of 444 Turkish undergraduate students with an average age of 20.90. The sample was majority women (54.3%). No information on religious affiliation was reported, but Turkey is a predominantly Muslim country. The PIOS was translated into Turkish, and its factor structure was examined. Results displayed poor fit across all indices: GFI of .78, AGFI of .73, CFI of .87, NFI of .85, and RMSEA of 1.12. Then, the authors completed an exploratory factor analysis (EFA) with an oblimin rotation, and eigenvalues and the scree plot suggested a two-factor solution. However, certain items loaded more strongly on the opposite factor to the
Abramowitz et al. (2002) factors. Specifically, items 4 and 6 loaded more strongly onto FOG than FOS. Additionally, items 10 and 15 loaded similarly on both factors. The authors tested a two-factor model in which items 10 and 15 were removed and items 4 and 6 were respecified to load onto FOG. Indices improved, but still showed relatively poor fit: GFI of .87, AGFI of .83, CFI of .93, NFI of .91, and RMSEA of .09. The authors examined areas of localized strain, resulting in further modifications to the model, including freeing the error covariances between items 17 and 19 and items 8 and 11. This resulted in improved fit, with indices that met or were close to meeting cutoffs for good fit: GFI of .90, AGFI of .86, CFI of .95, NFI of .93, and RMSEA of .07. The authors accepted this as the final measurement model. Although it is a two-factor model, it involves substantial revisions from the Abramowitz et al. (2002) solution and has more similarities to the PIOS-R than to the PIOS, with the removal of items 10 and 15 from the scale. Of course, the properties of a translated measure cannot be compared with the properties of the original without using great caution. This is because the process of translation can influence the meaning and nature of the measure, and thus the structure can be different across any two translations.

The findings across the factor analyses offer mixed support for the two-factor model of the PIOS. While the Inozu, Keser, and Karanci’s (2017) results support a two-factor model, the final measurement model involved substantial changes to the PIOS, including the removal of multiple items from the measure entirely. For the purposes of examining the structure of the full 19-item PIOS, results from their study indicated poor fit of the two-factor model. Additionally, the two-factor model showed worse fit indices than a bifactor model (Huppert & Fradkin, 2016; Pirutinsky & Rosmarin, 2018). However, analyses revealed good CFI and TLI values (≥ .95) in
two studies (Huppert & Fradkin, 2016; Olatunji et al., 2007) and it was the best solution in the exploratory analysis conducted by Abramowitz et al. (2002). However, Brown (2015) noted that comparative fit indices such as the CFI and TLI are more likely to appear favorable than other indices because they evaluate model fit compared to a null model that proposes no relationships among the variables, which is essentially an extremely poor model. It is argued that this criterion is too liberal and may lead to overly generous estimations of fit. Other fit indices were within the acceptable range in the Olatunji et al. (2007) study, but this finding was not replicated in the other samples and using a WLSMV estimator. Together, these findings may suggest that the two-factor model is more appropriate in unselected US samples or the model may not be the most appropriate, given that confirmatory analyses using the recommended estimator showed poor fit.

**Bifactor Model.** A bifactor model is a model in which there is one general factor onto which all indicators or items are expected to load. In addition, the indicators also directly load onto specific factors. This is different from a higher order model, in which the specific factors mediate the relationships between the indicators and the general factor. In the case of the PIOS, the general factor would be scrupulosity and the specific factors would be the two subscales.

Given that the two-factor showed poor fit on two out of four fit indices in addition to a nonsignificant chi-square test, Huppert and Fradkin (2016) turned to exploratory methods to investigate the factor structure of the PIOS. The authors employed an EFA with parallel analysis and comparison data extraction methods. The parallel analysis, which used a polychoric correlation matrix given that the data were ordinal, indicated a one-factor model. However, the comparison data method, which used Spearman’s rank-order correlation matrix given that the data were ordinal, indicated a three-factor solution. The authors interpreted these findings as
suggestive of a bifactor structure. This explanation was tested further by conducting an EFA with a bifactor goemin oblimin rotation with a WLSMV estimator. Results indicated acceptable to good fit, with an RMSEA of .078, CFI and TLI of .99, and SRMR of .03. For the general factor, all factor loadings were significant and salient (> .68). For the FOG-revised factor (FOG-R), four loadings were significant, salient (> .30), and positive, and for the Fear of Immodesty (FOI) factor (revised from the FOS subscale), six loadings were significant, salient (> .34), and positive. The authors concluded that this bifactor measurement model was the best-fitting model for the PIOS.

Pirutinsky and Rosmarin (2018) tested Huppert and Fradkin’s (2016) bifactor model in their sample of Jewish community members. They tested this model in the same way as they tested the two-factor model: using a CFA with a WLSMV estimator. The fit indices ranged from poor to good, with an RMSEA of .066 and almost perfect CLI (.996) and TLI (.995), but a WRMR of 1.6. The authors concluded that the bifactor model showed satisfactory fit overall despite that the WRMR was above 1.0, given that three out of four indices were in the acceptable range and given that the model appeared to fit better when compared to the one- and two- factor models (see One-Factor Model and Two-Factor Model sections).

The bifactor model appears to be best-fitting in terms of fit indices. It also has been shown to have significant and salient factor loadings (Huppert & Fradkin, 2016). It has been tested in two specific populations (i.e., OCD outpatients and Jewish community members) and using the recommended estimator (e.g., WLSMV estimator). The WLSMV estimator is recommended by Kline (2016) because it allows for ordinal indicators and is robust to nonnormality. The bifactor model has not yet been tested in an unselected sample. Additionally,
a one-factor model fit was very similar to – although technically better than – the bifactor model in one study (Pirutinsky & Rosmarin, 2018).

Only four studies were available in the literature that examined the factor structure of the PIOS and results across the four studies were mixed. Given that one of the studies only conducted exploratory analyses, only two of the studies used the same CFA estimator, and three very different types of samples were used across the four studies, it is difficult to draw definite conclusions about the factor structure of the PIOS. The one-factor model was not the best-fitting model in any of the four studies. A two-factor solution showed acceptable fit in the studies that involved unselected samples of undergraduate students (i.e., Abramowitz et al., 2002; Olatunji et al., 2007), but it had poorer fit in the studies that used selected samples of OCD outpatients and Jewish community members, respectively, and that used the WLSMV estimator for ordinal variables (i.e., Huppert & Fradkin, 2016; Pirutinsky & Rosmarin, 2018). Two studies found the bifactor solution had the best fit (i.e., Huppert & Fradkin, 2016; Pirutinsky & Rosmarin, 2018), but in one study, the fit may not have been substantially different from the one-factor model (i.e., Pirutinsky & Rosmarin, 2018). Given these findings, an argument can be made for the bifactor model as the best-fitting measurement model for the PIOS. However, there was a significant amount of “mud in the water” at this stage to exclude the other two models entirely. Thus, the current study examined all three models in the current sample before moving on to measurement invariance analyses. This adds to the literature by using a CFA with a WLSMV estimator to investigate the factor analysis of the PIOS in a nonclinical sample of multiple religious affiliations.
PIOS-R

To the author’s knowledge, the factor structure of the PIOS-R only has been examined in four studies in its English form (Huppert & Fradkin, 2016; Olatunji et al., 2007; Phillips & Fisak, 2022; Shapiro et al., 2013) and in one study in its Spanish translation (Gallegos et al., 2018). Olatunji et al. (2007) first investigated the factor structure in their PIOS-R development study (see section PIOS-R Development). In their study, a two-factor model similar to the original two-factor model of the PIOS was retained. Subsequent studies have been conducted to examine two-factor, one-factor, and bifactor models of the PIOS-R using confirmatory and exploratory analyses. There is some support for both two-factor and one-factor models across these studies.

Two-Factor Model. Olatunji et al. (2007) used modification indices to explore respecifications to the two-factor model of the PIOS that led to testing the removal of four items with redundant content, resulting in the PIOS-R. Their final model, a two-factor model, resulted in good fit, as evidenced by multiple fit indices: RMSEA of .058, CFI of .97, TLI of .97, and SRMR of .047. Huppert and Fradkin (2016) examined the two-factor model from Olatunji et al. (2007) and they concluded that the model had inadequate fit: RMSEA of .12 (with a confidence interval of .10-.14), CFI of .98, TLI of .97, and WRMR of 1.036. Although fit was good when compared to a null model, the absolute fit appeared to be poor. An argument can be made for the RMSEA being adequate given that its confidence interval included the cutoff for acceptable fit (.10) and given that strict use of fit index thresholds is inappropriate (Brown, 2015; Kline, 2016), and Hu and Bentler (1999), whose guidelines are often followed, deliberately use language indicating that values close to the cutoffs can be considered acceptable. Finally, when compared
with a nested one-factor model, the two-factor model had better fit, according to a chi-square difference test, $\chi^2(1) = 32.85, p < .0001$. The authors concluded poor factorial validity for the PIOS-R, but more critical examination of the fit indices, along with a lack of significant modification indices, suggests that the fit may be close to adequate. However, the absolute fit indices clearly were not strong.

Shapiro et al. (2013) tested the two-factor model in a sample of 417 participants who were diagnosed with OCD and were in partial hospitalization or residential treatment programs. The participants included 95 individuals who experienced scrupulous symptoms and 322 individuals who did not experience scrupulous symptoms. Age, gender, race/ethnicity, and religious affiliation of the sample were not reported. The authors conducted a CFA using a maximum likelihood (ML) estimator, which is not considered appropriate for ordinal data, so the results should be interpreted with caution. Three separate CFAs were conducted for the full sample, the scrupulous sample, and the non-scrupulous sample. In the full sample, all factor loadings were significant and salient ($>.62$), and the reported fit indices were mediocre to good: RMSEA of .09, CFI of .98, and SRMR of .04. Fit also was acceptable in the scrupulous sample (CFI of .97 and SRMR of .06) and in the non-scrupulous sample (RMSEA of .10, CFI of .97, and SRMR of .05). The authors retained the two-factor measurement model of the PIOS-R.

Phillips and Fisak (2022) tested the two-factor model in two undergraduate student samples: one sample of individuals who identified as Christian ($N = 1158$), and another sample of individuals who identified as atheist ($N = 189$). The authors’ goal was to examine and compare the factor structure of the PIOS-R in the two samples. The Christian sample had a mean age of 20.14 ($SD = 4.8$) and was majority female (67.1%). Participants primarily were White
(52.6%), and others identified as Hispanic/Latino (22.6%), Black (4.2%), Asian/Pacific Islander (4.2%), Native American (0.3%), or a different race/ethnicity (4.0%). The atheist sample had a mean age of 19.91 ($SD = 3.23$) and was 46.3% female. Participants identified as White (69.8%), Hispanic/Latino (14.8%), Black (4.2%), Asian/Pacific Islander (6.9%), Native American (0.5%), or a different race/ethnicity (3.7%). For both samples, CFAs were conducted, but the authors did not report the estimator used or the latent factor scaling method. There were mixed results in the Christian sample, with indices suggesting poor to adequate fit: RMSEA of .098, CFI of .912, and TLI of .897. Although fit did not appear to be “good,” all indices met or were very close (.003 away) to meeting acceptable thresholds. For the atheist sample, fit indices suggested poor fit: RMSEA of .125, CFI of .886, and TLI of .865. However, $\chi^2$/df was 3.93, which indicates good model fit. The authors then conducted an EFA with principal axis extraction and a promax rotation in the atheist sample. Based on the scree plot and factor loadings, the authors found a two-factor solution. However, two items were dropped due to nonsalient loadings on both factors (items 3 and 11). One of the items (item 3) contained the word “sins” whereas most other items on the subscale (FOS) used more secular terms (e.g., immoral). It was noted that the FOG factor included all of the items from the original Olatunji et al. (2007) subscale, and the factor was likely due in part to a floor effect in this sample. To illustrate, the mean FOG subscale score was only 0.77 ($SD = 2.28$). This indicates that this scale should not be applied to an atheist sample. It also suggests that the items on the FOG subscale, along with item 3 and 11, are potential areas of religious bias in the measure. Nevertheless, the results for the Christian sample were not very strong, so the authors did not retain the two-factor model as the best-fitting model. The authors did retain a revised two-factor model for the atheist sample.
Gallegos et al. (2018) created a Spanish translation of the PIOS-R and examined its factor structure in a sample of 361 Mexican undergraduate students with an average age of 20.59 (SD = 1.92). The sample was majority female (79.4%) and predominantly Hispanic (91.2%). Participants identified as Christian (80.5%), atheist/agnostic (15.5%), or a different religion (4.0%). A CFA was conducted, but the authors did not report the estimator used. The factor analysis revealed that the fit of the two-factor model had mediocre to poor fit: RMSEA of .099, NFI of .87, and CFI of .90. The authors did not retain the two-factor model and instead moved to EFA (see One-Factor Model section).

Together, the findings across the five studies provide mixed support for the two-factor model. In two of the studies, neither of which used a WLSMV estimator, overall model fit was good and factor loadings were significant and salient (Olatunji et al., 2007; Shapiro et al., 2013). Both used sufficient sample sizes (N = 352 and 417), but were selected from different populations (i.e., undergraduate students and treatment-seeking individuals with OCD). Fit was generally poor or mixed in the other studies. Two showed indices that were adequate or fairly close to adequate (Huppert & Fradkin, 2016; Phillips & Fisak, 2022), and an argument could be made for the two-factor model in these studies, but the overall fit is not strong and cannot be said to be “good.” Examination of the Spanish version of the PIOS-R also revealed poor fit overall, with no indices in the “good” range and two indices barely meeting the thresholds for acceptable fit (Gallegos et al., 2018). Given the mix of findings, it will be important to continue testing the two-factor model in different samples and using the appropriate CFA methods for ordinal data.

One-Factor Model. A simple one-factor solution was examined or found in four studies (Gallegos et al., 2018; Huppert & Fradkin, 2016; Phillips & Fisak, 2022; Shapiro et al., 2013).
Huppert and Fradkin (2016) did not report fit indices for the one-factor model, but they did report poorer fit compared to the two-factor model, $\chi^2(1) = 32.85, p < .001$. Shapiro et al. (2013) examined the one-factor model in their three samples. For the full sample, fit indices were poor to good: RMSEA of .17, CFI of .95, and SRMR of .07. The results suggested that comparative fit was good, absolute fit was acceptable to poor. The RMSEA indicated that the absolute fit corrected for parsimony of the model is not acceptable. Fit indices were poorer for the non-scrupulous sample: RMSEA of .19, CFI of .92, and SRMR of .09. The CFI and SRMR were in the acceptable range but not in the good range, and the RMSEA value was poor. For the scrupulous sample, however, fit was somewhat better, with a CFI of .93 and SRMR of .08. However, the fit indices were all in the acceptable range for the two-factor model (see previous section). The authors did not retain a one-factor model.

Phillips and Fisak (2022) used separate EFAs with principal axis extraction and a promax rotation to explore the factor structure of the PIOS-R in the Christian sample. The scree plot indicated a one-factor solution, and all factor loadings were significant and salient ($>.56$). The authors concluded that a single-factor solution fit best in their sample of Christian participants. Gallegos et al. (2018) also retained a one-factor solution in an EFA conducted using principal axis extraction. All factor loadings were significant and salient ($>.49$).

In two studies, the one-factor model was found to fit significantly worse than a two-factor model and showed questionable fit (Huppert & Fradkin, 2016; Shapiro et al., 2013). In two different exploratory studies, a one-factor model was retained (Gallegos et al., 2018; Phillips & Fisak, 2022). Importantly, both of these studies employed a sample of exclusively or primarily Christians. Neither reported the estimator used for their CFAs, but both concluded that the two-
factor model had poor fit. This may be especially relevant given that many of the studies used samples of participants with a mix of religious affiliations or did not report religious affiliation. It is possible that differences in religion is a cause of the “mud” in the water regarding factor structure of both the PIOS and the PIOS-R. Given the mixed results across multiple studies, the one-factor model was examined in the current study.

**Bifactor Model.** Only one study examined a bifactor model in the PIOS-R (Phillips & Fisak, 2022). The authors cited Huppert and Fradkin (2016) when they reported their plan to test a bifactor model, but Huppert and Fradkin (2016) established a bifactor model for the original 19-item PIOS, not for the PIOS-R. No other information about the exact structure of the bifactor model tested by Phillips and Fisak (2022) was provided in their published article, limiting readers’ understanding of the results. The bifactor model that was tested showed poor fit in both the Christian (RMSEA of .11; CFI of .901, and TLI of .866) and atheist (RMSEA of .147, CFI of .861, and TLI of .813) samples. However, since the bifactor model has only been tested in one previous study, the bifactor model tested by Phillips and Fisak (2022) was also tested in the current study.

**Psychometric Properties**

**Reliability**

Several aspects of scale reliability of the PIOS and PIOS-R have been examined in the literature, including measures of internal consistency. Different authors have used different methods for this, including Cronbach’s alpha and McDonald’s omega. Across many studies involving the PIOS that reported Cronbach’s alpha, values for the full scale ranged from .93
(e.g., Abramowitz et al., 2002) to .98 (Inozu, Eremsoy, et al., 2017), indicating excellent internal consistency. Huppert and Fradkin (2016) also examined the PIOS using an omega hierarchical, which resulted in a value of .91, indicating excellent internal consistency. For the FOS subscale of the 19-item PIOS, Cronbach’s alphas ranged from .88 (Inozu et al., 2020) to .93 (Inozu, Clark, & Karanci, 2012). Alpha values for the FOG subscale ranged from .86 (Bruce et al., 2011) to .95 (Inozu, Keser, & Karanci, 2017). The PIOS and its subscales consistently show good internal consistency reliability across diverse samples.

For the PIOS-R, results in the literature have been similar, with Cronbach’s alphas for the full scale ranging from .92 (Gallegos et al., 2018) to .95 (e.g., Fergus & Rowatt, 2014a). The FOS subscale has excellent internal consistency, as indicated by Cronbach’s alphas ranging from .91 (Fergus, 2014) to .95 (Shapiro et al., 2013). Alpha values for the FOG subscale ranged from .86 (Shapiro et al., 2013) to .91 (Olatunji et al., 2007). Thus, the PIOS-R displays good internal consistency reliability across many studies with diverse samples.

The correlations between the FOS and FOG subscales provide information on the extent to which the constructs fit in the same scale and the extent to which they both provide unique information. For the original 19-item PIOS, subscale intercorrelation coefficients ranged from .67 ($p < .001$; Abramowitz et al., 2002) to .85 ($p < .01$; Nelson et al., 2006), indicating that the subscales are highly related. However, since at least one study has found a correlation of .85, the subscales may have poor discriminant validity. Generally, a correlation coefficient of .85 or higher between two factors indicates that the factors are not sufficiently unique and can be combined (Brown, 2015). For the PIOS-R, correlations between the two subscales ranged from .82 ($p < .01$; Fergus, 2014) to .99 ($p < .05$; Olatunji et al., 2007). The almost perfect correlation
in at least one study indicates that the subscales of the PIOS-R are even less likely to be representative of unique factors. This further illustrates the potential differences between the two versions of the PIOS on their factor structures. Item-total correlations also are often examined as evidence of a scale’s reliability. Few studies have reported item-total or corrected item-total correlations, to the author’s knowledge. One study showed PIOS corrected item-total correlations ranging from .55 to .80, with the majority above .60 (Olatunji et al., 2007). Another study calculated item-total correlations for the PIOS in two samples (Pirutinsky & Rosmarin, 2018). In a sample of Jewish community members, item-total correlations ranging from .53 to .80, again with the majority above .60. In a small ($N = 34$) sample of outpatients at an anxiety center who identified as Orthodox Jewish, item-total correlations ranged from .47 to .87, with the vast majority above .70. In sum, all evidence suggests excellent reliability of the PIOS and the PIOS-R, even when tested in samples differing in culture, religion, education, and age.

**Validity**

Several aspects of the validity of the PIOS and PIOS-R have been examined in the literature, including known-group, convergent, discriminant, and predictive validity. There is little research on known-groups validity for the PIOS-R, as no studies on this were found during the course of the author’s literature review.

**Known-Group Validity.** Known-group validity of the PIOS was examined in a few studies. Nelson et al. (2006) compared total scores on the PIOS between groups of people whose primary OCD symptom dimension was either contamination ($n = 15$), harming ($n = 20$), symmetry ($n = 9$), and unacceptable thoughts with religious, violent, and sexual content ($n = 22$).
It was found that those with primary unacceptable thoughts \((M = 35.55, SD = 16.88)\) had higher PIOS scores than those with primary contamination symptoms \((M = 15.13, SD = 11.41; F(3, 62) = 3.95, p < .05, \text{partial } \eta^2 = .16)\). However, there were no significant differences between the unacceptable thoughts group and the symmetry and harming groups. It is not surprising that the harming group would score similarly on the PIOS, given that there are often moral themes in fear of harming OCD. However, it would be expected that the symmetry group would score lower on the PIOS. It should be noted that participants did not appear to have been screened for overlap of symptom domains and seemed to have been placed in the different groups based on primary symptom dimension only. Thus, there likely were participants in the contamination, harming, and symmetry groups who also experienced symptoms in the unacceptable thoughts domain as well, even though it was not their primary symptom theme. This potential crossover in symptoms may explain the lack of significant differences between the unacceptable thoughts group and the other two groups because participants in those groups may also have experienced significant scrupulous symptoms.

In another study, participants were placed into four diagnostic groups based on the Y-BOCS symptom checklist and chart reviews (Huppert & Fradkin, 2016). These included participants with OCD and scrupulous obsessions \((n = 46)\), those with OCD and other repugnant obsessions (but not scrupulous obsessions; \(n = 43\)), those with OCD who did not experience scrupulous or other repugnant obsessions \((n = 42)\), and those with an anxiety disorder (but not OCD; \(n = 28\)). On PIOS total scores, those with scrupulous OCD symptoms scored higher than the other three groups, and those with other repugnant obsessions scored higher than those with no repugnant obsessions and those with an anxiety disorder \((F[3, 155] > 16, p < .001, \text{partial}\)
Those with scrupulous obsessions also scored higher than the other three groups on FOG scores, and the scrupulous and repugnant obsessions groups scored higher than the other two groups on FOI scores. Additionally, the authors found that the PIOS and its subscales were higher for OCD patients with scrupulous obsessions compared to OCD patients without scrupulosity and those with anxiety disorders ($F_{[3, 155]} > 11, ps < .001$, partial $\eta^2 > .18$). These results suggest that the PIOS distinguishes between OCD patients with and without scrupulous obsessions, but the PIOS may not distinguish between OCD patients with scrupulosity and patients with other repugnant obsessions, at least when using the FOI subscale.

Another study involved giving the PIOS to 34 Orthodox Jewish individuals with OCD (27%), anxiety disorders (46%), or mood disorders (26%; Pirutinsky & Rosmarin, 2018). It was found that PIOS total means were highest in the OCD group and lowest in the mood disorders group, although the results were not statistically significant ($F_{[2, 31]} = .26, p = .78$, partial $\eta^2 = .02$). However, the authors argue that statistical nonsignificance, in this case, likely reflects very small sample sizes of the groups (e.g., < 16). The effect size does suggest that the difference was meaningful, albeit small. Overall, known-groups analyses of the PIOS revealed that the PIOS discriminates well between individuals diagnosed with OCD and those diagnosed with anxiety disorders or other problems. It also reveals differences between OCD patients with scrupulous obsessions and those with primarily contamination concerns or those who do not experience scrupulous obsessions. However, the Fear of Sin/Immorality subscale may not differentiate well between OCD patients with scrupulous obsessions and OCD patients with other repugnant obsessions. The Fear of God subscale may do a better job of this.
Convergent and Discriminant Validity. Many studies have examined convergent and discriminant validity of the PIOS and PIOS-R. Of the available data, I will review a selected portion focused primarily on convergent validity with constructs frequently associated with scrupulosity and on discriminant validity with general distress, positive affect, and OC symptom dimensions that do not include scrupulous symptoms. There are some variables that should conceptually show stronger relationships with scrupulous OCD than with other types of OCD. This would include stronger correlations with moral thought-action fusion (moral TAF; the belief that thinking about an action is morally equivalent to performing the action), religiosity, and guilt. Thus, convergent validity with these variables is described below.

The full length PIOS has been correlated with many measures of OC symptomatology, including the OCI-R, the Y-BOCS, the Clark-Beck Obsessive-Compulsive Inventory (CBOCI; Clark & Beck, 2002), and the MOCI. Some studies have found nonsignificant correlations between the PIOS and general OC symptoms, as measured by the OCI-R (e.g., rs of .01 and -.01, ns; Nelson et al., 2006). However, most researchers have found significant, very small to large sized, positive correlations with general OC symptoms, as measured by total OCI-R scores (rs = .34-.61; Huppert & Fradkin, 2016; Inozu et al., 2020; Pirutinsky & Rosmarin, 2018; Witzig & Pollard, 2013), total Y-BOCS scores (r = .17; Buchholz et al., 2019), CBOCI-Obsessions scores (rs = .32-.62; Inozu, Clark, & Karanci, 2012; Inozu, Keser, & Karanci, 2017), CBOCI-Compulsions scores (rs = .32-.43; Inozu, Clark, & Karanci, 2012; Inozu, Keser, & Karanci, 2017), MOCI scores (r = .36; Abramowitz et al., 2002). Further, both PIOS subscales correlated significantly and substantively with measures of obsessions and compulsions, even after
controlling for depression and anxiety (partial $rs = .43-.53$ for CBOCI-Obsessions and $.21-.25$ for CBOCI-Compulsions; Inozu, Clark, & Karanci, 2012).

As expected, correlations with measures of contamination OC symptoms (i.e., the Washing subscale of the OCI-R, the Washing subscale of the MOCI, and the Contamination subscale of the Dimensional Obsessive-Compulsive Scale [DOCS]; Abramowitz et al., 2010) in the literature often were small in size and sometimes nonsignificant (significant $rs = .17-.22$, and nonsignificant $r = .11$; Buchholz et al., 2019; Huppert & Fradkin, 2016; Abramowitz et al., 2002). However, one study revealed a significant, large correlation with DOCS-Contamination ($r = .54$; Pirutinsky & Rosmarin, 2018). All correlations with OCI-R Checking, Hoarding, and Symmetry/Ordering subscales were relatively small ($rs \leq .21$) and often nonsignificant (Huppert & Fradkin, 2016; Nelson et al., 2006). One study found a significant correlation with the OCI-R Neutralizing subscale that was medium in size ($r = .34$; Huppert & Fradkin, 2016), but another study revealed a small and nonsignificant negative correlation ($r = -.12$, ns; Nelson et al., 2006).

There are stronger and more consistent correlations between the PIOS and repugnant obsessions, as measured by the DOCS Unacceptable Thoughts subscale ($r = .52$; Buchholz et al., 2019) and the OCI-R Obsessing subscale ($rs = .40-.68$; Huppert & Fradkin, 2016; Nelson et al., 2006; Witzig & Pollard, 2013). One study did find a low but nonsignificant relationship between the PIOS total score and DOCS-Unacceptable Thoughts in a sample of outpatient Orthodox Jews ($r = .25$, ns; Pirutinsky & Rosmarin, 2018). Notably, in this study, correlations between repugnant obsessions and PIOS total and FOG subscale scores were relatively small and nonsignificant, but the correlation with PIOS Fear of Immorality scores was significant and relatively large in size ($r = .40$). Overall, the results show an expected pattern of results between
the PIOS and general and specific OC symptoms. Specifically, the PIOS displays small to large relationships with general OC symptoms and is more strongly related to repugnant obsessions (a symptom dimension that often includes scrupulous symptoms) than to other symptom themes.

The PIOS should also be correlated with measures of certain cognitive features that are often associated with OCD, such as TAF. TAF is an underlying belief that thoughts are intimately connected or equivalent to actions. There are two types of TAF. As discussed previously in this section, moral TAF is highly conceptually related to scrupulous OCD. Likelihood TAF is the other type; it is the belief that thinking about an event makes it more likely to occur. It also is involved in OCD, but its conceptual link to scrupulosity is less strong. As expected, the PIOS was correlated with the Thought-Action Fusion Scale (TAFS; Shafran et al., 1996) Moral subscale ($r = .36-.44$), but not with the TAFS Likelihood subscale ($r = .04, ns$; Nelson et al., 2006).

Correlations between general measures of religiosity (a term for strength of religious faith) and PIOS total scores vary widely in statistical significance, size, and direction. Some studies (e.g., Pirutinsky & Rosmarin, 2018) have found nonsignificant and very low ($rs = -.02-.09, ns$) or small ($r = -.16, ns$; and $r = .20, p < .05$) correlations between the PIOS and measures of religiosity, including the Duke University Religion Index (DURI; Koenig et al., 1997), the Brief Multidimensional Measure of Religiousness/Spirituality (Fetzer Institute, 1999), and a measure of 7 items created for the Pirutinsky and Rosmarin (2018) study. However, other studies discovered significant relationships that were relatively large (e.g., $r = .41$, with religiosity measured by two items created by Inozu and colleagues for their study; Inozu et al., 2020). Relatively large correlations have also been found using a single item created for the study
(\(r = .31-.36\); Abramowitz et al., 2002; Huppert & Fradkin, 2016), and using the PIOS subscales and the Anxiety subscale of the Attachment to God Inventory (AGI-Anxiety; Beck & McDonald, 2004; \(rs = .53-.66\); Bruce et al., 2011). In one study involving a sample of Anabaptists, there was a significant, small, negative correlation between the PIOS and the Spiritual Well-Being Scale (Paloutzian & Ellison, 1982; \(r = -.32\); Witzig & Pollard, 2013). Importantly, the correlations between scrupulosity and religiosity seem to depend in part on religious affiliation of the sample, with Jewish samples showing lower or more nonsignificant correlations than unselected samples and Muslim samples (see Religion and Scrupulosity section). The wide variation of findings likely also depends on how religiosity is measured and whether one specific aspect of religiosity is focused on more in the measure that is used (e.g., focus on extrinsic vs. intrinsic religiosity).

Convergent validity with a measure of guilt has also been established. The Guilt Inventory (Kugler & Jones, 1992) has shown significant, medium to large correlations with the PIOS total and subscale scores (\(rs = .42-.63\); Inozu, Clark, & Karanci, 2012; Inozu et al., 2020). Since guilt is a common problem for scrupulous individuals, moderately sized correlations are expected.

Correlations also have been calculated with measures of general anxiety and depression. The PIOS and its subscales have shown small to large correlations with general anxiety, as measured by the Beck Anxiety Inventory (BAI; Beck & Steer, 1990; \(rs = .17-.38\), with most coefficients greater than .30; Inozu, Clark, & Karanci, 2012; Inozu, Keser, & Karanci, 2017) and the STAI-T (significant \(r = .49\); Witzig & Pollard, 2013; and nonsignificant \(r = .32\); Nelson et al., 2006). Considering that anxiety plays a strong role in OCD and scrupulosity, it is expected that there would be some relatively strong correlations found between them. There were small sized, typically significant, correlations with general depression, as measured by the Beck
Depression Inventory (BDI; Beck et al., 1979; \(rs = .23-.36\); Inozu, Clark, & Karanci, 2012; Inozu, Keser, & Karanci, 2017; Nelson et al., 2006; Witzig & Pollard, 2013). In one study, PIOS subscales were not significantly correlated with depression after controlling for OC symptoms (partial \(rs = -.07-.06, ns\); Inozu, Clark, & Karanci, 2012). These results suggest that, although the constructs are related, they are unique, as expected.

Fewer studies have examined the PIOS-R and thus there is less convergent and discriminant validity data compared to the PIOS. However, the available data include correlations between the PIOS-R and several constructs, including general OC symptomatology, specific OC symptom themes, religiosity, general anxiety, negative affect, positive affect, and moral TAF. The PIOS-R has displayed moderate correlations with general OC symptoms, as measured by the DOCS \((r = .57; \text{Fergus & Rowatt, 2014a})\) and the OCI-R \((r = .43; \text{Olatunji et al., 2007})\). The PIOS-R showed good convergent validity with OCI-R-Obsessing \((r = .45)\), and a stronger relationship with Obsessing than with other subscales such as Checking \((r = .25)\), Ordering \((r = .23)\), Hoarding \((r = .35)\), and Neutralizing \((r = .32; \text{Olatunji et al., 2007})\). The PIOS-R has been correlated with the new DOCS-Scrupulous and Religious Thoughts (DOCS-SR), showing strong correlations in a student sample \((r = .69)\) and an OCD patient sample \((r = .74; \text{Wetterneck et al., 2021})\). A correlation with mental contamination as measured by the Vancouver Obsessive-Compulsive Inventory-Mental Contamination (Radomsky et al., 2014) was quite large \((r = .69; \text{Fergus, 2014})\), and correlations with contamination symptoms were generally small to moderate when measured by DOCS-Contamination \((r = .50; \text{Fergus, 2014})\) and by OCI-R-Washing \((r = .27; \text{Olatunji et al., 2007})\). These correlations overall were in the expected direction and showed an expected pattern of strength of relationships. Correlations also
were significant and positive between the PIOS-R and TAFS-Moral \( (r = .41; \text{Gallegos et al., 2018}) \), a construct conceptually related to scrupulosity. The relationship between the PIOS-R and religiosity also is in the expected direction, with coefficients ranging from \( .21-.45 \) when using the General Religiousness Scale (GRS; Rowatt et al., 2009; Fergus, 2014; Fergus & Rowatt, 2014a; Lau & Ramsay, 2019) and \( .40 \) when using the Santa Clara Strength of Religious Faith Questionnaire (SCSRFQ; Plante & Boccaccini, 1997a; Gallegos et al., 2018).

In addition to the expected relative differences in strength of relationship to OC symptom themes, discriminant validity has been established with other constructs. The PIOS-R and its subscales have displayed small, positive correlations with general anxiety, as measured by the STAI-T \( (rs = .23-.38; \text{Olatunji et al., 2007; Stewart et al., 2020}) \). Negative affect, as measured by the PANAS-NA has had small to medium, positive correlations with the PIOS-R \( (rs = .30-.46; \text{Fergus, 2014; Fergus & Rowatt, 2014b; Olatunji et al., 2007}) \). Although it is expected that those high in scrupulosity experience more negative affect, the largest correlation coefficient \( (r = .46) \) is on par with the PIOS-R’s relationship with OCI-R-Obsessing \( (r = .45, \text{see previous paragraph}) \). However, these correlations are from separate samples, and the correlation between the PIOS-R and the PANAS-NA was smaller than that between the PIOS-R and OCI-R-Obsessing when considering a single study sample \( (i.e., rs = .30 \text{ and } .45, \text{respectively; Olatunji et al., 2007}) \). Furthermore, in this study, the correlation between the PIOS-R and OCI-R total scores was still significant and medium in size after controlling for trait anxiety \( (\text{partial } r = .33) \) and negative affect \( (\text{partial } r = .37; \text{Olatunji et al., 2007}) \). Additionally, for the study in which the relationship between the PIOS-R and the PANAS-NA was \( .46 \), the correlation between the PIOS-R and the DOCS total scores was \( .57 \) (Fergus & Rowatt, 2014a). A statistical comparison
between the two correlation coefficients reveals that they are significantly different, \( z = 2.94, p < .01 \). The PIOS-R was uncorrelated to PANAS-PA \((r = -.06, ns; Olatunji et al., 2007)\). Overall, expected relationships with various constructs were in the expected direction and showed appropriate relative levels of strength. Both the PIOS and the PIOS-R appear to have acceptable convergent and discriminant validity.

Two studies have reported receiver operator characteristic (ROC) analyses to determine the sensitivity and specificity of the PIOS and PIOS-R. Huppert and Fradkin (2016) examined the PIOS total scores and revised subscales based on their bifactor model, FOG-Revised (FOG-R) and FOI, using ROC. They found that the PIOS discriminated scrupulosity from other OC symptom dimensions and anxiety disorders (AUC = .773). However, it did not discriminate between those with scrupulosity and those with other types of repugnant obsessions (AUC = .649). When examining the subscales, it was revealed that the FOG-R showed somewhat higher AUC values than the FOI subscale when identifying scrupulous patients from a full sample of individuals with OCD and anxiety disorders (AUC = .768 for FOG-R and .722 for FOI), from a sample of those with OCD (AUC = .751 for FOG-R and .694 for FOI), and from those with non-scrupulous repugnant obsessions (AUC = .685 for FOG-R and .588 for FOI). Although the subscales are different from the original subscales, these data provide some evidence that the PIOS has good validity, although it is poorer when discriminating between those with scrupulosity and those with other repugnant obsessions. Shapiro et al. (2013) completed a similar analysis using the PIOS-R. They found that the PIOS-R total score discriminated well between those with scrupulous OCD and those with non-scrupulous OC symptom dimensions (AUC = .78). The PIOS-R FOG and FOS subscales also discriminated between the two groups well.
(AUC = .78 and .76, respectively). Since the subscales did not provide any increased discriminative ability, the authors concluded that the total score would be most appropriate to use for cutoffs or discriminatory purposes.

Religion and Scrupulosity

Based on the extant literature, there is evidence that suggests that scrupulosity may be connected to religiosity differently based on religious affiliation. This can be seen when examining differences on the PIOS/PIOS-R between individuals identifying with different religious affiliations.

When choosing the religious groups to be studied in the scrupulosity literature, researchers often categorize Protestant Christians separate from Catholics. In one sample of adults diagnosed with OCD, PIOS scores were not significantly different between Catholics and Protestants ($p = .86$, Hedges’ $g = .14$; Buchholz et al., 2019). In another sample of individuals diagnosed with OCD, there also was no significant difference, ($F[4, 122] = 1.65, p = .17$, partial $\eta^2 = .05$; Huppert & Fradkin, 2016). In yet another OCD sample, the Protestant mean PIOS score was higher – but not significantly different from – the Catholic mean PIOS score (Nelson et al., 2006). However, the effect size was medium (Hedges’ $g = .68$), indicating it is possible that the study may have been underpowered, given that sample sizes were somewhat small (32 Protestants and 19 Catholics). In an undergraduate sample, Abramowitz et al. (2002) found that, when compared to Catholics, Protestants also scored higher on the FOS subscale ($F[3, 116] = 4.81, p < .01$, partial $\eta^2 = .11$) but not on the FOG subscale ($F[3, 116] = 6.14, p < .01$, partial $\eta^2 = .14$). Importantly, only one study appears to have made comparisons on the PIOS-R between
religious groups: Fergus (2014) examined differences between Catholics and Protestants in an otherwise unselected sample. No differences were found on the PIOS-R total score (ns, Hedges’ $g = .12$) or on either subscale score (ns, Hedges’ $g = .06-.19$). Another study used a regression analysis to compare Protestants and Catholics on the PIOS after controlling for related variables (Gonsalvez et al., 2010). After controlling for age, gender, depression, obsessive beliefs, religiosity, and belief in a punitive God, Protestant or Catholic affiliation was not related to PIOS scores ($\beta = .00$, ns). Although the effect size was relatively strong in one study and a significant difference on one of the two subscales was found in another study, the majority of the data indicate that there are no differences between the groups on the PIOS. Thus, there is not strong evidence for true differences in levels of scrupulosity between Catholics and Protestants. This conclusion provides rationale for collapsing the two groups into one Christian group. It should be noted that, if differences are to be found between Christian denominations, there is more evidence that there are differences between liberal and conservative groups (Bruce et al., 2011) than between Protestants and Catholics as categorized. In a sample of adults over the age of 50, Bruce et al. (2011) found that Baptists (considered a more conservative denomination) scored significantly higher on the FOS subscale ($F[1, 82] = 19.58, p < .01, \text{partial } \eta^2 = .19$) and the FOG subscale ($F[1, 82] = 9.36, p < .01, \text{partial } \eta^2 = .10$) when compared to Methodists (a more liberal denomination). Although it may be a ripe area for future research, comparing liberal and conservative denominations of Christianity is outside the scope of the current study. Thus, Christians will be collapsed into a single religious group (see Current Study section, p. 41).

Jewish groups also have been included in comparisons between religious affiliations using the PIOS. Buchholz et al. (2019) found that Jewish individuals scored lower on the PIOS
compared to Catholics ($p < .05$, Hedges’ $g = .83$), but not compared to Protestants or individuals with no religious affiliation ($ns$, Hedges’ $g = .60$ and .19, respectively). Notably, the effect size for Protestants was relatively strong, and there were not significant differences between Protestants than Catholics. Thus, it is possible that the study was merely underpowered to detect differences between Jewish ($n = 19$) and Protestant ($n = 48$) participants. In another study, highly religious Jews scored significantly lower on FOS compared to highly religious Catholics, Protestants, and those of “other” religions ($F[3, 70] = 8.97$, $p < .01$, partial $\eta^2 = .28$; Abramowitz et al., 2002). Additionally, Jews scored significantly lower on the FOG subscale compared to Protestants and Catholics ($F[3, 116] = 6.14$, $p < .01$, partial $\eta^2 = .14$), but not compared to “other” religions. Huppert and Fradkin (2016), however, found no differences between Jewish individuals and Catholics, Protestants, those of no religion, and those of “other” religions ($F[4, 122] = 1.65$, $p = .17$, partial $\eta^2 = .05$). Notably, Huppert and Fradkin (2016) found no differences between any of the five religious categories they examined on PIOS total scores, in their sample of OCD patients. Pirutinsky and Rosmarin (2018) compared Orthodox and non-Orthodox Jews on PIOS scores and found that the Orthodox group scored significantly higher, $p < .001$, $d = .53$. Across studies, Jewish individuals tend to have lower scrupulosity scores compared to Christian groups, but, similar to Christians, may show differences among different branches of the religion.

Muslims also have been studied in the scrupulosity literature. One study used the PIOS to compare Turkish individuals and Canadian individuals (Inozu, Clark, & Karanci, 2012), who were further split into highly religious and low-religious groups. Although groups were not selected based on religion, all of the highly religious Turkish participants identified as Muslim,
and all of the highly religious Canadian participants identified with some form of Christianity (i.e., Catholicism, evangelical, or mainline Protestant). Results showed no differences between the groups on both of the PIOS subscales among low-religious participants. On the FOS subscale, there also were no differences between the groups for high religious participants. On the FOG subscale, among the highly religious groups, Turkish participants scored higher than Canadian participants, $F(2, 330) = 7.56, p < .001$, partial $\eta^2 = .04$. However, the effect was small.

One study conducted in Israel compared 34 Muslims and 43 Jews on PIOS scores (Siev et al., 2017). Muslim participants scored significantly higher than Jewish participants, and the effect was quite large, $t(74) = 4.04, p < .001, d = .94$. Based on the minimal available data involving Muslim individuals, this group seems to score higher on the PIOS compared to Jewish groups, and there may be minimal differences between Muslims and Christians, but only when both are also highly religious, and only on the FOG subscale. Thus, the current data showing differences between Muslims and Christians is weak.

These comparisons between religious groups on scrupulosity levels are important for understanding whether certain aspects of one’s religion might influence scrupulous OCD symptoms or whether individuals of certain religions might be more vulnerable to scrupulous symptom themes in OCD. For example, Abramowitz and Hellberg (2020) proposed that certain religions might “cultivate” scrupulous symptoms due to teachings on the moral equivalence of thoughts and actions (i.e., moral TAF) and fear of punishment. This is important information to consider when working with individuals with scrupulosity in treatment settings. Much work is still to be done in making comparisons among many different religious groups. However, even the data we currently have may be biased since it has not been established that the PIOS is
invariant across even the most common, broadly defined religious groups (e.g., Christianity, Judaism, and Islam). The extant research suggests that there are differences on the PIOS between groups of individuals who identify with different religions, but it is presently unclear as to whether these are true differences on the construct being measured or if they are an artifact of differences in the meaning of the measure.

**Religion and PIOS content**

Whereas Christianity, Judaism, and Islam are all Abrahamic religions, there are clear differences in religious laws and practices that translate to differences in the presentation of scrupulous OCD. In order for an instrument to properly measure scrupulosity in each religion, these differences should be carefully considered in its creation. Although the developers of the PIOS did not report religious affiliation of their initial sample, there is a strong probability that the sample was similar in religion to the sample in the second study within their published article (Abramowitz et al., 2002). Recall that their study 2 sample (see PIOS Development section) identified primarily as Christian (49.2%) and 21.3% Jewish, with the remaining either unaffiliated (12.7%) or identified with some “other” religion (16.8%). This suggests that the process of selection from the original item pool may have been biased by the relative lack of representation from Jewish and Muslim individuals in the sample. Further, the developers did not report in detail on the original item pool, leaving readers uncertain as to whether the original items may have represented scrupulous symptoms in one religion over others.

There are both conceptual and evidence-based differences among Christianity, Islam, and Judaism that may influence scrupulosity presentation. Whereas each religion has its own rituals
that must be performed in a certain way, in Christianity, the performance of rituals is largely left to clergy members who are specially trained. Thus, most lay members with scrupulosity are less likely to experience symptoms related to performing rituals correctly. However, practicing Muslims and Jews have multiple daily or weekly rituals that they must perform correctly in accordance with their religious laws. For example, Muslims have washing rituals (wudu, ghusl) that are performed in order for prayer rituals to be valid. This is important given that Muslim law requires practitioners to pray five times each day. Wudu typically does not need to be completed before every prayer and can last up to one day, but certain activities or occurrences break wudu (e.g., after using the bathroom, after menstruation, after sleeping). Ghusl is a more intensive ritual that must be performed after larger impurities such as sexual intercourse or menstruation. In both rituals, there are specific steps and requirements for the ritual to be accepted by God. Thus, many Muslims with scrupulosity have symptoms related to these rituals (Besiroglu et al., 2014). Jewish law involves many rituals and practices that involve seemingly strict rules. Jewish law also involves cleansing rituals that differ depending on the branch to which one belongs. Another example of Jewish practice is Shabbat, which occurs once a week during a period of a full day in which Jews are required to observe a day of rest. During Shabbat, there are very specific activities that Jewish individuals are not allowed to engage, including making repairs, cooking, writing, and other categories of work. Given the focus in Judaism on rituals and laws that emphasize behaviors, Jewish individuals are likely to have more symptoms related to following these laws (Horwitz et al., 2019). There also is evidence that a focus on morality in thinking is emphasized more in Christianity than in Judaism. Specifically, moral TAF (see Psychometric Properties section) consistently has been found to be lower in Jewish groups.
compared to Christian groups (e.g., Cohen, 2003; Siev & Cohen, 2007). One study also found that Muslims were higher on moral TAF compared to Jews ($t = 4.99, p < .001, d = 1.16$; Siev et al., 2017). Although Muslims and Christians have not differed on moral TAF (Yorulmaz et al., 2009), Muslims tended to be higher than Christians on CBOCI-Compulsions (Inozu, Karanci, & Clark, 2012), which reflects a difference in emphasis on moral behavior versus moral thinking across the two religions.

Although the current PIOS/PIOS-R have only two factors—Fear of God and Fear of Sin—it is possible that inclusion of more Muslim and Jewish participants and related items in the item pool would have resulted in different factors. For example, a factor focused on performing rituals correctly would seem highly relevant to Muslims and Jews with scrupulous OCD, given the focus on rituals and behavior in those religions. There are seven items on the PIOS that specifically describe fear of immoral thoughts (items 1, 8, 11, 14, 16, 18, and 19; e.g., “I am afraid of having immoral thoughts”), with all but one of those items on the FOS subscale. However, fear of acting immorally is described in six items (items 3, 6, 9, 10, 12, and 13; e.g., “I fear I will act immorally”), and those items are included in both subscales. It is entirely possible that, in a Jewish sample, for example, the PIOS may have had a subscale specifically related to fear of behaving immorally. Additionally, some aspects of Christianity are represented in the PIOS despite being less salient in the other religions. Specifically, item 4, “I feel urges to confess sins over and over again,” may reflect some aspect of reassurance-seeking for Jews and Muslims, but for Christians, confession is often an important part of their religious practice. Thus, this item is likely to be less salient for non-Christians. In conclusion, there are several reasons why the PIOS may be invariant across Christians, Jews, and Muslims.
There is a clear paucity of research examining the measurement invariance of the PIOS/PIOS-R. To the author’s knowledge, only one published study has examined measurement invariance of the PIOS across religious groups (Pirutinsky & Rosmarin, 2018). No other studies have tested measurement invariance of the PIOS among any other groups, religious or otherwise, and none have examined measurement invariance of the PIOS-R. No studies have even examined differential item functioning using item response theory, another method for examining aspects of measurement invariance. However, there is minimal evidence that the PIOS/PIOS-R may not be invariant among certain religions.

Pirutinsky and Rosmarin (2018) tested measurement invariance of the full length PIOS across Orthodox and non-Orthodox Jewish groups. They used a bifactor model similar to Huppert and Fradkin (2016) and found evidence for scalar invariance, $\Delta \chi^2(80) = 15.45, p = .61, \Delta RMSEA = .004, \Delta CFI = .001, \Delta TLI < .001, \Delta WRMR = .15$. This suggests that the PIOS has the same factor structure and likely has the same meaning in Orthodox and non-Orthodox Jewish groups. This is valuable information that provides a basis for making comparisons between those two groups based on PIOS scores. However, it is only a start. Information is still needed on the invariance among several other religious groups.

Huppert and Fradkin (2016) conducted a study in which they completed an ROC analysis (see Validity section above) on the PIOS. They conducted separate ROC analyses for Christian and non-Christian participants, all of whom had OCD, and found that, for Christians, the PIOS total score adequately discriminated between those with scrupulous symptoms and those with non-scrupulous symptoms, AUC = .765. For non-Christians, however, the PIOS did not
discriminate between scrupulous and non-scrupulous OCD, AUC = .554. They also conducted a 2 (Christians vs. non-Christians) x 2 (repugnant vs. scrupulous obsessions) ANOVA. A significant difference was found between the repugnant and scrupulous groups on the PIOS for Christians ($F[1, 121] = 15.56, p < .001$, partial $\eta^2 = .11$), but not for non-Christians ($F[1, 121] = 0.54, p = .46$, partial $\eta^2 = .00$). The authors concluded that the PIOS’s validity may depend on religious affiliation. This study offers preliminary evidence that the PIOS may function differently in Christians compared to those of other religious affiliations.

Another study offered some evidence that the factor structure of the PIOS-R differs based on religious beliefs (Phillips & Fisak, 2022). However, the groups examined were Christian and atheist. Although this was important work because researchers often have included atheists in their research using the PIOS/PIOS-R, it is not a group with which the PIOS/PIOS-R were likely meant to be used. Nevertheless, the study offered relevant information, including that the FOS/FOI subscale of the PIOS-R may be more relevant for use with individuals who do not have religious beliefs but who may have moral concerns outside of the context of religion (excluding, of course, two items that did not perform well in the atheist sample). Overall, an examination of the extant literature reveals a distinct gap in information which the current study aims to fill.

Emic and Etic Approaches

One way to frame measurement invariance examinations across cultures or in this case, across a very important aspect of culture (one’s religion) is by considering emic and etic approaches to scrupulous OCD. There are benefits to emic research, which is from the perspective of the subject, because there is a better understanding of the inner complexities and
true meaning behind cultural phenomena. Although it is important that research be done from an emic perspective, such as OCD researchers studying scrupulosity within groups of their own religious background, there is also need for cross-cultural research, which is typically etic in nature. That is why we need an instrument that can translate across religious tradition. However, etic approaches can be biased by the culture of the observer. As discussed in the Religion and PIOS Content section above, the PIOS’s development may have been influenced by the use of primarily Christian participants. Measurement invariance studies can clear up some of the concerns around cross-cultural research methods because it takes into consideration the emic perspective, or the meaning, of the measure.

Current Study

For this project, measurement invariance testing of the PIOS and the PIOS-R was conducted using MGCFA. The analyses examined whether the scales were invariant across three religious affiliations: Christianity, Judaism, and Islam. This is crucial given that these three affiliations often are included in studies using the PIOS/PIOS-R and conclusions about differences on the PIOS/PIOS-R among these groups have been offered. This was done even though it was unknown whether these differences represent true differences on the construct of interest (i.e., scrupulosity) or if they may be an artifact of some kind of measurement bias. There has been some research indicating that the measure may have better discriminatory properties – as tested using ROC analyses by Huppert and Fradkin (2016) – among individuals who identify as Christian than among individuals who identify with other religions. Further, differences on the PIOS/PIOS-R have been found in the literature between groups of different religions. By
completing MGCFAs to test configural, metric, and scalar invariance, researchers will have a better understanding of how they should or should not be using the test and of the inferences that can or cannot be made based on the measure.

Prior to measurement invariance analyses, it is necessary to determine a best-fitting measurement model for both the PIOS and PIOS-R. Thus, I first examined multiple extant measurement models of the PIOS and the PIOS-R, respectively, in the full sample. The literature indicated three potential models for the PIOS (i.e., a one-factor, two-factor, and bifactor model) and two potential models for the PIOS-R (i.e., a one-factor and two-factor model). Additionally, although the bifactor model showed poor fit for the PIOS-R, it was examined only in one previous study and thus it also was tested in the current study to confirm or challenge previous results. Given that results from previous studies have found multiple different best-fitting models, it is important to continue to examine the factor structure in different samples. The current project accomplished this by being the first to test the factor structure of the PIOS and the PIOS-R separately among individuals of Christian, Jewish, and Muslim affiliations. Further, the current study’s findings offer guidance on the relevance of the PIOS compared to the PIOS-R. Many researchers continue to use the PIOS despite the evidence that the PIOS-R is superior in terms of factorial validity and equivalent on other markers of reliability and validity. Although it is impossible to directly compare model fit indices between the two versions because they have a different set of indicators, it is possible that the PIOS-R may show a clearer and more consistent picture of factorial validity than the PIOS. If that is the case, and if the two versions are equivalent on other tests of validity and reliable, as has been seen in the literature, then a strong argument can be made for discontinuing the use of the 19-item PIOS in future research.
Differences among all three religious groups on PIOS/PIOS-R scores had not previously been examined together in a single study. Additionally, to the author’s knowledge, latent mean differences had not yet been examined in the literature, and latent mean comparisons typically are more accurate than composite score comparisons because latent factors are free of measurement error (Ployhart & Oswald, 2004). The current study also examined other aspects of population heterogeneity of the PIOS/PIOS-R. This included testing the equality of factor variances and covariances among the three groups.

Finally, other aspects of reliability and validity of the PIOS/PIOS-R were examined. This included internal consistency estimates, convergent validity with a measure of OC symptomatology, and discriminant validity with measures of religiosity and general distress.

**Hypotheses**

**Hypothesis 1**

Given some evidence that the concurrent validity of the PIOS may be weaker in non-Christians compared to Christians (Huppert & Fradkin, 2016), it was hypothesized that the PIOS and PIOS-R would be noninvariant across the religious groups and between Christians and the other two groups. There was a lack of data on the PIOS/PIOS-R to guide hypotheses regarding pairwise analyses between Jewish and Muslim individuals. However, they have been shown to have different scores on the PIOS (Siev et al., 2017), despite evidence that similar rates of scrupulosity are found among individuals with OCD in both groups (Greenberg & Shefler, 2002; Mahgoub & Abdel-Hafeiz, 1991). Therefore, it was hypothesized that the PIOS/PIOS-R also would be noninvariant across those groups.
Hypothesis 2

For the PIOS, it was hypothesized that the best-fitting model for the full sample would be a bifactor model. This is because the two-factor model was best in US college samples (Abramowitz et al., 2002; Olatunji et al., 2007), which tend to be irreligious (Eagan et al., 2016) compared to the general population (Pew Research Center, 2015), whereas the bifactor model was best in a nonclinical religious sample (Pirutinsky & Rosmarin, 2018). Since the current study used a nonclinical religious sample, it was expected that the bifactor model would show best fit. For the PIOS-R, the best-fitting model from the literature using a sample closest to the current sample (i.e., a mix of individuals of different religious affiliations as opposed to an exclusively Christian sample; Huppert & Fradkin, 2016) was a two-factor model. Thus, it was hypothesized that the two-factor model would fit best in the current study.

Hypothesis 3

Given that the PIOS/PIOS-R were hypothesized to be noninvariant across the three religious groups, it was not expected that analyses of population heterogeneity would be indicated. However, in the case that the measures were invariant, it was hypothesized that latent means would be significantly different (at $p < .05$), given differences in composite score means in the literature among Jews, Christians, and Muslims. Specifically, it was hypothesized that Christians and Muslims would display higher latent means on both measures compared to Jewish individuals and that Muslims would be higher on the latent variable than Christians.
Hypothesis 4

Although models were tested for both the PIOS and the PIOS-R, it was expected that the factorial validity of the PIOS-R would be superior to the factorial validity of the PIOS. Although model fit indices cannot be directly compared, it was possible that the PIOS-R could show overall better fit (e.g., if the overall fit of the PIOS-R is determined to be “good” but the overall fit of the PIOS is “poor”).

Hypothesis 5

Finally, given that the PIOS/PIOS-R generally have shown good markers of reliability and validity in previous research, it was hypothesized that the internal consistency would be good ($\omega > .80$) and that a meaningful pattern of convergent and discriminant correlations would be observed in the current sample. Specifically, internal consistency estimates were hypothesized to be in the “good” range or better. The correlations between the PIOS/PIOS-R and DOCS total score (a measure of OC symptoms) were hypothesized to be significant ($p < .05$), positive, and medium in size. The correlations between the PIOS/PIOS-R and the SCSRFQ (a measure of strength of religious faith) were hypothesized to be ($p < .05$), positive, and small to medium in size. The correlations between the PIOS/PIOS-R and the Depression and Anxiety subscales of the Depression, Anxiety, and Stress Scale - 21 (DASS-21; Lovibond & Lovibond, 1995) were hypothesized to be significant ($p < .05$), positive, and small to medium in size. The correlations between the PIOS/PIOS-R and PANAS-NA were hypothesized to be significant ($p < .05$), positive, and medium in size. Further, correlations between the PIOS/PIOS-R and each SCSRFQ, DASS-21 Depression and Anxiety subscales, and PANAS-NA were expected to be
significantly smaller ($p < .05$) than correlations with the DOCS-SR subscale, which were expected to be significant ($p < .05$), positive, and medium to large in size.
CHAPTER 2

METHOD

Participants

Participants were recruited through the crowdsourcing website Amazon Mechanical Turk (MTurk; see following section). Participants included MTurk workers who (a) identified as either Christian, Jewish, or Muslim; (b) were from the United States; (c) had approval ratings above 95%; and (d) successfully completed a CAPTCHA, which is a challenge-response test in computing designed to determine whether the user is a human. Initial sample sizes included a total $N$ of 844 participants, 280 of whom were in the Christian group, 299 in the Muslim group, and 265 in the Jewish group. Of the 844 individuals who participated in the study, 112 were excluded from the analyses due to disparate responses to religious affiliation items between the screener and the full study. These included 3 participants who initially identified (i.e., in their response to the screener) as Christian, 48 who initially identified as Muslim, and 61 who initially identified as Jewish. From the 732 remaining participants, 14 were excluded due to 89% or more missing data. This process left a final total $N$ of 718, with 274 in the Christian group, 243 in the Muslim group, and 201 in the Jewish group.

Participants were asked several demographic questions, including age, sex assigned at birth, gender, education level, race/ethnicity, and religious affiliation. The overall mean age was 36.11 years; the mean age was 37.98 for the Christian group, 32.02 for the Muslim group, and
38.49 for the Jewish group. The Muslim group was significantly younger than the Christian and Jewish groups, \( F(2) = 22.72, p < .001 \). Though statistically significant, the mean differences were only 5.96 and 6.47 years, respectively. For the full sample, 54.5% of participants were assigned female at birth, 45.2% were assigned male at birth, and 0.3% identified as intersex.

There were no significant differences on sex assigned at birth among the groups, \( \chi^2(4) = 9.35, p = .053 \), Cramer’s \( V = .081 \). In the overall sample, 54.1% identified as women, 44.8% as men, 0.1% as transgender women/trans feminine, 0.1% as transgender men/trans masculine, 0.6% as non-binary/genderqueer/gender fluid, and 0.3% prefer to self-describe. There was a significant difference among groups on gender (\( \chi^2(10) = 20.23, p = .027 \), Cramer’s \( V = .119 \)), and the Jewish group had a greater proportion of self-identified women (58.2% of the Jewish group) compared to men (38.8% of the Jewish group). Additionally, 42.1% of the self-identified men were in the Christian group, as compared to 33.6% who were in the Muslim group and 24.3% who were in the Jewish group. The uneven proportions of gender among groups may be taken into consideration for interpretive purposes.

Of the total sample, the highest level of education was some high school for 1.1%, high school diploma or equivalent for 11.6%, bachelor’s degree for 60.0%, master’s degree for 23.4%, PhD or higher for 2.9%, and trade school for 0.8%. Notably, Pearson’s chi-square was significant (\( \chi^2(10) = 19.20, p = .038 \), Cramer’s \( V = .116 \)), suggesting that the percentage of participants with a bachelor’s degree is significantly higher than the percentage of participants with lower or higher levels of education. For self-identified race, the full sample identified as 14.9% African American/Black, 18.4% Asian/Asian-American, 12.2% Arab/Arab-American, 2.7% American Indian or Alaskan Native, 1.4% Native Hawaiian or Other Pacific Islander, and
50.5% European American. There was a significant and large difference among groups on race, $\chi^2(10) = 208.11, p < .001$, Cramer’s $V = .398$. Given that religion is a part of culture, and that culture and race are interconnected, it was expected that race would be related to religious affiliation. In the Muslim group, there was a greater proportion of Asian/Asian Americans (27.8%) and a greater proportion of Arab/Arab-Americans (31.6%) compared to Christian (17.9% and 2.1%, respectively) and Jewish participants (7.4% and 0.5%, respectively).

Additionally, there was a higher proportion of European Americans in the Christian (64.3%) and Jewish (69.8%) groups compared to the Muslim group (20.9%). As for ethnicity, there appeared to be significantly different proportions among the religious groups, $\chi^2(2) = 8.56, p = .014$, Cramer’s $V = .112$. There appeared to be a lesser proportion of those identifying as Hispanic/Latino in the Muslim group (13.9%) compared to the Christian (22.8%) and Jewish (24.1%) groups. These differences were moderate in size, but some difference among groups would make sense for the same reason as the observed differences based on race.

**MTurk**

MTurk is a crowdsourcing website used to match requesters (i.e., individuals looking for people to complete tasks such as responding to surveys) and workers (i.e., individuals who are willing to complete requested tasks for a small monetary benefit). The number of psychology research articles using MTurk data that were published in journals with an impact factor above 2.5 has grown exponentially since 2007 (Chandler & Shapiro, 2016). It is a fast and relatively cost-effective way to recruit online convenience samples. Although US MTurk samples are not fully representative of the US population as a whole (primarily due to differences between
Internet users and the full US population; Chandler & Shapiro, 2016), they have been found to be more representative than college samples and community samples from a college town (Berinsky et al., 2012; Buhrmester et al., 2011), which are commonly used in psychology research. Additionally, since there is so much interest in MTurk-based research, the characteristics of MTurk samples are well understood, making it easier to interpret findings, including limitations.

There are, of course, several limitations of the use of MTurk for data collection. However, many of the relevant limitations are also present in other data collection methods, and researchers have offered methodological guidelines to reduce the impact of these limitations (e.g., Chandler & Shapiro, 2016). One important issue involves inattention; one review found that about 15% of workers fail attention and compliance checks (Aguinis et al., 2021). However, some researchers have found that MTurk workers were equally likely to answer an instructional manipulation check correctly compared to a community sample (Goodman et al., 2013). On the other hand, some researchers argue that MTurk workers are incentivized to pay attention and to pass attention checks, since many researchers specify that only workers with approval ratings above 95% may participate in their studies. Of note, researchers have argued that MTurk workers learn to expect attention checks and easily recognize checks that are frequently used by researchers. Thus, passing attention checks may not represent a worker’s attention on the other items on a survey, rendering the checks ineffective (Chandler & Shapiro, 2016). Peer et al. (2014) found that implementing an inclusion criterion of approval ratings above 95% (i.e., using only high-reputation workers) works just as well as, if not better than, attention check questions for ensuring data quality when using MTurk. Specifically, they found that only 2.6% of high-reputation workers failed at least one check question, compared to 33.9% of low-reputation
workers (i.e., workers with below 95% approval ratings). Additionally, no high-reputation workers failed more than one check. High-reputation workers also had significantly higher reliability scores on three questionnaires ($\chi^2$s\_[1] = 13.75, 19.86, and 62.60, $ps < .001$) and significantly lower scores on social desirability ($F[4, 689] = 2.52, p = .04, \eta^2 = .01$) compared to low-reputation workers. Attention check questions did improve data quality, but only for low-reputation workers, and not for the reliability scores. For one of the three questionnaires given in the study, the reliability score for low-reputation workers who passed check questions was not better compared to low-reputation workers who failed checks. Similarly, all low-reputation workers showed more central tendency bias than high-reputation workers, regardless of whether they passed or failed checks. These data suggest that the use of high-reputation workers is slightly preferred over the use of attention check questions and that attention check questions are redundant for high-reputation workers. Thus, the current study included the use of participants with above a 95% approval rating and not attention check questions.

Another issue is the vulnerability to web robots, or “bots” – software programs that can be designed to imitate human users in online studies to receive compensation. They can be difficult to differentiate from human respondents. Thus, guidelines for research using MTurk typically involve implementing a CAPTCHA that can prevent bots from participation (Aguinis et al., 2021), and this guideline was followed in the current study. Worker self-misrepresentation also is a relevant issue to the current study. This study required participants who identified with one of three broad religious groups: Christians, Jews, and Muslims. If this criterion is stated in the description of the task on MTurk, then workers may misrepresent themselves in order to participate and gain compensation. Thus, it is recommended that this information is kept out of
the task description (Chandler & Shapiro, 2016). Instead, researchers typically implement prescreening tasks in which a smaller reward (e.g., $0.01) is given to complete the prescreen. Then, those that pass will be invited to complete the full survey, which will have a more substantial reward (e.g., $0.70), reflecting the relative time taken to complete the surveys. Thus, workers were not informed in the task description that the inclusion criterion for the current study was identification with Christianity, Judaism, or Islam. Instead, it informed workers that the task was an initial screening survey and that those who pass the screening would be invited to participate in another survey. The compensation and expected time commitment for both the prescreen and the longer survey were included in the description.

Measures

**Religious Affiliation**

Among the other demographic questions, each participants’ religious affiliation was measured. This involved two items. The first was used to screen out participants who did not fall into the broad umbrellas of Christianity, Judaism, and Islam. This was a forced choice item listing several religious affiliations (see Appendix A). Participants were asked to choose the response option that most closely aligns with their religious affiliation. The second item was an open-ended question asking participants to type in their exact religious affiliation, including any specific denomination (see Appendix A). This item was for descriptive purposes only and to get a clearer picture of the religious make-up of the sample.
Penn Inventory of Scrupulosity

The PIOS (Abramowitz et al., 2002) is a 19-item self-report questionnaire measuring scrupulous OCD symptoms in a dimensional manner. Response options are on a 5-point scale, ranging from 0 (never) to 4 (constantly). The revised version of the PIOS (PIOS-R; Olatunji et al., 2007) also was utilized in the current study. Scores for the PIOS-R were derived from the reduced item set fully contained within the PIOS. Response options are identical to the PIOS options. The psychometric properties of the PIOS and the PIOS-R are reviewed in length in the Introduction section. Generally, both versions show good internal consistency, and meaningful patterns of convergent and discriminant validity.

Dimensional Obsessive-Compulsive Scale

The DOCS (Abramowitz et al., 2010) is a 20-item self-report questionnaire measuring OC symptomatology. Individuals are instructed to respond to each item question based on their experiences regarding the content of the item in the past month. Response options differ depending on the item, but are all on scales ranging from 0 to 4. The DOCS has four subscales with 5 items each: Contamination, Responsibility for Harm, Unacceptable Thoughts, and Symmetry, each corresponding to a prominent OC symptom dimension. The factor structure of the DOCS has been shown to correspond to its four subscales in multiple samples (Abramowitz et al., 2010). Internal consistency was good for the full scale (αs = .90-.93) and for the subscales (αs = .83-.96). Twelve-week test-retest reliability for total and subscale scores was acceptable (rs = .55-.66). The DOCS has been correlated with other measures of OCD, including the Y-BOCS (r = .43-.54; Abramowitz et al., 2010; Wetterneck et al., 2021) and the OCI-R (r = .65-
It has shown discriminant validity against depression (as measured by DASS-21 Depression subscale and the BDI) in multiple samples, as evidenced by significant differences between correlations with measures of depression and with measures of OC symptomatology ($p < .05$; Abramowitz et al., 2010).

In 2021, a new study expanded on the original DOCS, adding a Scrupulous or Religious Thoughts subscale, DOCS-SR (Wetterneck et al., 2021). Factor analyses on the new subscale alone revealed a one-factor solution, as expected. Factor analyses on the full DOCS, including the new subscale, revealed a five-factor solution, corresponding to the five subscales. Internal consistency was acceptable ($\alpha = .79$) in a student sample and excellent ($\alpha = .96$) in a sample of individuals diagnosed with OCD. Test-retest reliability has yet to be examined for the DOCS-SR. The new subscale has been shown to be significantly correlated with the PIOS-R ($r_s = .69-.74$), the OCI-R ($r = .32$), OC-related beliefs as measured by the Obsessive Beliefs Questionnaire-44 (Obsessive Compulsive Cognitions Working Group, 2005; $r = .45$), and with a measure of guilt (Perceived Threat from Emotions Questionnaire-Guilt subscale; McCubbin & Sampson, 2006; $r = .39-.50$). The DOCS-SR was uncorrelated with depression, as measured by the Patient Health Questionnaire-9 (Kroenke et al., 2001; $r = .11$, $ns$) and the BDI-II (Beck et al., 1996; $r = .17$, $ns$). It was negatively correlated with the Self-Compassion Scale Short-Form (Raes et al., 2011) in a student sample ($r = -.27$) and uncorrelated in a sample of individuals diagnosed with OCD ($r = -.08$, $ns$). The DOCS-SR subscale was used in the current study, along with the original DOCS, to examine convergent validity with the PIOS/PIOS-R.
The SCSRFPQ (Plante & Boccaccini, 1997a) is a 10-item self-report questionnaire measuring strength of religious faith, termed religiosity. This measure was developed to be used with almost all religious traditions, making it appropriate for use in the current study. Respondents are instructed to indicate their level of agreement with each item, with options ranging from 1 (strongly disagree) to 4 (strongly agree). The measure is dimensional, with low scores representing low religious faith, and higher scores representing stronger religious faith. A number of studies have confirmed a one-factor model of the SCSRFPQ (e.g., Freiheit et al., 2006; Plante et al., 2002), so a total score appears to be most appropriate to represent an individual’s religiosity. The SCSRFPQ has good internal consistency reliability ($\alpha = .95$; Plante & Boccaccini, 1997a). Three-week test-retest reliability was excellent ($r = .93$; Sherman et al., 2001). The SCSRFPQ displayed good convergent validity with other measures of religiosity, such as the revised Spiritual Experience Index (Genia, 1997; $r = .76$; Freiheit et al., 2006) and the Age Universal Religious Orientation Scale (Gorsuch & Venable, 1983) Intrinsic Religiousness subscale ($rs = .87-.90$) and Extrinsic Religiousness subscale ($rs = .64-.73$; Plante & Boccaccini, 1997b). It also has shown discriminant validity with depression, as measured by the Symptom Check List-90-Revised (Derogatis, 1977; $r = -.20$; Plante & Boccaccini, 1997a). It was unrelated to a measure of negative religious coping, the Brief Religious Coping Scale (Pargament et al., 1998; $r = .00$, ns; Freiheit et al., 2006), and to a measure of social desirability, the Marlowe Crowne Social Desirability Scale (Crowne & Marlowe, 1960; $rs = -.02-.09$, ns; Plante et al., 1999). The SCSRFPQ was used in the current study to examine convergent validity with the PIOS/PIOS-R.
Depression, Anxiety, and Stress Scale - 21

The DASS-21 (Henry & Crawford, 2005) is an abbreviated 21-item version of the original 42-item DASS (Lovibond & Lovibond, 1995). Individuals are instructed to respond to each item based on how much it applied to them during the past week. Response options range from 0 (did not apply to me at all) to 3 (applied to me very much or most of the time). The measure includes three subscales, Depression, Anxiety, and Stress, each of which include 7 items. Only the Depression and Anxiety subscales were used in the present study. In lieu of the Stress subscale, the PANAS-NA was used to measure negative affect (see following section). Scores on each subscale are calculated by summing the item responses for each subscale. Although the DASS-21 scores originally were calculated by summing items on each subscale and doubling the resulting scores to compare to the full DASS norms, norms have since been established for the DASS-21 subscales without doubling the scores (Henry & Crawford, 2005). Importantly, just like the other measures used in this study, the DASS-21 subscales measure depression and anxiety on a dimensional scale. Factor analyses of the DASS-21 have found either three factors, corresponding with the three subscales (Antony et al., 1998; Gloster et al., 2008) or four factors, with one general factor and three specific factors that correspond to the subscales (e.g., Henry & Crawford, 2005; Szabó, 2010). The DASS-21 subscales have shown good internal consistency reliability (α = .94 and .87 for Depression and Anxiety, respectively; Antony et al., 1998).

Convergent validity has been established for the two subscales being implemented in this study. The DASS-21 Depression subscale (DASS-21-D) has shown convergent validity with the BDI (r = .79), and the Anxiety subscale (DASS-21-A) has good convergent validity with the
BAI ($r = .85$; Antony et al., 1998). Discriminant validity among the two subscales was established as well. In the Antony et al. (1998) study, it was found that DASS-21-D correlated most highly with the BDI, and this correlation was significantly higher than correlations with the BAI ($r = .51; z = 5.52, p < .001$) and the STAI-T ($r = .71; z = 2.32, p < .05$). The DASS-21-A correlated most strongly with the BAI, and this correlation was significantly higher than correlations with the BDI ($r = .62; z = 6.56, p < .001$) and the STAI-T (which was conceptualized to be more strongly related to stress; $r = .55; z = 7.31, p < .001$). Known-groups validity for the DASS-21 subscales also has been examined by Antony et al. (1998). On the DASS-21-D, individuals diagnosed with major depressive disorder scored higher than those with anxiety disorders, and the nonclinical group scored lower than the anxiety disorders and depression groups ($F = 42.30, p < .001, \eta^2 = .41$). On the DASS-21-A, individuals diagnosed with panic disorder scored highest, followed by those diagnosed with social anxiety disorder and depression, followed by those with OCD and specific phobias, with the nonclinical group scoring lowest ($F = 24.86, p < .001, \eta^2 = .29$). This indicates that the Anxiety subscale may best measure physiological arousal. These scores suggest adequate convergent and discriminant validity of the Depression and Anxiety subscales.

The DASS-21-D and DASS-21-A were used to establish discriminant validity with the PIOS/PIOS-R in the current study. Correlations between the PIOS and the two DASS-21 subscales were examined in a previous study by Huppert and Fradkin (2016) in a sample of individuals with OCD or anxiety disorders. There were significant, medium-sized correlations between the PIOS total score and DASS-21-D ($r = .26$) as well as DASS-21-A ($r = .26$). However, after Bonferroni corrections, neither of the correlations remained significant. The
current study examined discriminant validity using the DASS-21 subscales for both the PIOS and the PIOS-R. Additionally, these correlations were examined in a group of nonclinical, religious participants.

Positive and Negative Affect Schedule – Negative Affect

The PANAS (Watson et al., 1988) is a 20-item self-report questionnaire measuring positive and negative affect in a dimensional manner. Individuals are instructed to respond to each item, consisting of affect terms, based on the extent to which they generally experience each affect. Response options range from 1 (very slightly or not at all) to 5 (extremely). The measure includes two subscales PANAS-PA and PANAS-NA, each of which includes ten items. Only the PANAS-NA was included in the current study. This is because PANAS-NA was used in lieu of the DASS-21 Stress subscale (DASS-21-S) to measure negative affect. Although there have been reasonable correlations between the DASS-21-S and PANAS-NA (e.g., $r = .64$; Henry & Crawford, 2005), the correlation is not high enough to permit strong confidence that they measure the same construct. Generally, the PANAS-NA is a more widely accepted measure of negative affect.

Factor analyses have validated a two-factor solution that corresponds with the two subscales (Crawford & Henry, 2004; Watson et al., 1988). Internal consistency of the PANAS-NA using the “general” time instructions is good ($\alpha = .87$) and 8-week test-retest reliability is acceptable ($r = .71$; Watson et al., 1988). In the development study by Watson et al. (1988), convergent validity was initially established through strong correlations with measures of related constructs, such as the BDI ($r = .58$), the STAI-S ($r = .51$), and the Hopkins Symptom Checklist.
(Derogatis et al., 1974; \( r = .74 \)), a measure of general distress and dysfunction. Other studies have offered more strong evidence for convergent validity. Crawford and Henry (2004) found that PANAS-NA correlated strongly with measures of depression, such as DASS-21-D \( (r = .60) \), and the Depression subscale of the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983; \( r = .44 \)), and measures of anxiety, such as DASS-21-A \( (r = .60) \) and the Anxiety subscale of the HADS \( (r = .65) \). PANAS-NA has also been negatively correlated with self-esteem \( (r = -.43) \) and was uncorrelated with locus of control \( (r = .22, ns) \) and with social desirability \( (r = -.03, ns; \) Huebner & Dew, 1995), providing evidence for discriminant validity. One study showed convergence with peer ratings by examining self-peer correlations among the PANAS-NA items and found significant correlations for all items \( (r = .19-.41) \) and for the full scale \( (r = .36; \) Watson & Clark, 1991). Furthermore, self-rated PANAS-NA was uncorrelated with peer-rated PANAS-PA \( (r = -.05) \) and vice versa \( (r = -.13) \). Thus, the PANAS-NA shows adequate internal consistency, retest reliability, convergent validity, and discriminant validity.

Procedure

Selection

Participants were recruited using Amazon Mechanical Turk (MTurk). Individuals were screened prior to participation for four criteria. All participants were required to be from the United States. This criterion was meant to increase the probability that participants have a sufficient English language reading level so as to read the questionnaires and other study materials, which were provided only in English. According to the Flesch-Kincaid readability test, the reading grade level of the questionnaires range from grade 4.2 to grade 5.8, indicating that
they are generally easy to read – especially given the high level of education in most of the current sample. All included participants indicated identification as either Christian, Jewish, or Muslim, since these are the religions that were examined in this study. These religious groups were selected because they are commonly studied in the scrupulosity literature and because these are the three largest religious groups, not including atheists/agnostics, in the United States (Pew Research Center, 2015). Although Protestants and Catholics often are separated in the literature, findings for the two groups on PIOS scores were statistically equivalent in five out of six studies, and the singular statistical difference in one study (i.e., Abramowitz et al., 2002) was only on one of the two subscales (see Religion and Scrupulosity section). Thus, a strong majority of the data suggests that Protestants and Catholics, as broad groups, do not have different levels of scrupulosity symptoms and thus were combined in the current study. Given the lower rates of Jewish and Muslim MTurk workers compared to Christian workers (Burnham et al., 2018), recruitment was stratified by time. Specifically, 50 assignments per religious group were open until 50 responses from each group were made. Once the 50 assignments were completed, 50 more assignments per group were opened. This continued until the sample reached 250 initial participants per group. In the cases that workers identified one religion in the prescreen and listed a completely different religious affiliation in the full survey (e.g., “Muslim” is chosen in the prescreen, but “Catholic” is written in the full survey), those participants were excluded from analyses. Additionally, only MTurk workers with a 95% approval rating were allowed to complete the study. This criterion was in place to ensure quality of data. Finally, participants were required to successfully complete a CAPTCHA, which was in place for the purpose of increasing the probability that humans (i.e., not computer programs) responded to the study.
Tasks

MTurk workers first were presented with an electronic informed consent document. Participants indicated their informed consent via keystroke before proceeding to the initial screening questionnaire, which included a CAPTCHA and a question about religious affiliation. Workers were paid $0.01 for completing the screening question and were informed that it was a screener and that those who pass the screener would be invited to participate in the longer survey for $0.70. Participants who pass the screener were given permission to complete the longer survey, in which they were asked to complete another CAPTCHA, demographic questions, and five questionnaires. The order of presenting the five questionnaires was counterbalanced across participants. Once the questionnaires were completed, participants were provided with an electronic debriefing form and thanked for their participation. Participants who completed the longer survey were paid $0.70 for their time.
CHAPTER 3
DATA ANALYSIS PLAN

Confirmatory Factor Analyses

Mplus (Version 8.8) was employed to conduct all CFAs and measurement invariance analyses. Since the PIOS data are ordinal (i.e., each item is on a 5-point Likert-type scale), a WLSMV estimator and polychoric matrix were used for all CFAs. Per Kline (2016), delta parameterization was used, and the metric of the latent variable(s) was defined using the marker indicator approach, with the first indicator on each factor as the marker. The raw data file was used as input. Since no one model has consistently borne out in the literature as the best-fitting model for the PIOS, and the literature on the PIOS-R suggests two competing models, preliminary CFAs were conducted to determine the best-fitting model for the current overall sample. Then the model was examined using separate CFAs in each religious group.

Examination of model fit was based on approximate fit indices, including RMSEA, CFI, TLI, and SRMR; localized areas of strain; and the interpretability, size, and statistical significance of parameter estimates. Model $\chi^2$ was calculated, but due to its sensitivity to sample size, it will not be used as a primary indicator of model fit (Brown, 2015). Brown (2015) and Kline (2016) argue that the strict use of fit index thresholds is inappropriate, but general guidelines such as those recommended by Hu and Bentler (1999) can aid in the interpretation of the indices. RMSEA values less than .05 are considered to indicate good fit, values between .05-
.08 indicate reasonable fit, values between .08-.10 indicate mediocre fit, and values above .10 indicate poor fit (Browne & Cudeck, 1993). CFI and TLI values less than or equal to .90 are considered to indicate poor fit, values greater than .90 indicate adequate fit (Bentler, 1990), and values greater than .95 indicate good fit (Hu & Bentler, 1999). SRMR values less than .08 are considered to indicate acceptable fit (Hu & Bentler, 1999).

Localized areas of strain were evaluated using modification indices. Modification indices provide information on how much the model $X^2$ may decrease if a parameter is respecified to be freely estimated. A modification index can be considered to indicate that the model can be significantly improved if the value is 4.00 or greater (Brown, 2015). This information was used to determine whether or not there are areas of localized strain in the models being examined. Modification indices were not used on their own to make changes in the models being tested in this study.

Interpretability, size, and statistical significance of the parameter estimates were examined for each CFA model tested. The parameter estimates should make statistical and substantive sense. In other words, the estimates were reviewed to ensure that there were no Heywood cases and that the directions of the parameter estimates reflected predictions based on substantive reasoning. Statistically significant parameter estimates reflect a better-fitting model and provide reasoning for each item to load on the factors. The researcher also reviewed the standard errors of the parameter estimates to check for outlying values or values approaching zero. The parameter estimates’ magnitudes also were reviewed to ensure that they were substantively meaningful. Specifically, factor loadings equal to or greater than .30 are considered salient (Brown, 2015).
Along with the aforementioned aspects of fit, model comparisons were based on the DIFFTEST function in Mplus and information criterions, including AIC and BIC. DIFFTEST is somewhat less sensitive to sample size compared to the commonly used chi-square difference test, and a significant DIFFTEST suggests that the more restrictive model has poorer fit (Muthén & Muthén, 1998-2017). Information criterion indices were calculated. The AIC and BIC statistics both take into account model fit and complexity. Lower AIC and BIC values indicate a better-fitting model in relation to competing nested or non-nested models (Brown, 2015). These two indices were computed using robust maximum likelihood estimation, since they are unable to be calculated using WLSMV in Mplus.

CFA Models Tested With Full Sample

PIOS

For the 19-item PIOS, three models from the literature were tested in the full sample: a simple one-factor model (Figure 1), the two-factor model established by Abramowitz et al. (2002; Figure 2), and the bifactor model established by Huppert and Fradkin (2016; Figure 3). These models have each been supported in the literature. The use of the PIOS in research on scrupulosity (i.e., whether it is scored using a total score, subscales, or both) is inconsistent and may correspond with any of the three models (e.g., Bruce et al., 2011; Buchholz et al., 2019; Gonsalvez et al., 2010), suggesting some substantive support for each.

One-Factor Model. The most parsimonious model to be tested was a simple one-factor model in which all indicators (P1-P19) load onto one factor. The one-factor model corresponds with the developers’ conceptualization that the total score of the questionnaire can be used to
Figure 1: The one-factor model of the PIOS.

Note. This figure illustrates the one-factor model of the full-length PIOS. The individual PIOS items are labeled P1 through P19.
Figure 2: The two-factor model of the PIOS.

Note. This figure illustrates the two-factor model of the full-length PIOS, with Fear of Sin (FOS) and Fear of God (FOG) as the latent factors. The individual PIOS items are labeled P1 through P19.
Figure 3: The Bifactor Model of the PIOS.

Note. This figure illustrates the bifactor model of the full-length PIOS, with scrupulosity as the general latent factor and Fear of Immorality (FOI) and Fear of God-Revised (FOG-R) as the specific latent factors. The individual PIOS items are labeled P1 through P19.
measure scrupulosity broadly (Abramowitz et al., 2002). For this model, the marker indicator was P1. The factor loading of the marker indicator was fixed to 1.0, thus the freely estimated parameters included 18 factor loadings and 1 factor variance. Since there are 5 response options for each indicator, there are 4 thresholds per item, with a total of 76 freely estimated thresholds. Error variances are not freely estimated when using delta parameterization in WLSMV.

**Two-Factor Model.** The two-factor model of the PIOS is the original model posited by the developers of the measure (Abramowitz et al., 2002). In this model, factor one, referred to as “FOS,” includes the indicators P1, P3, P4, P6-8, P10-12, P14, P16, and P18. The second factor, referred to as “FOG,” includes the indicators P2, P5, P9, P13, P15, P17, and P19. This model reflects proposed substantive differences between items on the separate factors. For the two-factor model, the marker indicator for FOS was P1 and the marker for FOG was P2. Both of the factor loadings for these indicators were fixed to 1.0, and thus the free parameters included 17 factor loadings, 2 factor variances, and 1 factor correlation. As in the one-factor model, there were 76 freely estimated thresholds, and error variances were not freely estimated.

**Bifactor model.** The bifactor model that was first espoused by Huppert and Fradkin (2016) was tested. In this model, a general factor was specified as well as two specific factors that together include a subset of the total indicators. Specific factor one, labelled “FOI,” includes indicators P3, P8, P11, P14, P16, and P18. Specific factor two, labelled “FOG-R,” includes indicators P9, P13, P17, and P19. While the bifactor model is the least parsimonious, it may more accurately reflect a conceptual understanding of scrupulosity as a broad concept with different facets (i.e., fear of God and sin/immorality). Although a few studies have scored the PIOS using both its total score and subscale scores together in a single study (e.g., Gonsalvez et
al., 2010), the bifactor model has not been widely used to guide how the PIOS is scored in the literature, especially in terms of which items constitute each subscale. The subscales that are used in the literature correspond to the Abramowitz et al. (2002) subscales, not the Huppert and Fradkin (2016) subscales. Generally, this newer model is being considered more in the literature (e.g., Pirutinsky & Rosmarin, 2018) and is important to test along with the others. Originally, the marker indicator approach was used to define the metric of the latent variables, the same as with the other models tested. However, this needed to be changed (see Results) so that the variances were fixed to 1. The free parameters included 29 factor loadings, 0 factor variances, and 0 factor correlations, as well as the 76 thresholds.

**PIOS-R**

For the PIOS-R, a simple unidimensional model (Figure 4) and the two-factor model from Olatunji et al. (2007; Figure 5) were tested and compared. The bifactor model tested by Phillips and Fisak (2022) displayed poor fit, but to my knowledge, has not been tested for the PIOS-R by other researchers. Thus, the bifactor model (Figure 6) was tested in this study to confirm or challenge their results. Recall that the PIOS-R includes all PIOS items except P2, P6, P10, and P15. For clarity, PIOS-R indicators were labelled the same as the PIOS indicators. Both models were tested in the same way as the PIOS models: using a CFA with a WLSMV estimator and delta parameterization.

**One-Factor Model.** In the one-factor model, all indicators (P1-P19) load onto a single factor. This model represents one substantive factor for scrupulosity. Researchers have used the PIOS-R by summing or averaging all items into one score (e.g., Fergus & Rowatt, 2014a;
Figure 4: The one-factor model of the PIOS-R.

*Note.* This figure illustrates the one-factor model of the PIOS-R, with scrupulosity as the latent factor. The individual PIOS items are labeled P1 through P19, although P2, P6, P10, and P15 are not included in this revised version of the PIOS.
Figure 5: The two-factor model of the PIOS-R.

Note. This figure illustrates the two-factor model of the PIOS-R, with Fear of Sin (FOS) and Fear of God (FOG) as the latent factors. The individual PIOS items are labeled P1 through P19, excluding P2, P6, P10, and P15 from the original PIOS.
Figure 6: The bifactor model of the PIOS-R.

*Note.* This figure illustrates the bifactor model of the PIOS-R, with scrupulosity as the general latent factor and Fear of Immorality (FOI) and Fear of God-Revised (FOG-R) as the specific latent factors. The individual PIOS items are labeled P1 through P19 excluding P2, P6, P10, and P15 from the original PIOS.
Mauzay & Cuttler, 2018). The marker indicator in this model was P1. The freely estimated parameters included 14 factor loadings, 1 factor variance, and 60 thresholds. As stated previously, error variances are not freely estimated when using delta parameterization in WLSMV.

**Two-Factor Model.** The two-factor model was the one established by Olatunji et al. (2007). In this model, factor one, labelled “FOS,” includes indicators P1, P3, P4, P7, P8, P11, P12, P14, P16, and P18. Factor two, labelled “FOG,” includes indicators P5, P9, P13, P17, and P19. Researchers have used this model to define PIOS-R subscales (e.g., Fergus, 2014; Stewart et al., 2020). The marker was P1, which was fixed to 1.0. Freely estimated parameters included 13 factor loadings, 2 factor variances, 1 factor correlation, and 60 thresholds.

**Bifactor Model.** The bifactor model was the one established by Huppert and Fradkin (2016), excluding the items not included in the PIOS-R. A general factor representing scrupulosity was specified as well as two specific factors that together included a subset of the total indicators. Specific factor one, labeled “FOI,” includes indicators P3, P8, P11, P14, P16, and P18. Specific factor two, labeled “FOG-R,” includes indicators P9, P13, P17, and P19. Note that the marker approach to defining the metric of the latent variables also needed to be adjusted once analyses were run (see Results). The metric of the latent variable instead was defined by fixing the variances to 1. The freely estimated parameters included 25 factor loadings, 0 factor variances, 0 factor correlations, and 60 thresholds.
Measurement Invariance Analysis

Measurement invariance analyses were conducted using multiple-groups CFA following the order recommended by Brown (2015): testing configural, metric, and then scalar invariance. Again, the WLSMV estimator and polychoric matrix were used, and delta parameterization was specified. Model fit used the same criteria used for the CFAs. However, model comparison involved examining the DIFFTEST function in Mplus, examination of the overlap in RMSEA 90% confidence intervals, and $\Delta$CFI (Putnick & Bornstein, 2016; Reis & Judd, 2000). Notably, chi-square difference testing is strongly influenced by sample size, even when computed using DIFFTEST in Mplus. Specifically, when sample sizes are over 200, any differences based on DIFFTEST are likely to be trivial (Meade et al., 2008). Thus, a CFI reduction equal to or less than .002 was considered to represent a nonsignificant change (Meade et al., 2008).

The first step was to examine configural invariance, or equality of model form. In this step, the same CFA model was specified in each group such that the number of factors and the pattern of factor loadings were the same, but all parameters were freely estimated in each group. This tested whether the items measure the same constructs, FOS and FOG. In other words, the general factor structure is equivalent in each group. In Mplus, the configural invariance test fixes scale factors to 1.0 in all groups for identification purposes (i.e., so that all thresholds and loadings can be freely estimated in all groups). The factor means were fixed to 0 in all groups.

Next, the analysis proceeded to the next step, testing for metric invariance (i.e., equality of factor loadings). In this step, factor loadings were constrained to be equal across groups (excluding the loading for the marker indicator(s), which was fixed to 1.0), and fit was compared to the configural model. In this way, the equality of factor loadings across groups was tested.
This tests whether each item contributes to the latent constructs to a similar degree in each group. In other words, metric invariance provides evidence that no one item is more strongly related to FOS or FOG, respectively, in one group compared to the others. All thresholds were not yet constrained to equality, but certain thresholds had to be constrained for identification purposes (Bowen & Masa, 2015). Specifically, the first two thresholds of the marker indicators were constrained to be equal across groups, as well as the first indicator threshold of each other indicator on the factors. Additionally, the mean of the first factor was fixed to 0.

The third step involved testing for equality of thresholds, or scalar invariance. In this step, all factor loadings (again excluding the marker indicators) and thresholds were constrained to equality. The resulting model was examined and compared to the metric model. This step tests whether the three groups use the response scale in the same way. In other words, scalar invariance provides evidence that a person from each group with the same level of FOS or FOG should have the same score on the indicator. Factor means remained fixed to 0 in one group and freed in the other groups. Additionally, the scale factors were fixed to 1.0 in one group, and freely estimated in the others.
CHAPTER 4

RESULTS

Data Screening

Missing Data

The researcher examined the data for missing values. The data were tested using Little’s MCAR, which was found to be nonsignificant ($\chi^2[71] = 50.34, p = .970$), indicating that the data were MCAR. Thus, the data were left as is because analyses using WLSMV estimation in Mplus (Version 8.8) deal with the missing data automatically using a pairwise present approach (Muthén & Muthén, 1998-2017).

Outliers

To check for univariate outliers among total scores on the five measures, IBM SPSS Statistics (Version 26) was used to review box plots and histograms of the data. Specifically, values were identified as outliers if they fell more than $3\times$(interquartile range) below the first quartile or above the third quartile of the box plot (i.e., those values identified by SPSS as “extreme” outliers). No outliers were found in the data for the PIOS, DOCS, DASS-21, or PANAS-NA. Five outliers were identified for the SCSRFQ data and involved data from one participant in the Christian group and four in the Muslim group. The researcher first checked the data for errors (e.g., a typing error) and found no identifiable errors. Since the sources of the
outliers were not clear, the values were winsorized to three standard deviations from the group means (i.e., to 14 for the Christian group, and 15 for the Muslim group).

**Distribution**

Total scores on all questionnaires (including the PIOS/PIOS-R) were examined for normality by examining histograms and Q-Q plots, calculating the skew and kurtosis indices, and conducting a Shapiro-Wilk test. Kline (2016) asserted that, for skewness, an absolute value above 3.0 is cause for concern; for kurtosis, an absolute value above 10.0 indicates a problem and a value above 20.0 indicates a serious problem. The Shapiro-Wilk test was significant for all measures for all groups. However, this test is sensitive to large sample sizes. The skew values for each measure for each group were all below 3 (0.01 – 1.32), and the kurtosis values were all well below 10 (0.06 – 2.32). It was determined that no serious violations of normality were found.

**Descriptives**

Means and standard deviations for each questionnaire and for each PIOS item were calculated both for the full sample and separately for each religious group (see Table 1).

**Confirmatory Factor Analyses**

**CFA With Full Sample**

**PIOS**

The aforementioned one-factor, two-factor, and bifactor models for the PIOS in the
Table 1

Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total Sample</th>
<th>Christian</th>
<th>Muslim</th>
<th>Jewish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Ns = 713-718)</td>
<td>(Ns = 272-274)</td>
<td>(Ns = 240-243)</td>
<td>(Ns = 299-201)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>PIOS</td>
<td>31.30</td>
<td>19.31</td>
<td>61</td>
<td>32.29</td>
</tr>
<tr>
<td>PIOS-R</td>
<td>24.67</td>
<td>15.36</td>
<td>58</td>
<td>25.55</td>
</tr>
<tr>
<td>DOCS ††</td>
<td>27.81</td>
<td>16.44</td>
<td>66</td>
<td>27.95</td>
</tr>
<tr>
<td>DOCS-C</td>
<td>7.36</td>
<td>4.40</td>
<td>68</td>
<td>7.29</td>
</tr>
<tr>
<td>DOCS-UT</td>
<td>6.88</td>
<td>4.60</td>
<td>64</td>
<td>6.96</td>
</tr>
<tr>
<td>DOCS-H</td>
<td>7.01</td>
<td>4.57</td>
<td>62</td>
<td>7.06</td>
</tr>
<tr>
<td>DOCS-JR</td>
<td>6.57</td>
<td>4.79</td>
<td>61</td>
<td>6.64</td>
</tr>
<tr>
<td>DOCS-SR</td>
<td>6.04</td>
<td>4.91</td>
<td>63</td>
<td>6.34</td>
</tr>
<tr>
<td>DASS-21-D</td>
<td>8.44</td>
<td>6.26</td>
<td>64</td>
<td>8.08</td>
</tr>
<tr>
<td>DASS-21-A</td>
<td>7.56</td>
<td>5.95</td>
<td>69</td>
<td>7.55</td>
</tr>
<tr>
<td>PANAS-NA</td>
<td>24.22</td>
<td>10.05</td>
<td>60</td>
<td>23.94</td>
</tr>
<tr>
<td>SCSRFSQ</td>
<td>30.56</td>
<td>6.81</td>
<td>55</td>
<td>31.30</td>
</tr>
</tbody>
</table>

*Note.* Means with a common superscript are statistically equivalent (Tukey’s HSD p < .05).

PIOS = Penn Inventory of Scrupulosity; PIOS-R = Penn Inventory of Scrupulosity – Revised; DOCS = Dimensional Obsessive-Compulsive Scale; DOCS-C = DOCS Contamination subscale; DOCS-H = DOCS Harm subscale; DOCS-UT = DOCS Unacceptable Thoughts subscale; DOCS-JR = DOCS Just Right subscale; DOCS-SR = DOCS Scrupulous and Religious Thoughts subscale; DASS-21 = Depression, Anxiety, and Stress Scale – 21; PANAS-NA = Positive and Negative Affect Schedule, Negative Affect; SCSRFSQ = Santa Clara Strength of Religious Faith Questionnaire.

† T scores calculated using published norms from PIOS to SCSRFSQ, in order, were taken from: Abramowitz et al. (2002), Olatunji et al. (2007), Abramowitz et al. (2010; student group data), Henry and Crawford (2005), Watson et al. (1988), and Plante and Boccaccini (1997a).

†† Total DOCS scores were computed using only the original Abramowitz et al. (2010) subscales.
full sample were tested and fit was compared (see Table 2). See the Confirmatory Factor Analyses in the Data Analysis Plan section for the criteria used to determine goodness of fit and to compare models. These CFAs were conducted to examine Hypothesis 2 (see Current Study section of the Introduction).

One-Factor Model. As was expected due to the large sample size, the chi-square test was significant, $\chi^2(152) = 1,020.99, p < .001$. The RMSEA value indicated mediocre model fit in terms of absolute fit. The CFI and TLI values both indicated good fit in comparison to the worst possible (i.e., null) model. The SRMR value also indicated a well-fitting model. The modification indices all were below 4.0, indicating no evidence for localized areas of strain. Results revealed no Heywood cases. The factor loadings all were statistically significant and positive, in line with substantive reasoning. The size of factor loadings ranged from .76 to .88, indicating that all loadings were salient.

Overall, the one-factor model showed acceptable fit, although the RMSEA value indicated room for improvement. The RMSEA represents absolute fit and is more stringent than incremental fit indices. However, the SRMR also is an absolute fit index and its value for this model, paired with no large residuals, indicated a well-fitting model. This may be related to the parsimony correction in RMSEA, which penalizes models with many variables, such as this one (Brown, 2015).

Two-Factor Model. The two-factor model of the PIOS was examined. The chi-square test also was significant, as expected, $\chi^2(151) = 873.04, p < .001$. The RMSEA value was modestly better ($\Delta$RMSEA = .007) than the one-factor model but technically remained in the “mediocre”
Table 2

Goodness-of-Fit Statistics of the Penn Inventory of Scrupulosity in Full Sample (N = 718)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$\Delta \chi^2$</th>
<th>RMSEA</th>
<th>LL</th>
<th>UL</th>
<th>CFI</th>
<th>TLI</th>
<th>SRMR</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PIOS</strong></td>
<td>One-Factor</td>
<td>1,020.99*</td>
<td>152</td>
<td>-</td>
<td>0.089</td>
<td>0.084</td>
<td>0.094</td>
<td>0.979</td>
<td>0.977</td>
<td>0.032</td>
<td>33,820.38</td>
<td>34,080.84</td>
</tr>
<tr>
<td></td>
<td>Two-Factor</td>
<td>873.04*</td>
<td>151</td>
<td>76.25*</td>
<td>0.082</td>
<td>0.076</td>
<td>0.087</td>
<td>0.983</td>
<td>0.980</td>
<td>0.030</td>
<td>33,647.08</td>
<td>33,912.11</td>
</tr>
<tr>
<td></td>
<td>Modified Two-Factor</td>
<td>673.22*</td>
<td>151</td>
<td>101.19*</td>
<td>0.069</td>
<td>0.064</td>
<td>0.075</td>
<td>0.987</td>
<td>0.986</td>
<td>0.025</td>
<td>33,487.63</td>
<td>33,752.66</td>
</tr>
<tr>
<td></td>
<td>Bifactor</td>
<td>812.27*</td>
<td>142</td>
<td>251.82*</td>
<td>0.081</td>
<td>0.076</td>
<td>0.087</td>
<td>0.984</td>
<td>0.981</td>
<td>0.028</td>
<td>33,562.14</td>
<td>33,868.29</td>
</tr>
<tr>
<td><strong>PIOS-R</strong></td>
<td>One-Factor</td>
<td>705.75*</td>
<td>90</td>
<td>-</td>
<td>0.098</td>
<td>0.091</td>
<td>0.104</td>
<td>0.982</td>
<td>0.979</td>
<td>0.029</td>
<td>26,677.87</td>
<td>26,883.50</td>
</tr>
<tr>
<td></td>
<td>Two-Factor</td>
<td>356.26*</td>
<td>89</td>
<td>101.99*</td>
<td>0.065</td>
<td>0.058</td>
<td>0.072</td>
<td>0.992</td>
<td>0.991</td>
<td>0.019</td>
<td>26,402.27</td>
<td>26,612.46</td>
</tr>
<tr>
<td></td>
<td>Bifactor</td>
<td>433.97*</td>
<td>80</td>
<td>249.99*</td>
<td>0.079</td>
<td>0.071</td>
<td>0.086</td>
<td>0.990</td>
<td>0.986</td>
<td>0.022</td>
<td>26,426.14</td>
<td>26,677.46</td>
</tr>
</tbody>
</table>

Note. Models computed using mean- and variance-adjusted weighted least squares (WLSMV) estimation. 
$\Delta \chi^2$ computed using Mplus 8.8 DIFFTEST function. 
PIOS = Penn Inventory of Scrupulosity; PIOS-R = Penn Inventory of Scrupulosity – Revised. 
$^a = \Delta \chi^2$ comparing each model with the one-factor model; $^b$ = fit statistic computed using robust maximum likelihood estimation. 
*p < .001.
range. The CFI and TLI values both indicated good model fit and the SRMR indicated a well-fitting model. Unlike the one-factor model, several modification indices were above 4.00, indicating potential localized areas of strain. Specifically, the modification indices suggested that respecifying five indicators (P2, P9, P13, P17, and P18) to load onto FOS and five indicators (P1, P6, P8, P11, and P12) to load onto FOG would improve model fit. Notably, P2 on FOS had a modification index of 257.93, suggesting that model fit would greatly improve if P2 were allowed to load onto FOS. The researcher ran the model again, respecifying P2 to load onto FOS and not on FOG. This resulted in improved fit. This same change was made by Olatunji et al. (2007), who also found that respecifying item 2 on the PIOS to load onto FOS improved model fit. Item 2 is “I fear that I might be an evil person.” Substantively, it makes sense that this item should load onto a factor representing fear of sin rather than a factor representing fear of God. The item appears to be concerned with being an evil person (i.e., a person who has sinned) and not necessarily fearing God or punishment from God. Although the goal of this study was not to find a new model or to modify existing models for the PIOS, the researcher decided to retain this modified model for use in the remaining analyses. This reasoning was based on (a) previous research by Olatunji et al. (2007) who found the same respecification improved the two-factor model in their sample, suggesting that the issue is not merely an artifact of the current sample; (b) the strong substantive basis for the respecification, including the semantic relationship between this item and the other FOS items; and (c) the degree to which the model improved by this one change. For reference, Table 2 includes model fit statistics for both the original and modified two-factor models.
The modified two-factor model revealed a significant chi-square, $\chi^2(151) = 673.22, p < .001$. The RMSEA value noticeably improved from the one-factor ($\Delta$RMSEA = .020) and original two-factor ($\Delta$RMSEA = .013) models and represented reasonable fit. The CFI and TLI values remained in the “good” range of fit. The SRMR value also improved by .005 from the original two-factor model and indicated a well-fitting model, given that there were no large residuals. There were 9 modification indices above 4.0, indicating some evidence for localized areas of strain. Specifically, the modification indices suggested that respecifying one indicator (P5) onto FOS and eight indicators (P1, P2, P3, P6, P8, P11, P16, and P18) onto FOG would improve model fit. Results revealed no Heywood cases. The factor loadings all were statistically significant and positive, in line with substantive reasoning. The size of factor loadings ranged from .77 to .91, indicating that all loadings were salient. Overall, the modified two-factor model appeared to be a well-fitting model, with fit indices reflecting reasonable to good fit.

Bifactor model. The bifactor model from Huppert and Fradkin (2016) was tested. When run using Mplus, the model would not converge. Output revealed that the maximum number of iterations had been reached, so the researcher increased the number of iterations to 100,000 and still there was no convergence and the number of iterations exceeded. Next, the researcher ran the model without the general factor (i.e., a two-factor model using the specific factors only from Huppert & Fradkin [2016]). This did converge and revealed acceptable fit across fit indices. The chi-square test was significant ($\chi^2[34] = 112.52, p < .001$), but the RMSEA value was in the “reasonable” range (RMSEA = .057); the CFI and TLI values indicated good fit (CFI = .997; TLI = .996), and the SRMR indicated a well-fitting model (SRMR = .013). Thus, it was unlikely that nonconvergence was simply due to misfit because the marker variable loadings were fixed to
1 for identification purposes (i.e., which could have been vastly different from the true loading). Upon further examination, the factor estimates for FOI were not able to be estimated by Mplus, although the factor loadings for the other factors were able to be estimated. To examine whether there was misfit in the model due to use of the marker indicator approach (e.g., that the first factor loading of FOI is not estimated close to one), the model was respecified such that the first factor loading of each factor was freed and the factor variances were fixed to 1. This resolved the convergence problem.

Results indicated that the bifactor model showed acceptable fit, with global indices suggesting mediocre to good fit. The chi-square test was significant ($\chi^2[142] = 812.27, p < .001$), and the RMSEA suggested mediocre to reasonable fit (it was ~.001 from the “reasonable” range). The CFI and TLI values both indicated good model fit, and the SRMR indicated a well-fitting model, given there were no large residuals. However, there was evidence for localized areas of strain, such as many modification indices above 4.00. These suggested that respecifying 16 factor loadings and allowing the two specific factors to correlate may improve model fit. There were no Heywood cases, but the loadings of P3 and P8 on FOI were negative, which is not in line with substantive reasoning. Additionally, the loading of P8 on FOI was nonsignificant ($\lambda = -0.016, p = .639$), and all but two factor loadings on the specific factors (i.e., P9 and P13 on FOG-R) were not salient. The size of factor loadings on the general factor ranged from .76 to .86, indicating that all loadings were significant and salient. Overall, whereas global fit indices suggest an acceptable fit, the bifactor model had many areas of localized strain and nonsalient factor loadings, as well as loadings that were not in the expected direction. Thus, the bifactor model cannot be said to be a well-fitting model for the current sample.
Model Comparisons. When comparing the one- and original two-factor models, the DIFFTEST function in Mplus revealed that the two-factor model fit the data better. The RMSEA 90% confidence intervals did overlap slightly, however, suggesting a potentially nonsignificant difference in absolute model fit. The change in CFI was .004, suggesting a significantly better fit. Further, the AIC and BIC values both decreased from the one- to two-factor models. Thus, the two-factor model was retained. Notably, as discussed above, one substantively meaningful respecification of the original two-factor model resulted in significantly better model fit. This can be seen when examining the DIFFTEST results, RMSEA confidence interval, ΔCFI, and comparison of AIC and BIC values. Specifically, the DIFFTEST result was significant, indicating significantly better fit for the modified two-factor model. The RMSEA confidence interval displayed lower values and no overlap with the confidence interval of the original two-factor model. The change in CFI was .004 and was significant, and AIC and BIC values were lower for the modified model than for the original model. Therefore, the modified two-factor model was retained. As for the bifactor model, DIFFTEST results indicated that it fit better than the one-factor model. However, model comparison with the modified two-factor model indicated the bifactor model had worse fit. The RMSEA value was lower for the modified two-factor model, and the confidence intervals did not overlap. The CFI worsened from the bifactor to the modified two-factor model by .003, and the AIC and BIC values decreased from the bifactor to the modified two-factor model. Thus, it was concluded that the modified two-factor model was the best-fitting model in the current sample, although there remained multiple areas of localized strain, and it was retained for the subsequent analyses. This result was not consistent with Hypothesis 2.
PIOS-R

The aforementioned one-factor, two-factor, and bifactor models for the PIOS-R were tested in the full sample and fit was compared between models (see Table 2). These CFAs were conducted to test Hypothesis 2.

**One-Factor Model.** Similar to each of the PIOS models, the chi-square test was significant, $\chi^2(90) = 705.75, p < .001$. The RMSEA value indicated mediocre fit. The CFI and TLI values both indicated good fit, and the SRMR value also indicated a well-fitting model, given that there were no large residuals. The modification indices all were below 4.0, indicating no evidence for localized areas of strain. Results revealed no Heywood cases. The factor loadings all were positive, in line with substantive reasoning, and statistically significant. The size of factor loadings ranged from .76 to .89, indicating that all loadings were salient.

Overall, the one-factor model showed acceptable fit, except for mediocre absolute fit based on the RMSEA value. However, the SRMR, another absolute fit index, indicated a well-fitting model. Thus, the less favorable RMSEA value may have been related to the parsimony correction.

**Two-Factor Model.** The chi-square test again was significant, $\chi^2(89) = 356.26, p < .001$. The RMSEA value indicated reasonable fit. The CFI and TLI values both indicated good fit, and the SRMR value also indicated a well-fitting model, given that there were no large residuals. There were 6 modification indices above 4.0, indicating some evidence for localized areas of strain. Specifically, the modification indices suggested that respecifying several indicators (P1, P3, P8, P11, P16, and P18) onto FOG would improve model fit. Results revealed no Heywood
cases. The factor loadings all were positive, in line with substantive reasoning, and statistically significant. The size of factor loadings ranged from .76 to .92, indicating that all loadings were salient.

**Bifactor Model.** As with the full PIOS, the bifactor model for the PIOS-R would not converge when ran using Mplus. Output revealed that the maximum number of iterations had been reached and increasing the number of iterations did not result in convergence. Next, the researcher ran the model without the general factor (i.e., a two-factor model using the specific factors only). Similar to the PIOS bifactor model, the factor estimates for FOI were not able to be estimated by Mplus. To examine whether there was misfit in the model due to use of the marker indicator approach, the model was respecified such that the first factor loading of each factor was freed and the factor variances were fixed to 1. This resolved the convergence problem.

The chi-square test was significant, $\chi^2(80) = 433.97, p < .001$. The RMSEA value indicated reasonable fit. The CFI and TLI values both indicated good fit. The SRMR value indicated a well-fitting model, given there were no large residuals. There were no Heywood cases. Fit issues were found, however, in examination of localized areas of strain and the factor loadings themselves. Results revealed 18 modification indices with a value above 4.00. These suggested that model fit could be improved by respecifying 15 factor loadings, in allowing the two specific factors to correlate, or in respecifying the two specific factors to load on or regress on the other (in both directions). Additionally, the loadings of P3 and P8 on FOI were negative, which is not in line with substantive reasoning; all loadings were expected to be positive. All factor loadings were statistically significant, but none of the loadings on FOI were salient.
Model Comparisons. The two-factor model was compared to the one-factor model. The DIFFTEST function in Mplus revealed that the two-factor model fit the data better, $p < .001$. The RMSEA 90% confidence intervals had no overlap, with plenty of room between the two separate intervals. The change in CFI was .010, suggesting a significantly better fit. Furthermore, the AIC and BIC values both decreased from the one- to two-factor models. Thus, the two-factor model was retained. The bifactor model was compared to the other two models. DIFFTEST results indicated that it fit better than the one-factor model. However, the bifactor model appeared to have worse fit compared to the two-factor model. The RMSEA value was lower for the two-factor model, and the confidence intervals showed minimal overlap. The CFI increased from the bifactor to the two-factor model by .002, which is the threshold for statistical significance and indicates a nonsignificant difference. The AIC and BIC values decreased from the bifactor to the two-factor model, although only slightly. Given that the bifactor model was more complex and that the two-factor model seemed to fit slightly better, the two-factor model was retained for the subsequent analyses. This result was consistent with Hypothesis 2.

CFAs With Each Group

The modified two-factor model for the PIOS was tested separately in each religious group. Fit statistics can be found in Table 3. Factor loadings can be found in Table 4. The fit indices were not markedly different among the different groups. This is illustrated in the strong RMSEA 90% confidence interval overlap between the Muslim and Jewish groups, and slight overlap between those two groups and the Christian group. Whereas the Christian group had
<table>
<thead>
<tr>
<th>Measure (Model)</th>
<th>Group</th>
<th>$\chi^2$</th>
<th>$df$</th>
<th>RMSEA</th>
<th>LL</th>
<th>UL</th>
<th>CFI</th>
<th>TLI</th>
<th>SRMR</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PIOS (Modified Two-Factor)</strong></td>
<td>Christian ($n = 274$)</td>
<td>278.44*</td>
<td>151</td>
<td>.055</td>
<td>.045</td>
<td>.066</td>
<td>.992</td>
<td>.991</td>
<td>.024</td>
<td>12,679.18</td>
<td>12,888.32</td>
</tr>
<tr>
<td></td>
<td>Muslim ($n = 243$)</td>
<td>342.34*</td>
<td>151</td>
<td>.072</td>
<td>.062</td>
<td>.082</td>
<td>.978</td>
<td>.975</td>
<td>.036</td>
<td>11,669.50</td>
<td>11,871.38</td>
</tr>
<tr>
<td></td>
<td>Jewish ($n = 201$)</td>
<td>323.74*</td>
<td>151</td>
<td>.075</td>
<td>.064</td>
<td>.087</td>
<td>.991</td>
<td>.990</td>
<td>.034</td>
<td>9,014.72</td>
<td>9,206.31</td>
</tr>
<tr>
<td><strong>PIOS-R (Two-Factor)</strong></td>
<td>Christian ($n = 274$)</td>
<td>180.96*</td>
<td>89</td>
<td>.061</td>
<td>.049</td>
<td>.074</td>
<td>.993</td>
<td>.992</td>
<td>.021</td>
<td>9,974.60</td>
<td>10,140.47</td>
</tr>
<tr>
<td></td>
<td>Muslim ($n = 243$)</td>
<td>161.45*</td>
<td>89</td>
<td>.058</td>
<td>.043</td>
<td>.072</td>
<td>.990</td>
<td>.988</td>
<td>.028</td>
<td>9,225.09</td>
<td>9,385.20</td>
</tr>
<tr>
<td></td>
<td>Jewish ($n = 201$)</td>
<td>219.79*</td>
<td>89</td>
<td>.086</td>
<td>.071</td>
<td>.100</td>
<td>.992</td>
<td>.991</td>
<td>.031</td>
<td>7,077.53</td>
<td>7,229.48</td>
</tr>
</tbody>
</table>

*Note.* Models computed using mean- and variance-adjusted weighted least squares (WLSMV) estimation. PIOS = Penn Inventory of Scrupulosity; PIOS-R = Penn Inventory of Scrupulosity – Revised. *a* = fit statistic computed using robust maximum likelihood estimation. *p* < .001.
Table 4

Standardized Factor Loadings for the Penn Inventory of Scrupulosity

<table>
<thead>
<tr>
<th>Item</th>
<th>PIOS Modified Two-Factor</th>
<th></th>
<th></th>
<th>PIOS-R Two-Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Sample</td>
<td>Christian</td>
<td>Muslim</td>
<td>Jewish</td>
</tr>
<tr>
<td>Factor 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>.80</td>
<td>.86</td>
<td>.76</td>
<td>.79</td>
</tr>
<tr>
<td>P2</td>
<td>.82</td>
<td>.84</td>
<td>.83</td>
<td>.82</td>
</tr>
<tr>
<td>P3</td>
<td>.83</td>
<td>.83</td>
<td>.81</td>
<td>.86</td>
</tr>
<tr>
<td>P4</td>
<td>.86</td>
<td>.85</td>
<td>.85</td>
<td>.88</td>
</tr>
<tr>
<td>P6</td>
<td>.80</td>
<td>.79</td>
<td>.73</td>
<td>.85</td>
</tr>
<tr>
<td>P7</td>
<td>.77</td>
<td>.78</td>
<td>.76</td>
<td>.76</td>
</tr>
<tr>
<td>P8</td>
<td>.86</td>
<td>.86</td>
<td>.83</td>
<td>.88</td>
</tr>
<tr>
<td>P10</td>
<td>.82</td>
<td>.84</td>
<td>.81</td>
<td>.82</td>
</tr>
<tr>
<td>P11</td>
<td>.84</td>
<td>.83</td>
<td>.75</td>
<td>.92</td>
</tr>
<tr>
<td>P12</td>
<td>.85</td>
<td>.85</td>
<td>.85</td>
<td>.85</td>
</tr>
<tr>
<td>P14</td>
<td>.85</td>
<td>.83</td>
<td>.80</td>
<td>.91</td>
</tr>
<tr>
<td>P16</td>
<td>.87</td>
<td>.85</td>
<td>.84</td>
<td>.90</td>
</tr>
<tr>
<td>P18</td>
<td>.88</td>
<td>.86</td>
<td>.84</td>
<td>.94</td>
</tr>
<tr>
<td>Factor 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P5</td>
<td>.81</td>
<td>.81</td>
<td>.68</td>
<td>.90</td>
</tr>
<tr>
<td>P9</td>
<td>.85</td>
<td>.84</td>
<td>.80</td>
<td>.91</td>
</tr>
<tr>
<td>P13</td>
<td>.89</td>
<td>.89</td>
<td>.82</td>
<td>.94</td>
</tr>
<tr>
<td>P15</td>
<td>.83</td>
<td>.84</td>
<td>.77</td>
<td>.88</td>
</tr>
<tr>
<td>P17</td>
<td>.88</td>
<td>.89</td>
<td>.82</td>
<td>.93</td>
</tr>
<tr>
<td>P19</td>
<td>.91</td>
<td>.90</td>
<td>.89</td>
<td>.95</td>
</tr>
</tbody>
</table>

*Note.* All factor loadings are statistically significant at $p \leq .001$ (two-tailed).
slightly better fit indices, the overlap in RMSEA confidence intervals indicates that fit generally was similar across all three groups.

For the two-factor model of the PIOS-R, fit across groups was examined. As with the full PIOS, there was overlap in RMSEA 90\% confidence intervals among all three groups. Although the Jewish group had a somewhat lower RMSEA value, and less overlap in confidence interval with the other two groups, the CFA and TLI values were comparable to or better than that of the other groups. Thus, it did not appear that fit was markedly different among the groups. Fit statistics for the PIOS-R two-factor model in each group are included in Table 3. Factor loadings for each group can be found in Table 4.

Measurement Invariance Analysis

**PIOS**

Support for PIOS configural invariance was indicated by acceptable fit across indices (see Table 5). Specifically, the RMSEA indicated reasonable absolute fit of the model, and the CFI and TLI values represented good fit. The SRMR, in conjunction with a lack of large residuals across groups, indicated a well-fitting model. There were a few modification indices above 10, including: factor loadings P1 (14.26) and P11 (10.97) on FOG for the Christian group; factor loading P6 (12.45) on FOG for the Muslim group; and factor loadings P5 (14.00) on FOS as well as P3 (11.37), P6 (20.05), and P14 (21.36) on FOG. Despite some localized areas of strain, overall fit was deemed acceptable, and the solution served as a baseline model for subsequent invariance analyses.
Table 5

Goodness-of-Fit Statistics for Measurement Invariance Analyses

<table>
<thead>
<tr>
<th>Measure (Model)</th>
<th>Invariance Test</th>
<th>χ²</th>
<th>df</th>
<th>Δχ²</th>
<th>RMSEA</th>
<th>90% CI</th>
<th>LL</th>
<th>UL</th>
<th>CFI</th>
<th>TLI</th>
<th>SRMR</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIOS (Modified Two-Factor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural</td>
<td></td>
<td>938.41*</td>
<td>453</td>
<td>-</td>
<td>.067</td>
<td>.061</td>
<td>.073</td>
<td>.989</td>
<td>.988</td>
<td>.031</td>
<td>33,365.33</td>
<td>34,158.43</td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td></td>
<td>1,046.13*</td>
<td>487</td>
<td>139.15* a</td>
<td>.069</td>
<td>.063</td>
<td>.075</td>
<td>.987</td>
<td>.987</td>
<td>.034</td>
<td>33,334.97</td>
<td>33,974.70</td>
<td></td>
</tr>
<tr>
<td>Scalar</td>
<td></td>
<td>1,181.07*</td>
<td>597</td>
<td>179.84* b</td>
<td>.064</td>
<td>.059</td>
<td>.069</td>
<td>.987</td>
<td>.989</td>
<td>.035</td>
<td>33,413.99</td>
<td>33,898.35</td>
<td></td>
</tr>
<tr>
<td>Equal factor variances</td>
<td></td>
<td>1,223.67*</td>
<td>601</td>
<td>26.47* c</td>
<td>.066</td>
<td>.061</td>
<td>.071</td>
<td>.986</td>
<td>.988</td>
<td>.038</td>
<td>33,419.27</td>
<td>33,885.36</td>
<td></td>
</tr>
<tr>
<td>Equal factor covariances</td>
<td></td>
<td>1,192.65*</td>
<td>603</td>
<td>1.65 d</td>
<td>.064</td>
<td>.059</td>
<td>.069</td>
<td>.987</td>
<td>.989</td>
<td>.038</td>
<td>33,417.19</td>
<td>33,874.13</td>
<td></td>
</tr>
<tr>
<td>Equal latent means</td>
<td></td>
<td>1,488.02*</td>
<td>605</td>
<td>40.33* e</td>
<td>.078</td>
<td>.073</td>
<td>.083</td>
<td>.980</td>
<td>.983</td>
<td>.045</td>
<td>33,455.02</td>
<td>33,902.83</td>
<td></td>
</tr>
<tr>
<td>PIOS-R (Two-Factor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural</td>
<td></td>
<td>569.21*</td>
<td>267</td>
<td>-</td>
<td>.069</td>
<td>.061</td>
<td>.077</td>
<td>.992</td>
<td>.990</td>
<td>.027</td>
<td>26,277.22</td>
<td>26,907.81</td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td></td>
<td>648.16*</td>
<td>293</td>
<td>91.65* a</td>
<td>.071</td>
<td>.064</td>
<td>.079</td>
<td>.990</td>
<td>.990</td>
<td>.029</td>
<td>26,257.04</td>
<td>26,768.82</td>
<td></td>
</tr>
<tr>
<td>Scalar</td>
<td></td>
<td>775.54*</td>
<td>379</td>
<td>161.49* b</td>
<td>.066</td>
<td>.059</td>
<td>.073</td>
<td>.989</td>
<td>.991</td>
<td>.031</td>
<td>26,318.51</td>
<td>26,711.48</td>
<td></td>
</tr>
<tr>
<td>Equal factor variances</td>
<td></td>
<td>827.67*</td>
<td>383</td>
<td>27.78* c</td>
<td>.070</td>
<td>.063</td>
<td>.076</td>
<td>.988</td>
<td>.990</td>
<td>.035</td>
<td>26,322.33</td>
<td>26,697.02</td>
<td></td>
</tr>
<tr>
<td>Equal factor covariances</td>
<td></td>
<td>805.97*</td>
<td>385</td>
<td>2.83 d</td>
<td>.068</td>
<td>.061</td>
<td>.074</td>
<td>.989</td>
<td>.991</td>
<td>.035</td>
<td>26,319.09</td>
<td>26,684.65</td>
<td></td>
</tr>
<tr>
<td>Equal latent means</td>
<td></td>
<td>1,085.45*</td>
<td>387</td>
<td>42.78* e</td>
<td>.087</td>
<td>.081</td>
<td>.093</td>
<td>.981</td>
<td>.985</td>
<td>.045</td>
<td>26,358.75</td>
<td>26,715.16</td>
<td></td>
</tr>
</tbody>
</table>

Note. Models computed using mean- and variance-adjusted weighted least squares (WLSMV) estimation. Δχ² computed using Mplus 8.8 DIFFTEST function.

= Δχ² comparing metric and configural; = Δχ² comparing scalar and configural; = Δχ² comparing equal factor variances and scalar; = Δχ² comparing equal factor covariances and equal factor variances; = Δχ² comparing equal latent means and equal factor covariances;

= fit statistic computed using robust maximum likelihood estimation.

* p < .001.
For metric invariance, support was indicated by acceptable fit across indices. The RMSEA indicated reasonable fit, and CFI and TLI values both indicated good fit. The SRMR result represented a well-fitting model. When compared to the configural model, the CFI decreased by .002, indicating invariance. The AIC value decreased by only 28.43, indicating potential invariance. The change in BIC was -183.79 and may indicate minimal improvement in model fit. Finally, the RMSEA 90% confidence intervals almost completely overlapped, indicating invariance across groups. There were no large (i.e., > 10) modification indices for the Christian or Muslim groups, but the following modification indices were found for the Jewish group: factor loadings P2 (10.68), P5 (20.23), and P11 (12.19) on FOS; loadings P1 (14.88), P2 (17.53), P3 (16.27), P6 (25.10), P11 (14.87), and P14 (23.37); and the first thresholds for P2 (10.67) and P11 (12.18). Overall, results suggested that metric invariance was supported, and analysis proceeded to the next step.

Scalar invariance was supported by acceptable fit across indices. The RMSEA indicated reasonable fit, and the CFI and TLI values both indicated good fit. The SRMR represented a well-fitting model. There was no change in CFI from the metric to scalar model, indicating strong support for invariance. The AIC increased from the metric model by only 79.02, and the change in BIC was -76.35, both representing a relatively minimal change in model fit. Finally, the RMSEA 90% confidence intervals showed clear overlap with the confidence intervals from both the metric and scalar models. Again, there were no large modification indices for the Christian or Muslim groups, but the following modification indices were found for the Jewish group: factor loadings P18 (10.66) on FOS; factor loadings P1 (10.33), P7 (17.02), P14 (26.96), and P18 (14.21) on FOG; and the first threshold for P1 (13.15). Overall, the results supported
scalar invariance of the PIOS, conflicting with Hypothesis 1.

**PIOS-R**

Support for PIOS-R configural invariance was indicated by acceptable fit across fit indices (see Table 5). Specifically, the RMSEA indicated reasonable absolute fit of the model, and the CFI and TLI values represented good fit. The SRMR, in conjunction with a lack of large residuals across groups, indicated a well-fitting model. There were a few modification indices above 10, including: factor loadings P1 (12.80) and P11 (14.34) for the Christian group; and factor loadings P5 (16.14) on FOS, and loadings P3 (11.10), P7 (10.05), and P14 (19.91) on FOG for the Jewish group. Overall, fit was deemed acceptable, and the solution served as a baseline model for subsequent invariance analyses.

For metric invariance, support was indicated by acceptable fit across indices. The RMSEA indicated reasonable fit, and CFI and TLI values both indicated good fit. The SRMR result represented a well-fitting model. When compared to the configural model, the CFI decreased by .002, indicating invariance. The AIC value decreased by only 20.18, indicating potential invariance. The change in BIC was -138.99 and may indicate minimal improvement in model fit. Finally, the RMSEA 90% confidence intervals showed considerable overlap, indicating invariance across groups. There were no large modification indices for the Christian or Muslim groups, but the following modification indices were found for the Jewish group: factor loadings P1 (12.71), P5 (18.61), and P11 (13.28) on FOS; loadings P1 (15.03), P3 (16.40), P11 (18.02), P14 (18.60), and P18 (10.91) on FOG; and the second threshold for P1 (20.10) as well
as the first threshold for P11 (13.28). Overall, results suggested that metric invariance was supported, and analysis proceeded to the next step.

Scalar invariance was supported by acceptable fit across indices. The RMSEA indicated reasonable fit, and the CFI and TLI values both indicated good fit. The SRMR represented a well-fitting model. There was a decrease of .001 in CFI from the metric to scalar model, indicating support for invariance. The AIC increased from the metric model by only 61.47, and the change in BIC was -57.34, both representing a relatively minimal change in model fit. Finally, the RMSEA 90% confidence intervals overlapped with the confidence intervals from both the metric and scalar models. There were no large modification indices for the Christian or Muslim groups, but the following modification indices were found for the Jewish group: factor loadings P7 (13.14) and P18 (10.40) on FOS; loadings P1 (13.30), P7 (22.11), P11 (10.30), P14 (26.04), and P18 (15.89) on FOG; and the first threshold of P1 (18.24), P3 (12.42), and P7 (11.91). Overall, the results supported scalar invariance of the PIOS-R, conflicting with Hypothesis 1.

Population Heterogeneity

Tests of population heterogeneity were conducted for the PIOS (see Table 5). First, equality of factor variances was examined and compared to the scalar model. The change in CFI was in the positive direction and was .001, indicating no significant change in fit. Additionally, although the RMSEA increased, the RMSEA 90% confidence interval overlapped with the scalar solution. Change in AIC and BIC values both were small (5.28 and 12.99, respectively), and the
ΔBIC was negative. These data suggest that equality of factor variance across groups was established and factor covariances could be examined. Support was strong for equality of factor covariances. When compared to the equal factor variances model, the Δχ² was nonsignificant, the CFI increased by .001, and the RMSEA confidence intervals showed some overlap. Further, change in AIC and BIC values was small (-2.08 and -11.23, respectively).

Lastly, equality of latent means was examined and compared to the equal factor covariances model. All fit indices worsened. The change in CFI was .007, indicating significantly worse fit. The RMSEA confidence interval showed no overlap with the equal factor covariances solution, and there was an increase in both AIC and BIC values. Thus, there is evidence to suggest that the latent means in the two-factor model of the PIOS were significantly different across groups. Further analyses of latent means were conducted, using the Christian group as the reference group. The Muslim group’s latent mean for FOS was not significantly different (M_F1 = 0.10, z = 1.31, p = .190), but the FOG latent mean was significantly greater at α = .05 (M_F2 = 0.16, z = 2.07, p = .038). Jewish group means both were significantly lower (M_F1 = -0.28, z = -3.37, p = .001, and M_F2 = -0.36, z = -4.43, p < .001). When using the Muslim group as the reference group, the Jewish group means both were significantly lower (M_F1 = -0.39, z = -4.46, p < .001, and M_F2 = -0.51, z = -6.11, p < .001). These results provided partial support for Hypothesis 3.

PIOS-R

Tests of population heterogeneity were conducted for the PIOS-R and also can be found in Table 5. The equality of factor variances model was compared to the scalar solution, and the
change in CFI was .001. The RMSEA confidence intervals did overlap, and changes in AIC and BIC values were small (3.82 and 14.46, respectively). Furthermore, the BIC value decreased. Based on these results, equality of factor variance was established and factor covariances could be examined. When compared to the equal factor variances model, the equal factor covariances model was not significantly different based on DIFFTEST results ($\Delta \chi^2[2] = 2.83, ns$). Change in CFI was .001, RMSEA confidence intervals overlapped, and changes in AIC and BIC values were small (3.24 and 12.37, respectively). Thus, there was strong support for equality of factor covariances.

The equal latent means model was then computed and compared to the equal factor covariances model. All fit indices worsened, and the change in CFI was .008. The RMSEA confidence interval was well above and outside of the confidence interval of the equal factor covariances model. There was an increase in both AIC and BIC values. This suggested that latent means of the two-factor model of the PIOS-R were significantly different across groups. Therefore, latent means were examined further, at first using the Christian group as the reference group. For the Muslim group, the FOS latent mean was not significantly different ($M_{F1} = 0.09, z = 1.19, p = .235$), but the FOG mean was significantly greater at $\alpha = .05$ ($M_{F2} = 0.17, z = 2.18, p = .030$). The Jewish group means were significantly lower ($M_{F1} = -0.29, z = -3.47, p = .001$, and $M_{F2} = -0.37, z = -4.63, p < .001$). When using the Muslim group as the reference group, the Jewish group latent means both were significantly lower ($M_{F1} = -0.38, z = -4.45, p < .001$, and $M_{F2} = -0.54, z = -6.38, p < .001$). These results provided partial support for Hypothesis 3.
Internal Consistency

Cronbach’s alpha is a common measure of internal consistency reliability reported in the psychology literature (McNeish, 2018). However, many contemporary authors suggest against using Cronbach’s alpha to test scale internal consistency (Dunn et al., 2014; Ferketich, 1990; McNeish, 2018; Revelle & Zinbarg, 2009; Savalei & Reise, 2019; Trizano-Hermosilla & Alvarado, 2016). One reason for this is the assumption of essentially tau-equivalence must be met to use Cronbach’s alpha. Authors have argued that scales in psychology rarely are truly unidimensional (e.g., Dunn et al., 2014; Trizano-Hermosilla & Alvarado, 2016), and the PIOS and PIOS-R have been shown to be multidimensional in this and other studies (e.g., Abramowitz et al., 2002; Olatunji et al., 2007). This multidimensionality makes it inappropriate to use Cronbach’s alpha. For these and other reasons (see McNeish, 2018), use of McDonald’s omega coefficient is recommended for evaluating internal consistency (Dunn et al., 2014; Ferketich, 1990; Revelle & Zinbarg, 2009). For interpretation purposes, McDonald’s omega should meet the same standards as Cronbach’s alpha (Watkins, 2017). In other words, an omega coefficient of .70 or higher generally is considered acceptable (McNeish, 2018).

Following recommendations by Savalei and Reise (2019), omega total (ωt) was calculated for the PIOS and PIOS-R. Additionally, the internal consistency of the subscales was examined by calculating ωt for each subscale. These calculations were completed using JASP Version 0.16.4 (JASP Team, 2022). Results revealed that the ωt for the full PIOS was .97, indicating excellent reliability. For the subscales of the PIOS, FOS had good reliability (ωt = .96), and FOG also had excellent reliability (ωt = .93). For the PIOS-R, the overall reliability was excellent (ωt = .96). For the subscales of the PIOS-R, FOS had excellent reliability (ωt = .95),
and FOG had excellent reliability ($\omega_t = .92$). Correlations between subscales were fairly high for both the PIOS ($r = .87, p < .001$) and the PIOS-R ($r = .85, p < .001$). Thus, Hypothesis 5 was supported.

**Convergent and Discriminant Validity**

A correlation matrix including all study measures is provided in Table 6. Pairwise deletion was employed for missing values. For the full-length PIOS, convergent validity was established by strong, significant correlations with the DOCS total ($r = .69, p < .001$) and DOCS-SR ($r = .71, p < .001$) scores. The PIOS also was strongly correlated with the other DOCS subscales, but the strongest correlation was with DOCS-SR, and the weakest was with DOCS Contamination ($r = .57, p < .001$). This pattern of correlations was expected given that the PIOS is meant to measure scrupulosity. The same pattern was found for correlations with the PIOS-R. For example, the PIOS-R was strongly correlated with the DOCS total ($r = .68, p < .001$) and the DOC-SR ($r = .71, p < .001$), and the weakest DOCS subscale correlation was with DOCS Contamination ($r = .56, p < .001$).

Discriminant validity was examined using the DASS-21-D, DASS-21-A, PANAS-NA, and SCSRFQ. The sizes of the correlations between the PIOS/PIOS-R and the DOCS-SR subscale were compared to the correlations with the aforementioned measures. Significance was tested using Steiger’s $z$ tests. For the PIOS, the DOCS-SR correlation ($r = .71, p < .001$) was significantly larger than that of the DASS-21-D ($r = .60, p < .001; z = 4.74, p < .001$), DASS-21-A ($r = .67, p < .001; z = 2.11, p < .05$), and SCSRFQ ($r = .35, p < .001; z = 10.61, p < .001$). Notably, the correlation between the PIOS and the DOCS-SR ($r = .71$) was slightly smaller than
Table 6

Zero-order Correlation Matrix for Study Variables in Full Sample (Ns range from 713-718)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. PIOS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. PIOS-R</td>
<td></td>
<td>.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. DOCS &lt;sup&gt;a&lt;/sup&gt;</td>
<td>.69</td>
<td>.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. DOCS-Contamination</td>
<td>.57</td>
<td>.56</td>
<td>.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. DOCS-Harm</td>
<td>.61</td>
<td>.60</td>
<td>.91</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. DOCS-Unacceptable Thoughts</td>
<td>.64</td>
<td>.64</td>
<td>.89</td>
<td>.69</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. DOCS-Just Right</td>
<td>.63</td>
<td>.62</td>
<td>.91</td>
<td>.70</td>
<td>.79</td>
<td>.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. DOCS-Scrupulosity</td>
<td>.71</td>
<td>.71</td>
<td>.85</td>
<td>.69</td>
<td>.74</td>
<td>.79</td>
<td>.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. DASS-21-D</td>
<td>.60</td>
<td>.58</td>
<td>.66</td>
<td>.52</td>
<td>.62</td>
<td>.63</td>
<td>.60</td>
<td>.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. DASS-21-A</td>
<td>.67</td>
<td>.66</td>
<td>.78</td>
<td>.65</td>
<td>.73</td>
<td>.70</td>
<td>.72</td>
<td>.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. PANAS-NA</td>
<td>.74</td>
<td>.72</td>
<td>.67</td>
<td>.54</td>
<td>.62</td>
<td>.63</td>
<td>.62</td>
<td>.65</td>
<td>.64</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>12. SCSRFQ</td>
<td>.35</td>
<td>.36</td>
<td>.25</td>
<td>.23</td>
<td>.23</td>
<td>.20</td>
<td>.23</td>
<td>.27</td>
<td>.11</td>
<td>.18</td>
<td>.13</td>
</tr>
</tbody>
</table>

<sup>Note.</sup> PIOS = Penn Inventory of Scrupulosity; PIOS-R = Penn Inventory of Scrupulosity – Revised; DOCS = Dimensional Obsessive-Compulsive Scale; DASS-21-D = Depression, Anxiety, and Stress Scale – 21, Depression subscale; DASS-21-A = Depression, Anxiety, and Stress Scale – 21, Anxiety subscale; PANAS-NA = Positive and Negative Affect Schedule, Negative Affect; SCSRFQ = Santa Clara Strength of Religious Faith Questionnaire.

All correlations significant at <i>p < .01</i>.

<sup>a</sup>Total DOCS scores were computed using only the original Abramowitz et al. (2010) subscales.
the correlation between the PIOS and the PANAS-NA ($r = .74, p < .001$), but the difference was not significant, $z = -1.54, p = .123$. Results suggest that the PIOS has relative discriminant validity with measures of psychological distress and religiosity, but not with negative affect.

For the PIOS-R, the DOCS-SR correlation ($r = .71, p < .001$) was significantly larger than that of the DASS-21-D ($r = .58, p < .001; z = 5.53, p < .001$), DASS-21-A ($r = .66, p < .001; z = 2.61, p < .01$), and SCSRFQ ($r = .36, p < .001; z = 10.36, p < .001$). The DOCS-SR correlation was slightly smaller than the correlation with the PANAS-NA ($r = .72, p < .001$), but the difference was not significant, $z = -0.50, p = .615$. These data suggest that the PIOS-R, like the original PIOS, has relative discriminant validity with measures of psychological distress and religiosity, but not with negative affect. This provides some support for Hypothesis 5.

Further review of Table 6 reveals a pattern of generally strong correlations among all measures, excluding the SCSRFQ, which had weak to small correlations with the other measures, $rs = .11-.36, ps < .01$. By way of comparison, the lowest correlation among any measure except the SCSRFQ was .52 (the correlation between DASS-21-D and DOCS-Contamination) and most were in the .60-.80 range. Conversely, the mean for all values involving the SCSRFQ was .23. Notably, after a Bonferroni correction, the two weakest correlation coefficients with the SCSRFQ were nonsignificant.
CHAPTER 5
DISCUSSION

The primary aim of the present study was to examine whether the PIOS and the PIOS-R measure the construct of scrupulosity equivalently across samples of individuals who identified as Christian, Jewish, and Muslim. This was particularly important to establish given the problem of factorial validity of the PIOS/PIOS-R in the current literature, and given that research studies examining factor structure have involved the use of samples of different religious groups (e.g., Abramowitz et al., 2002; Pirutinsky & Rosmarin, 2018; Phillips & Fisak, 2022). Religion is strongly relevant to scrupulosity and, given findings that scrupulosity may be more prevalent in certain religious groups than others (e.g., Abramowitz et al., 2002; Bruce et al., 2011; Buchholtz et al., 2019), it was prudent to ensure that the problems with factorial validity of the PIOS/PIOS-R were not the result of measurement noninvariance across religious affiliations. Establishing measurement invariance across these broad religious groups—notably, the primary groups studied in the extant literature—is necessary to support the use of the PIOS/PIOS-R in these groups and to draw conclusions about potential differences found (Kline, 2016). The current project was the first to examine the measurement invariance of the PIOS/PIOS-R across the three Abrahamic religions. Secondary aims of the study were to find the best-fitting measurement models of the PIOS/PIOS-R in the current sample, to examine other aspects of reliability and validity, and to compare the psychometric properties of the PIOS and the PIOS-R. This study also was the first to examine latent mean differences across the three groups.
PIOS-R Findings

Measurement Invariance

The results of the MGCFAs supported configural invariance of the PIOS-R, indicating that the pattern of factor loadings was not significantly different among Christian, Jewish, and Muslim groups. The metric model did not have significantly poorer fit compared to the configural model, indicating that the strength of the factor loadings (i.e., the degree to which each indicator loaded onto the scrupulosity latent factor) was equivalent across the three groups. In other words, no one item was more closely related to scrupulosity in one group than in the others. When compared to the metric model, the scalar model did not have significantly poorer fit, providing evidence for scalar invariance. This suggests that the thresholds are equivalent across the three groups. In other terms, the different groups use the response scale (i.e., 0 = never; 1 = almost never; 2 = sometimes; 3 = often; 4 = constantly; Abramowitz et al., 2002) in the same way. Thus, three individuals from the different groups with the same level of scrupulosity should have the same score on the indicators (Kline, 2016). Further, any mean difference among the three religious groups on any PIOS-R item can be attributed to a difference in scrupulosity and not to measurement error.

It was hypothesized that the PIOS-R would be noninvariant across Christian, Jewish, and Muslim samples and between each pair of samples. However, the current findings suggest that the PIOS-R measures scrupulosity equivalently across groups. This study provides substantial support for the continued use of PIOS-R in these groups and for comparison of scrupulosity as measured by the PIOS-R among these populations. Although the Jewish group showed somewhat worse absolute fit when comparing the fit indices for the two-factor model against the
indices for the Christian and Muslim groups, these differences were not significant. Additionally, it is unknown whether this was due to a slightly smaller sample size (Jewish group $n = 201$ versus Christian and Muslim group $n = 274$ and 243, respectively) or of somewhat lower level of religiosity. Regardless, the differences were not large enough to forego measurement invariance analyses, and the PIOS-R was found to be invariant across the three groups.

**Best-Fitting Model**

Prior to conducting measurement invariance analyses, a measurement model for the PIOS-R had to be identified. In order to determine the best-fitting model for the PIOS-R in the overall sample, the two most-supported models in the literature—and a third candidate model—were tested. The hypothesis that the two-factor model would display best fit was upheld by the results. This finding was consistent with Olatunji et al. (2007) and Shapiro et al. (2013), who concluded that the two-factor model of the PIOS-R had the best fit. Findings were inconsistent with Huppert and Fradkin (2016), who concluded that the PIOS-R had poor factorial validity due to poor absolute fit indices, and with Phillips and Fisak (2022), who also found mixed results, with a poor TLI (.897), a mediocre RMSEA (.098), and an adequate CFI (.912).

In the current sample, the one-factor model displayed mediocre to good fit across indices. Additionally, the upper limit of the RMSEA 95% confidence interval was .104, a value which would indicate poor absolute global fit. The other fit indices indicated good fit for the one-factor model, but there was room for improvement. When compared to the one-factor model, the two-factor model had better fit according to multiple forms of comparison, and the indices for the two-factor model showed a more consistent picture of fit. The finding was consistent with
Huppert and Fradkin (2016), Shapiro et al. (2013), and the current hypothesis. The result was inconsistent with Gallegos et al. (2018) and Phillips and Fisak (2022). However, Gallegos et al. (2018) used a Spanish translation, and adapting the scale to another language and culture can result in changes to the meaning of items (Auer et al., 2000). As discussed in the Factor Structure section under Measurement of Scrupulosity in the Introduction (under the Two-Factor Model heading of the PIOS-R subsection), it is unknown whether the correct estimator was used during data analysis of both studies, since the estimator was not reported in their published article. Additionally, both studies used a primarily or exclusively Christian sample, which may have limited generalizability.

As expected, the bifactor model also had poorer fit compared to the two-factor model, although it did fit better than the one-factor model. This finding was consistent with Phillips and Fisak (2022), who concluded poor fit for a bifactor model. In the current study, initial testing of this model resulted in the statistical software exceeding the maximum number of iterations, even when the number of iterations was greatly increased. Convergence problems are known to be prevalent in bifactor CFAs, due in part to factor collapse. Factor collapse occurs when much of the indicator variance is shifted from the specific factor(s) to the general factor, resulting in specific factors that no longer have significant loadings (Robertson, 2019). It is possible that the original result was a result of the FOS factor collapsing. This may suggest that the items on FOS (at least, those included on FOS in the bifactor model used in this study) are more directly related to scrupulosity itself, as pure indicators of the general factor (Robertson, 2019). If this is the case, then future research can involve exploration of a bifactor model of the PIOS-R that has one general and only one specific factor. The idea that FOS items are pure indicators of scrupulosity

makes some sense, conceptually. Triggers for individuals with scrupulous OCD typically are more directly related to fearing sin and can come from a core fear of being a bad person or of being punished by God (e.g., sent to Hell). However, nonreligious individuals also are known to experience scrupulous OCD, and often are primarily fearful of being a bad person. In this case, being punished by God is not involved. To illustrate, EFAs conducted by Phillips and Fisak (2022) showed that the FOS subscale was relevant to a sample of atheists, although the FOG subscale was not (as evidenced by floor effects on this factor). Thus, items about fearing immorality may be more directly related to scrupulosity, whereas items about fearing God specifically may be a distinct factor that is highly related to scrupulosity. More research is needed to test this idea.

Considering the evidence from the current findings and past research, our current knowledge suggests that the most appropriate measurement model for the PIOS-R is the two-factor model put forth by Olatunji et al. (2007), especially in a mixed religious sample. This was the model used to test measurement invariance and population heterogeneity in the current study. However, there still are other possibilities, such as the bifactor model with only FOG as a specific factor, that should be tested against the two-factor model, and the two-factor model should continue to be examined in a variety of samples and contexts.

The practical value for two subscales in the literature is not clear. In creating the initial subscales on the original PIOS, Abramowitz et al., (2002) argued that

Inspection of the items indicated that factor I broadly related to fears of acting or thinking sinfully and the interference in functioning caused by such fears. Therefore, we named this factor, which contained 12 items, Fear of Sin. Factor II, which included seven items, pertained to fears related to God and punishment. We therefore named this factor Fear of God. (p. 829)
The items on each subscale fit these descriptions well. However, no rationale was offered for each of these constructs having practical value. Although many researchers will calculate both subscales and use them in analyses, it is not always clear how the two subscales are differently related to other constructs. In some studies, the subscales are ignored entirely in favor of using a total score (e.g., Lau & Ramsay, 2019). However, some researchers have found differential results between the subscales. This includes the Phillips and Fisak (2022) finding that the FOS subscale had relevance for a group of atheist participants (after dropping two items specifically using the word “sin”), while the FOG subscale showed floor effects in this sample. This may indicate that a “fear of immorality” subscale may be more practical for atheists than the FOG subscale, for example. There are at times minor differences such as findings from the current study that Muslims had higher latent means than Christians on FOG but not FOS. This indicates that there may be practical relevance for the subscales, but the interpretation of disparate findings such as this can be difficult to interpret, given the closeness of the two constructs represented by the subscales.

**Population Heterogeneity**

Tests of population heterogeneity typically involve examination of equality of factor variances, equality of factor covariances, and equality of latent means. Each of these parameters was constrained to equality across groups, one at a time, and the resulting models were compared to ensure that model fit did not worsen with each step. This process is similar to the process of testing measurement invariance and starts with adding the constraint of equal factor variances to the scalar model. The current findings indicated that factor variances and covariances were
equivalent across the three religious groups. Equality of factor variances mainly is relevant in that it allows for testing of equality of factor covariances (Brown, 2015). The finding of equal factor covariances indicates that the two factors, FOS and FOG, are not more strongly correlated in one group than in the others. Additionally, equality of factor covariances permitted a comparison of latent means. The equality of latent means model showed significantly worse fit compared to the equality of factor covariances model. This was consistent with the hypothesis that the latent means would be significantly different across Christian, Jewish, and Muslim groups.

**Latent Mean Differences**

The latent means were compared between each group pair, and results indicated that the Jewish group had lower latent means for both FOG and FOS compared to both the Christian and Muslim groups. This finding was consistent with my hypothesis and with the literature, which suggests that Jewish individuals had lower scores overall compared to Christian and (often) Muslim groups (Abramowitz et al., 2002; Buchholtz et al., 2019; Siev et al., 2017). Since measurement invariance has been established among the three groups, readers can be confident that this finding is representative of differences in levels of scrupulosity and not of differences in the meaning of the PIOS-R items.

Given the differences in religious practice and in the presentation of scrupulosity among Christian, Muslims, and Jewish individuals (see Religion and PIOS Content section in Introduction), there may be some moderating factor that would explain lower scrupulosity scores in Jewish samples. For example, there has been ongoing research on the relationships among
religion, scrupulosity, and moral TAF. Past research revealed differences in levels of moral TAF between those affiliated with Judaism and those affiliated with Christianity or Islam (Siev et al., 2017; A. D. Williams et al., 2013), but no differences between those affiliated with Christianity and Islam (Yorulmaz et al., 2009). There are clear differences in the focus on internal moral behavior (e.g., thoughts), and on intent behind actions, both of which are features of Christian and Islamic beliefs, whereas the practice of Judaism typically is focused on external behavior (see PIOS Content and Religion section of Introduction). Additionally, it is possible that the PIOS-R is missing the more behavioral aspects of scrupulosity typically involved in its presentation for Jewish individuals—for example, concerns about whether one has performed a ritual correctly or whether one has followed dietary laws. Although the PIOS-R may have the same basic meaning for all three religions, that does not necessarily mean that it has equal content validity. If the PIOS-R were to include items involving concern about performing rituals correctly, it is possible that the measure would show noninvariance because rituals are typically not performed by laypeople. Future research should explore these and other possible factors that may explain the lower levels of scrupulosity in Jewish individuals.

It was expected that the latent mean for the Muslim group would be higher than that of the Christian group, and this was the current finding only for FOG. Others have found that Christians and Muslims were not significantly different on PIOS scores (Inozu, Clark, & Karanci, 2012), although comparisons between these two groups on PIOS-R total or subscale scores or latent means have not been reported. Thus, the current study is the first to establish differences between a Jewish sample and Christian and Muslim samples on PIOS-R latent means.
Reliability and Validity

Results provided evidence for internal reliability and convergent and discriminant validity. The PIOS-R full scale and subscales all displayed excellent internal reliability. This finding was consistent with Cronbach’s alpha estimates from the literature that indicated excellent internal reliability for the full scale ($\alpha = .92-.95$; Gallegos et al., 2018, and Fergus & Rowatt, 2014a, respectively), for the FOS subscale ($\alpha = .91-.95$; Fergus, 2014, and Shapiro et al., 2013, respectively), and for the FOG subscale ($\alpha = .86-.91$; Shapiro et al., 2013, and Olatunji et al., 2007, respectively). Notably, the correlation between subscales was fairly high ($r = .85$). No previous research had reported subscale intercorrelations for the PIOS-R, but for the full-length PIOS, coefficients have ranged from $.67-.85$ (Abramowitz et al., 2002; Nelson et al., 2006), the higher end of which is congruent with the current findings. Given that the subscales are meant to tap into two aspects of the same construct, it may be permissible that they are highly interrelated. Brown (2015) offers a criterion of a factor correlation of .85 or greater as representing problematic discriminant validity between two factors. When two factors are found in factor analyses but factor intercorrelations are high, it is necessary to establish that all of the factors are meaningful conceptually. This is in part because, in those situations, a second factor may be an artifact of method effects. For example, research on the Penn State Worry Questionnaire has involved factor analyses that revealed a two-factor solution (e.g., Fresco et al., 2002), yet the factors were highly correlated ($r = .87$; Brown, 2003). Other researchers discovered this two-factor solution appeared to be due to a method effect resulting from negatively worded items, and that an Absence of Worry dimension had no substantive basis (Hazlett-Stevens et al., 2004). In such situations, it is advisable to drop the method factor and its indicators. However, in the
literature on the PIOS and PIOS-R, researchers have asserted that there is a substantive basis for
FOS and FOG, and there does not appear to be evidence of a method effect in play. Future
research may be needed to engender full confidence that both factors are justified.

The pattern of correlations between the PIOS-R and the additional scales generally was
consistent with good convergent and discriminant validity. The PIOS-R was strongly correlated
with a measure of OCD symptomatology. Further, of the DOCS subscales, the PIOS-R was most
strongly related to the subscale measuring scrupulosity, providing a clear picture of convergent
and discriminant validity among different OCD symptom themes. Correlations with the DASS-
21-D and DASS-21-A were both quite strong as well, although they were significantly weaker
than the correlation with the DOCS-SR. The correlation between the PIOS-R and the DASS-21-
D was lower than that between the PIOS-R and the DASS-21-A. This would be expected given
that OCD is considered an anxiety-related disorder (Abramowitz et al., 2019) and thus anxiety is
more strongly associated with OCD than depression (Moore & Howell, 2017). The PIOS-R
showed convergent validity with a measure of religiosity, and the strength of the correlation was
on par with previous findings ($r = .40$; Plante & Boccaccini, 1997a). The correlation was small to
medium in size, and it was significantly smaller than the correlation with the DOCS-SR. Of note,
the PANAS-NA had a stronger relationship with the PIOS-R than expected ($r = .72$). The size of
the correlation was much higher than findings from previous studies ($r_s = .30-.46$; Fergus, 2014;
Fergus & Rowatt, 2014b; Olatunji et al., 2007) and on par with the correlation between the
PIOS-R and DOCS-SR. This raises concerns about discriminant validity of the PIOS-R with
negative affect.
Researchers have been able to distinguish OCD symptomatology from negative affect in the past (Moore & Howell, 2017), and negative affect is theoretically distinct from OCD. The correlation between the PANAS-NA and the PIOS-R may be unexpectedly large due in part to the wording of item stems in the PIOS-R. Specifically, the item stems begin with phrases such as “I worry that...”; “I fear...”; “I am afraid...”; and “I feel guilty...” Thus, those who are prone to negative affectivity overall, and who identify with a religious affiliation, also may be prone to experiencing negative affectivity in relation to their religious life and experience. It is possible that the PIOS-R may better measure negative affect in relation to religion than true levels of scrupulosity. Although the two concepts are similar, scrupulosity is more distinct as a symptom theme of OCD, and a good measure of scrupulosity should be able to discriminate between the two. Notably, the correlation between the DOCS-SR and the PANAS-NA was strong but lower than between the PIOS-R and the PANAS-NA. Although the DOCS-SR is only beginning to be established, and further research is needed to determine psychometric adequacy, the new subscale is a promising addition to the scrupulosity and OCD literature. This is especially true given potential superiority over the PIOS-R in this aspect of discriminant validity with negative affect.

PIOS Findings and Relevance

A secondary aim of the current study was to examine whether the PIOS-R is psychometrically superior to the original 19-item version. The full-length PIOS overall appeared to perform similarly to the PIOS-R, or slightly worse. Testing of the various models put forth in the literature (i.e., the one-factor, original two-factor, and bifactor models) revealed generally
acceptable but mediocre fit. Although most measures of fit were not categorically worse than those of the PIOS-R models, the AIC and BIC values were appreciably higher for the PIOS models (indicating a generally worse-fitting model). The PIOS model from the literature that fit the current sample best may have been the bifactor model, as hypothesized, although it showed very similar fit to the original two-factor model. However, when examining the two-factor model, there was a single modification index that was too high to ignore, and I decided to test a modified two-factor model. This resulted in significantly better fit, and categorically improved the RMSEA value. Thus, none of the models from the literature was retained for further analysis. These results suggest that, if researchers are to continue to use the full-length PIOS, any subscale scores should be calculated using the modified factor structure (i.e., item 2 should be a part of FOS).

Although the modification to the two-factor PIOS model certainly improved fit, the PIOS-R showed better factorial validity. Specifically, comparison of information criterion indices suggested that the best-fitting model of the PIOS-R was superior to that of the PIOS. Other studies involving modifications to the PIOS factor structure have led to similar conclusions. Indeed, Olatunji et al. (2007) created the PIOS-R after finding poor factorial validity for the PIOS and sought to improve it. The modification to the two-factor PIOS model in the current study also was made by Olatunji et al. (2007) in their initial phases of creating the PIOS-R. Moreover, when creating a Turkish version, Inozu, Keser, and Karanci (2017) found that the original two-factor model had poor fit and performed an EFA which resulted in a different two-factor solution and removal of certain items. As a result, their version and its proposed subscales ended up more similar to the PIOS-R than to the original PIOS. There is
clear evidence, now from three different studies in varied samples, that the factorial validity of the PIOS-R is superior to that of the PIOS.

Not only did the PIOS-R display slightly better factorial validity, but all other aspects of validity and reliability examined in this study were equivalent to the PIOS. Fewer items on a measure tend to result in worsened internal reliability. However, despite containing fewer items, the PIOS-R omega value was comparable to that of the PIOS ($\omega = .96$ for the PIOS-R, .97 for the PIOS). A near perfect correlation between the two versions ($r = .99, p < .001$) suggests that no content was sacrificed with the removal of items. Finally, on all other analyses, including measurement invariance analyses, latent mean analysis, and the pattern of correlations with other study measures, the performances of the PIOS and the PIOS-R were essentially equivalent. There appears to be no psychometric reason to continue to use the full-length length PIOS in future research, especially given that, as a slightly shorter measure, the PIOS-R has the added benefit of decreasing burden on research participants and patients.

Limitations

Despite its many strengths, the current study had multiple limiting factors involving data collection method, characteristics of the sample, and measurement of religious affiliation. To begin, there were a few major sources of potential bias. During the data cleaning process, 112 participants were removed due to disparate religious affiliation responses between the screener and full surveys. The issue lies in the fact that only a few of the participants removed during this step originally identified on the screener as Christian ($n = 3$); almost all originally identified as either Jewish ($n = 61$) or Muslim ($n = 48$). This may have introduced bias by removing more
individuals with certain responses than others. However, it is unclear with what religion these individuals were truly affiliated. For example, it is possible that all of those removed were Christian, in which case, the bias would be in the opposite direction. Nonetheless, it is an important matter to keep in mind. It illustrates some of the limitations inherent in using self-report data and in collecting via MTurk since it is always possible that individuals are not being accurate with their responses. Of course, the likelihood of attention and compliance problems in MTurk samples has been found to be comparable to research using community samples (Goodman et al., 2013), as noted in the Method section (see “MTurk”). Beyond potential issues with attention, other sources of inaccuracy may be that self-report requires participants to both understand what the content of the item is asking and be able to apply that content to themselves. This requires some amount of insight on the part of participants, and it requires participants to interpret the items in the way we mean them to, which is never a guarantee.

There also were significant differences among the three groups on a few demographic characteristics and on measure scores (apart from the PIOS/PIOS-R). Specifically, the Muslim group was on average slightly younger than the other two groups, with means differences of approximately 6 years. The Jewish group had a greater proportion of women compared to the other samples and was significantly different from the other groups on multiple measures. The Jewish sample was significantly lower than the other two groups on DOCS total scores, DOCS-Harm, DOCS-Unacceptable Thoughts, DOCS-Just Right, and DOCS-SR, as well as on SCSRFQ scores. Differences on religiosity in particular make it unclear as to whether latent mean differences on scrupulosity were confounded by religiosity. Past research suggested that differences among religious groups on constructs such as scrupulosity may depend on whether
individuals are high or low on religiosity (e.g., Abramowitz et al., 2002). Additionally, scrupulosity was moderately correlated with religiosity in the current study. This limitation must be taken into account when interpreting group differences in the current study.

In the current study, broadly-defined religious groups were studied in order to determine measurement invariance across the Abrahamic religions. However, within each religion, there are many denominations or sects, and there are even more specific branches or subgroups beyond that. Although they may belong to the same general religion (e.g., Christianity), there can be significant differences on many relevant constructs among subgroups and subdivisions of subgroups. For example, researchers have found that clergy of conservative versus liberal subgroups of the same denomination within the same religion displayed different characteristics in terms of moral thought-action fusion, belief in a “micromanaging” God (e.g., a God who is very concerned about every little thing you think or do), and potential responses to scrupulous parishioners that reinforce OCD (Deacon et al., 2013). Others have found that one denomination within Christianity (Baptist) was higher on scrupulosity compared to another denomination (Methodist; Bruce et al., 2011). Furthermore, there is often disagreement as to what groups are considered part of the overarching religion. For example, many Christians do not consider Mormons to be Christians, whereas Mormons may identify themselves as Christian. Similarly, there is significant conflict between many Sunni and Shi’ite Muslims. Given that there are so many different subgroups of religious affiliations and so many possible reasons for differences on levels of scrupulosity or moral thought-action fusion, much more research is required to tease apart these factors and bring about a clearer picture of the relations between religion and scrupulosity.
Since there are so many denominations or sects within the broad affiliations identified in this study, it did not capture the heterogeneity of religious faith and belief. It is possible that results may have been different if participants had been separated into more specific religious groups—or perhaps separated by other characteristics such as belief in moral thought-action fusion or position on the political spectrum. Additionally, no other broad religious group was studied, and no evidence regarding measurement invariance in non-Abrahamic religions was gathered. However, one may speculate that many non-Abrahamic religions likely are too different in nature for the PIOS-R to have the same meaning. However, further study of the FOI subscale of the PIOS, which has been found to be relevant in even atheist samples (Phillips & Fisak, 2022), could be of interest for use in populations of any kind of religious or spiritual belief.

**Broad Implications**

The current findings may help to clarify the factor structure of the PIOS/PIOS-R. It is likely that previous issues and inconsistency relate in part to loading item 2 on FOG rather than FOS, as it is in the modified model. Another reason may be that the samples tested in previous studies (including studies examining the PIOS-R) were not all religious or not all identifying with Abrahamic religions. In one study, researchers found that the factor structure of the PIOS-R involved a two-factor model in an atheist group and a one-factor model in a Christian group, primarily due to floor effects of the FOG factor (Phillips & Fisak, 2022). Another study revealed that factor analyses using a mixed sample of religious and non-religious participants returned inadequate absolute fit for both the PIOS and the PIOS-R (Huppert & Fradkin, 2016).
Additionally, in that study, ROC analyses showed that the PIOS had better discriminatory properties in a Christian than a non-Christian sample. Finally, Shapiro et al. (2013) found that both one-factor and two-factor models of the PIOS-R showed worse fit in non-scrupulous OCD patients than in scrupulous OCD patients. More research is needed to confirm whether specification of item 2 or nonreligious samples may be causing differences in factor structure findings.

This project was the first to examine the factor structure of the PIOS/PIOS-R separately among Christian, Jewish, and Muslim samples. It was the first to examine measurement invariance of the PIOS/PIOS-R across different broad religious affiliations in which the measure is commonly applied. It is the first, to my knowledge, to test latent mean differences on the PIOS/PIOS-R, and to test differences on PIOS/PIOS-R scores between all three of the Christian, Jewish, and Muslim groups in the same study. Findings of this project have established measurement invariance of the PIOS/PIOS-R among the Abrahamic religions, allowing researchers to confidently use this measure in all three groups and to ensure that differences between groups on the measure are interpretable. Findings have provided strong evidence that the original factor structure of the full-length PIOS may be discontinued. If the PIOS is to be used, the modified measurement model should be applied. However, current findings suggest that there likely remains no persuasive evidence for continued use of the full-length PIOS in favor of the PIOS-R. Finally, findings suggest that those with a Jewish affiliation may be lower on scrupulosity than those who identify as Christian or Muslim. Although this finding is consistent with past research, it may be confounded in the current study by religiosity. In sum,
the PIOS-R appears to be an adequately reliable and valid measure of religiosity that has the same meaning for Christians, Jews, and Muslims.
REFERENCES


JASP Team (2022). JASP (Version 0.16.4) [Computer software].


APPENDIX A

RELIGIOUS AFFILIATION ITEMS
Religious Affiliation Items

Item 1: Please select the response that most closely reflects your current religion, if any:

1. Protestant Christian (e.g., Mainline, Evangelical)
2. Catholic
3. Mormon
4. Orthodox Christian
5. Jehovah’s Witness
6. Jewish
7. Muslim
8. Buddhist
9. Hindu
10. Unitarian
11. Atheist
12. Agnostic

Item 2: In the box below, please type in your current religious affiliation, including any specific denomination or sect.
APPENDIX B

PENN INVENTORY OF SCRUPULOSITY ORIGINAL AND REVISED VERSIONS
Instructions: The following statements refer to experiences that people sometimes have. Please indicate how often you have these experiences using the following key: 0 = never; 1 = almost never; 2 = sometimes; 3 = often; 4 = constantly (Abramowitz et al., 2002).

1. I worry that I might have dishonest thoughts*
2. I fear that I might be an evil person
3. I fear I will act immorally*
4. I feel urges to confess sins over and over again*
5. I worry about heaven and hell*
6. I worry I must act morally at all times or I will be punished
7. Feeling guilty interferes with my ability to enjoy things I would like to enjoy*
8. Immoral thoughts come into my head and I can’t get rid of them*
9. I am afraid my behavior is unacceptable to God*
10. I fear I have acted inappropriately without realizing it
11. I must try hard to avoid having certain immoral thoughts*
12. I am very worried that things I did may have been dishonest*
13. I am afraid I will disobey God’s rules/laws*
14. I am afraid of having sexual thoughts*
15. I worry I will never have a good relationship with God
16. I feel guilty about immoral thoughts I have had*
17. I worry that God is upset with me*
18. I am afraid of having immoral thoughts*
19. I am afraid my thoughts are unacceptable to God*

*Note. *These items are included in the PIOS-R.
APPENDIX C

DIMENSIONAL OBSESSIVE-COMPULSIVE SCALE WITH
SCRUPULOSITY/RELIGIOUS THOUGHTS SUBSCALE
Dimensional Obsessive-Compulsive Scale with Scrupulosity/Religious Thoughts subscale

Instructions: This questionnaire asks you about 4 different types of concerns that you might or might not experience. For each type there is a description of the kinds of thoughts (sometimes called obsessions) and behaviors (sometimes called rituals or compulsions) that are typical of that particular concern, followed by 5 questions about your experiences with these thoughts and behaviors. Please read each description carefully and answer the questions for each category based on your experiences in the last month (Abramowitz et al., 2010).

**Category 1: Concerns about Germs and Contamination**

*Examples...*
- Thoughts or feelings that you are contaminated because you came into contact with (or were nearby) a certain object or person.
- The feeling of being contaminated because you were in a certain place (such as a bathroom).
- Thoughts about germs, sickness, or the possibility of spreading contamination.
- Washing your hands, using hand sanitizer gels, showering, changing your clothes, or cleaning objects because of concerns about contamination.
- Following a certain routine (e.g., in the bathroom, getting dressed) because of contamination.
- Avoiding certain people, objects, or places because of contamination.

The next questions ask about your experiences with thought and behaviors related to contamination over the last month. Keep in mind that your experiences might be different than the examples listed above. Please select the number next to your answer:

1. About how much time have you spent each day thinking about contamination and engaging in washing or cleaning behaviors because of contamination?
   
   0  None at all  
   1  Less than 1 hour each day  
   2  Between 1 and 3 hours each day  
   3  Between 3 and 8 hours each day  
   4  8 hours or more each day

2. To what extent have you avoided situations in order to prevent concerns with contamination or having to spend time washing, cleaning, or showering?
   
   0  None at all  
   1  A little avoidance  
   2  A moderate amount of avoidance  
   3  A great deal of avoidance  
   4  Extreme avoidance of nearly all things
3. If you had thoughts about contamination but could not wash, clean, or shower (or otherwise remove the contamination), how distressed or anxious did you become?

0  Not at all distressed/anxious  
1  Mildly distressed/anxious  
2  Moderately distressed/anxious  
3  Severely distressed/anxious  
4  Extremely distressed/anxious

4. To what extent has your daily routine (work, school, self-care, social life) been disrupted by contamination concerns and excessive washing, showering, cleaning, or avoidance behaviors?

0  No disruption at all.  
1  A little disruption, but I mostly function well.  
2  Many things are disrupted, but I can still manage.  
3  My life is disrupted in many ways and I have trouble managing.  
4  My life is completely disrupted and I cannot function at all.

5. How difficult is it for you to disregard thoughts about contamination and refrain from behaviors such as washing, showering, cleaning, and other decontamination routines when you try to do so?

0  Not at all difficult  
1  A little difficult  
2  Moderately difficult  
3  Very difficult  
4  Extremely difficult
Category 2: Concerns about being Responsible for Harm, Injury, or Bad Luck

Examples...

− A doubt that you might have made a mistake that could cause something awful or harmful to happen.
− The thought that a terrible accident, disaster, injury, or other bad luck might have occurred and you weren’t careful enough to prevent it.
− The thought that you could prevent harm or bad luck by doing things in a certain way, counting to certain numbers, or by avoiding certain “bad” numbers or words.
− Thought of losing something important that you are unlikely to lose (e.g., wallet, identify theft, papers).
− Checking things such as locks, switches, your wallet, etc. more often than is necessary.
− Repeatedly asking or checking for reassurance that something bad did not (or will not) happen.
− Mentally reviewing past events to make sure you didn’t do anything wrong.
− The need to follow a special routine because it will prevent harm or disasters from occurring.
− The need to count to certain numbers, or avoid certain bad numbers, due to the fear of harm.

The next questions ask about your experiences with thoughts and behaviors related to harm and disasters over the last month. Keep in mind that your experiences might be slightly different than the examples listed above. Please select the number next to your answer:

1. About how much time have you spent each day thinking about the possibility of harm or disasters and engaging in checking or efforts to get reassurance that such things do not (or did not) occur?

   0 None at all
   1 Less than 1 hour each day
   2 Between 1 and 3 hours each day
   3 Between 3 and 8 hours each day
   4 8 hours or more each day

2. To what extent have you avoided situations so that you did not have to check for danger or worry about possible harm or disasters?

   0 None at all
   1 A little avoidance
   2 A moderate amount of avoidance
   3 A great deal of avoidance
   4 Extreme avoidance of nearly all things
3. When you think about the possibility of harm or disasters, or if you cannot check or get reassurance about these things, how distressed or anxious did you become?

0  Not at all distressed/anxious  
1  Mildly distressed/anxious  
2  Moderately distressed/anxious  
3  Severely distressed/anxious  
4  Extremely distressed/anxious

4. To what extent has your daily routine (work, school, self-care, social life) been disrupted by thoughts about harm or disasters and excessive checking or asking for reassurance?

0  No disruption at all.  
1  A little disruption, but I mostly function well.  
2  Many things are disrupted, but I can still manage.  
3  My life is disrupted in many ways and I have trouble managing.  
4  My life is completely disrupted and I cannot function at all.

5. How difficult is it for you to disregard thoughts about possible harm or disasters and refrain from checking or reassurance-seeking behaviors when you try to do so?

0  Not at all difficult  
1  A little difficult  
2  Moderately difficult  
3  Very difficult  
4  Extremely difficult
**Category 3: Unacceptable Thoughts**

*Examples...*

- Unpleasant thoughts about sex, immorality, or violence that come to mind against your will.
- Thoughts about doing awful, improper, or embarrassing things that you don’t really want to do.
- Repeating an action or following a special routine because of a bad thought.
- Mentally performing an action or saying prayers to get rid of an unwanted or unpleasant thought.
- Avoidance of certain people, places, situations or other triggers of unwanted or unpleasant thoughts.

The next questions ask about your experiences with unwanted thoughts that come to mind against your will and behaviors designed to deal with these kinds of thoughts over the last month. Keep in mind that your experiences might be slightly different than the examples listed above. Please select the number next to your answer:

1. **About how much time have you spent each day with unwanted unpleasant thoughts and with behavioral or mental actions to deal with them?**

   0  None at all  
   1  Less than 1 hour each day  
   2  Between 1 and 3 hours each day  
   3  Between 3 and 8 hours each day  
   4  8 hours or more each day

2. **To what extent have you been avoiding situations, places, objects and other reminders (e.g., numbers, people) that trigger unwanted or unpleasant thoughts?**

   0  None at all  
   1  A little avoidance  
   2  A moderate amount of avoidance  
   3  A great deal of avoidance  
   4  Extreme avoidance of nearly all things

3. **When unwanted or unpleasant thoughts come to mind against your will how distressed or anxious did you become?**

   0  Not at all distressed/anxious  
   1  Mildly distressed/anxious  
   2  Moderately distressed/anxious  
   3  Severely distressed/anxious  
   4  Extremely distressed/anxious
4. To what extent has your daily routine (work, school, self-care, social life) been disrupted by unwanted and unpleasant thoughts and efforts to avoid or deal with such thoughts?

0 No disruption at all.
1 A little disruption, but I mostly function well.
2 Many things are disrupted, but I can still manage.
3 My life is disrupted in many ways and I have trouble managing.
4 My life is completely disrupted and I cannot function at all.

5. How difficult is it for you to disregard unwanted or unpleasant thoughts and refrain from using behavioral or mental acts to deal with them when you try to do so?

0 Not at all difficult
1 A little difficult
2 Moderately difficult
3 Very difficult
4 Extremely difficult
Category 4: Concerns about Symmetry, Completeness, and the Need for Things to be “Just Right”

Examples...
- The need for symmetry, evenness, balance, or exactness.
- Feelings that something isn’t “just right.”
- Repeating a routine action until it feels “just right” or “balanced.”
- Counting senseless things (e.g., ceiling tiles, words in a sentence).
- Unnecessarily arranging things in “order.”
- Having to say something over and over in the same way until it feels “just right.”

The next questions ask about your experiences with feelings that something is not “just right” and behaviors designed to achieve order, symmetry, or balance over the last month. Keep in mind that your experiences might be slightly different than the examples listed above. Please select the number next to your answer:

1. About how much time have you spent each day with unwanted thoughts about symmetry, order, or balance and with behaviors intended to achieve symmetry, order or balance?
   0 None at all
   1 Less than 1 hour each day
   2 Between 1 and 3 hours each day
   3 Between 3 and 8 hours each day
   4 8 hours or more each day

2. To what extent have you been avoiding situations, places or objects associated with feelings that something is not symmetrical or “just right?”
   0 None at all
   1 A little avoidance
   2 A moderate amount of avoidance
   3 A great deal of avoidance
   4 Extreme avoidance of nearly all things

3. When you have the feeling of something being “not just right,” how distressed or anxious did you become?
   0 Not at all distressed/anxious
   1 Mildly distressed/anxious
   2 Moderately distressed/anxious
   3 Severely distressed/anxious
   4 Extremely distressed/anxious
4. To what extent has your daily routine (work, school, self-care, social life) been disrupted by the feeling of things being “not just right,” and efforts to put things in order or make them feel right?

0  No disruption at all.
1  A little disruption, but I mostly function well.
2  Many things are disrupted, but I can still manage.
3  My life is disrupted in many ways and I have trouble managing.
4  My life is completely disrupted and I cannot function at all.

5. How difficult is it for you to disregard thoughts about the lack of symmetry and order, and refrain from urges to arrange things in order or repeat certain behaviors when you try to do so?

0  Not at all difficult
1  A little difficult
2  Moderately difficult
3  Very difficult
4  Extremely difficult
Category 5: Immoral and Scrupulous Thoughts (Wetterneck et al., 2021)

Examples...
− Blurring out obscenities
− Blasphemous thoughts or immoral thoughts related to a religious figure
− Repeating an action or following a special routine because of an immoral thought.
− Mentally performing an action or saying prayers to get rid of an immoral thought.
− Avoidance of certain people, places, situations or other triggers of immoral thoughts

The next questions ask about your experiences with thoughts and feelings that something is immoral and behaviors designed to deal with them over the last month. Keep in mind that your experiences might be slightly different than the examples listed above. Please select the number next to your answer:

1. About how much time have you spent each day with immoral thoughts and with behavioral or mental actions to deal with them?

   0 None at all
   1 Less than 1 hour each day
   2 Between 1 and 3 hours each day
   3 Between 3 and 8 hours each day
   4 8 hours or more each day

2. To what extent have you been avoiding situations, places, objects and other reminders (e.g., numbers, people) that trigger immoral thoughts?

   0 None at all
   1 A little avoidance
   2 A moderate amount of avoidance
   3 A great deal of avoidance
   4 Extreme avoidance of nearly all things

3. When immoral thoughts come to mind against your will how distressed or anxious did you become?

   0 Not at all distressed/anxious
   1 Mildly distressed/anxious
   2 Moderately distressed/anxious
   3 Severely distressed/anxious
   4 Extremely distressed/anxious
4. To what extent has your daily routine (work, school, self-care, social life) been disrupted by immoral thoughts and efforts to avoid or deal with such thoughts?

0  No disruption at all.
1  A little disruption, but I mostly function well.
2  Many things are disrupted, but I can still manage.
3  My life is disrupted in many ways and I have trouble managing.
4  My life is completely disrupted and I cannot function at all.

5. How difficult is it for you to disregard immoral thoughts and refrain from using behavioral or mental acts to deal with them when you try to do so?

0  Not at all difficult
1  A little difficult
2  Moderately difficult
3  Very difficult
4  Extremely difficult
APPENDIX D

SANTA CLARA STRENGTH OF RELIGIOUS FAITH QUESTIONNAIRE
Santa Clara Strength of Religious Faith Questionnaire

Instructions: Please answer the following questions about religious faith using the scale below. Indicate the level of agreement (or disagreement) for each statement. 1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree (Plante & Boccaccini, 1997a).

1. My religious faith is extremely important to me.
2. I pray daily.
3. I look to my faith as a source of inspiration.
4. I look to my faith as providing meaning and purpose in my life.
5. I consider myself active in my faith or church.
6. My faith is an important part of who I am as a person.
7. My relationship with God is extremely important to me.
8. I enjoy being around others who share my faith.
9. I look to my faith as a source of comfort.
10. My faith impacts many of my decisions.
APPENDIX E

DEPRESSION, ANXIETY, AND STRESS SCALE – 21,

DEPRESSION AND ANXIETY SUBSCALES
Instructions: Please read each statement and circle a number 0, 1, 2 or 3 which indicates how much the statement applied to you over the past week. There are no right or wrong answers. Do not spend too much time on any statement. The rating scale is as follows: 0 = Did not apply to me at all; 1 = Applied to me to some degree, or some of the time; 2 = Applied to me to a considerable degree or a good part of time; 3 = Applied to me very much or most of the time (Lovibond & Lovibond, 1995).

1. (A) I was aware of dryness of my mouth
2. (D) I couldn’t seem to experience any positive feeling at all
3. (A) I experienced breathing difficulty (e.g., excessively rapid breathing, breathlessness in the absence of physical exertion)
4. (D) I found it difficult to work up the initiative to do things
5. (A) I experienced trembling (e.g., in the hands)
6. (A) I was worried about situations in which I might panic and make a fool of myself
7. (D) I felt that I had nothing to look forward to
8. (D) I felt down-hearted and blue
9. (A) I felt I was close to panic
10. (D) I was unable to become enthusiastic about anything
11. (D) I felt I wasn’t worth much as a person
12. (A) I was aware of the action of my heart in the absence of physical exertion (e.g., sense of heart rate increase, heart missing a beat)
13. (A) I felt scared without any good reason
14. (D) I felt that life was meaningless

Note. A = Anxiety subscale items; D = Depression subscale items.
APPENDIX F

POSITIVE AND NEGATIVE AFFECT SCHEDULE, NEGATIVE AFFECT SUBSCALE
Instructions: This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you generally feel this way, that is, how you feel on the average. Use the following scale to record your answers: 1 = Very slightly or not at all; 2 = A little; 3 = Moderately; 4 = Quite a bit; and 5 = Extremely (Watson et al., 1988).

1. Distressed
2. Upset
3. Guilty
4. Scared
5. Hostile
6. Irritable
7. Ashamed
8. Nervous
9. Jittery
10. Afraid