




2019

ASSESSING THE MODEL FIT OF MULTIDIMENSIONAL ITEM RESPONSE THEORY MODELS WITH POLYTOMOUS RESPONSES USING LIMITED-INFORMATION STATISTICS

Caihong Rosina Li

University of Kentucky, caihong.li@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0002-7790-5436>

Digital Object Identifier: <https://doi.org/10.13023/etd.2019.006>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Li, Caihong Rosina, "ASSESSING THE MODEL FIT OF MULTIDIMENSIONAL ITEM RESPONSE THEORY MODELS WITH POLYTOMOUS RESPONSES USING LIMITED-INFORMATION STATISTICS" (2019). *Theses and Dissertations--Education Science*. 45.

https://uknowledge.uky.edu/edsc_etds/45

This Doctoral Dissertation is brought to you for free and open access by the College of Education at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Education Science by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Caihong Rosina Li, Student

Dr. Michael D. Toland, Major Professor

Dr. Margaret Bausch, Director of Graduate Studies

ASSESSING THE MODEL FIT OF MULTIDIMENSIONAL ITEM RESPONSE
THEORY MODELS WITH POLYTOMOUS RESPONSES USING LIMITED-
INFORMATION STATISTICS

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Education
at the University of Kentucky

By

Caihong Rosina Li

Lexington, Kentucky

Director: Dr. Michael D. Toland, Associate Professor of Educational, School, and
Counseling Psychology

Lexington, Kentucky

Copyright © Caihong Rosina Li 2018

ABSTRACT OF DISSERTATION

ASSESSING THE MODEL FIT OF MULTIDIMENSIONAL ITEM RESPONSE THEORY MODELS WITH POLYTOMOUS RESPONSES USING LIMITED-INFORMATION STATISTICS

Under item response theory, three types of limited information goodness-of-fit test statistics – M_2 , M_{ord} , and C_2 – have been proposed to assess model-data fit when data are sparse. However, the evaluation of the performance of these GOF statistics under multidimensional item response theory (MIRT) models with polytomous data is limited. The current study showed that M_2 and C_2 were well-calibrated under true model conditions and were powerful under misspecified model conditions. M_{ord} were not well-calibrated when the number of response categories was more than three. $RMSEA_2$ and $RMSEA_{C_2}$ are good tools to evaluate approximate fit.

The second study aimed to evaluate the psychometric properties of the Religious Commitment Inventory-10 (RCI-10; Worthington et al., 2003) within the IRT framework and estimate C_2 and its RMSEA to assess global model-fit. Results showed that the RCI-10 was best represented by a bifactor model. The scores from the RCI-10 could be scored as unidimensional notwithstanding the presence of multidimensionality. Two-factor correlational solution should not be used. Study two also showed that religious commitment is a risk factor of intimate partner violence, whereas spirituality was a protecting factor from the violence. More alcohol was related with more abusive behaviors. Implications of the two studies were discussed.

Key words: multidimensional item response theory, limited-information goodness of fit statistics, M_2 , M_{ord} , C_2

Caihong Rosina Li

December 3, 2018

Date

ASSESSING THE MODEL FIT OF MULTIDIMENSIONAL ITEM RESPONSE
THEORY MODELS WITH POLYTOMOUS RESPONSES USING LIMITED-
INFORMATION STATISTICS

By

Caihong Rosina Li

Michael D. Toland

Director of Dissertation

Margaret Bausch

Director of Graduate Studies

December 3, 2018

Dedicated to my beloved parents, family, and friends

ACKNOWLEDGEMENTS

I would like to thank all my committee members who have supported me throughout my time in graduate school and as I was working on my dissertation. I would like to thank Dr. Toland, who led me into the world of methodology and built my confidence as a researcher in this area. I will always be grateful for your mentorship and generous help. Secondly, I would like to express my sincere thanks to my committee members, Dr. Xin Ma, Dr. Diane Follingstad, Dr. Katherine Thompson, and Dr. Claire Renzetti. They have graciously sacrificed their time and attention to attend my committee meetings, shared their knowledge, proof-read my dissertation, and provided great advices both on my research and my career.

I give my special thanks to the Applied Psychometric Strategies (APS) Lab, the P20 Motivation and Learning Lab, the Robinson Scholarship Program, the Center for Research on Violence Against women, the Quantitative and Psychometric Methods (QPM) program, the Educational Psychology (EDP) program, and College of Education, who not only provided me financial support but also offered me copious social capital. I was inspired by some great purpose, some extraordinary project, and all their thoughts about breaking bounds and becoming excellence.

I would like to thank my family and friends who accompanied me through this long journey. Those people include Adelyn, Zhaoshuai, Wenjin, Angela, Qiwen Kang, Hao Zhou, and Yan Xu. Also, special thanks to my friend, Xian Wu, who introduced God to me. I would give my deepest gratitude to my beloved parents, Zhuoxiang Li and Xiuying Zhang. Your words gave me the strength to persist whenever I met any setbacks. I would not be where I am today without the love and support from all of you.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iii
LIST OF TABLES.....	vi
LIST OF FIGURES.....	vii
SYMBOLES.....	ix
Chapter 1 – Introduction.....	1
Organization of Dissertation.....	2
Full-Information IRT Model-Data Fit Statistics.....	2
Limited-Information Goodness-of-fit Model-Fit Statistics.....	4
Purpose of the Current Dissertation.....	10
Chapter 2 – Study one.....	12
Limited-information GOF statistics.....	12
Approximate Fit Statistics.....	19
Purpose of Current Study.....	23
Method.....	24
Simulation design.....	24
Calibrations.....	25
Type I error rates.....	28
Power to detect model misspecification.....	28
Performance of RMSEA.....	28
Results.....	29
Convergence results and data cleaning.....	29

Type I error rates.....	31
Power to detect model misspecification..	39
Performance of RMSEA.....	44
Discussion	52
Chapter 3 – Study two	57
The RCI: Empirical Studies of its Psychometric Properties.....	57
Psychometric Concerns over the Dimensionality and Scoring of the RCI-10.....	61
Using a Bifactor IRT Model to Assess the Dimensionality of the RCI-10.	62
Religiosity, Spirituality, and Perpetration of Intimate Partner Violence (IPV)..	
.....	64
Purpose of Current Study.....	65
Method	65
Participants.....	65
Measures.....	66
Data Analysis Plan.....	70
Results	77
Descriptive statistics of item responses and response process assessment.....	77
Local independence assessment.....	77
Evaluation of item-level model-data fit.....	78
Global model-data fit and comparison.....	78
Comparison of item parameters across the unidimensional GR model and the	
bifactor GR model.....	79

Comparison of the unidimensional GR model trait scores and precision with the bifactor GR model general trait scores..	82
Correlational evidence.	84
Measurement Invariance.	85
SEM models.	85
Discussion	89
Appendix A	91
Appendix B	93
Appendix C	94
Appendix D	95
Appendix E	97
Appendix F	102
Appendix G	103
References	104
Vita	117

LIST OF TABLES

Table 2.1; Generating Item Parameters.....	27
Table 2.2; Convergence Rate for All conditions.....	31
Table 2.3; Means of M_2 , M_{ord} , and C_2 and the Associated Degrees of Freedom by Condition.....	34
Table 2.4; Type I Error Rates by Conditions.....	35
Table 2.5; Mean M_2 , M_{ord} , and C_2 Values and Rejection Rates of M_2 , M_{ord} , and C_2 by Fitting a Unidimensional Model to Multidimensional Data.....	40
Table 2.6; Rejection Rates by Conditions Under Alternative Conditions.....	41
Table 2.7; Mean RMSEA ₂ , RMSEA _{ord} , and RMSEA _{C2} Values by Condition.....	47
Table 2.8; RMSEA Mean and Standard Deviation by Condition for the Misspecified Models.....	48
Table 2.9; RMSEA Cut-Off Criterion Based on the Number of Categories and the Level of Correlation.	51
Table 3.1; Global Model Fit Results.....	83
Table 3.2; Unidimensional Graded Response Model Item Parameters Estimates for the RCI-10.	84
Table 3.3; Bifactor Model Item Parameter Estimates for the RCI-10.....	85
Table 3.4; Means, Standard Deviations, and Correlations for the Variables in the Study ($N = 392$).....	88
Table 3.5; Fit indices of the Measurement Invariance Tests for the 10-item Religious Commitment Inventory Scale (RCI-10) and the 42-item Severity of Violence Against Women Scale (SVAWS).....	90

Table A1; Descriptive Information of M_{ord} With Conditions Including Two Factors and Five Response Categories Under Null Condition.....	116
Table A2; Descriptive Information of M_{ord} With Conditions Under Alternative Condition.....	117
Table E1; Frequency and Descriptive Statistics of the SVAWS Used to Measure Intimate Partner Violence..	122
Table E2; Correlation Among the Psychological Abuse, Physical Abuse, and the Sexual Abuse Factors Using the Multidimensional Correlational Model (N = 392)...	123
Table E3; Goodness of Fit Statistics for All Tested Measurement Models (N = 392)...	124
Table E4; Confirmatory Factor Analysis Standardized Loadings, Relative Parameter Bias, and Individual Explained Common Variance.....	125
Table E5; Factor-Level Bifactor Indices.....	126

LIST OF FIGURES

Figure 2.1; Quantile-quantile plots of observed M2, Mord, and C2 values and their reference χ^2 distributions.....38

Figure 2.2; Empirical rejection rates at different RMSEA cut-off criteria under the misspecified model condition49

Figure 3.1; Hypothesized model testing the main effect of alcohol consumption, religiosity, and spirituality on intimate partner violence.....80

Figure 3.2; Total informationa function (TIF) for the RCI-10 fit by the unidimensional graded response (GR) model and marginal TIF (MTIF) for the RCI-10 data fit by the bifactor GR (bifac-GR) model.....87

Figure 3.3; Structural equation model testing joint effect of alcohol consumption, religiosity, and spirituality on the intimate partner violence.91

Figure 3.4; Structural equation model testing the moderation effect of the alcohol consumption on the relationship between religiosity, spirituality, and the intimate partner violence.....92

SYMBOLS

N number of participants	$i = 1, \dots, N$
D number of dimensions of a scale	$d = 1, \dots, D$
m number of items in a scale	$m = 1, \dots, M$
K number of categories in each item	$k = 1, \dots, K$
C number of cells in a contingency table/number of response patterns	$c = 1, \dots, C$
π_c probability of each response pattern in the population	$c = 1, \dots, C$
π probabilities of response patterns for a population	
q dimension of parameters to be estimated by a certain model	
θ vector of parameters that need to be estimated under a certain model, one for each item	
$\pi(\theta)$ item response theory model; restrictions forced on the probabilities based on a model	
p_c probability of each response pattern based on the sample	$c = 1, \dots, C$
p probabilities of response patterns for a sample	
\mathbf{q} vector of parameters θ to be estimated from the data (i.e., in the case of a three-parameter logistic model with a standard normal distributed trait, $q = 3n$ and θ is the vector of intercepts, slopes, and the guessing parameters, one for each item)	

Chapter 1 – Introduction

Item response theory (IRT) models need to be evaluated by model-data fit statistics so that the inferences drawn from the IRT results are valid. In the IRT world, the purpose of fitting a model to sample data is to reproduce the population probability of each response pattern. The null hypothesis for a model-data fit statistic is that the population probability of each response pattern equals to the corresponding model-estimated probability. On the contrary, the alternative hypothesis is that these two parts do not equal. Researchers examine the model-data fit statistics in the hope that it does not reject the null hypothesis so that the considered model could be interpreted meaningfully. Among extensive literature investigating model-data fit statistics, the field of global goodness-of-fit (GOF) statistics within the IRT framework has been relatively stagnant in the sense that classical GOF statistics such as Pearson's χ^2 (Pearson, 1900) and the likelihood ratio statistics G^2 (Wilks, 1938) are inaccurate when the response frequencies become sparse for some response patterns, especially when there are many items and/or response categories (Koehler & Larntz, 1980; Thissen & Steinberg, 1997). The sparseness problem is even more severe when it comes to multidimensional IRT (MIRT) models which often contain a relatively larger number of items compared to unidimensional IRT models.

Upon the need of handling sparseness problem, limited-information GOF statistics such as M_2 (Maydeu-Olivares & Joe, 2005, 2006), M_2^* (also known as M_{ord} ; Cai & Hansen, 2013), and C_2 (Cai & Monroe, 2014) was developed and have promoted the application of global GOF statistics in IRT models. The root mean square error approximation (RMSEA) corresponding to these GOF statistics have also been

introduced to examine the approximate fit of IRT models. However, the evaluation of M_2 , M_{ord} , and C_2 and related RMSEA indices are mostly limited to unidimensional 1PL, 2PL, 3PL, or graded response (GR; Samejima, 1969) models, correlated multidimensional models with binary responses, or bifactor GR models (e.g., Cai & Hansen, 2013; Jurich, 2014; Maydeu-Olivares & Joe, 2005, 2006, 2014). Few studies have examined the performance of these statistics and indices under non-bifactor MIRT models with polytomous data. The purpose of the current study is to assess the performance of M_2 , M_{ord} , and C_2 and their corresponding approximate indexes under various MIRT conditions and also apply C_2 and corresponding RMSEA to C_2 to an empirical study evaluating the Religious Commitment Inventory-10 scale (RCI-10; Worthington et al., 2003).

Organization of Dissertation. This dissertation is organized into several chapters. In the first chapter, full-information and limited information statistics is elaborated with a deeper understanding of the problem that was studied. The second chapter is a complete report of study one which investigated the performance of limited-information statistics and their RMSEAs under MIRT conditions. The third chapter is a complete report of study two which examined the psychometric properties of the RCI-10.

Full-Information IRT Model-Data Fit Statistics. Pearson's χ^2 and the likelihood ratio statistic G^2 are two full-information GOF statistics traditionally used in evaluating global model-data fit. Both Pearson's χ^2 and the likelihood ratio statistic G^2 are computed from all possible response patterns using the full contingency table. Pearson's χ^2 is defined as

$$\chi^2 = N \sum_c (p_c - \hat{\pi}_c)^2 / \hat{\pi}_c, \quad (1)$$

where N is the number of participants, c is the number of response patterns (if there are m items with k categories, then c ranges from 1 to k^m), p_c is the observed probability of each response pattern, and $\hat{\pi}_c$ is the estimated probability of each response pattern. When the model perfectly fits the data, $\chi^2 = 0$. The likelihood ratio statistic G^2 is defined as

$$G^2 = 2N \sum_c p_c \ln(p_c / \hat{\pi}_c). \quad (2)$$

When the model perfectly fits the data, $G^2 = 0$. Researchers have indicated that when the model holds and maximum likelihood estimation is used, these two statistics approximately follow a χ^2 distribution and are asymptotically equivalent to each other (e.g., Agresti, 1990). When maximum likelihood (ML) estimation is used, an asymptotic p value related with both statistics with a degrees of freedom (df) of $k^m - q - 1$ could be estimated. Here q is defined as the number of parameters to be estimated from the data [i.e., $q = 2 \times$ the number of items when a two-parameter logistic (2PL) model is fit to the data]. The underlying assumptions of these two statistics include that the latent trait to be measured should be normally distributed and items should be multinomial.

Both χ^2 and G^2 are computed using probabilities of all possible response patterns, so they are also called full-information GOF statistics. However, these two statistics could be useless when the probabilities of certain response patterns become too small or nonexistent. Sample size and the number of response patterns both influence the accuracy of the asymptotic χ^2 . So, when the sample size is small, the probabilities of some response patterns become poorly estimated (close to 0), which further influence the accuracy of the sampling distribution estimated from these inaccurate probabilities. Also, when the number of response patterns goes up, the probabilities of response patterns become smaller or negligible (also known as sparse), which could cause a large

discrepancy between the observed χ^2 distribution and the asymptotic χ^2 distribution. Typically, under sparseness, the asymptotic χ^2 is bigger than the observed χ^2 , which could suggest a larger Type I error. Thissen and Steinberg (1997) have shown that the asymptotic p value obtained for χ^2 and G^2 are invalid with any IRT model that has more than 6 items with 5 or more categories.

Researchers have developed several different approaches to conquer the aforementioned problems related to the asymptotic χ^2 distribution, such as using resampling methods (Stone, Ankenmann, Lane, & Liu, 1993; Stone, 2000; Tollenaar & Mooijart, 2003) and limited information methods. Resampling methods have been criticized as being cumbersome in computation (Stone et al., 1993) and inaccurate in p values calculated for χ^2 and G^2 statistics (Tollenaar & Mooijart, 2003). On the other hand, limited-information methods (Reiser, 1996; Reiser & Lin, 1999) have been found to be efficient and powerful in practice (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2005, 2006, 2014).

Limited-Information Goodness-of-fit Model-Fit Statistics. Limited-information GOF statistics, just as the name suggests, are computed via limited information of the probabilities of response patterns. Instead of using all possible response patterns, limited-information GOF statistics use only part of the contingency table. Limited information procedures have been a booming area recently because of their power to test the global fit of IRT models, especially when the traditional full information procedures cannot be used when the data are sparse (Maydeu-Olivares & Joe, 2005, 2006). The origin of the limited information statistics came from the area of factor analyses. Specifically, Christoffersson (1975) proposed a statistic with an asymptotic χ^2 distribution for binary

data with multiple factors using marginal distributions of single and paired items in factor analysis. Muthén (1978) presented another factor analytic statistic which is more efficiently computed than the one proposed by Christoffersson (1975) using the first-order and second-order marginal probabilities. These two statistics share in common that they both originated from factor analytic approach and aim to build statistics by reference to a χ^2 distribution with asymptotic p values using only first-order and second-order marginal probabilities. Thus, when IRT researchers need to address the sparseness problem, a limited information approach becomes a natural go-to solution because lower-order marginal tables are better filled.

The application of limited information GOF statistics within the IRT framework started in 2005 and 2006 when Maydeu-Olivares and Joe developed a family of M_r statistics based on lower margins probabilities to test the absolute model fit when items are dichotomous or polytomous. The M_r statistics, especially M_2 based on the univariate and bivariate margins, have been found to be asymptotically more powerful than the traditional full information statistics with sparse or non-sparse data (Maydeu-Olivares & Joe, 2005, 2006; Joe & Maydeu-Olivares, 2010). To deal with the sparseness problem that occurs at the bivariate and/or univariate margins, Cai and Hansen (2013) proposed a more condensed GOF statistic, M_2^* (also known as M_{ord}), which assumes the data is ordinal instead of nominal. Cai and Monroe (2014) presented a hybrid GOF statistics, C_2 , motivated by the fact that M_{ord} cannot be used when the degrees of freedom is negative.

Along with the aforementioned absolute model-data fit statistic based on the limited information approach, theories on approximate fit indexes related to limited information fit statistics also have progressed. For instance, Maydeu-Olivares (2013)

provides a general review of the newly developed GOF statistics and corresponding approximate fit indexes including bivariate RMSEA ($RMSEA_2$), RMSEA for ordinal data based on M_{ord} ($RMSEA_{ord}$), and the supplemental index standardized root mean square residual (SRMSR; Kline, 2016). Meanwhile, Cai and Monroe (2014) proposed the RMSEA corresponding to C_2 ($RMSEA_{c_2}$). To date, the research related to the limited information approach in the field of IRT has been quite fruitful, both theoretically and empirically. Theoretically, many researchers are working on examining the behavior (Type I error rate, power, cutoff values for practice, and asymptotic relative efficiency) of such fit statistics and approximate fit indexes under various conditions. Empirically, more and more researchers have adopted M_2 with $RMSEA_2$, or M_{ord} with $RMSEA_{ord}$ in their own research when using IRT methods. Along with the theoretical and empirical applications is the development of software available for computing the fit statistics and approximate fit indexes mentioned.

Until now, several studies have investigated the performance of M_2 and $RMSEA_2$ (i.e., Maydeu-Olivares & Joe, 2005, 2006, 2014), two studies have investigated the performance of M_{ord} and $RMSEA_{ord}$ (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2014), and only one paper has assessed C_2 (Cai & Monroe, 2014) under limited simulation conditions. M_2 has been reported to have good Type I error control for overall model-data fit and is powerful to detect model misspecification (i.e., dimensionality misspecification) when the attribute(s) is normally distributed. Researchers have investigated the Type I error rate and power of M_2 with binary data under unidimensional one-parameter logistic (1PL) model, two-parameter logistic (2PL) model, three-parameter logistic (3PL) model, four-parameter logistic (4PL; Barton & Lord, 1981) model, log-

linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), and compensatory IRT models (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2005; Xu, Paek, & Xia, 2017). Evaluations of M_2 have also been extended to polytomous data (3-5 categories) under unidimensional GR IRT models (Cai & Hansen, 2013; Cai & Monroe, 2014; Maydeu-Olivares & Joe, 2006). Also, researchers have evaluated M_2 under correlational multidimensional GR and LCDM models using binary data (2-5 dimensions; Jurich, 2014; Liu, Tian, & Xin, 2016; Xu, Paek, & Xia, 2017). Cai and Hansen (2013) further evaluated the performance of M_2 in bifactor GR models with binary and polytomous data. However, studies have also shown M_2 does not have enough power to detect nonnormality of ability distribution (Hansen, Cai, Monroe, & Li, 2014; Li & Cai, 2012). A recent paper by Paek, Xu, and Lin (2018) examined the performance of M_2 under 2PL and 3PL unidimensional models when the attributes were normally distributed, positively skewed, and negatively skewed and reported the power of M_2 to detect nonnormality is poor.

Although the investigation of M_2 is quite fruitful, there still is a gap in the evaluation of M_2 under MIRT models with multiple categories (3-5 categories). This is also the case for M_{ord} and C_2 . Researchers have investigated the Type I error rate and power of M_{ord} under unidimensional and bifactor GR models using binary and polytomous data (2-5 categories; Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2014). Cai and Hansen found that M_{ord} was better calibrated with polytomous data and more powerful than M_2 in detecting misspecified bifactor models. Maydeu-Olivares and Joe (2014) also reported M_{ord} has more power to reject misspecified unidimensional models compared to M_2 when the items are polytomous (3-4 categories). Similar to M_2 , few

studies have examined the performance of M_{ord} when the test includes many dimensions and items are polytomous. As for C_2 , only Cai and Monroe (2014) have examined the performance of C_2 and reported C_2 is more powerful than M_2 and M_{ord} to detect model misspecification under a unidimensional GR IRT model with four categories.

The study of RMSEA based on limited-information statistics are also gaining more and more attention. Maydeu-Olivares and Joe (2014) examined the rejection rate for $RMSEA_2$ by fitting two-dimensional 3PL or GR models to binary and polytomous (3-4 categories) data that were simulated from unidimensional IRT models and provided the following recommendations: $RMSEA_2 \leq .089$ as adequate fit, $RMSEA_2 \leq .05$ as close fit, and $RMSEA_2 \leq .05/(K-1)$ as excellent fit. Note here when the data is binary, $.05/(K-1) = .05$ and thus $.05$ serves both the close and excellent fit when there are two response categories (Maydeu-Olivares & Joe, 2014). Jurich (2014) evaluated the rejection rate of $RMSEA_2$ for multidimensional models (2-4 correlated dimensions) when the null model was bidimensional for binary data. However, studies have not yet examined the rejection rate for $RMSEA_2$ for polytomous data when the null model is multidimensional. For $RMSEA_{ord}$, Maydeu-Olivares and Joe (2014) did not find a clear relationship between $RMSEA_{ord}$, number of items, and number of items when detecting misspecified unidimensional IRT models and thus did not offer any suggested cutoff values for when to consider adequate global model data fit. In addition, Jurich (2014) evaluated the performance of $RMSEA_2$ for misspecified correlated MIRT models (2-4 dimensions) using binary data and suggested $RMSEA_2$ values above $.04$ as a cutoff for dimension misspecification if the intercorrelation is low ($\rho = .50$) and $.035$ to $.04$ be used if the intercorrelation is high ($\rho = .80$).

Along with the theoretical development of limited-information GOF statistics, there is a thriving trend to apply limited information statistics and their associated RMSEAs in empirical studies in the past several years. M_2 and $RMSEA_2$ have been applied to different item response data (e.g. dichotomous, polytomous), dimensionality (e.g. unidimensional models, correlational models, bifactor models), and different types of IRT models (e.g., Rasch, 2PL, GR). For example, Maulana, Helms-Lorenz, and van de Grift (2015) used M_2 statistics and $RMSEA_2$ for the overall model-data fit of a Rasch model (Rasch, 1960) in a study of a measure that assesses pupil's perceptions of teaching behavior. M_{ord} and $RMSEA_{ord}$ have also been increasingly applied to evaluate global GOF fit within the IRT framework. For instance, Fossati, Widiger, Borroni, Maffei, and Somma (2015) applied M_{ord} and $RMSEA_{ord}$ by fitting a five-factor confirmatory IRT model to the data. Yost, Waller, Lee, and Vincent (2017) also used M_{ord} and $RMSEA_{ord}$ to assess the measurement properties of the Patient-Reported Outcome Measurement Information System (PROMIS) fatigue item bank (FIB) using bifactor GR IRT model. Furthermore, although not exhaustive, published studies have used C_2 and $RMSEA_c$ to evaluate model-data fit under IRT models. For example, Toland et al. (2017) in their paper introducing bifactor polytomous IRT analysis used C_2 and $RMSEA_{c2}$ in their empirical example of evaluating global model-data fit of a unidimensional GR model, a multi-factor GR model, and a bifactor GR model. Overall, applications of the limited information statistics and the associated approximate fit RMSEA will certainly go to increase as more researchers become aware of this statistic for evaluating global model-data fit.

Purpose of the Current Dissertation. The current dissertation firstly evaluated the performance of limited-information GOF statistics and their related RMSEAs under various MIRT conditions and then applied C_2 and its associated RMSEA in an empirical study.

The first study investigated the Type I error rate and power of M_2 , M_{ord} , and C_2 under different data structures using a Monte Carlo simulation approach. Four variables were manipulated: the number of dimensions, the number of response categories, sample size, and the magnitude of the interfactor correlation. We also compared M_2 , M_{ord} , and C_2 to each other under conditions when these three limited-information GOF statistics could all be obtained. In addition to the investigation of M_2 , M_{ord} , and C_2 , their corresponding approximate fit indices — the RMSEAs — were examined. The current paper attempted to provide a guideline for the use of RMSEAs when MIRT models are used to study the construct validity and reliability/precision of an instrument that contains several correlational dimensions. The current study showed that M_2 and C_2 were well-calibrated under true model conditions and were powerful under misspecified model conditions. M_{ord} were not well-calibrated when the number of response categories are more than four. $RMSEA_2$ and $RMSEA_{C_2}$ are good tools to evaluate approximate fit. Findings from the first study benefit practitioners and researchers who will use limited information GOF statistics to assess global-level (scale-level) model-data fit and misfit in unidimensional and multidimensional IRT applications with dichotomous or polytomous data.

The second study assessed the psychometric properties of the Religious Commitment Inventory-10 scale (RCI-10; Worthington et al., 2003) using three competing IRT models. Religiosity was defined as the level of adherence to one's

religious values, beliefs, and practices that are used in one's daily life by Worthington et al. (1988). Corresponding to this definition, Worthington and his colleagues developed a series of religious commitment inventory (RCI) measures that are purported to be related to motivational and behavioral commitment to one's religious beliefs. To date, only one study has examined the psychometric properties of the RCI-10, the most updated version of the RCI series. Although scores from the RCI-10 was concluded as two-dimensional, Worthington and his colleagues (2003) suggested it should be treated as unidimensional due to the high correlation between the two factors. However, inconsistencies have been found between how researchers have scored the RCI-10 and how scoring has been suggested. The second study revisited the dimensionality of the RCI-10 by testing responses from a national community sample of 392 adults who had diverse religious affiliations. Specifically, three competing IRT models were used: the bifactor graded response (GR) model was examined in contrast to the two-factor correlated GR model and a unidimensional GR model. In this empirical study, C_2 and related RMSEA was used for global model-data fit. Also, structural equation modeling was used to examine the explanatory capability of religiosity, spirituality, and alcohol consumption on intimate partner violence. Results showed that the RCI-10 was best represented by a bifactor model. The scores from the RCI-10 could be scored as a unidimensional scale with the presence of multidimensionality. Two-factor correlational solution should be rejected. Study two also showed that religious commitment is a risk factor of intimate partner violence, whereas spirituality was a protecting factor from the violence. More alcohol was related with more abusive behaviors.

Chapter 2 – Study one

Within the IRT framework, three types of limited-information goodness-of-fit (GOF) statistics – M_2 , M_{ord} , and C_2 – have been proposed to assess global model-data fit when the data are sparse (Cai & Hansen, 2013; Cai & Monroe, 2014; Maydeu-Olivares & Joe, 2005, 2006). This simulation study aims to investigate the power and Type I error rate of M_2 , M_{ord} , and C_2 for the overall model-data fit among different data structures under multidimensional item response theory (MIRT) framework. The performance of RMSEA corresponding to M_2 , M_{ord} , and C_2 were also examined. Findings from the current study benefited practitioners and researchers who will use limited-information GOF statistics to evaluate global (scale-level) model-data fit and misfit in unidimensional and multidimensional IRT applications with dichotomous or polytomous data. In the following sections, the technical details related to the computation of the limited-information GOF statistics (M_2 , M_{ord} , and C_2) and their related RMSEAs were firstly introduced, followed by a literature review of the simulation studies that have evaluated the performance of the aforementioned GOF statistics and RMSEAs.

Limited-information GOF statistics. Limited-information GOF statistics, just as the name suggests, are computed from probabilities of certain response patterns. Instead of using all possible response patterns, limited-information GOF statistics use only part of the contingency table (e.g., the first-order probabilities and the second-order probabilities), a summary of part of the contingency table (e.g., sample means and cross-products), or a hybrid of the two (e.g., first-order probabilities and means and cross-products for the second-order probabilities).

To clarify how a limited-information GOF statistic works, an example is described. Suppose we have data collected from three items, each with two response categories coded 0 and 1. We can obtain $2^3 = 8$ possible response patterns. If we use $\pi_{i,j,k}$ as the probability of each response pattern, where $i (= 0, 1)$, $j (= 0, 1)$, and $k (= 0, 1)$ denote the categories for the 1st, 2nd and 3rd items respectively, then the 8 probabilities of all response patterns could be organized using a column vector as follows

$$\begin{pmatrix} \pi_{0,0,0} \\ \pi_{0,0,1} \\ \pi_{0,1,0} \\ \pi_{1,0,0} \\ \pi_{0,1,1} \\ \pi_{1,0,1} \\ \pi_{1,1,0} \\ \pi_{1,1,1} \end{pmatrix}. \quad (3)$$

Maydeu-Olivares and Joe (2005) suggested the population probabilities could also be organized using marginal probabilities, like first-order probabilities $\boldsymbol{\pi}_1$, second-order probabilities $\boldsymbol{\pi}_2$, and up-to-m-order probabilities $\boldsymbol{\pi}_m$, where m is the total number of items. In other words, $\boldsymbol{\pi}_1$ symbolizes the marginal probabilities when the participants correctly answered one item in a scale, $\boldsymbol{\pi}_2$ denotes the marginal probabilities when the participants correctly answered two items spontaneously, and $\boldsymbol{\pi}_m$ represents the marginal probabilities when the participants correctly answered all items. Considering the 3-item example from above, the probabilities of 8 possible response patterns could then be transformed into the following

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \\ \boldsymbol{\pi}_3 \end{pmatrix} = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \dot{\pi}_{1,2} \\ \dot{\pi}_{1,3} \\ \dot{\pi}_{2,3} \\ \dot{\pi}_{1,2,3} \end{pmatrix}, \quad (4)$$

where $\boldsymbol{\pi}$ (in bold) denotes all marginal probabilities, $\boldsymbol{\pi}_1$ (in bold) denotes all first-order marginal probabilities ($\dot{\pi}_1$, $\dot{\pi}_2$, and $\dot{\pi}_3$), $\boldsymbol{\pi}_2$ (in bold) denotes all second-order probabilities ($\dot{\pi}_{1,2}$, $\dot{\pi}_{1,3}$, and $\dot{\pi}_{2,3}$), and $\boldsymbol{\pi}_3$ denotes the third-order probability ($\dot{\pi}_{1,2,3}$). The column vector $\boldsymbol{\pi}$ could be calculated by multiplying the column vector of the response pattern probabilities by a $(2^m - 1) \times 2^m$ matrix with zeros and ones (Maydeu-Olivares & Joe, 2005, 2006).

The above calculation could also be generalized to items with more than two response categories. Suppose one scale includes a set of items with the same numbers of response categories, K , then the marginal probability vector $\boldsymbol{\pi}$ could be computed as the product of the column vector of the response pattern probabilities and a $(K^m - 1) \times K^m$ matrix containing zeros and ones. As such, the formula for the full-information GOF statistic Pearson's χ^2 using all marginal probabilities is

$$\chi^2 = N (\mathbf{P}_m - \hat{\boldsymbol{\pi}}_m)' \hat{\boldsymbol{\Sigma}}_m^{-1} (\mathbf{P}_m - \hat{\boldsymbol{\pi}}_m), \quad (5)$$

where N is the sample size (number of subjects), \mathbf{P}_m is a column vector of the observed marginal probabilities, $\hat{\boldsymbol{\pi}}_m$ is a column vector of the estimated marginal probabilities, and $N \hat{\boldsymbol{\Sigma}}_m^{-1}$ is the asymptotic covariance matrix of the observed marginal probabilities. In contrast, a limited-information GOF statistic uses only part of the marginal probabilities

generated from the items with a multivariate multinomial distribution.

M_2 . Maydeu-Olivares and Joe (2005, 2006) suggested the use of first-order marginal probabilities and second-order marginal probabilities for the fit of IRT models in order to get more accurate p values of the asymptotic χ^2 distribution under the null hypothesis and commonly larger power under the alternative hypothesis. When only the first- and second-order marginal probabilities are included in the estimation, the statistic used for detecting model-data fit is defined as

$$M_2 = N (\mathbf{P}_2 - \hat{\boldsymbol{\pi}}_2)' \hat{\mathbf{C}}_2 (\mathbf{P}_2 - \hat{\boldsymbol{\pi}}_2), \quad (6)$$

$$\hat{\mathbf{C}}_2 = \boldsymbol{\Xi}_2^{-1} - \boldsymbol{\Xi}_2^{-1} \Delta_2 (\Delta_2' \boldsymbol{\Xi}_2^{-1} \Delta_2)^{-1} \Delta_2' \boldsymbol{\Xi}_2^{-1}, \quad (7)$$

where Δ_2 is the matrix of all first-order partial derivatives of the marginal probabilities corresponding to the parameters of the model, and $N \hat{\boldsymbol{\Xi}}_2^{-1}$ is the asymptotic covariance matrix of the first- and second-order marginal probabilities. Equation 7 is a quadratic form of the subset of marginal probabilities. Maydeu-Olivares and Joe (2005, 2006) suggested that when all items share the same number of categories, M_2 asymptotically follows a χ^2 distribution with a degrees of freedom (df) equal to $m(K - 1) + \frac{m(m-1)}{2} (K - 1)^2 - q$, where q is defined as the number of parameters to be estimated by the model. M_2 belongs to the family of M_r ($M_1, M_2, M_3, \dots, M_m$). M_2 is calculated from first- and second-order marginal probabilities, whereas M_r is calculated from first- and up-to-order- r marginal probabilities. Degrees of freedom for M_r is the total number of multivariate marginal probabilities used for testing minus the number of estimated parameters. When ML is used, M_m equates with χ^2 .

M_{ord} . Cai and Hansen (2013) proposed M_2^* , also known as M_{ord} (Maydeu-Olivares

& Joe, 2006), to deal with the sparseness problem occurring in the second-order probabilities, which cannot be handled by M_2 , especially within MIRT models when the numbers of dimensions, items, and/or response categories are large. M_{ord} was also developed because estimating M_2 requires a large amount of computing capacity and is quite time-consuming. Although M_{ord} shares a similar quadratic form with M_2 , M_{ord} differs from M_2 in that it employs the means and cross-products of the multinomial items in the quadratic form, assuming all the categories are measured at an ordinal level. This permits M_{ord} to be more estimable since it avoids the sparseness problem in the second-order marginal probabilities. The following paragraph discusses the technical details of computing M_{ord} when the data is polytomous.

Suppose we have m items with K categories ranging from 0 to $K-1$, and we use \mathbf{k} to denote the sample vector of sample means and cross-products and $\boldsymbol{\kappa}$ as the population counterpart (in other words, $\boldsymbol{\kappa}$ is the mathematical expectation, or the expected value of all m items), then the mean and cross-product for a single item is

$$\boldsymbol{\kappa}_i = E [Y_i] = 0 \times \Pr(Y_i = 0) + 1 \times \Pr(Y_i = 1) + \dots + (K - 1) \times \Pr(Y_i = K - 1), \quad (8)$$

and the mean and cross-product for a pair of items is

$$\boldsymbol{\kappa}_{ij} = E [Y_i Y_j] = 0 \times \Pr(Y_i = 0) \times \Pr(Y_j = 0) + \dots + (K - 1) \times \Pr(Y_i = K - 1) \times \Pr(Y_j = K - 1). \quad (9)$$

Computation of the means and cross-products assumes items should be measured at an ordinal level. The matrix $\boldsymbol{\kappa}$ contains means and cross-products of single items and pairs of items in a scale. Reducing the marginal probabilities to just mean and cross-products dramatically aggregates the information from single items and pairs of items,

which in a point resolves the computational burden and also the sparseness problem occurring at the second-order marginal probability level when the contingency table is large. M_{ord} is defined as

$$M_{ord} = N(\mathbf{k} - \hat{\boldsymbol{\kappa}})' \hat{\mathbf{C}}_{ord} [\mathbf{k} - \hat{\boldsymbol{\kappa}}], \mathbf{C}_{ord} = \boldsymbol{\Xi}_{ord}^{-1} - \boldsymbol{\Xi}_{ord}^{-1} \Delta_{ord} (\Delta_{ord}' \boldsymbol{\Xi}_{ord}^{-1} \Delta_{ord})^{-1} \Delta_{ord}^{-1} \boldsymbol{\Xi}_{ord}^{-1}, \quad (10)$$

where N is the sample size, $\hat{\boldsymbol{\Xi}}_{ord}^{-1}$ is the asymptotic covariance matrix of the means and cross-products divided by N , and Δ_{ord} is the matrix of partial derivatives of parameters θ , which is similar to that in M_2 . M_{ord} also follows an asymptotically χ^2 distribution with $df = m(m+1)/2 - q$. Of note, when the data is binary for all items, $M_2 = M_{ord}$.

C_2 . Cai and Monroe (2014) proposed a new GOF statistic C_2 for ordinal data as a remedy to a problem that plagues M_{ord} in that oftentimes it is impossible to compute M_{ord} due to a lack of df . For example, for a unidimensional model, if the items are scored using five response categories, the minimum number of items needed to compute M_{ord} is 10. As such, in applied settings, M_{ord} cannot be used to estimate the overall model-data fit for a scale such as the Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985; Pavot & Diener, 1993; Pavot, Diener, Colvin, & Sandvik, 1991), which includes five items with seven categories. Instead, the C_2 statistic is calculated using the first-order probabilities and means and cross-products of the second-order probabilities. As a result, Cai and Monroe called C_2 as a hybrid of M_2 (which uses first- and second-order marginal probabilities) and M_{ord} (which only uses means and cross-products for both first- and second-order probabilities). After a careful examination of sparseness, Cai and Monroe concluded that most sparseness issues don't happen in the first-order

probabilities. Therefore, aggregating both first- and second-order probabilities into means and cross-products is too aggressive. By releasing the first-order probabilities, C_2 perfectly solves the problem of lack of degrees of freedom.

C_2 is defined as

$$C_2 = N(\mathbf{r})' \hat{\boldsymbol{\pi}}(\mathbf{r}), \quad \boldsymbol{\pi} = \boldsymbol{\Xi}^{-1} - \boldsymbol{\Xi}^{-1} \Delta (\Delta' \boldsymbol{\Xi}^{-1} \Delta)^{-1} \Delta^{-1} \boldsymbol{\Xi}^{-1}, \quad (11)$$

where \mathbf{r} is a $m^{(K-1)} + m \times (m-1)/2$ dimensional matrix which contains the first-order residual marginal probabilities and the aggregated second-order residual marginal probabilities, $\boldsymbol{\Xi}^{-1}$ is the covariance matrix divided by N , and Δ is the matrix of the partial derivatives of \mathbf{r} with respect to the parameters, $\boldsymbol{\theta}$. Using ML, C_2 approximately follows the χ^2 distribution, with a $df = m(K-1) + m(m-1)/2 - q$.

Akin to M_{ord} , C_2 also assumes the data include ordinal categories. The number assigned to each category is the weight to be used for aggregation of the second-order marginal probabilities in Equation 11. Thus, C_2 is recommended for data with ordinal response categories, small number of items, and large number of categories.

To date, the research related to the limited information approach in the field of IRT has been quite fruitful, both theoretically and empirically. Theoretically, many researchers are working on examining the behavior (Type I error rate, power, cutoff values for practice, and asymptotic relative efficiency) of such fit statistics and approximate fit indexes under various conditions. Empirically, more and more researchers have adopted M_2 or M_{ord} in their own research when using IRT methods.

Researchers have investigated the Type I error rate and power of M_2 under: (a) unidimensional IRT models with binary data (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2005; Xu, Paek, & Xia, 2017), (b) unidimensional IRT models with polytomous data

(3-5 categories; Cai & Hansen, 2013; Cai & Monroe, 2014; Maydeu-Olivares & Joe, 2006), (c) correlated MIRT models using binary data (2-5 dimensions; Jurich, 2014; Xu, Paek, & Xia, 2017), and (d) bifactor IRT models using binary (Xu, Paek, & Xia, 2017; Cai & Hansen, 2013) and polytomous data (3-5 categories; Cai & Hansen, 2013). However, none of the aforementioned studies examined M_2 under correlated MIRT models with multiple categories (3-5 categories).

Researchers have investigated the Type I error rate and power of M_{ord} under unidimensional and bifactor GR models using binary and polytomous data (2-5 categories; Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2014). Cai and Hansen found that M_{ord} was better calibrated with polytomous data and more powerful than M_2 in detecting misspecified bifactor models. Maydeu-Olivares and Joe (2014) also reported M_{ord} has more power to reject misspecified unidimensional models compared to M_2 when the items are polytomous (3-4 categories). Similar to M_2 , few studies have examined the performance of M_{ord} when the scale includes many dimensions and items are polytomous. As to C_2 , only Cai and Monroe (2014) have examined the performance of C_2 and reported C_2 is more powerful than M_2 and M_{ord} to detect model misspecification under a unidimensional IRT model when there are four response categories.

Approximate Fit Statistics. In reality, few IRT models can perfectly reproduce the observed data, especially when the df is larger than two. Thus, indicators of goodness/badness of models is quite necessary since all models are imperfect. Maydeu-Olivares and Joe (2014) suggested the use of RMSEA and the standardized root mean squared residual (SRMSR) as indicators of approximate fit. Study one primarily focused on the corresponding RMSEAs as they relate to the aforementioned GOF statistics.

RMSEA. In practical terms, RMSEA measures the approximate error of discrepancy per df. As Maydeu-Olivares (2013) put it, “an imperfect model doesn’t indicate a useless model”. Although the above-mentioned GOF statistics could provide us a general idea whether the model fit the data or not, they could not inform us how good or bad the model is. This encourages researchers to use the RMSEA and the corresponding GOF statistics jointly to obtain a better understanding of the overall model-data fit.

Historically, RMSEA is defined as

$$\hat{\varepsilon} = \max\left(\sqrt{\frac{\hat{F}}{df}}, 0\right) = \max\left(\sqrt{\frac{\chi^2 - df}{N \times df}}, 0\right), \quad (12)$$

where $\hat{F} = F - df/N$ (Browne & Cudeck, 1993). \hat{F} is an estimate of the unbiased discrepancy between the population and the null model (Steiger & Lind, 1980).

Depending on different cutoff values, in general, larger values of RMSEA is an indication of possible misspecification and lower values or those below the cutoff values are indications that not enough evidence is presented to refute the current IRT model to the data.

Maydeu-Olivares and Joe (2014) also introduced how to compute population RMSEA in their paper for both binary data and polytomous data. The population RMSEAs could be obtained by selecting the IRT parameter vectors that minimize the F function in Equation 12. The population RMSEA is computed as follows,

$$\varepsilon_0 = \sqrt{\frac{F}{df}}. \quad (13)$$

In reality, population RMSEA is usually obtained through the multinomial ML

discrepancy function between the population probability vector and the misspecified model under the null model. For example, if we simulate the data using a 2PL model and we fit the model with a 1PL model using IRT models estimated by ML, we could obtain the population RMSEA by choosing the 1PL model parameter vector that minimize the F function.

Since RMSEA is a point estimate and the 90% confidence interval (CI) of RMSEA has been taken as a solid supplement of RMSEA in the structural equation modeling (SEM) literature suggested by Steiger (2007), Maydeu-Olivares and Joe (2014) extended the 90% CI of RMSEA to IRT. A 90% CI for RMSEA based on χ^2 is calculated as

$$\left(\sqrt{\frac{\mathcal{L}}{N \times df}}, \sqrt{\frac{\mathcal{U}}{N \times df}} \right). \quad (14)$$

Here \mathcal{L} and \mathcal{U} are the roots of the non-central distribution of χ^2 with the df used for the test. Specifically,

$$F_{\chi^2}(\chi^2; df, \mathcal{L}) = .95, \quad (15)$$

and

$$F_{\chi^2}(\chi^2; df, \mathcal{U}) = .05. \quad (16)$$

The df for Equations 13 through 16 is $K^m - q - 1$, where K is defined as number of categories, m is the number of items, and q is the number of parameters to be estimated by the model. We can substitute the χ^2 and df in Equations 13 through 16 by any type of the limited information GOF statistic and associated df to get the corresponding RMSEA and 90% CI. In particular, when M_2 is used, we replace χ^2 in the above formulae by M_2 to compute the sample bivariate RMSEA ($RMSEA_2$). $RMSEA_2$ is calculated as follows

$$\text{RMSEA}_2 = \sqrt{\frac{M_2 - df}{N \times df}} \quad (17)$$

and the associated 90% CI is computed as

$$\left(\sqrt{\frac{\mathcal{L}}{N \times df}}, \sqrt{\frac{\mathcal{U}}{N \times df}} \right), \quad (18)$$

where $F_{M_2}(M_2; df, \mathcal{L}) = .95$ and $F_{M_2}(M_2; df, \mathcal{U}) = .05$. Here df is $m(K - 1) + \frac{m(m-1)}{2} (K - 1)^2 - q$.

Similarly, when M_{ord} is being used, RMSEA_{ord} and its 90% CI are calculated as

$$\text{RMSEA}_{ord} = \sqrt{\frac{M_{ord} - df}{N \times df}} \quad (19)$$

and the associated 90% CI is computed as,

$$\left(\sqrt{\frac{\mathcal{L}}{N \times df}}, \sqrt{\frac{\mathcal{U}}{N \times df}} \right), \quad (20)$$

where $F_{M_{ord}}(M_{ord}; df, \mathcal{L}) = .95$, $F_{M_{ord}}(M_{ord}; df, \mathcal{U}) = .05$, and $df = m(m+1)/2 - q$.

Following the same logic, by replacing M_{ord} by C_2 in Equations 19 and 20, RMSEA and its 90% CI for C_2 could be obtained (Cai & Monroe, 2014) as follows

$$\text{RMSEA}_{C_2} = \sqrt{\frac{C_2 - df}{N \times df}} \quad (21)$$

and the associated 90% CI is computed as

$$\left(\sqrt{\frac{\mathcal{L}}{N \times df}}, \sqrt{\frac{\mathcal{U}}{N \times df}} \right), \quad (22)$$

where $F_{C_2}(C_2; df, \mathcal{L}) = .95$, $F_{C_2}(C_2; df, \mathcal{U}) = .05$, and $df = m(K-1) + m(m-1)/2 - q$.

Researchers have proposed and investigated the use of RMSEA associated with M_2 and M_{ord} under misspecified unidimensional IRT models (Mayedu-Olivares & Joe, 2014) and MIRT models (Jurich, 2014). In particular, Mayedu-Olivares and Joe recommend the use of $RMSEA_2 \leq .05$ as a cutoff for detecting misspecified unidimensional GR models with binary data and $RMSEA_2 \leq .05 / (K-1)$ as a cutoff for misspecified unidimensional models when the data is polytomous. However, when it comes to $RMSEA_{ord}$ under wrongly specified unidimensional IRT models with polytomous data, the relationship between $RMSEA_{ord}$, number of items, and number of categories were not clear. For example, RMSEA increases when the number of categories increases from 2 to 3, but decreases when the number of categories increases from 3 to 4. Also, RMSEA decreases if the number of items goes up with $K = 2$ or 4, but when there are 3 categories, RMSEA first increase if the number of items goes up and then decrease when the number of items goes up). Thus, Mayedu-Olivares and Joe did not offer any benchmarks for the use of $RMSEA_{ord}$. Jurich (2014) examined the performance of $RMSEA_2$ for misspecified correlated MIRT models (2-4 dimensions; $\rho = .50, .80$) using binary data and suggested .04 as a cutoff for dimension misspecification if the intercorrelation is .50 and .035 to .04 be used if the intercorrelation is .80. Currently, no studies have examined the rejection rate for RMSEAs associated with M_2 or M_{ord} with polytomous data to detect misspecified MIRT models. Nor have studies examined the performance of $RMSEA_{C2}$ of misspecified unidimensional models with polytomous data.

Purpose of Current Study. A growing body of literature has evaluated the performance of the aforementioned limited-information GOF statistics and related approximate fit indices. However, studies have not yet examined such statistics for

polytomous data when the null model is a correlated MIRT model. The main purpose of this simulation study aims to investigate the Type I error rate and power of M_2 , M_{ord} , and C_2 and associated RMSEAs for polytomous data under MIRT models. In particular, the aims of the current study were to evaluate the: (a) Type I error rates of M_2 , M_{ord} , and C_2 under correctly specified unidimensional and multidimensional IRT models, (b) power of these three statistics under incorrect model specifications, and (c) the rejection rate of RMSEAs when the null model is correct/incorrect unidimensional or multidimensional with categorical data. The study also explored if a general guideline could be established for evaluating the RMSEAs under MIRT models that functions adequately under a variety of conditions.

Method

Simulation design. To study Type I error rate, the following six independent variables were manipulated: (1) number of dimension ($D = 1, 2, 3$); (2) number of response categories ($K = 2-5$); (3) sample size ($N = 300, 1000, 3000, 5000$); (4) interfactor correlation ($\rho = .20, .80$). Crossing conditions results in a 4×4 factorial design for unidimensional IRT models and $2 \times 4 \times 4 \times 2$ factorial design for correlated IRT models. The total number of unique conditions is 80. To maintain manageability, two constants were included: number of items for each dimension (5) and number of replications (1000). A 5-item per dimension test was selected to better reflect the length of typical instruments used in educational psychology (e.g., the 10-item Rosenberg Self-Esteem Scale; Rosenberg, 1965) and public health (e.g., the 10-item Religious Commitment Inventory; Worthington et al., 2003). A unidimensional IRT model was included in the simulation in order to have a complete evaluation of C_2 and $RMSEA_{C_2}$

under various dimensions, response categories, and sample sizes. Sample sizes were examined at four levels to represent smaller sample sizes ($N = 300$) and relevantly moderate sample sizes ($N = 5000$) for MIRT models. Sample sizes of 300, 1000, and 3000 were selected to demonstrate Type I error and power of the GOF statistics and RMSEAs with short tests, extending the studies by previous researchers to MIRT models (Cai & Hansen, 2013; Cai & Monroe, 2014; Maydeu-Olivares & Joe, 2005, 2006, 2014; Mayedu-Olivares & Montãno, 2013; $N = 100, 300, 500, 750, 1000, 1200, 1500, 3000$). The correlations among latent dimensions were manipulated at two different levels to better represent the correlations among dimensions, to represent an ideal situation when the dimensions are distinct ($\rho = .20$) or hard to differentiate ($\rho = .80$).

To study power, model misspecification were introduced by misspecifying dimensionality. Particularly, the null unidimensional models were fitted to data generated from the alternative two-dimensional or three-dimensional models. The following factors were manipulated to study power: (1) misspecified dimensions/factors ($F = 2, 3$); (2) number of response categories ($K = 2-5$); (3) sample size ($N = 300, 1,000, 3000, 5000$); (4) interfactor correlation ($\rho = .20, .80$). A total of 64 conditions were examined. The same configuration of slopes and intercepts used in the first simulation that examined the Type I error rates of the three statistics were used.

Calibrations. After generating data, item parameters were estimated by fitting either a unidimensional or multidimensional IRT model using the Bock-Aitkin EM algorithm (Bock & Aitkin, 1981). Although Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2010) algorithm could dramatically improve the calculation efficiency for multidimensional models, Bock-Aitkin EM algorithm was selected to overcome the

confounding effect of different algorithms on the model estimation. The default maximum number of EM cycles (500) and the default convergence criterion (0.0001) used in FlexMIRT were employed. The FlexMIRT default quadrature points (21) and the range over which the points could be spread (-4.0 to 4.0) are selected. A same dataset was repeated fitted three times separately to obtain the M_2 , M_{ord} , and C_2 statistics from FlexMIRT output files. Given that many times RMSEA values are close to 0 and FlexMIRT only printed RMSEA with two decimal places which is not precise enough to provide cutoff criterion, RMSEA values were computed using the M_2 , M_{ord} , and C_2 statistics, their corresponding dfs and samples sizes based on Equations 18, 20, and 22.

Table 2.1

Generating Item Parameters

Item	Slope β	$K=2$	$K=3$		$K=4$			$K=5$			
		α_1	α_1	α_2	α_1	α_2	α_3	α_1	α_2	α_3	α_4
1	0.882	0.095	-1.532	-2.186	1.518	1.378	-2.076	0.460	-0.519	-0.607	-2.604
2	1.037	-0.62	2.967	1.285	2.745	2.307	-2.268	2.715	0.027	-0.356	-0.379
3	0.846	0.033	0.208	0.116	0.558	0.111	-0.722	0.869	0.835	0.079	-0.132
4	1.376	2.59	1.384	-1.918	-2.289	-3.106	-3.501	2.696	0.387	-1.847	-4.004
5	1.068	-2.098	-0.967	-1.235	-0.633	-1.787	-2.952	1.041	-0.003	-0.049	-1.633
6	0.849	-0.858	1.232	0.114	0.333	-1.514	-2.425	1.625	0.235	-0.664	-0.931
7	1.102	-1.946	0.142	-2.389	0.296	-1.406	-1.533	-0.074	-0.196	-0.672	-2.253
8	1.159	2.727	0.431	-1.852	0.996	0.696	0.625	2.954	2.034	1.816	-2.477
9	1.122	-1.506	2.799	1.718	1.176	0.461	-2.093	1.827	1.628	1.496	0.731
10	0.941	0.501	2.423	-2.119	1.986	-0.592	-1.451	1.622	0.675	-0.54	-1.294
11	1.353	-2.606	3.252	1.306	3.953	2.414	-1.256	1.747	0.385	-0.608	-3.21
12	1.081	-0.954	1.192	-2.202	0.952	-1.369	-1.398	2.743	2.107	-0.936	-2.563
13	0.883	-1.499	0.812	-0.099	2.103	2.005	1.217	2.461	0.285	-1.277	-1.307
14	0.642	-0.204	0.640	-0.624	0.981	0.207	-1.898	1.522	-0.406	-0.55	-1.079
15	1.252	-0.223	0.700	0.178	2.680	-1.003	-1.053	-2.739	-2.86	-3.029	-3.221

Note. The same intercepts were used for all conditions with a same number of response categories.

Type I error rates. Convergence rates were calculated and unconverged replications were excluded from subsequent analysis. Mean and standard deviations of the three statistics were reported. Since M_2 , M_{ord} , and C_2 are theoretically approximately follow χ^2 distribution, the means of these three statistics could be examined with respect to the degrees of freedom. A large discrepancy between the mean and the degrees of freedom indicates a high probability of rejecting the null. The observed Type I error rates of M_2 , M_{ord} , and C_2 statistics were also examined at three α nominal levels: .01, .05, and .10. The comparison between the observed Type I error rates and the nominal levels could determine how well these statistics approximate the χ^2 distribution across the manipulated simulation conditions. The two-sided Kolmogorov-Smirnov (KS) tests were utilized to examine if any of the statistics fail to follow a reference χ^2 distribution. A $p > .05$ for the KS test indicate a specific GOF statistics are well-calibrated (Cai & Hansen, 2013; Cai & Monroe, 2014). Once M_2 , M_{ord} , and C_2 were determined to be approximately χ^2 distributed, the power of M_2 , M_{ord} , and C_2 then was compared.

Power to detect model misspecification. Means and dfs related to the M_2 , M_{ord} , and C_2 statistics were reported. The observed rejection rates of M_2 , M_{ord} , and C_2 statistics were examined at three different significant levels: .01, .05, and .10 to determine how powerful these statistics are when the models were misspecified.

Performance of RMSEA. Performance of RMSEAs was firstly examined when the generating and fitted models were the same. Means of $RMSEA_2$, $RMSEA_{ord}$, and $RMSEA_{C_2}$ were reported. When the alternative model was true, mean RMSEA and a population RMSEA values were reported for each condition. Next, the discrepancy between the mean of observed (sample) RMSEA values and the population RMSEA

values were examined. If these two numbers are close to each other, then the sample estimation of the RMSEA is consistent with its population RMSEA. Performance of $RMSEA_2$ was also assessed by comparing with recommendations given in Maydeu-Olivares and Joe (2014) and Jurich (2014). Following strategies used in Jurich (2014), this study compared model rejection rates at various RMSEA cut-off values in an attempt to establish general guidelines for evaluating model fit.

Results

Convergence results and data cleaning. Across 80 conditions, 71 conditions had acceptable convergence rates ranging from 97.6% to 100% and seven conditions had poor convergence rates that were less than 67.0% (Table 2.2). Conditions with poor convergence rates were connected to extremely small sample size ($N = 300$) and a large number of categories (i.e., unidimensional model with five categories, two-dimensional model with five categories, or three-dimensional model with four or five categories). A data frequency check showed that some nonconverged data sets occurred within conditions when an item had fewer response categories than the model expected (e.g., the population model consisted of a five-response category format, but the simulated data set resulted in zero frequencies for one category). Consequently, the nonconverged data sets were excluded from subsequent analyses.

In addition, when the number of categories was four or five and the generating model was multidimensional, negative M_{ord} values were observed. In fact, 0.1% to 5.5% of the M_{ord} values were negative when evaluating Type I error rates of M_{ord} and 0.1% to 15.4% of the M_{ord} values were negative in the when evaluating the power of M_{ord} (see Appendix A). Because negative M_{ord} values are not theoretically possible (i.e., the M_{ord}

statistic is expected to be positive and follow a χ^2 sampling distribution), these inadmissible solutions were excluded from subsequent data analyses.

Table 2.2

Convergence Rates Across Simulation Conditions

Dimension	Correlation	No. of Categories	Convergence rate				
			$N = 300$	$N = 1000$	$N = 3000$	$N = 5000$	
1	-	2	100	100	100	100	
	-	3	99.8	100	100	100	
	-	4	99.7	100	100	100	
	-	5	59.1	99.5	100	100	
2	0.2	2	100	100	100	100	
		3	99.7	100	100	100	
		4	97.8	100	100	100	
		5	60.2	98.6	100	100	
		2	100	100	100	100	
	0.8	3	99.4	100	100	100	
		4	97.6	100	100	100	
		5	58.1	99.3	100	100	
		3	2	100	100	100	100
			0.2	3	99.6	100	100
4	62.8			98.7	100	100	
5	42.1			97.8	100	100	
2	100			100	100	100	
0.8	3	100	100	100	100		
	4	66.8	99.1	100	100		
	5	40.6	98.2	100	100		

Type I error rates. Tables 2.3 and 2.4 present the simulation results for M_2 , M_{ord} , and C_2 under the null model, when the generating model and analysis model match. When working with the null model it is important to determine that the test statistics follow the expected sampling distribution and that the rejection rate (Type I error rate) matches the nominal alpha level. First, the p values associated with the KS tests for examining the distribution of the M_2 and C_2 statistics were greater than the nominal alpha level (See Appendix B), which means the observed M_2 and C_2 statistics followed the expected χ^2 distribution and the mean of the statistic can be meaningfully compared to its

associated df . However, this is not true for M_{ord} when there were five response categories. Thus, from this point onward, the results of M_{ord} for conditions with five response categories under the correct model specification were not summarized. Second, a comparison between the mean M_2 , M_{ord} , and C_2 and respective df shows that the discrepancies were negligible. That is, the relative discrepancy for the M_2 , M_{ord} , and C_2 statistics (i.e., $\frac{GOF-df}{df} = \frac{\bar{M}_2-df}{df}$) across all study conditions ranged from $< .001$ to $.05$ with means of $.005$, $.007$, and $.009$, respectively. Across all conditions, the M_2 and C_2 statistics maintained the nominal Type I error rates (Table 2.4). Similarly, Type I error rates associated with the M_{ord} statistics followed closely with the nominal alpha levels for conditions with two to four categories. When comparing the Type I error rates of M_2 and C_2 , C_2 maintained a better Type I error rates compared to M_2 with more categories (i.e., four or five response categories) under multidimensional models.

Figure 2.1 displays the differences between the observed and expected quantiles of the M_2 , M_{ord} , and C_2 statistics. Across all conditions both M_2 and C_2 aligned closely with the expected χ^2 distribution. However, for conditions with five response categories, M_2 began to drift away from the diagonal (i.e., the degrees of freedom for that specific condition) of the quantile-quantile plots, indicating that the sparseness (i.e., lack of marginal probabilities) problem occurring at higher-order marginal probabilities might start to influence the performance of M_2 . As to M_{ord} , it performed well across conditions with three response categories. However, for some conditions with four or five response categories, M_{ord} fell under the diagonal of the quantile-quantile plots, indicating many of the M_{ord} values were larger than the expected χ^2 value. To summarize, when the

generating model and analysis model matched (null model), all three GOF statistics followed a χ^2 distribution and maintained the Type I error rate except for M_{ord} when the number of categories was four or five.

Table 2.3

Means of M_2 , M_{ord} , and C_2 and Associated Degrees of Freedom by Condition: Null Model

D	ρ	K	M_2					M_{ord}					C_2						
			300	1000	3000	5000	<i>df</i>	300	1000	3000	5000	<i>df</i>	300	1000	3000	5000	<i>df</i>		
1	-	2	4.75	5.04	5.00	4.90	5	4.75	5.04	5.00	4.90	5	4.75	5.04	5.00	4.90	5		
	-	3	34.80	34.53	34.97	34.84	35	-	-	-	-	0	4.80	4.97	5.15	4.92	5		
	-	4	84.93	84.89	84.96	85.43	85	-	-	-	-	-	4.88	4.97	5.08	4.97	5		
	-	5	156.11	153.89	155.44	155.14	155	-	-	-	-	-	5.01	5.01	5.02	4.76	5		
2	0.2	2	34.14	33.53	33.98	34.47	34	34.14	33.53	33.98	34.47	34	34.14	33.53	33.98	34.46	34		
		3	170.23	168.52	167.57	169.04	169	24.18	24.00	23.72	23.80	24	34.14	34.07	33.71	33.60	34		
		4	395.42	393.44	392.58	394.95	394	14.12	14.12	14.46	15.46	14	33.91	33.78	33.87	34.05	34		
		5	710.95	709.26	709.97	712.11	709	12.51	11.49	11.37	11.09	4	34.70	34.09	33.49	33.78	34		
		2	34.16	33.57	33.83	34.52	34	34.16	33.57	33.83	34.52	34	34.16	33.57	33.83	34.52	34		
	0.8	3	171.26	169.40	169.76	169.47	169	24.40	24.03	24.00	24.07	24	34.56	34.15	34.20	33.95	34		
		4	395.69	394.80	394.13	394.01	394	14.11	14.03	14.34	14.21	14	34.26	34.16	33.90	34.03	34		
		5	713.98	711.99	708.47	709.06	709	6.92	7.02	6.69	6.45	4	34.65	34.40	33.40	33.80	34		
		3	0.2	2	86.66	87.43	86.70	87.78	87	86.66	87.43	86.70	87.78	87	86.66	87.43	86.70	87.78	87
				3	404.21	402.76	402.49	403.01	402	72.73	71.82	71.91	72.38	72	87.53	86.90	86.71	87.19	87
4	930.03			928.06	926.21	927.89	927	57.62	57.29	57.86	57.01	57	87.74	87.72	87.18	86.81	87		
5	1666.28			1662.20	1663.14	1661.35	1662	49.99	54.59	54.69	51.67	42	87.33	87.48	87.16	87.32	87		
2	87.44			88.07	86.96	86.88	87	87.44	88.07	86.96	86.88	87	87.44	88.07	86.96	86.88	87		
0.8	3		404.17	403.61	402.34	401.63	402	72.60	71.94	71.97	71.88	72	87.58	86.91	86.88	86.99	87		
	4		929.62	926.79	929.03	929.76	927	57.81	56.36	57.23	57.31	57	88.28	86.71	87.25	86.84	87		
	5		1671.47	1666.70	1661.94	1662.27	1662	43.27	45.08	42.04	41.91	42	88.09	87.85	87.15	87.32	87		

Note. *df* = degrees of freedom. *D* = dimension. ρ = correlation among dimensions. *K* = response category. Samples sizes are in bold. M_{ord} could not be obtained for data generated from unidimensional models when the categories were 3, 4, and 5 due to lack of *df*.

Table 2.4

*Type I Error Rates by Conditions: Null Model**a) Compared to .01 Nominal Alpha Level*

Dimension	Correlation	No. of Categories	M_2				M_{ord}				C_2			
			300	1000	3000	5000	300	1000	3000	5000	300	1000	3000	5000
1	-	2	.012	.008	.010	.006	.012	.008	.010	.006	.012	.008	.010	.006
	-	3	.014	.010	.013	.010	-	-	-	-	.008	.008	.018	.006
	-	4	.023	.012	.009	.009	-	-	-	-	.009	.012	.015	.009
	-	5	.039	.035	.026	.014	-	-	-	-	.017	.009	.010	.010
2	0.2	2	.006	.009	.010	.010	.006	.009	.010	.010	.006	.009	.010	.010
		3	.009	.006	.011	.013	.009	.004	.012	.008	.008	.010	.011	.007
		4	.021	.009	.013	.018	.011	.013	.019	.022	.006	.009	.013	.011
		5	.040	.014	.011	.014	.272	.283	.287	.271	.017	.009	.005	.008
	0.8	2	.014	.008	.020	.009	.014	.008	.020	.009	.014	.008	.020	.009
		3	.008	.008	.014	.010	.013	.010	.019	.011	.013	.006	.020	.007
		4	.030	.009	.014	.011	.006	.007	.013	.011	.008	.010	.011	.005
		5	.041	.022	.015	.013	.053	.048	.048	.047	.005	.008	.013	.009
3	0.2	2	.003	.014	.012	.012	.003	.014	.013	.012	.003	.014	.013	.012
		3	.016	.012	.006	.007	.014	.008	.007	.011	.016	.011	.005	.012
		4	.024	.025	.013	.009	.013	.009	.008	.013	.013	.013	.011	.009
		5	.043	.018	.013	.005	.074	.098	.100	.093	.012	.011	.004	.013
	0.8	2	.009	.012	.011	.007	.009	.012	.011	.007	.009	.012	.011	.007
		3	.010	.007	.009	.010	.012	.007	.012	.011	.014	.010	.011	.016
		4	.030	.017	.014	.022	.007	.006	.009	.010	.013	.009	.019	.008
		5	.067	.027	.014	.009	.020	.012	.016	.010	.015	.013	.009	.013

Note. df = degrees of freedom. D = dimension. ρ = correlation among dimensions. K = response category. Samples sizes are in bold.

Table 2.4 (continued)

b) Compared to .05 Nominal Alpha Level

Dimension	Correlation	No. of Categories	M_2				M_{ord}				C_2			
			300	1000	3000	5000	300	1000	3000	5000	300	1000	3000	5000
1	-	2	.046	.047	.046	.036	.046	.047	.046	.036	.046	.047	.046	.036
	-	3	.057	.033	.057	.041	-	-	-	-	.046	.048	.062	.035
	-	4	.066	.057	.046	.050	-	-	-	-	.044	.051	.053	.044
	-	5	.107	.090	.075	.055	-	-	-	-	.061	.054	.037	.040
2	0.2	2	.047	.049	.048	.055	.047	.049	.048	.055	.047	.049	.048	.055
		3	.056	.050	.036	.053	.052	.045	.054	.038	.041	.045	.052	.031
		4	.072	.045	.045	.050	.043	.051	.073	.062	.056	.066	.043	.056
		5	.098	.061	.061	.056	.553	.525	.561	.533	.065	.045	.037	.055
	0.8	2	.047	.033	.053	.047	.047	.033	.053	.047	.047	.033	.053	.047
		3	.072	.048	.055	.044	.058	.039	.057	.050	.053	.052	.066	.056
		4	.073	.053	.060	.053	.054	.041	.064	.055	.052	.053	.054	.045
		5	.119	.073	.058	.057	.140	.128	.134	.145	.055	.044	.051	.051
3	0.2	2	.045	.047	.052	.061	.045	.047	.052	.061	.045	.047	.052	.061
		3	.062	.051	.055	.043	.058	.041	.048	.064	.057	.043	.039	.051
		4	.083	.072	.056	.042	.051	.053	.056	.049	.072	.044	.041	.042
		5	.109	.075	.069	.049	.189	.247	.249	.259	.050	.063	.048	.054
	0.8	2	.050	.053	.045	.050	.050	.053	.045	.050	.050	.053	.045	.050
		3	.056	.053	.047	.054	.058	.042	.050	.049	.062	.043	.049	.051
		4	.072	.061	.050	.055	.064	.047	.056	.061	.072	.055	.053	.044
		5	.126	.084	.049	.052	.074	.070	.057	.047	.076	.057	.061	.048

Note. D = dimension. ρ = correlation among dimensions. K = response category. Samples sizes are in bold.

Table 2.4 (continued)

c) Compared to .10 Nominal Alpha Level

Dimension	Correlation	No. of Categories	M_2				M_{ord}				C_2			
			300	1000	3000	5000	300	1000	3000	5000	300	1000	3000	5000
1	-	2	.086	.092	.106	.095	.086	.092	.106	.095	.086	.092	.106	.095
	-	3	.107	.075	.011	.087	-	-	-	-	.077	.100	.109	.090
	-	4	.108	.107	.102	.098	-	-	-	-	.075	.099	.101	.089
	-	5	.139	.141	.129	.109	-	-	-	-	.103	.110	.084	.080
2	0.2	2	.098	.089	.105	.100	.098	.089	.105	.100	.098	.089	.105	.100
		3	.012	.096	.074	.086	.100	.100	.101	.079	.112	.097	.110	.072
		4	.119	.095	.086	.096	.097	.096	.123	.116	.101	.095	.095	.097
		5	.146	.118	.121	.110	.667	.691	.703	.682	.106	.086	.094	.105
	0.8	2	.093	.092	.104	.103	.093	.092	.104	.103	.093	.092	.104	.103
		3	.013	.095	.107	.095	.109	.089	.103	.106	.115	.096	.112	.106
		4	.128	.111	.118	.095	.101	.098	.119	.111	.114	.095	.096	.093
		5	.165	.122	.100	.104	.200	.196	.195	.223	.102	.101	.094	.099
3	0.2	2	.088	.101	.098	.108	.088	.101	.098	.108	.088	.101	.098	.108
		3	.110	.100	.115	.091	.105	.087	.100	.112	.103	.080	.096	.111
		4	.135	.113	.096	.102	.121	.098	.118	.091	.132	.099	.097	.099
		5	.162	.139	.126	.102	.267	.331	.373	.377	.121	.126	.098	.106
	0.8	2	.091	.122	.090	.101	.091	.122	.090	.101	.091	.122	.090	.101
		3	.118	.091	.107	.106	.110	.086	.106	.087	.103	.098	.096	.107
		4	.120	.102	.106	.106	.129	.098	.103	.115	.114	.091	.101	.088
		5	.177	.144	.096	.103	.124	.119	.096	.100	.121	.108	.102	.106

Note. D = dimension. ρ = correlation among dimensions. K = response category. Samples sizes are in bold.

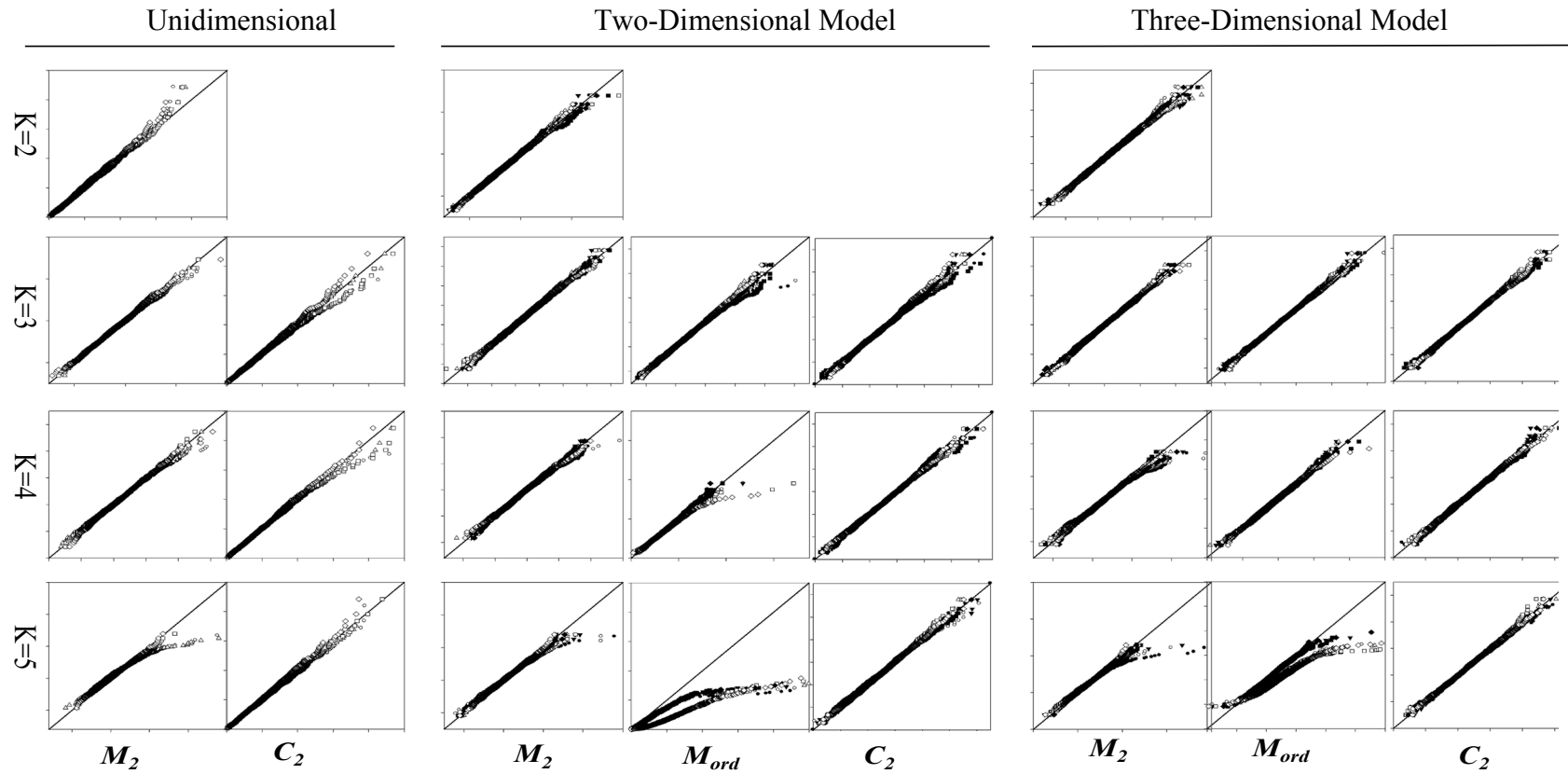


Figure 2.1. Quantile-quantile plots of observed M_2 , M_{ord} , and C_2 values and their reference χ^2 distribution (degrees of freedom shown in Table 3). X-axis is the observed quantiles and Y-axis is the expected χ^2 quantiles. The 45 degree (diagonal) line is the expected df.

Power to detect model misspecification. Tables 2.5 and 2.6 present the simulation results of M_2 , M_{ord} , and C_2 detecting dimensionality misspecification by fitting unidimensional models to data generated from multidimensional models. An immediate finding was that M_2 , M_{ord} , and C_2 tended to be more powerful at detecting misspecification when the sample size increased. When sample sizes were 300, M_2 , M_{ord} , and C_2 were not powerful enough at detecting misspecified models. When sample sizes increased to 1000, M_2 , M_{ord} , and C_2 became more powerful, especially for conditions involving a high level of misspecification (i.e., when the interdimension correlation of the generating model was 0.2). When sample size increased to 3000 and 5000, M_2 , M_{ord} , and C_2 showed close to perfect power (i.e., power = 1.00) for detecting the wrong model across all conditions.

As expected, the level of misspecification also influences the performance of M_2 , M_{ord} , and C_2 . M_2 , M_{ord} , and C_2 have high statistical power to detect misspecification when the interdimension correlation of the generating model was 0.2, but when the interdimensional correlaton of the generating model was 0.8, all three GOF statistics had lower statistical power.

Across M_2 , M_{ord} , and C_2 , C_2 appears to be the most powerful GOF statistic compared to either M_2 or M_{ord} . For instance, C_2 had a 0.80 rejection rate across all conditions when the interdimension correlation of the generating model was 0.2. In contrast, M_2 were less powerful to detect dimensional misspecification under the same interdimension correlation (0.2) when sample size was 300 and the response categories were four or five. M_{ord} appeared to be less stable when the number of dimensions increased from two to three for conditions with four or five response categories.

Table 2.5

Mean M_2 , M_{ord} , and C_2 Across Conditions when Fitting a Unidimensional Model to Multidimensional Data

D	ρ	K	M_2					M_{ord}					C_2						
			300	1000	3000	5000	<i>df</i>	300	1000	3000	5000	<i>df</i>	300	1000	3000	5000	<i>df</i>		
2	0.2	2	73.47	180.98	501.52	817.73	35	73.47	180.98	501.52	817.73	35	73.47	180.98	501.52	817.73	35		
		3	244.90	443.15	1025.16	1616.90	170	96.39	287.95	847.21	1411.93	25	107.45	298.76	858.25	1423.36	35		
		4	477.44	681.73	1269.98	1861.25	395	71.03	166.60	552.79	655.26	15	105.97	279.98	779.67	1279.88	35		
		5	826.95	1124.66	1986.93	2850.40	710	64.36	199.88	453.23	657.05	5	144.03	415.92	1192.10	1965.61	35		
		2	43.88	65.67	130.69	193.02	35	43.88	65.66	130.69	193.02	35	43.88	65.67	130.69	193.02	35		
	0.8	3	189.74	232.25	356.39	481.36	170	41.32	80.65	190.75	301.35	25	53.09	95.60	215.79	336.74	35		
		4	415.78	462.87	600.15	734.73	395	20.17	30.15	58.99	86.94	15	53.14	96.29	220.96	341.71	35		
		5	742.45	810.76	1006.07	1204.21	710	10.75	19.39	33.35	52.51	5	62.00	126.02	306.62	488.50	35		
		3	0.2	2	177.86	418.94	1104.78	1795.68	90	177.86	418.94	1104.78	1795.68	90	177.86	418.94	1104.78	1795.68	90
				3	569.63	1004.57	2238.24	3476.24	405	229.24	640.52	1806.58	2972.91	75	249.73	668.91	1851.13	3035.56	90
4	1121.17			1606.94	3007.40	4425.63	930	228.38	496.18	1317.40	2084.86	60	258.56	684.81	1903.38	3128.33	90		
5	1889.32			2448.16	4053.79	5651.61	1665	199.95	609.92	1478.69	2294.62	45	297.17	807.76	2259.26	3707.71	90		
2	109.24			155.54	285.07	413.78	90	109.24	155.54	285.07	413.78	90	109.24	155.54	285.07	413.78	90		
0.8	3		442.00	530.20	776.52	1024.63	405	106.33	181.62	391.09	601.95	75	125.24	211.05	450.06	691.07	90		
	4		973.68	1071.27	1361.26	1653.55	930	82.81	129.90	274.49	419.42	60	129.26	218.13	478.21	740.02	90		
	5		1723.69	1842.02	2177.31	2524.92	1665	71.15	132.89	295.92	469.73	45	137.19	250.67	562.44	881.22	90		

Note. D = dimension. ρ = correlation among dimensions. K = response category. df = degrees of freedom. Samples sizes are in bold.

Table 2.6

Statistical Power of M_2 , M_{ord} and C_2 by Conditions Under Alternative Conditions

a) Compared to .01 Nominal Alpha Level

D	ρ	K	M_2				M_{ord}				C_2			
			300	1000	3000	5000	300	1000	3000	5000	300	1000	3000	5000
2	0.2	2	.860	1.000	1.000	1.000	.860	1.000	1.000	1.000	.860	1.000	1.000	1.000
		3	.852	1.000	1.000	1.000	.997	1.000	1.000	1.000	.993	1.000	1.000	1.000
		4	.618	1.000	1.000	1.000	.829	1.000	.994	1.000	.991	1.000	1.000	1.000
		5	.693	1.000	1.000	1.000	.826	.932	.980	.989	1.000	1.000	1.000	1.000
		2	.104	.688	1.000	1.000	.104	.688	1.000	1.000	.104	.688	1.000	1.000
	0.8	3	.119	.735	1.000	1.000	.346	.987	1.000	1.000	.357	.978	1.000	1.000
		4	.085	.483	1.000	1.000	.094	.428	.971	1.000	.338	.981	1.000	1.000
		5	.119	.581	.999	1.000	.243	.465	.669	.809	.590	.999	1.000	1.000
		2	.990	1.000	1.000	1.000	.990	1.000	1.000	1.000	.990	1.000	1.000	1.000
		3	.991	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.2	4	.941	1.000	1.000	1.000	.998	.999	1.000	1.000	1.000	1.000	1.000	1.000
		5	.846	1.000	1.000	1.000	.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		2	.185	.912	1.000	1.000	.185	.912	1.000	1.000	.185	.912	1.000	1.000
		3	.150	.924	1.000	1.000	.472	1.000	1.000	1.000	.504	1.000	1.000	1.000
		4	.109	.764	1.000	1.000	.320	.979	1.000	1.000	.570	.999	1.000	1.000
	0.8	5	.148	1.000	1.000	1.000	.461	.995	1.000	1.000	.724	1.000	1.000	1.000

Note. D = dimension. ρ = correlation among dimensions. K = response category. Samples sizes are in bold.

Table 2.6 (continued)

b) Compared to .05 Nominal Alpha Level

D	ρ	K	M_2				M_{ord}				C_2			
			300	1000	3000	5000	300	1000	3000	5000	300	1000	3000	5000
2	0.2	2	.947	1.000	1.000	1.000	.947	1.000	1.000	1.000	.947	1.000	1.000	1.000
		3	.952	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	1.000	1.000	1.000
		4	.821	1.000	1.000	1.000	.896	.973	.995	.996	.997	1.000	1.000	1.000
		5	.846	1.000	1.000	1.000	.875	.957	.987	.996	1.000	1.000	1.000	1.000
		2	.263	1.000	1.000	1.000	.263	1.000	1.000	1.000	.263	1.000	1.000	1.000
	0.8	3	.297	.879	1.000	1.000	.582	.998	1.000	1.000	.561	.997	1.000	1.000
		4	.200	.719	1.000	1.000	.227	.650	.992	1.000	.545	.996	1.000	1.000
		5	.241	.784	1.000	1.000	.386	.600	.780	.889	.769	1.000	1.000	1.000
		2	.999	1.000	1.000	1.000	.999	1.000	1.000	1.000	.999	1.000	1.000	1.000
		3	.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.2	4	.984	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
		5	.952	1.000	1.000	1.000	.995	1.000	1.000	1.000	1.000	1.000	1.000	
		2	.379	.979	1.000	1.000	.379	.979	1.000	1.000	.379	1.000	1.000	1.000
		3	.340	.976	1.000	1.000	.678	1.000	1.000	1.000	.700	1.000	1.000	1.000
	0.8	4	.266	.903	1.000	1.000	.552	.990	1.000	1.000	.762	.999	1.000	1.000
		5	.283	.870	1.000	1.000	.707	.997	1.000	1.000	.867	1.000	1.000	1.000

Note. D = dimension. ρ = correlation among dimensions. K = response category. Samples sizes are in bold.

Table 2.6 (continued)

c) Compared to .10 Nominal Alpha Level

D	ρ	K	M_2				M_{ord}				C_2			
			300	1000	3000	5000	300	1000	3000	5000	300	1000	3000	5000
2	0.2	2	.976	1.000	1.000	1.000	.976	1.000	1.000	1.000	.976	1.000	1.000	1.000
		3	.976	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	1.000	1.000	1.000
		4	.882	1.000	1.000	1.000	.918	.977	.996	.998	1.000	1.000	1.000	1.000
		5	.890	1.000	1.000	1.000	.900	.969	.990	.997	1.000	1.000	1.000	1.000
		2	.386	.908	1.000	1.000	.386	.908	1.000	1.000	.386	.908	1.000	1.000
	0.8	3	.410	.921	1.000	1.000	.710	.999	1.000	1.000	.682	.999	1.000	1.000
		4	.298	.822	1.000	1.000	.345	.746	.996	1.000	.673	.999	1.000	1.000
		5	.361	.864	1.000	1.000	.461	.688	.823	.911	.835	1.000	1.000	1.000
		2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.2	4	.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
		5	.974	1.000	1.000	1.000	.995	1.000	1.000	1.000	1.000	1.000	1.000	
		2	.513	.991	1.000	1.000	.513	.991	1.000	1.000	.513	.991	1.000	1.000
		3	.498	.989	1.000	1.000	.788	1.000	1.000	1.000	.805	1.000	1.000	1.000
	0.8	4	.386	.940	1.000	1.000	.687	.993	1.000	1.000	.855	.999	1.000	1.000
		5	.389	.930	1.000	1.000	.805	.998	1.000	1.000	.921	1.000	1.000	1.000

Note. D = dimension. ρ = correlation among dimensions. K = response category. Samples sizes are in bold.

Performance of RMSEA. A unidimensional model is considered an approximate fit to the data if $RMSEA_2 \leq .05$ (binary data) or $RMSEA_2 \leq .05 / (K-1)$ (polytomous data) (Maydeu-Olivares & Joe, 2014). Jurich (2014) suggested $RMSEA_2 \leq .04$ as a cutoff for approximate fit if the intercorrelation is .50 and .035 to .04 be used if the intercorrelation is .80. As expected, under null model conditions, mean RMSEAs were generally small and fell below the aforementioned benchmarks. Mean $RMSEA_2$ ranged from .001 to .014 with a mean of .006, mean $RMSEA_{ord}$ ranged from .002 to .014 with a mean of .006, and $RMSEA_{C2}$ ranged from .002 to .016 with a mean of .006 (Table 2.7).

The effect of sample size, level of interdimension correlation, and the number of categories on the three types of RMSEAs were also inspected. When sample size increased, all three types of RMSEA statistics slightly decreased. This is as expected because as sample size approaches infinity, the RMSEA values are expected to approach 0 when the generating model and analysis model are the same. Also, the magnitude of the interdimension correlation had negligible influence on the three types of RMSEAs. Finally, the number of categories seems to have a different influence on the three types of RMSEAs. Specifically, when the number of categories increased, $RMSEA_2$ decreased whereas $RMSEA_{ord}$ increased. $RMSEA_{C2}$ stayed unchanged when the number of categories increased. However, none of these values became alarming (i.e., exceeded the aforementioned benchmarks).

When the model was misspecified, mean $RMSEA_2$, $RMSEA_{ord}$, and $RMSEA_{C2}$ were compared to their respective population values to determine whether the sample estimation of the RMSEA was consistent with its population RMSEA. Mean $RMSEA_2$, $RMSEA_{ord}$, and $RMSEA_{C2}$ tended to approximate their population RMSEAs regardless of

sample size, but the number of dimensions, the size of the interdimension, and the number of response categories did have an influence. The three types of RMSEAs decreased when the interdimension correlation and the number of dimensions increased. However, the relationship between the RMSEA and the number of response categories is inconsistent across the three types of RMSEAs. For $RMSEA_2$, its mean value decreased when the number of categories increased. However, when it comes to $RMSEA_{ord}$ and $RMSEA_{c2}$, the relationship with the number of categories fluctuated. That is, $RMSEA_{ord}$ was larger for conditions with an odd number of response categories (i.e., three or five) than those with even number of response categories (i.e., two or four). The influence of the sample size, the interdimension correlation, and the number of response categories on the population RMSEAs followed similar patterns.

Given that the interdimension correlation, the number of dimensions, and the number of categories all contributed to the variation of RMSEAs for detecting dimensionality misspecification, to obtain a general guideline for the cut-off values of RMSEAs, a series of plots were created to illustrate the empirical rejection rates at different RMSEA values (Figure 2.2). Since sample size did not have an influence on the RMSEAs, empirical rejection rates were collapsed across the four sample sizes, which means, 4,000 rejection rate points were used for every line in Figure 2.2. As expected, when the interdimension correlation was high, more stringent cut-off values should be used to reach a reasonable rejection rate (i.e., the dark lines were closer to 0 compared to the grey lines). Also, although the number of dimensionality influenced the rejection rates, such influence was minor across $RMSEA_2$, $RMSEA_{ord}$ ($K = 2, 3$), and $RMSEA_{c2}$ (the dash lines were adjacent to the solid lines and when the lines approaching the Y-axis

at .95, the dash line and the solid line overlapped with each other). The number of response categories seems to have a greater influence on $RMSEA_2$, as more stringent cut-offs should be used when the number of categories increase. To summarize, $RMSEA_{C2}$ is least influenced by the number of categories, the number of dimensions, and the sample size.

Table 2.7

Mean $RMSEA_2$, $RMSEA_{ord}$, and $RMSEA_{C2}$ Values by Condition: Null Model

D	ρ	K	$RMSEA_2$					$RMSEA_{ord}$					$RMSEA_{C2}$				
			300	1000	3000	5000	<i>df</i>	300	1000	3000	5000	<i>df</i>	300	1000	3000	5000	<i>df</i>
1	-	2	.014	.009	.005	.004	5	.014	.009	.005	.004	5	.014	.009	.005	.004	5
	-	3	.011	.006	.004	.003	35	-	-	-	-	-	.015	.009	.006	.004	5
	-	4	.009	.005	.003	.002	85	-	-	-	-	-	.016	.009	.005	.004	5
	-	5	.008	.005	.003	.002	155	-	-	-	-	-	.016	.009	.005	.004	5
2	0.2	2	.011	.006	.004	.003	34	.011	.006	.004	.003	34	.011	.006	.004	.003	34
		3	.008	.004	.002	.002	169	.013	.007	.004	.003	24	.012	.006	.003	.003	34
		4	.007	.003	.002	.002	394	.014	.007	.005	.004	14	.011	.006	.004	.003	34
		5	.006	.003	.002	.002	709	.070	.038	.022	.017	4	.012	.006	.003	.003	34
	0.8	2	.011	.006	.004	.003	34	.011	.006	.004	.003	34	.011	.006	.004	.003	34
		3	.009	.004	.003	.002	169	.013	.006	.004	.003	24	.012	.006	.004	.003	34
		4	.007	.004	.002	.002	394	.014	.007	.005	.003	14	.012	.006	.004	.003	34
		5	.007	.003	.002	.001	709	.028	.016	.009	.008	4	.012	.006	.003	.003	34
3	0.2	2	.009	.005	.003	.002	87	.009	.005	.003	.002	87	.009	.005	.003	.002	87
		3	.007	.003	.002	.002	402	.010	.005	.003	.002	72	.009	.005	.003	.002	87
		4	.006	.003	.002	.001	927	.011	.006	.003	.002	57	.010	.005	.003	.002	87
		5	.005	.003	.002	.001	1662	.020	.013	.008	.006	42	.009	.005	.003	.002	87
	0.8	2	.009	.005	.003	.002	87	.009	.005	.003	.002	87	.009	.005	.003	.002	87
		3	.007	.004	.002	.002	402	.010	.005	.003	.002	72	.009	.005	.003	.002	87
		4	.006	.003	.002	.001	927	.011	.005	.003	.003	57	.010	.005	.003	.002	87
		5	.005	.003	.001	.001	1662	.013	.006	.003	.003	42	.010	.005	.003	.002	87

Note. D = dimension. ρ = correlation among dimensions. K = response category. Samples sizes are in bold.

Table 2.8

RMSEA Mean and Standard Deviation by Conditions: Alternative Model

D	ρ	K	RMSEA ₂					RMSEA _{ord}					RMSEA _{C2}				
			300	1000	3000	5000	<i>PRMSEA</i> ₂	300	1000	3000	5000	<i>PRMSEA</i> _{ord}	300	1000	3000	5000	<i>PRMSEA</i> _{C2}
2	0.2	2	.059	.064	.067	.067	.067	.059	.064	.067	.067	.067	.059	.064	.067	.067	.067
		3	.038	.040	.041	.041	.041	.097	.102	.105	.105	.105	.082	.087	.089	.089	.089
		4	.026	.027	.027	.027	.027	.095	.096	.095	.087	.070	.081	.083	.0841	.084	.084
		5	.023	.024	.025	.025	.025	.181	.177	.152	.128	.026	.101	.104	.105	.105	.105
		2	.026	.029	.030	.030	.030	.026	.029	.030	.030	.030	.026	.029	.030	.030	.030
	0.8	3	.017	.019	.019	.019	.019	.043	.046	.047	.047	.047	.038	.041	.041	.041	.041
		4	.012	.013	.013	.013	.013	.029	.029	.031	.031	.031	.038	.041	.042	.042	.042
		5	.011	.012	.012	.012	.012	.050	.042	.035	.035	.032	.048	.051	.051	.051	.051
		2	.056	.060	.061	.062	.062	.056	.060	.061	.062	.062	.056	.060	.061	.062	.062
		3	.037	.038	.039	.039	.039	.082	.087	.088	.088	.088	.077	.080	.081	.081	.081
3	0.2	4	.026	.027	.027	.027	.027	.086	.085	.083	.081	.073	.079	.081	.082	.081	.082
		5	.021	.022	.022	.022	.022	.106	.108	.103	.099	.091	.087	.089	.090	.090	.090
		2	.024	.027	.027	.027	.027	.024	.027	.027	.027	.027	.024	.027	.027	.027	.027
		3	.016	.017	.017	.018	.018	.035	.037	.074	.037	.038	.034	.036	.036	.037	.037
	0.8	4	.011	.012	.012	.013	.012	.033	.034	.034	.035	.035	.036	.038	.038	.038	.038
		5	.009	.010	.010	.010	.010	.042	.044	.043	.043	.043	.041	.042	.042	.042	.042
		2	.056	.060	.061	.062	.062	.056	.060	.061	.062	.062	.056	.060	.061	.062	.062

Note. D = dimension. ρ = correlation among dimensions. K = response category. PRMSEA₂ = population RMSEA₂. PRMSEA_{ord} = population RMSEA_{ord}. PRMSEA_{C2} = population RMSEA_{C2}. Samples sizes are in bold.

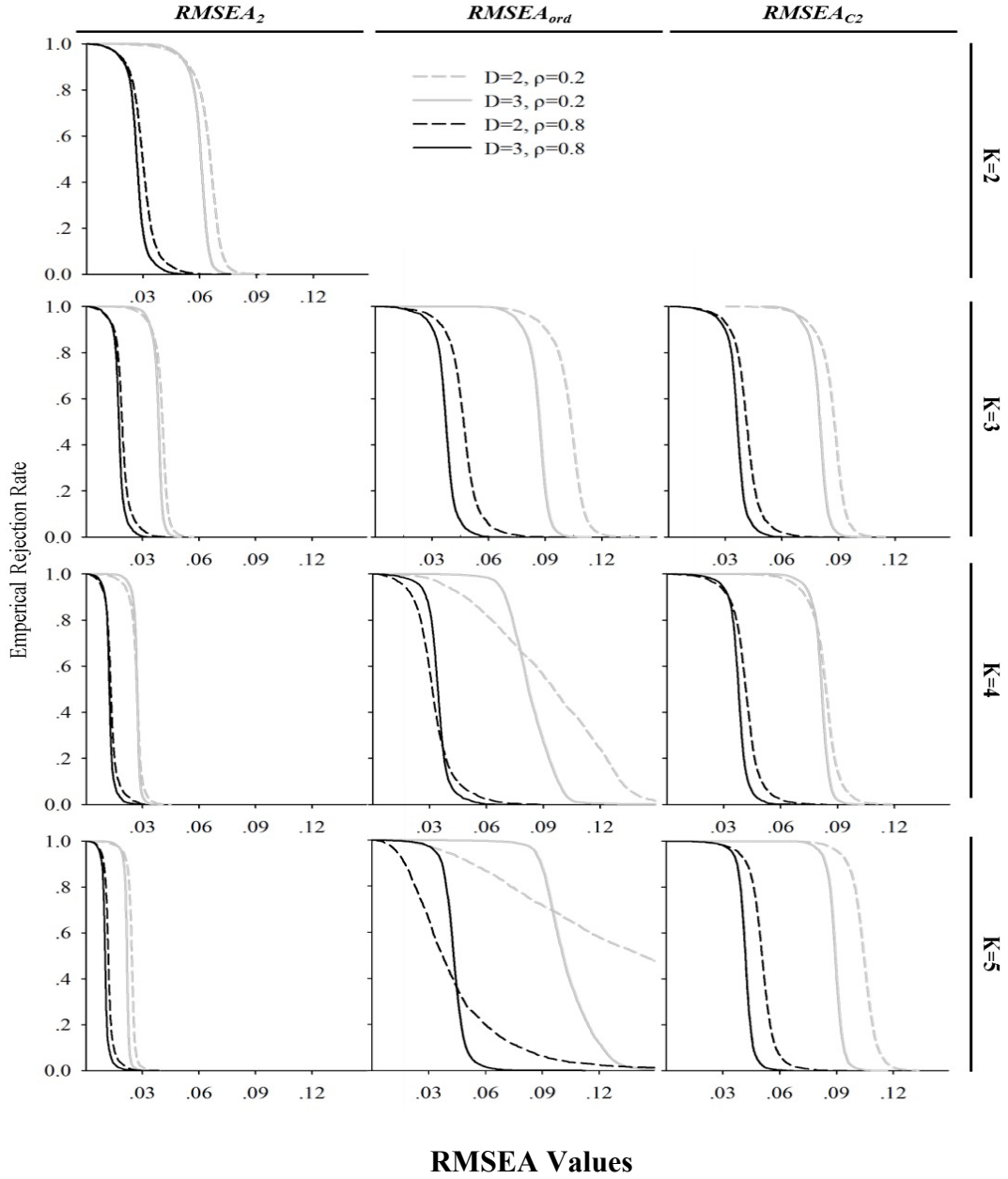


Figure 2.2. Empirical rejection rates for RMSEA for the collapsed dimension misspecification by interdimension correlation and the number of dimensions. Since sample size did not have an influence on the RMSEAs, empirical rejection rates were collapsed across the four sample sizes, which means, 4,000 rejection rate points were used for every line.

According to Figure 2, we created the following cut-off criterion according to the number of categories and the levels of correlation (Table 2.9). A first interesting observation is that the cut-off values of $RMSEA_{C2}$ seems very stable when the level of misspecification is low ($RMSEA_{C2}$ cutoff $\approx .030$) or high ($RMSEA_{C2}$ cutoff $\approx .070$). We explored the relationship between the cut-offs and the independent variables in Table 9 and found a linear relationship between $RMSEA_2$, ρ , and K , where $RMSEA_2 = .060 - .003 * \rho - .006 * K$, $F(2, 13) = 46.15$, $p < .001$, $R^2 = .877$. Using the same approach, we found a stronger linear relationship between $RMSEA_{C2}$, ρ , and K , where $RMSEA_{C2} = .053 - .007 * \rho + .009 * K$, $F(2, 13) = 133.101$, $p < .001$, $R^2 = .953$.

We firstly compared the cut-off with the recommendation by Jurich (2014) with binary responses. That is, an $RMSEA_2$ cut-off of around .035 to .04 was a reasonable guideline to reject model misspecifications including dimensionality misspecification for MIRT models with binary responses. Using 1,000 replications, instead of 100 in Jurich (2014), we found that his guideline were mostly useful to high level misspecification and when the number of categories are low. When the number of categories are large ($K = 4, 5$) or when the level of misspecification is low ($\rho = 0.8$), a $RMSEA_2$ cut-off around .035 to .04 become too stringent and might have a high rejection rate.

We then compared the cut-off with the recommendation by Maydeu-Olivares and Joe (2014). That is, a $RMSEA_2$ of .05 is a good criterion for close fit in binary IRT models. We could tell from Table 2.9 and Figure 2.2 that this criterion also applied to MIRT models with binary data when the level of misspecification is really high. Maydeu-Olivares and Joe (2014) also suggested that cutoff of excellent fit $.05/(\text{number of categories}-1)$ when the number of categories is equal or greater than two for

unidimensional models. We can tell that this criterion is more liberal to what we have observed in this paper when the level of misspecification is high.

Table 2.9

RMSEA Cut-Off Criterion Based on the Number of Categories and the Level of Correlation.

Correlation	No. of Categories	RMSEA ₂	RMSEA _{ord}	RMSEA _{C2}
0.2	2	.049-.050		
	3	.032-.033	.074-.084	.069-.071
	4	.021-.023	.038-.067	.068-.073
	5	.019-.020	.038-.086	.082-.091
0.8	2	.017-.018		
	3	.012	.027-.031	.026-.028
	4	.008-.009	.017-.025	.027-.029
	5	.007-.008	.013-.034	.035-.038

Note. The cut-off for RMSEA_{ord} with four or five response categories under the alternative model should be viewed with caution due to the limitation of M_{ord} .

Discussion

A growing body of literature has evaluated the performance of the family of limited-information GOF statistics and their associated approximate fit indices. However, studies have not yet examined such statistics using polytomous data with MIRT models. This simulation study investigated the Type I error rate and power of M_2 , M_{ord} , and C_2 and associated RMSEAs using polytomous data under MIRT models. Four important findings are shown by the current study and are discussed in the following section.

First, the estimation of the M_{ord} statistics was not only influenced by the lack of df, but also by the increasing of the number of categories. Previous studies by Maydeu-Olivares and Joe (2014) and Cai and Hansen (2013) both suggested the limitation of using M_{ord} when the number of items is small and the number of categories is large. Maydeu-Olivares and Joe (2014) developed an equation under unidimensional IRT models to indicate the lack of df issue. That is, M_{ord} could be approximately estimated when the difference between the number of items and the number of categories is larger than two, given that under such conditions the df is positive. A new finding from the current study showed that under multidimensional models, when the number of items is small and the number of categories is large, extreme M_{ord} values will be produced by FlexMIRT, even negative values, as indicated by Appendix A. When the null model is true, conditions with two dimensions and five categories response format involve 1.3% to 5.5% of the negative values. One explanation is that under such conditions the df is approaching zero (e.g., $df = 5$). The Jacobian matrix that is used to estimate M_{ord} might not be locally identified. The FlexMIRT software engineers suggested increasing the number and range of the quadrature points might help. However, a simple test (see

Appendix A lower panel) showed that when the number of quadrature points increased from 21 to 29 and the range increased from 4 to 6, the proportion of the negative M_{ord} values did not decrease. In addition, even more extreme M_{ord} values showed up. For example, for two-factor models with a high interdimension correlation (i.e., $\rho = .8$), five categories, and a sample size of 5,000, within 1,000 replications, the largest M_{ord} value is 31,3430.94, which will dramatically influence the mean M_{ord} if retained. As a contrast to the conditions when there were five response categories, for multidimensional models with four response categories not many negative M_{ord} observations were presented.

Although the results related to M_{ord} under the true model conditions were presented, all interpretations of the results should be viewed with caution. For example, in Table 3, the mean of M_{ord} ranged from .11.11 to .11.65 for condition with two dimensions, five response categories, and a low interdimension correlation (i.e., $\rho = 0.2$) when $N = 300 - 3000$, which showed a large discrepancy from its associated df (that is, 4). Under the same condition, in Table 4a), the Type I error rates of M_{ord} ranged from .286 to .303, showed an extreme large discrepancy compared to the nominal alpha level at .01.

Under the alternative model when unidimensional models were fitted to data generated from multidimensional models, again as shown in Appendix A2, M_{ord} has extremely large ranges and a considerable proportion of negative values. The negative M_{ord} values were again related with large number of categories when the numbers of categories are larger than three. In addition to that, it seems the size of the df, the level of misspecification, and the sample size all contributed to the occurrence of the negative M_{ord} values. The worst condition is when there were two dimensions, five response

categories, high interdimension correlation (i.e., $\rho = 0.8$), with a sample size of 3000 or 5000 ($df = 5$). With around 15% of the estimated M_{ord} values were negative, it bears the power of M_{ord} into question. More investigation and research should be continued to find out under what conditions M_{ord} becomes unstable and also to confirm whether it is a software problem or a theoretical problem.

Second, regardless of the M_{ord} results when the number of categories are larger than three, M_2 , M_{ord} , and C_2 statistics are all well-calibrated under the true model conditions. That is, the Type I error rates were close to the nominal alpha levels when the sample size is relatively medium to large (e.g., $N \geq 1000$). The M_2 , M_{ord} , and C_2 statistics all followed the χ^2 distribution. Different from Figure 3 in Cai and Hansen (2013) where the observed M_2 statistics were smaller than the expected M_2 statistics, the current study did not find this pattern. In fact, M_2 statistics did not have a large discrepancy compared to the expected M_2 statistics. This might be attributed to the small numbers of items used in this study (five items for unidimensional models and ten items in Cai and Hansen, 2013). In contrast, our findings resonated with the M_2 results showed in Jurich (2014) using dichotomous data under MIRT models. That is, M_2 approximately maintained nominal Type I error rates for the true model, although for some conditions the Type I error rates of M_2 slightly fluctuate around the nominal alpha levels. Of the three limited-information GOF statistics, C_2 is the most stringent statistics and M_2 is the most tolerant statistics under the true model conditions.

Third, the M_2 , M_{ord} , and C_2 were close to perfectly detecting misspecified models when sample size is above 3000 regardless of the level of misspecification, the number of categories, and the number of dimensions. As expected, when the level of

misspecification is low, a sample size of 1000 and above is sufficient to detect dimensionality misspecification. However, when the level of misspecification was high, more data will be needed to identify the dimensionality misspecification. Among M_2 , M_{ord} , and C_2 , C_2 is the most powerful. Larger sample sizes are required to be able to identify the incorrect models if the M_2 and M_{ord} statistics were used as the global fit GOF statistics. When sample size is extremely small, like 300, all three statistics failed to sufficiently identify the wrong models.

Fourth, we found that $RMSEA_2$ and $RMSEA_{C_2}$ are good tools to evaluate approximate fit. Similar to Jurich (2014), in this study, we also found that the RMSEA values are not influenced by sample size under misspecified model conditions, which made the RMSEAs better supplements to the statistical tests. Different from Jurich (2014), under true model specifications, we did find a slightly positive relationship between sample size and the mean RMSEA values. We expect this to happen under true model conditions because we expect when the sample size is large enough, RMSEA values should be approaching to zero. Perhaps the most exciting findings of the three types of RMSEA is that we were able to provide a guideline of RMSEA in relation to the correlation and the number of categories. Using mean RMSEA values and the plot of empirical rejection rates in the function of the RMSEA cut-off values, we found that the levels of the correlations and the number of categories both influence the performance of RMSEA. Based on this descriptive findings, we further found a linear relationship between $RMSEA_2$, ρ , and K , where $RMSEA_2 = .060 - .003 * \rho - .006 * K$ and a stronger linear relationship between $RMSEA_{C_2}$, ρ , and K , where $RMSEA_{C_2} = .053 - .007 * \rho + .009 * K$. Although these two equations should not be over interpreted as strict equations

for computing RMSEA cutoffs, we recommend more researchers to test these two equations in their simulations.

To summarize, there are three major takeaways from this simulation study. First, the limited-information statistics of M_2 and C_2 and corresponding RMSEAs are helpful tools to evaluate the global test of model fit in MIRT models. Second, M_2 and C_2 are powerful tools for detecting dimensionality misspecification. Third, the performance of M_{ord} and $RMSEA_{ord}$ were not stable in the current study and should involve more investigation in future. However, if M_{ord} and $RMSEA_{ord}$ are unstable and a researcher has C_2 available, then maybe future examination of the M_{ord} is unnecessary and more research on C_2 is needed.

Chapter 3 – Study two

Religion is a powerful social force that has potent influence on people's health-related, behavioral, and social outcomes (McCullough & Willoughby, 2009).

Worthington (1988) defined religiosity as “the degree to which a person adheres to his or her religious values, beliefs, and practices and uses them in daily living”. Consistent with this definition, Worthington, Hsu, Gowda, and Bleach (1988) developed the Religious Commitment Inventory (the RCI) as a measure of religiosity in Christians. Subsequent to considerable evaluations, the RCI was further refined to the RCI-17 (McCullough, Worthington, Maxey, & Rachal, 1997) and then the RCI-10 (Worthington et al., 2003; See the specific RCI-10 items in Appendix C). The RCI-10 has been employed in many substantive studies in well-being, mental health, and consumer behaviors (e.g., Frazier et al, 2013; Wade, Bailey, & Shaffer, 2005; Wade, Worthington, & Vogel, 2007). More importantly, the RCI-10 has been extensively used to study the role of religiosity in domestic violence, including hotspot issues such as sexual assault (e.g., Renzetti, DeWall, Messer, & Pond, 2015) and gun control (e.g., Follingstad, Coker, Chahal, Brancato, & Bush, 2016). To date, five studies have examined the dimensionality and score reliability of the RCI with different length and format using factor analyses or confirmatory factor analyses (CFA) techniques. Unfortunately, researchers have arrived at different conclusions regarding the factor structure, scoring, and interpretability of the measure, which complicates inferences made in substantive studies.

The RCI: Empirical Studies of its Psychometric Properties.

The RCI. Scholars who are interested in studying religion generally agree religion often positively influence mental health, well-being, and many other health-related outcomes (McCullough & Willoughby, 2009). Worthington (1988) proposed a theory to address the question that under what conditions religion has a positive influence on the counseling process and outcomes. The core hypothesis under his theory was that religiosity influences peoples' view of the world through religious values. Consistent with this proposal, Worthington et al. developed the Religious Values Scale (RVS) to measure seven major constructs of Worthington's theory. The RCI was one of the seven scales on the RVS. The RCI was used to measure the motivational and behavioral commitment to one's religious beliefs (e.g., "I am concerned that my behavior and speech reflect the teachings of my religion"). Worthington et al. intentionally wrote the scale in a generic way so it was appropriate to use for most faiths. The RCI contains 20 items with a five-point Likert response format ranging from 1 (not at all true of me) to 5 (totally true of me). The internal consistency reliability (coefficient α) estimate for the RCI was .92 (McCullough & Worthington, 1995).

The RCI-17. Given the need for using the RCI as an independent measure of religiosity (instead of as a subscale of the RVS), McCullough et al. (1997) assessed the RCI using principle components analysis (PCA) in a sample of 239 American Christian undergraduate students by forcing the scores to be unidimensional. Inconsistent with Worthington et al. (1988), three items did not load on the first component, indicating a lack of stability in structure. Total scores of the 17 items were used to represent religiosity with a reliability (α) estimate of .94. Worthington et al. also correlated total scores of the RCI-17 with scores from four other well-established religiosity measures

(two specifically for Christian beliefs). Their results provided for concurrent validity evidence for scores of the RCI-17 (r s ranged from .64 to .82).

The RCI-10. Given the need for a brief and psychometrically sound measure of religiosity during the counseling process and for counseling research, Worthington et al. (2003) developed the RCI-10 from the RCI-17 based on six studies that each used a unique American sample. Three of the studies involved dimensionality analyses on a sample of college students, adult Christians, and adult clients seeking help from diverse types of counseling agencies. In particular, Worthington et al. (2003) explored the internal structure of the RCI-17 using exploratory factor analysis with an orthogonal rotation in a sample of 155 U.S. college students and retained items with factor loadings of .60 or higher. Worthington et al.'s final solution retained 10 items across two factors. Specifically, the Intrapersonal Religious Commitment (Intrapersonal) factor includes six items (e.g., "I spend time trying to grow in understanding of my faith") and the Interpersonal Religious Commitment (Interpersonal) factor contains four items (e.g., "I enjoy spending time with others of my religious affiliation"). Given that the Intrapersonal and Interpersonal factors was observed to be highly correlated (r s ranging from .72 to .85), Worthington et al. (2003) re-examined the internal structure of the RCI-10 using CFA with maximum likelihood (ML) estimation in three additional samples ($n_{\text{Christian}} = 190$, $n_{\text{College student}} = 282$, and $n_{\text{Christian clients}} = 282$). Worthington et al. (2003) concluded that one factor should be used to score the RCI-10 due to the high correlation observed between the two factors across these three samples ($r = .75$ to $.89$), even though the two-factor model had better fit to the data. Reliability estimates (α) for the total scale

score, Intrapersonal scores, and Interpersonal scores were .92 to .98, .86 to .96, and .68 to .97, respectively.

It is worth mentioning that Worthington et al. (2003) also examined differences in general religiosity, Intrapersonal and Interpersonal religiosity using ANOVAs in a religiously diverse sample of 468 undergraduate students (52 Buddhist students, 278 Christian students, 10 Hindu students, 12 Muslim students, and 116 students that have no religious affiliation). However, no dimensionality analyses or measurement invariance tests were conducted before the comparison.

The RCI-A. The RCI-A is an 11-item scale designed to measure religiosity among adolescents. Miller, Shepperd, and McCullough (2013) modified the content of the RCI-10 to match the reading proficiency of ninth graders (e.g., Item 2 “I make financial contributions to my religious organization” was modified to be “I give money to my religious organization”). Since Item 10 (“I keep well informed about my local religious group and have some influence in its decisions”) was double-barreled, Miller et al. (2013) rewrote Item 10 into two items (“I am involved in my religious group” and “I have some influence on the decisions of my religious group”). Miller et al. evaluated the reliability, factor structure, and measurement invariance of the RCI-A among 1,419 American ninth graders who were mostly White (66.6%) and Christian/Protestant (60.1%).

Prior to dimensionality analysis, Miller et al. examined the reliability of the RCI-A scores in the whole sample and across gender and religious affiliation. They found scores from the RCI-A was reliable among most of the groups except for those identified as atheists ($\alpha = .59$) and suggested researchers should be cautious whether the RCI-A is appropriate to use with atheists. The dimensionality analysis confirmed a two-factor

structure of the data using CFA with ML estimation. Miller et al. concluded a one-factor structure should be used to score the RCI-A so as to be consistent with Worthington et al. (2003) and to avoid the multicollinearity problem in subsequent path analyses due to the high correlation between Intrapersonal and Interpersonal factors. By combining the students into White vs. Other and Christian/Protestant vs. Other, they found the RCI-A were partially invariant across gender and race and strictly invariant across religious affiliation.

The RCI-10-PL. The RCI-10-PL is a Polish version of the RCI-10 evaluated by Polak and Grabowski (2017) among 581 Polish adults using CFA with ML estimation. They randomly divided the sample into two subsamples, confirmed the one-factor model fit similarly with the two-factor model, and concluded the scale could be used either as unidimensional or two-dimensional.

Psychometric Concerns over the Dimensionality and Scoring of the RCI-10.

Researchers should continue to examine the dimensionality of the RCI-10. First, psychometric studies concluded differently regarding the dimensionality of the RCI-10. Although most of the aforementioned studies have suggested the RCI be used as unidimensional, Polak and Grabowski (2017) disagreed regarding the Polish version of the RCI-10 and concluded it can be used as two-dimensional. Second, although it is suggested to be unidimensional by Worthington et al. (2003), the RCI-10 was originally conceptualized to capture two aspects of the religiosity: motivational religiosity and behavioral religiosity. This might explain why in practice some researchers have scored the RCI-10 as multidimensional. For example, Tsang, McCullough, and Hoyt (2005) created Interpersonal and Intrapersonal subscale scores separately in their paper exploring

the relationship of religiosity and forgiveness. Similarly, Wighting and Liu (2009) used the RCI-10 as two-dimensional to investigate the relationship between religiosity and sense of belonging. Third, in addition to treating the RCI-10 as multidimensional, some researchers have used the Intrapersonal subscale as a representation of religiosity. For example, Frazier, Greer, Gabrielsen, Park, and Tomich (2013) have used the Intrapersonal subscale to represent religiosity in their paper testing the relationship between trauma exposure and prosocial behavior. However, inferences made from these studies might be complicated given that no psychometric evidence has supported the scoring and use of the RCI-10 subscales. Just as Worthington et al. (2003) suggested, limited evidence existed to suggest the scores from Intrapersonal or Interpersonal subscales are valid. Given these conflicting practices of the scoring and interpretation of the RCI-10, additional study of the dimensionality is needed. Moreover, the bifactor item response theory (IRT) model could be used to test the level of unidimensionality or multidimensionality of the RCI-10 scores. Fitting a bifactor IRT model could help diagnose whether the RCI-10 items are essentially unidimensional and should not be broken up into subscales or whether the items are essentially multidimensional and should be scored to represent this complexity.

Using a Bifactor IRT Model to Assess the Dimensionality of the RCI-10.

Measurement researchers have recommended the utility of the bifactor model to determine whether scores from a scale are sufficiently unidimensional for creating scores that can be used in item response theory (IRT) or structural equation modeling (SEM, e.g. Reise, 2012; Rodriguez et al., 2016; Toland et al., 2017). In a bifactor model, each item is an indicator of a general dimension and a secondary specific dimension simultaneously.

The general dimension has direct influence on all items and could capture the shared content of all items. The specific factors explain the response variation that is unique or particular to a set of items. Such uniqueness of the specific factor might be due to item content, wording, formatting, or other conceptual influences that make the item responses correlated with each other above and beyond the influence from the general dimension (Toland et al., 2017). In the case of the RCI-10, the general dimension is general religiosity and the specific dimensions include Intrapersonal religiosity and Interpersonal religiosity. All 10 items are indicators of general religiosity. Six of the 10 items are also indicators of Intrapersonal religiosity and the other four are indicators of Interpersonal religiosity. The general, Intrapersonal, and Interpersonal religiosities are orthogonal to each other.

The bifactor model differs from the unidimensional and two-dimensional models and such difference permits the advantages of the bifactor model. The bifactor model differs from the unidimensional model in that after the influence of the general dimension is extracted, specific factors are estimated to capture the residual influences on item responses. The bifactor model differs from the two-dimensional model in that there is a general dimension underlying all items and that all dimensions are orthogonal to each other. Given that the general and the specific dimensions are estimated simultaneously in the bifactor model, it could clarify the influence of the general/specific dimension on each item, which is not possible using unidimensional or two-dimensional models (Toland et al., 2017). Moreover, the bifactor model allows the researchers to determine the scoring and interpretation of the general and specific factors (DeMars, 2013; Reise,

Morizot, & Hays, 2007; Toland et al., 2017). Until now, no study has considered the bifactor model for examining the dimensionality of the RCI-10.

Religiosity, Spirituality, and Perpetration of Intimate Partner Violence (IPV).

According to Saltzman, Fanslow, McMahon, and Shelley (1999), IPV includes any potential or completed sexual, physical, or psychological abuses committed by current or former intimate partners. Religiosity and spirituality (defined as the search for meaning and purpose of life; Healy, 2005) are traditionally taken as coping strategies that victims can use to recover from IPV (Kreidler, 1995). However, recent findings showed that the effect of religiosity on IPV is paradoxical (Johnson & Stephens, 2015). On one hand, religiosity is a protective factor for IPV victimization. Studies have reported that higher religiosity (e.g., measured using church attendance, religious commitment) was associated with lower levels of IPV perpetration, victimization, and re-victimization (Cunradi, Caetano, & Schafer, 2002; Ellison & Anderson, 2001; Ellison, Trinitapoli, Anderson, & Johnson, 2007; Wang, Horne, & Levitt, 2009). On the other hand, religiosity (measured using the RCI-10) serves as a risk factor on physical and psychological IPV perpetration (Renzetti, DeWall, Messer, & Pond, 2015). That is, more religious participants were more likely to perpetrate. Moreover, when researchers put alcohol use together with religiosity, alcohol appear to associate more strongly with IPV perpetration than religiosity (Cunradi, Caetano, & Shafer, 2002) or its relationship with IPV perpetration is buffered by religiosity (DeWall, 2010). In light of these findings, it appears important to investigate the extent to which the role of religiosity, spirituality, and alcohol consumption relates to IPV perpetration.

Purpose of Current Study. The current study examined the psychometric properties of scores derived from the RCI-10 in a community sample of 392 adults who were affiliated to religious institutions and were currently in an intimate partner relationship. After the internal structure and dimensionality evidence of the RCI-10 was obtained, construct validity and consequential-related validity (AERA, APA, NCME, 1999) of the scores from the RCI-10 was investigated. A model comparison approach within the IRT framework was used to identify the best solution to the internal structure of the RCI-10 by comparing a unidimensional, two-dimensional, and bifactor GR models. The construct validity of the RCI-10 was tested via the polychoric correlation between religiosity measured by the RCI-10 and religiosity measured by a single religiosity item. The hypothesis was that the correlation between these two religiosity measures would be positive and high. The discriminant validity of the RCI-10 was examined through the polychoric correlation between the RCI-10 scores and the single-item spirituality measure. Previous studies have shown that religiosity and spirituality are two different constructs and have a moderate correlational relationship. Consequential validity of the RCI-10 was examined by studying the relationship between religiosity, gender, alcohol consumption, and IPV perpetration. It was hypothesized that there should be a negative relationship between religiosity and alcohol consumption, a negative relationship between religiosity and IPV perpetration, and a positive relationship between alcohol consumption and IPV perpetration. Such relationships were examined controlling for race and income.

Method

Participants. Data from a national community sample of 392 adults were collected using Amazon's Mechanical Turk (MTurk) which permits researchers to collect data with

greater sample diversity and higher response rate to sensitive items about IPV perpetration, compared to the typical convenience sampling approach (Buhrmester, Kwang, & Gosling, 2011). Individuals aged 18 and above and who were interested in this study was navigated to the online survey platform Qualtrics to participate in the study. Participants who did not sign the informed consent form or were not currently in a heterosexual intimate relationship were stopped from completing the survey. Individuals were compensated \$1.00 each for completion of the survey. Participants who self-identified as not a member of any religious affiliation were also excluded from the study given that the RCI-10 is not content-valid for people who have no religious belief.

Participants self-identified as 40.6% men and 59.4% women. We used a categorical response format to request for age information: 48.2% of the sample was between 25 and 34 and 25.3% was between 35 and 44. The majority of the participants were White (71.2%), had some college or college degree (76.0%), and worked full time (56.8%). The annual family income distributed normally across all categories, with \$40,000 - \$60,000 being the peak (34.2%). The majority of the participants identified themselves as Christian, with Catholic participants comprising 27.0%, Other Christian participants comprising 27.6%, and Evangelical Protestant participants comprising 19.6%. More detailed demographic information can be found in Appendix E.

Measures.

Pew Religious Landscape Survey (the RLS). The RLS surveys over 35,000 Americans about their religious affiliations, beliefs, and practices, and social and political

views across all 50 states. Four items from the 2014 RLS (Pew Research Center, 2018b) were used to measure religious identification and affiliation. The first item measured *religious affiliation*. Researchers asked the participants whether they belonged to a specific religion or church (1 = yes and 0 = no). The second question asked the respondents to choose the closest description of the religion or church they belong to (e.g., Evangelical Protestant, Catholic, Jewish, Muslim, etc.). The third question is a single question measuring *religiosity* – “Overall, to what extent do you consider yourself a religious person?” using four response categories ranging from 1 (*I do not consider myself religious at all*) to 4 (*I consider myself very religious*). The fourth question is a single question measuring *spirituality* – “Overall, to what extent do you consider yourself a spiritual person?” The third question also includes four response categories from 1 (*I do not consider myself religious at all*) to 4 (*I consider myself very religious*). The first and second questions were used to collect demographic information on religion and the third and fourth questions were used to examine the roles religiosity and spirituality played in intimate partner violence.

The Religious Commitment Inventory – 10 (RCI-10). This study used the 10-item RCI-10 developed by Worthington et al. (2003) to measure the intensity of religiosity using a five-point Likert-type response format ranging from 1 (not at all true of me) to 5 (totally true of me). Six of the 10 items measured the content area called intrapersonal religiosity (e.g., “I often read books about my faith.”) and the other four items measured the content area called interpersonal religiosity (e.g., “I enjoy working in the activities of my religious organization.”). The detailed dimensionality and scoring information of the RCI-10 was reported in the result section.

The Severity of Violence Against Women Scale (SVAWS). IPV perpetration was measured using a scale adapted from the 46-item SVAWS by Marshall (1992). Participants were asked to rate themselves on the frequency that they threatened to abuse their intimate partner in the past 12 months or actually abuse them physically. Scores from the SVAWS have been shown to have a second-order structure, with threats of physical violence (19 items) and actual physical violence (27 items) as the higher order dimensions. Threatening physical violence then includes four lower order factors: symbolic violence (4 items; e.g., “threw an object at your partner”), mild threats (4 items; e.g., “shook your fist at your partner”), moderate threats (4 items; e.g., “threatened to destroy property”), and serious threats (7 items; e.g., “threatened to kill your partner”). The actual violence dimension includes mild (4 items; e.g., “pushed or shoved your partner”), minor (5 items; e.g., “pulled your partner’s hair”), moderate (3 items; e.g., “slapped your partner with the back of your hand”), serious (9 items; e.g., “choked your partner”), and sexual violence (6 items; e.g., “physically forced your partner to have sex”). The items were modified to be gender-neutral (e.g., instead of “threatened to hurt her”, the item was modified as “threatened to hurt your partner”). Reliability estimates (α) ranged from .92 to .96 for a sample of 707 college women and from .66 to .89 for a sample of community women. The overall SVAWS scale adopted a four-point Likert response format ranging from 0 (never) to 3 (many times).

Frequency of the responses showed that the last two response categories were not used as expected and were collapsed before conducting the IRT analyses. However, the nine-factor GR model could not converge in FlexMIRT or Mplus. Then the psychological abuse, physical abuse, and sexual abuse subscale were examined separately. Mplus was

finally used to evaluate the internal structure of the SVAWS due to many extreme item parameters (> 4) in flexMIRT and huge local dependency (> 100). Confirmatory factor analysis showed that a bifactor model fit the SVAWS best and unidimensional model could be used for scoring purpose. Bifactor model suggested the general intimate partner violence trait was not a major source of variance in items 5, 6, 8, and 14. Thus these four items were excluded from the final scale. The final scale has excellent reliability, $\omega = .996$, 95% CI = [.995, .997]. Psychometric information about the SVAWS can be found under Appendix E.

Alcohol consumption. Alcohol consumption is measured using three items from the National Core Survey of College Alcohol Use (Presley, Meilman, & Leichter, 2002). Responses to the alcohol frequency question, “During the past 12 months, how often, on average, did you drink alcohol,” ranged from 0 (never/not at all) to 8 (about 4 or more times per day). Responses to the alcohol quantity question, “How many drinks did you usually have each time.” ranged from 0 (0/none; I did not drink any alcohol during the past 12 months) to 4 (4 or more). Response to the binge drinking question, “During the past 12 months, how many times have you been drunk or high from consuming alcohol?”, ranged from 0 (0/never) to 5 (about 4 or more times per week). Average alcohol consumption was calculated using the equation average alcohol consumption = $[(\text{total drinking days} - \text{binge-drinking days}) \times \text{quantity of drinks on a normal drinking day} + (\text{binge-drinking days} \times 5)] / \text{total drinking days}$ (Stahre, Naimi, Brewer, & Holt, 2006). Five people reported their daily alcohol consumption is 17.50 to 67.50, which was not reasonable for a normal person, so the scores from these five people were treated as 5

to represent a binge drinking habit. Finally, the average alcohol consumption per day was 2.20 (SD = 1.67, range = 0 – 5).

Data Analysis Plan.

To evaluate psychometric properties of the scores for the RCI-10, Samejima's (1969) GR model with the marginal maximum likelihood estimation (MML) based on IRT theory (De Ayala, 2009) was used. Forero and Maydeu-Olivares (2009) showed that when the number of items per dimension is at least 7, accurate estimates of item parameters could be obtained using FIML estimation even with a sample size of 200 with unidimensional GR models. Thus, the current study has a sufficient sample size to estimate the performance of the RCI-10 using a unidimensional GR models. The current sample has an acceptable sample size given that Jiang, Wang, and Weiss (2016) showed that a sample size of 500 will be sufficient to use with multi-unidimensional model to get accurate parameter estimates using the multi-unidimensional GR model. Item parameters was calibrated using flexMIRT 3.0 given that flexMIRT offers global model-data fit of C_2 and associated RMSEA for ordinal polytomous data (Cai, 2015; Toland et al., 2017).

Three competing IRT GR models were examined: a unidimensional GR model (uni-GR), a two-factor correlational model (two-GR), and a bifactor model (bifac-GR). A uni-GR model is considered since the conclusions from previous psychometric studies suggested RCI scale scores should be treated as unidimensional. The unidimensional model included all 10 items being explained by one latent construct called religiosity. Given that a two-dimensional model was always the statistically best model from previous research and researchers have used the RCI-10 as two-dimensional, a two-factor model was also included. The two-dimensional model allows a subset of six items

explained by Intrapersonal religiosity and the other four items explained by Interpersonal religiosity. The Intrapersonal religiosity dimension and the Interpersonal religiosity dimension were allowed to covary with each other. A bifactor model was also considered following the recommendations of Reise et al. (2007). In a bifactor model, all 10 items are explained by a general factor called general religiosity, the residual variances of a subset with six items are explained by Intrapersonal religiosity, and the residual variances of the rest four items are explained by Interpersonal religiosity. In the bifactor model, the general, Intrapersonal, and Interpersonal religiosity dimensions were not allowed to covary with each other.

After each model was fit to the data, conditional independence, item-level model-data fit, overall model-data fit, and model comparison using item parameters were evaluated. Conditional independence was examined via standardized local dependency (LD) χ^2 statistics (Chen & Thissen, 1997) and values ≤ 10 are acceptable (as cited in IRTPRO User Guide 4.2, p. 85; Toland, 2014). Item-level model-data fit was examined using the Orlando-Thissen-Bjorner item fit S- χ^2 statistics (Orlando & Thissen, 2003) at a p value corrected by the Benjamini-Hochberg (B-H) procedure (Benjamini & Hochberg, 1995). Global model-data fit was examined using C_2 , the limited-information goodness-of-fit statistic that is fast, accurate, and powerful to examine the global model-data fit with multidimensional polytomous data (Cai & Monroe, 2014). RMSEA based on C_2 (RMSEA $_{C_2}$) was used to examine the global misfit (e.g., model error or misspecification). The cutoff value of RMSEA $_{C_2}$ was determined using .035 according to the result from the first study. Given that the RCI-10 adopted a five-point response format and the correlation coefficient among the two subfactors was high, to reject the RCI-10 as a

unidimensional model, a $RMSEA_{C2}$ between .035 to .038 was suggested. Model comparison was assessed using the relative goodness of fit statistics including -2 log likelihood (-2LL) and the goodness-of-fit statistics including Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The differences in -2LL between two nested models were tested via χ^2 statistics with a degrees of freedom equal to the differences in the numbers of parameters to be estimated in the same two models. If the deviance statistic is significant, then the more complex model with more constraints is better than the relative simple model with fewer constraints. AIC difference greater than 6 or BIC difference greater than 10 between two models was treated as strong model-fit differences (Burnham & Anderson, 2002; Kass & Raftery, 1995; Symonds & Moussalis, 2011, p. 17). The model solution with smallest AIC and BIC was considered as a better model solution.

If a bifactor model is retained, ancillary indices including explained common variance (ECV; Reise et al., 2010; Ten Berge & Sočan, 2004; Toland et al., 2017) was examined to determine the severity of the multidimensionality and whether multidimensionality can be ignored. The common variances include variances explained by the general dimension and variances explained by the specific dimensions. The ECV for the general factor is the proportion of common variances that is explained by the general dimension. Similarly, the ECV for the specific factor is the proportion of common variances that are explained by the specific dimension. Stucky and Edelen (2014) suggested that ECV values higher than .85 suggest the multidimensionality could be ignored and the items could be treated as unidimensional. If ECV values are below .85 for the general dimension, item-level ECV (IECV; Stucky & Edelen, 2014; Stucky,

Thissen, & Edelen, 2013) could be examined to determine which item(s) is unidimensional. If our goal is to find a unidimensional measure of religiosity, IECV could help us to determine how much variance of each item is explained by the general dimension.

Once dimensionality is determined, item information functions (IIFs) were examined to determine the amount of precision across the broad range of -3 to +3 on the latent trait continuum of religiosity. By inspecting IIFs for all 10 items, we could identify whether there is any redundant item, or if any item provides less information to the RCI-10 scale and thus should be removed or modified. To understand how the RCI-10 works as a whole, the total information function (TIF) and the expected standard error of estimates (SEE; $SEE \cong 1/\sqrt{\text{information}}$) plot were examined (de Ayala, 2009; Toland et al, 2017). Marginal reliability was also reported for the RCI-10 (Green et al., 1984; Toland et al., 2017). It is noteworthy that flexMIRT is not able to provide the plot, but Toland et al. (2017) provided templates to draw the IIF, TIF, and SEE using Excel. Similar IRT procedures were repeated for the SVAWS scale measuring IPV perpetration. Instead of using three competing IRT GR models, only the multi-unidimensional GR model (multidimensional models where no cross-discrimination is allowed for each item) was used to confirm the internal structure of the SVAWS.

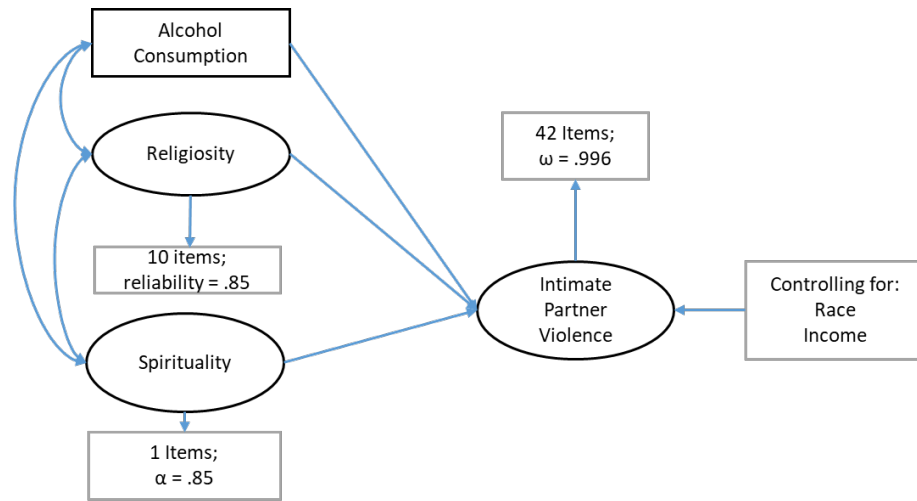
The relationships between the religiosity and similar/different variables were examined using SEM techniques. First, the relationships among the RCI-10 scores, the single-item religiosity scores, and the single-item spirituality scores were examined to indicate the construct validity of the scores from the RCI-10. Both single items were analyzed using the single indicator latent variable technique (Brown, 2006; Hayduk &

Littvay, 2012). The reliability of the single-item religiosity scale was set at .90 and the reliability of the single-item spirituality item was set at .85 based on the scholarship of religiosity and spirituality scales measuring similar constructs (e.g., Underwood & Teresi, 2002).

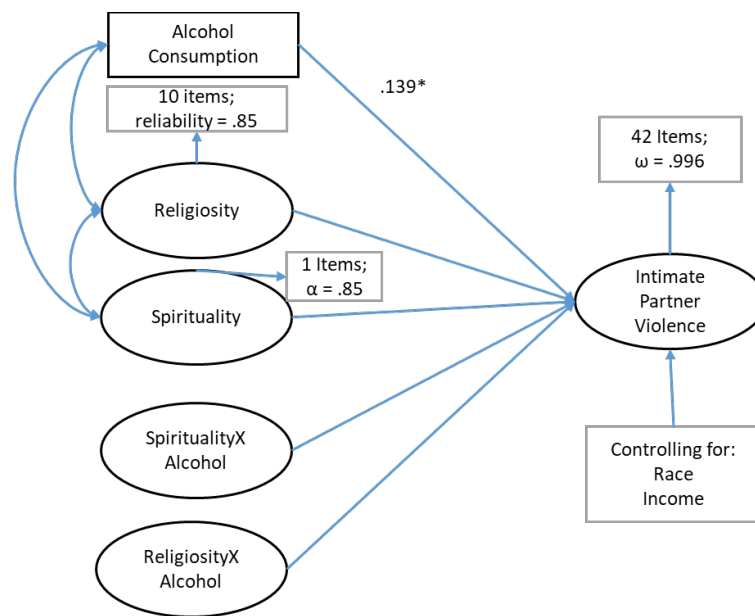
Given that there are gender differences in motivations for abusive behavior (e.g., Rajan & McCloskey, 2007), the final retained model of the RCI-10 and the SVAWS were subjected to a gender comparison. Two series of measurement invariance tests were conducted to examine the invariance of the RCI-10 and the SVAWS. Measurement invariance tests at three levels (i.e., configural, metric, and scalar) were conducted as a prerequisite for multiple group analyses, which were used to assess differences in latent means across gender (female vs. male). Stringent criteria were used for measurement invariance: significance of the change in χ^2 , -.01 change in CFI paired with changes in RMSEA of .015 (Chen, 2007; Reise, Widaman, & Pugh, 1993).

Second, the consequential relationship of the religious commitment with IPV was examined using SEM. The first SEM model of interest tested the influence of alcohol consumption, religiosity, and spirituality on the IPV controlling for race and income (Model 1; Figure 3.1). Given that alcohol consumption might buffer the relationship between religiosity and IPV perpetration, the second model added the interaction between alcohol consumption and religiosity and the interaction between alcohol consumption and spirituality in addition to Model 1 (Model 2). If none of the interaction terms had any influence on the three dependent variables, Model 1 that contained no interaction would be retained as the best model.

We used recommended cut-off values for other indices of fit, including the comparative fit index ($CFI > .90$; Bentler, 1990) and RMSEA $\leq .06$ (Hu & Bentler, 1999). Given that less than 1% of the data were missing, no special treatment was used to deal with missingness. Given that the dependent variables were ordinal in nature, WLSMV estimation was used. All SEM-related analyses were conducted in *Mplus* Version 7.11 (Muthén & Muthén, 2012).



Model 1



Model 2

Figure 3.1. Hypothesized model testing the main effect of alcohol consumption, gender, religiosity, and spirituality on intimate partner violence.

Results

Descriptive statistics of item responses and response process assessment. Prior to the dimensionality analyses, response frequency distributions were first inspected because low response frequencies might lead to extreme slopes and instability of threshold parameters in IRT analyses. Inspection of the data showed no evidence of floor or ceiling effects in item responses (see Appendix F). Also, inspection of the missing data showed that a negligible percentage (0.0% to 0.8%) of the data were missing, indicating no statistical method was needed to address missingness. Given the ordinal nature of the response options, Samejima (1969) GR models was used. Three different GR models including the unidimensional GR model, the two-factor GR model, and the bifactor GR model were fitted to the data.

Local independence assessment. Conditional independence was examined via standardized local dependency (LD) χ^2 statistics (Chen & Thissen, 1997) and values ≤ 10 are acceptable (as cited in IRTPRO User Guide 4.2, p. 85; Toland, 2014). Inspection of the LD χ^2 statistics (see Appendix G) showed that under the unidimensional model only one pair of items (Item 9 and Item 10) had large positive LD χ^2 statistics. This indicated that the assumption of the unidimensional model was tenable. As a contrast, it appears that fitting the two-factor GR model actually increased the local dependence: Two item pairs showed large LD χ^2 values — Items 6 and 9 and Items 6 and 10. Also, the correlation between the two factors was extremely high, $r = 0.88$, indicating that a single latent variable labeled as “religious commitment” could underlie all RCI-10 items and breaking this 10-item scale into two subscales measuring two different latent traits (intrapersonal religious commitment and interpersonal religious commitment) might

indicate redundancy from the general factor. When we inspected the bifactor GR model that accounted for the unique variance explained by the general factor and the unique variance explained by the subscales after controlling for the effect of the general factor, the LD reduced to only one pair of the items: Items 6 and 10. Thus, from the local independence assessment, the unidimensional GR model and bifactor GR model seemed to better represent the internal structure of the RCI-10.

Evaluation of item-level model-data fit. Orlando-Thissen-Bjorner item fit $S-\chi^2$ statistics together with the B-H procedure was used to examine how well a model predicts response behavior at the item level (Benjamini & Hochberg, 1995; Orlando & Thissen, 2000, 2003). Results showed that when a unidimensional model was fitted to the data, only Item 10 had a significant $S-\chi^2$ statistics result, indicating unidimensional model adequately fitted Items 1 to 9 but not Item 10. Also, both the two-factor GR model and bifactor GR model failed to fit Items 6 and 10. Thus, the item-level model-data fit suggested minor problems at the item-level existed when the unidimensional, two-factor, and bifactor models were fit to the data. Some questionability was raised regarding Item 10 if considering both the LD results and the item-fit results.

Global model-data fit and comparison. Once the assumption of the local independence and the item-level fit were found, we could compare the performance of the models using global model-data fit. Table 3.1 summarized the global model-data fit results of the three competing models together with the LD assessment result and the item-level fit results.

As a whole, the bifactor GR model was the champion. Note, it is possible that the bifactor GR model fits the data well relative to other models, but we should also be aware

that this might be due to the model complexity. Thus, to better understand the bifactor GR model results, item parameters, total information functions, and person parameter comparison with the GR model were examined.

Table 3.1

Global Model Fit Results

	Uni-GR	TwoFactor-GR	Bifactor-GR
# of positive LD pairs flagged	1	2	1
# of items fit by model	1	2	2
# of parameters	50	51	60
-2LL	9289.17	9164.21	9052.27
BIC	9389.17	9266.21	9172.27
AIC	9587.73	9468.74	9410.54
$C_2(df)$	260.21 (35)	193.25(34)	89.74(25)
RMSEA _{C2}	.13	.11	.08

Note. The differences in -2LL between two nested models were tested via χ^2 statistics with degrees of freedom equal to the differences in the numbers of parameters to be estimated in the same two models. For example, comparing the two-factor GR model with the uni-GR model, the $df = 1$. The deviance statistic $\Delta G^2 = 9289.17 - 9164.21 = 124.96$. $\Delta G^2(1) = 124.96, p < .001$. Using the same method, all deviance statistics between the nested models were significant at .001.

Comparison of item parameters across the unidimensional GR model and the bifactor GR model. Table 3.2 summarizes the item parameters for the unidimensional GR model. Table 3.3 summarizes the item parameters for the bifactor GR model. A comparison of the general factor conditional slopes showed that nine of the 10 item conditional slopes of the bifactor GR model were larger than the ones from the unidimensional model results. This indicated that in the bifactor GR model, the subscale latent traits might inflate the conditional slopes for the general factor. Marginal slopes of the 10 items that controlled for the effect of the subscale latent trait was then compared with their corresponding conditional slopes (Table 3.2). Results showed marginal slopes

of the general trait are much smaller in size for items with large conditional slopes: Items 4, 5, 9, and 10, indicating the subscale traits might have a big influence on these items. A comparison between the marginal slopes of the 10-items for the general factor in the bifactor GR model and the conditional slope of the 10 items in the unidimensional model showed that these two sets of the slopes were close to each other, indicating both the unidimensional GR solution and the bifactor GR solution reflected a general latent trait measuring the “religious commitment”.

Evaluating explained common variance in a bifactor model. Item-level ECV and factor-level ECV were both computed from the bifactor GR solution results (Table 3.3). Results showed that IECV for Items 1 to 8 are above .85 and IECV for Items 9 and 10 were below .75, indicating the general trait underlying every item and the effect of multidimensionality could be ignored. Also, ECV value for the general factor is .88, for the intrapersonal religiosity is .04, and for the interpersonal religiosity is .08, indicating the specific traits did not process much meaning.

Table 3.2

Unidimensional Graded Response Model Item Parameters Estimates for the RCI-10.

Item	a	c ₁	c ₂	c ₃	c ₄
RCI_intra_1	2.33	2.02	0.11	-1.61	-3.74
RCI_intra_2	3.65	5.51	2.09	-0.68	-3.56
RCI_intra_3	2.87	4.62	2.25	0.23	-2.14
RCI_intra_4	3.29	4.62	2.33	-0.16	-2.74
RCI_intra_5	3.92	4.91	1.83	-1.04	-4.06
RCI_intra_6	3.16	4.59	1.91	-0.21	-2.46
RCI_inter_1	1.98	1.72	0.14	-1.21	-2.63
RCI_inter_2	3.01	4.91	1.75	-0.33	-2.64
RCI_inter_3	2.82	2.90	0.69	-1.13	-3.30
RCI_inter_4	2.23	1.67	0.20	-1.30	-3.21

Note. a = slope. c₁ – c₄ = intercepts.

Table 3.3

Bifactor Model Item Parameter Estimates for the RCI-10

Item	Conditional Slope			Intercept				Factor Loading			IECV			Marginal Slopes		
	a ^G	a ^{Intra}	a ^{Inter}	c ₁	c ₂	c ₃	c ₄	*G	*Intra	*Inter	IECV _G	IECV _{Intra}	IECV _{Inter}	a ^{*G}	a ^{*Intra}	a ^{*Inter}
Intra1	2.88	-0.73		2.38	0.11	-1.95	-4.44	0.84	-0.21		0.94	0.06		2.65	-0.37	
Intra2	4.24	-0.21		6.27	2.37	-0.79	-4.03	0.93	-0.05		1.00	0.00		4.21	-0.08	
Intra3	3.11	1.06		5.13	2.54	0.30	-2.35	0.84	0.29		0.89	0.11		2.64	0.51	
Intra4	4.11	1.65		5.86	3.00	-0.17	-3.49	0.87	0.35		0.86	0.14		2.95	0.63	
Intra5	4.39	1.22		5.61	2.08	-1.16	-4.66	0.90	0.25		0.93	0.07		3.57	0.44	
Intra6	3.31	0.56		4.82	2.03	-0.19	-2.58	0.88	0.15		0.97	0.03		3.14	0.26	
Inter1	2.06		0.76	1.81	0.09	-1.35	-2.81	0.74		0.27	0.88		0.12	2.06		0.48
Inter2	2.93		0.72	4.93	1.73	-0.35	-2.65	0.85		0.21	0.94		0.06	2.93		0.36
Inter3	5.35		3.33	5.68	1.16	-2.40	-6.40	0.82		0.51	0.72		0.28	2.43		1.01
Inter4	2.88		1.81	2.18	0.12	-1.86	-4.36	0.76		0.48	0.71		0.29	1.97		0.92

Note. G = general religious commitment; Intra = intrapersonal religious commitment; Inter = interpersonal religious commitment; c₁ – c₄ = intercepts; a^{*G} = marginal slope for general trait; a^{*Intra} = marginal slope for the intrapersonal religious commitment; a^{*Inter} = marginal slope for the interpersonal religious commitment; IECVG = item explained common variance for the general trait; IECV_{intra} = item-level explained common variance for the intrapersonal religious commitment; IECV_{inter} = item-level explained common variance for the interpersonal religious commitment.

Comparison of the unidimensional GR model trait scores and precision with the bifactor GR model general trait scores. The correlation between the point estimates obtained from the unidimensional GR model trait scores and the bifactor GR model general trait scores were .99, indicating not much difference existed in the trait score estimates. The marginal reliability of the unidimensional solution is .94, whereas the marginal reliability of the general factor in the bifactor GR model is .91, indicating a minor difference existed in the overall score precision. As marginal reliability is only useful if the TIF function is constant, marginal TIF of the bifactor GR model and the TIF of the unidimensional model across the (-3, 3) range of latent trait were summarized (Figure 3.2). Results showed that the unidimensional GR model inflates the precision of the scores outside the latent trait interval [-.04, .04] and deflates the precision of the scores inside the latent trait interval [-0.4, .04]. Thus, although the overall reliability showed similar overall precision of the latent trait scores using these two model solutions, across the range of the latent trait, the latent trait scores from the bifactor GR model for the general factor are more precise. However, as Toland et al. (2017) indicated, if the analysis purpose is to obtain the IRT person estimates first and then use these person parameters in an analysis, then the latent trait scores from the unidimensional model could be used given that not much information was lost. Following model parsimony, we concluded a unidimensional model solution is the best, subscales should not be created and interpreted from the RCI-10, and latent trait scores from the unidimensional models are precise enough for single-level analysis such as regression.

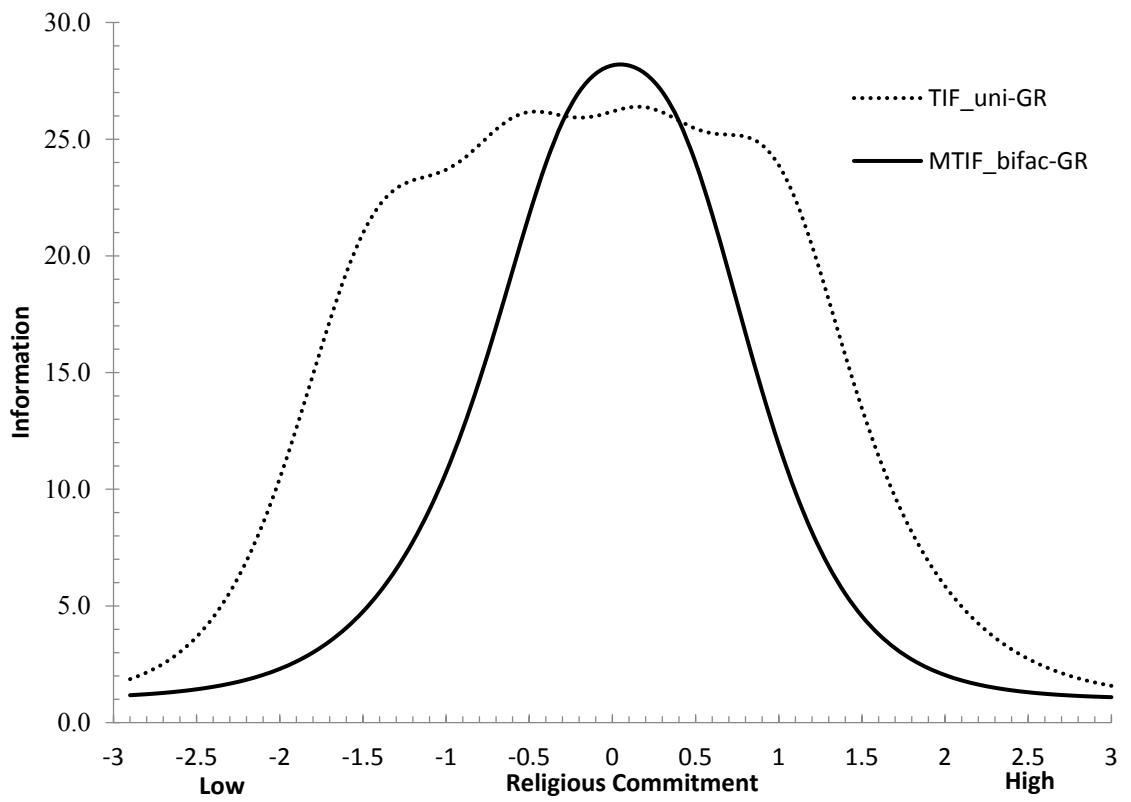


Figure 3.2. Total informational function (TIF) for the RCI-10 fit by the unidimensional graded response (GR) model and marginal TIF (MTIF) for the RCI-10 data fit by the bifactor GR (bifac-GR) model.

Correlational evidence. Correlational relationships of the scores from the RCI-10 were examined with the single-item religiosity measure, the single-item spirituality measure, alcohol consumption, and IPV in *Mplus*. Since the IRT scores of the SVAWS cannot be obtained, in order to keep the scoring method consistent with the scores from the RCI and the ones from the SVAWS, the RCI-10 were subjected to the CFA analyses in *Mplus*. Results showed the scores from the RCI-10 were adequately fitted, $\chi^2(35) = 446.25$, RMSEA = .17, CFI = .966, TLI = .956. Table 3.4 summarizes the correlation results. As expected, positive relationships were found between religiosity, using the RCI-10, with religiosity from the single-item religiosity measure, spirituality from the single-item spirituality measure, and IPV. The scores from the RCI-10 have a strong positive correlation with the scores from the single-item religiosity scale, $r = .65, p < .001$. The scores from the RCI-10 has a medium correlation with the single-item spirituality scale, $r = .49, p < .001$, indicating more religiosity is related with more spirituality. The scores from the RCI-10 also has a weak correlation with IPV, $r = .17, p < .01$.

Table 3.4

Means, Standard Deviations, and Correlations for the Variables in the Study (N = 392)

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5
1. Religiosity	2.08	1.05					
2. Single-item religiosity	1.87	0.80	.65***				
3. Single-item spirituality	3.16	0.83	.49***	.43***			
4. Intimate Partner Violence	1.14	0.35	.17**	.22***	-.11		
5. Alcohol	2.20	1.67	-.09	-.01	-.10	.17*	
6. Gender	-	-	.04	.04	.13*	-.20**	-.10

Note. Religiosity, single-item religiosity, single item spirituality, and intimate partner violence were treated as latent variables. The correlations with these variables were latent correlations. Alcohol and gender were treated as observed variables.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Measurement Invariance. Given that there are gender differences in motivations for abusive behavior (e.g., Rajan & McCloskey, 2007), two separate series tests for measurement invariance were examined first between men and women on the RCI-10 items and the SVAWS items. Results of the measurement invariance tests are summarized in Table 3.5. Results indicated that scalar invariance was reached for both latent variables across gender: p values for $\Delta\chi^2$ ranged from less than .001 to .470, ΔCFI ranged from .000 to .008, $\Delta RMSEA$ ranged from .001 to .031. Thus, multiple group latent mean were examined for religiosity and IPV. Descriptively, women scored higher than men on both latent traits. However, statistical results showed no differences were found between men and women on these two latent traits, $\Delta M_{religiosity} = .077, p = .433$, $\Delta M_{IPV} = .182, p = .508$.

SEM models. The first SEM model tested the effect of religiosity, alcohol consumption, and spirituality on IPV. Figure 3.3 summarizes the results of this prediction model. The effect of race and income were controlled. Alcohol consumption and religiosity were both positively associated with the IPV, indicating individuals who are more religious and consume more alcohol report more frequent violence against their partners. By contrast, spirituality was negatively associated with IPV, indicating individuals who are more spiritual are less likely to engage in intimate partner violence.

Table 3.5

Fit indices of the Measurement Invariance Tests for the 10-item Religious Commitment Inventory Scale (RCI-10) and the 42-item Severity of Violence Against Women Scale (SVAWS)

Model	df	χ^2	$\Delta \chi^2$	p	CFI	Δ CFI	RMSEA	Δ RMSEA
The RCI-10								
Configural Invariance	70	425.723		<.001	.971	-	.161	-
Metric Invariance	79	341.243	15.367	.081	.979	.008	.130	.031
Scalar Invariance	118	387.771	57.350	.029	.978	.001	.108	.022
The SVAWS								
Configural Invariance	1638	1688.862	-	.186	.999	-	.013	-
Metric Invariance	1679	1706.964	18.102	.312	1.000	.001	.009	.004
Scalar Invariance	1762	1794.860	84.896	.287	1.000	.000	.010	.001

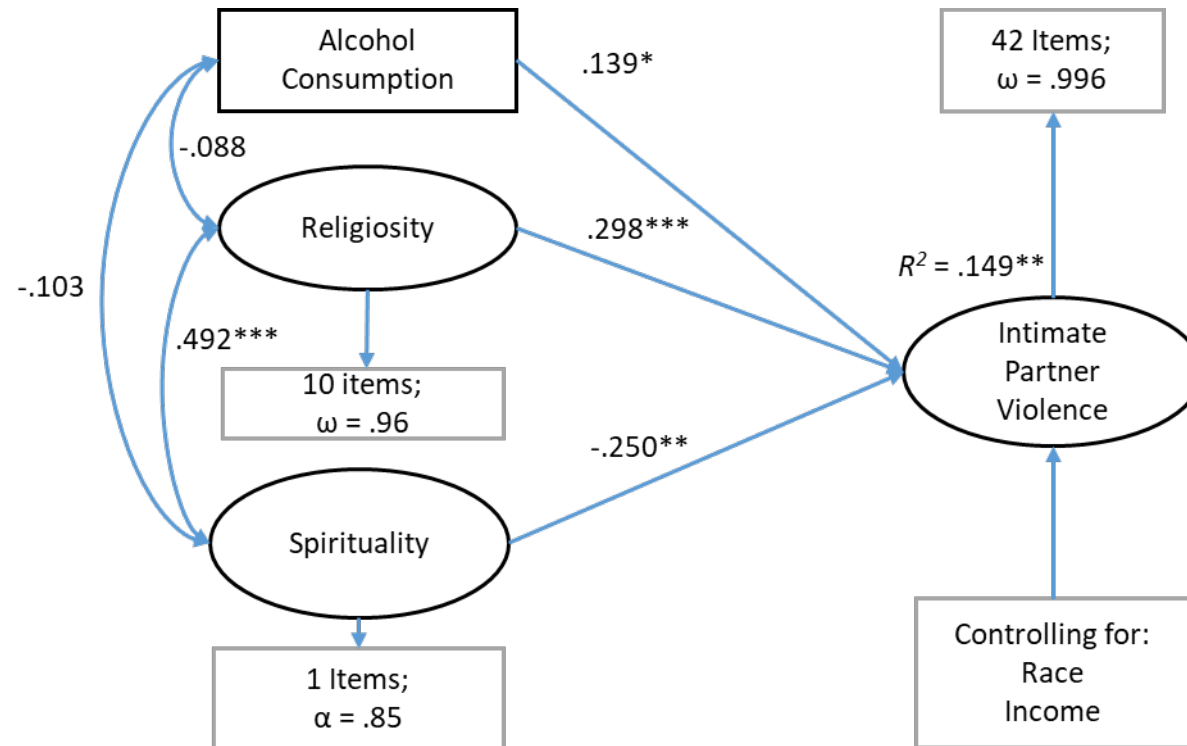


Figure 3.3. Structural equation model testing joint effect of alcohol consumption, religiosity, and spirituality on intimate partner violence, $\chi^2(1,699) = 2,022.252$, RMSEA = .022, 90% CI [.018, .026], CFI = .995, TLI = .995, $R^2 = .149$, $p = .01$. * $p < .05$. ** $p < .01$. *** $p < .001$.

The second SEM model tested the moderation effect of alcohol consumption on the relationship of the IPV on religiosity and spirituality. Figure 3.4 summaries the results of this prediction model. Given that type = random was used to estimate the interaction effect, all coefficients in Figure 3.4 are based on the unstandardized results. We could tell from the following figure that neither of the interaction terms were significant. Thus, Model 1 in Figure 3.3 was retained as the final model.

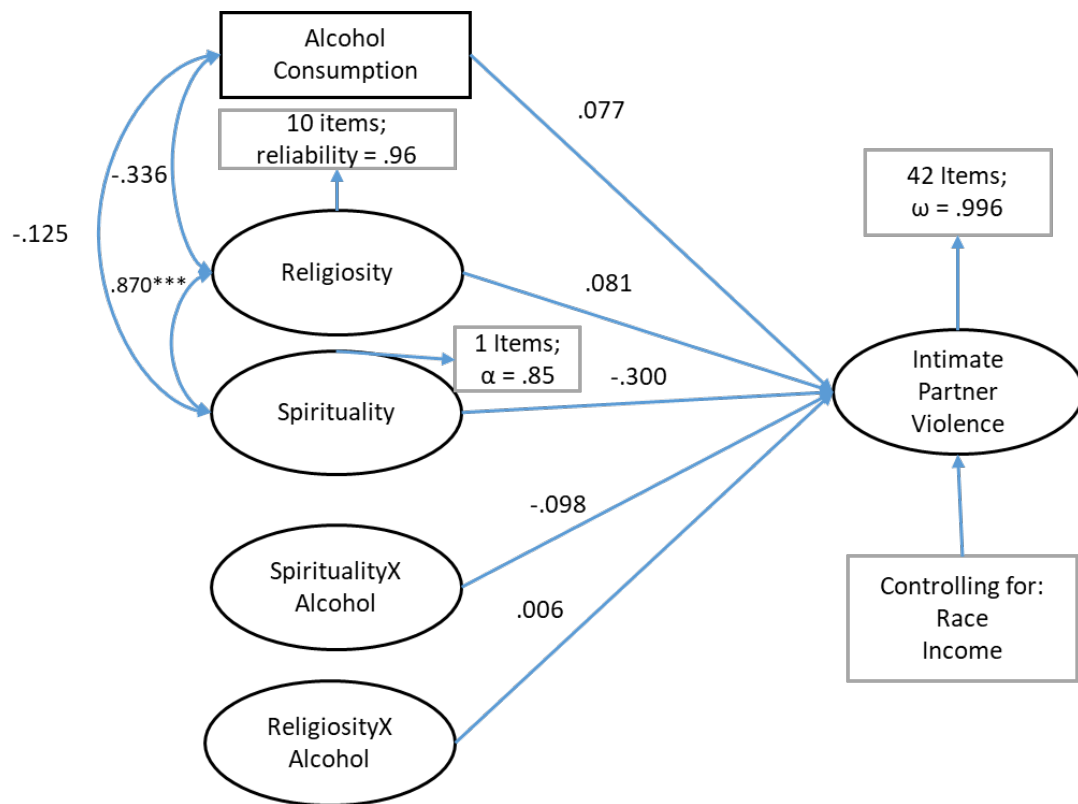


Figure 3.4. Structural equation model testing the latent interaction between alcohol consumption on the relationship between religiosity, spirituality, and the intimate partner violence. Model fit information was not provided due to the use of Type = random to obtain the interaction between a latent variable and an observed variable.

. *** $p < .001$.

Discussion

This study revisited the factor structure of the Religious Commitment Inventory-10 scale (RCI-10; Worthington et al., 2003) and extended our theoretical understanding of the religious commitment construct. It also explored the relationship between religious commitment and IPV in a community-based sample of 392 adults who were religiously affiliated.

By fitting three competing IRT models, the results showed the RCI-10 conformed to a bifactor structure. However, as the bifactor IRT scores correlated with the scores from the unidimensional solution at .99, the RCI-10 could be scored as a unidimensional scale. Our results showed that two-dimension correlational model did not fit the data well. Bifactor ancillary measures also showed that the intrapersonal religious commitment and the interpersonal religious commitment scales should not be created and interpreted as two meaningful concepts. Instead, only one general latent trait (religious commitment) ran through all ten items of the RCI-10. This suggested that the two “so-called” subscales of the RCI-10 are only two content areas of religious commitment. There is indeed only one latent trait underlying the RCI-10: religious commitment.

This study also explored the construct validity of the RCI-10. Results showed that scores from the RCI-10 correlated strongly with the item that measures religiosity and moderately with the item that measures spirituality. Using SEM approach, we also witnessed a positive relationship between religious commitment and IPV. Findings from the correlational results and the SEM results both resonated with the findings from Renzetti et al. (2015), that is, religiosity serves as a risk factor on IPV perpetration. We then explored whether alcohol consumption was a moderator between religious

commitment and IPV, as suggested by Cunradi et al. (2002) and DeWall (2010). However, alcohol consumption did not buffer the relationship between religious commitment and IPV in our sample, although alcohol consumption is found to have a positive relationship with IPV. As an interesting contrast, spirituality was negatively related with IPV, indicating if one is more spiritual, one is less likely to abuse their partner.

A third interesting finding from this study is that there was no gender difference in religious commitment nor IPV. Although researchers have shown motivational differences between men and women for abusive behavior, in our sample, we did not find that intimate partner violence was related with gender. Our finding is even more interesting when there is a big gender gap in the national data of the intimate partner violence, as Tjaden and Thoennes (2006) have reported that each year 7.6-11.5% of men and 12-25% of women are physically and/or sexually assaulted by an intimate partner. Perhaps more studies should be conducted to confirm if the findings from the current study are consistent.

As a conclusion, the RCI-10 should not be used as a multidimensional scale in a community-based sample of adults with religious beliefs.

Appendix A

Descriptive statistics of M_{ord}

Table A1

Descriptive Information of M_{ord} Under Null Conditions

(21,4)										Negative Value(s)	
D	ρ	K	N	df	M	SD	Min	Max	N	%	
2	0.2	4	300	14	14.08	5.46	-18.29	47.35	1	0.1	
2	0.2	4	1000	14	14.00	6.48	-106.11	38.28	1	0.1	
2	0.2	4	3000	14	13.71	23.40	-698.62	68.31	3	0.3	
2	0.2	4	5000	14	15.42	25.44	-8.43	754.54	2	0.2	
2	0.2	5	300	4	11.36	27.48	-417.58	329.12	11	1.8	
2	0.2	5	1000	4	10.84	16.26	-226.84	387.25	13	1.3	
2	0.2	5	3000	4	8.59	60.10	-1840.35	132.90	23	2.3	
2	0.2	5	5000	4	9.65	23.49	-670.30	51.70	21	2.1	
2	0.8	5	300	4	5.81	22.86	-85.25	493.43	34	5.5	
2	0.8	5	1000	4	4.51	63.32	-1714.27	801.19	33	3.3	
2	0.8	5	3000	4	5.52	29.69	-299.65	776.92	35	3.5	
2	0.8	5	5000	4	6.02	18.17	-86.17	536.62	26	2.6	
3	0.2	4	1000	57	57.23	10.84	-4.61	93.91	1	0.1	
3	0.2	4	5000	57	56.78	12.66	-170.07	110.97	1	0.1	
3	0.2	5	300	42	49.67	29.64	-19.09	591.03	2	0.5	
3	0.2	5	1000	42	52.79	114.93	-992.19	3396.82	9	0.9	
3	0.2	5	3000	42	53.53	51.92	-402.77	1025.11	6	0.6	
3	0.2	5	5000	42	49.64	31.32	-577.36	225.95	10	1.0	
3	0.8	5	300	42	37.18	85.59	-1163.27	207.91	3	0.7	
3	0.8	5	1000	42	44.90	82.91	-66.63	2619.61	2	0.2	
3	0.8	5	3000	42	41.92	11.74	-77.98	210.09	1	0.1	
3	0.8	5	5000	42	41.61	10.70	-53.58	110.75	4	0.4	

Note. All results were calculated based on raw M_{ord} results. Number of quadrature points = 21 and theta range was -4 to 4.

Table A2

Descriptive Information of M_{ord} With Conditions Under Alternative Condition

D	ρ	K	N	df	Mean	SD	(21, 4)		Negative Values	
							Min	Max	N	%
2	0.2	4	300	15	70.62	264.10	-173.32	5998.20	6	0.6
2	0.2	4	1000	15	165.99	90.22	-362.84	769.16	8	0.8
2	0.2	4	3000	15	538.97	2485.73	-1600.51	62002.19	23	2.3
2	0.2	4	5000	15	647.23	469.13	-3466.38	3686.78	13	1.3
2	0.2	5	300	5	62.97	50.75	-323.69	206.48	15	2.5
2	0.2	5	1000	5	195.18	228.92	-3416.51	3587.39	19	1.9
2	0.2	5	3000	5	394.7122	1017.43	-	9288	37	3.7
							16546.58			
2	0.2	5	5000	5	571.29	3070.55	-	61258.62	40	4.0
							30436.12			
2	0.8	5	300	5	5.50	80.64	-1642.36	538.74	28	4.8
2	0.8	5	1000	5	9.54	252.80	-7493.97	2163.38	75	7.5
2	0.8	5	3000	5	19.38	170.36	-2779.16	2606.36	145	14.5
2	0.8	5	5000	5	33.83	280.71	-3138.63	5018.71	154	15.4
3	0.2	4	300	60	226.59	728.80	-693.47	13065.44	3	0.5
3	0.2	4	1000	60	471.51	409.42	-8165.33	1106.34	10	1.0
3	0.2	4	3000	60	1298.77	620.14	-	3537.70	6	0.6
							14908.22			
3	0.2	4	5000	60	2048.84	1456.85	-	36812.43	5	0.5
							16102.54			
3	0.2	5	300	45	199.57	59.29	-160.15	605.15	1	0.2
3	0.2	5	1000	45	607.29	1209.50	-1937.93	37198.69	3	0.3
3	0.2	5	3000	45	1469.32	439.00	-2669.07	5314.11	5	0.5
3	0.2	5	5000	45	-	-	-	-	-	-
3	0.8	5	300	45	70.13	29.73	-415.30	211.90	1	0.2
3	0.8	5	1000	45	132.84	35.42	-55.27	615.91	1	0.1
3	0.8	5	3000	45	294.27	68.55	-1166.3	672.92	3	0.3
3	0.8	5	5000	45	469.52	69.71	-202.81	1052.91	1	0.1

Note. The degrees of freedom for conditions with two factors and five response categories are 4.

Appendix B

Results for Kolmogorov-Smirnov Test Under Null Conditions

D	ρ	K	M_2					M_{ord}					C_2				
			300	1000	3000	5000	df	300	1000	3000	5000	df	300	1000	3000	5000	df
1	-	2	.112	.836	.999	.973	5	.112	.836	.999	.973	5	.112	.836	.999	.973	5
	-	3	.502	.598	.886	.996	35	-	-	-	-	0	.641	.907	.756	.925	5
	-	4	.663	.930	.999	.480	85	-	-	-	-	-5	.843	.997	.999	.996	5
	-	5	.001	.243	.701	.999	155	-	-	-	-	-	.983	.935	.911	.476	5
10																	
2	0.2	2	.998	.150	.999	.396	34	.998	.150	.999	.396	34	.998	.150	.999	.396	34
		3	.547	.692	.503	.910	169	.868	.964	.587	.780	24	.915	.823	.428	.264	34
		4	.507	.734	.467	.315	394	.803	.332	.186	.622	14	.760	.864	.825	.980	34
		5	.380	.987	.385	.126	709	< .001	< .001	< .001	< .001	4	.212	.957	.360	.949	34
	0.8	2	.988	.745	.605	.309	34	.988	.745	.605	.309	34	.988	.745	.605	.309	34
		3	.026	.708	.611	.549	169	.451	.819	.936	.894	24	.106	.836	.970	.979	34
		4	.498	.810	.818	.745	394	.982	.992	.559	.717	14	.818	.835	.754	.955	34
		5	.091	.543	.795	.991	709	< .001	< .001	< .001	< .001	4	.288	.295	.069	.982	34
3	0.2	2	.669	.871	.670	.646	87	.669	.871	.670	.646	87	.669	.871	.670	.646	87
		3	.198	.980	.990	.505	402	.508	.977	.926	.986	72	.729	.966	.998	.948	87
		4	.480	.785	.996	.999	927	.511	.671	.227	.848	57	.464	.562	.907	.831	87
		5	.210	.311	.200	.933	1662	< .001	< .001	< .001	< .001	42	.972	.723	.836	.770	87
	0.8	2	.265	.471	.666	.914	87	.265	.471	.666	.914	87	.265	.471	.666	.914	87
		3	.286	.157	.879	.778	402	.897	.982	.999	.991	72	.763	.997	.941	.942	87
		4	.696	.948	.460	.068	927	.246	.534	.539	.690	57	.068	.995	.999	.850	87
		5	.114	.058	.993	.986	1662	.188	.841	.964	.999	42	.486	.631	.945	.492	87

Appendix C

The Religious Commitment Inventory – 10

Item	Content
1	I often read books about my faith.
2	I spend time trying to grow in understanding my faith.
3	Religion is especially important to me because it answers many questions about the meaning of life.
4	My religious beliefs lie behind my whole approach to life.
5	Religious beliefs influence all of my dealings in life.
6	It is important to me to spend periods of time in private religious thought and reflection.
7	I make financial contributions to religious organizations.
8	I enjoy spending time with others who share my religious affiliation.
9	I enjoy working in the activities of my religious organization.
10	I keep well informed about my local religious group and have some influence in its decisions.

Note. The response categories ranged from 1 (not at all true of me) to 5 (very true of me). The first six items belong to the intrapersonal religious commitment scale and the last four items belong to the interpersonal religious commitment scale.

Appendix D

Demographics of the Sample ($N = 392$)

Category	<i>n</i>	%
Age		
18-24	43	11.0
25-34	189	48.2
35-44	99	25.3
45-54	41	10.5
55-64	16	4.1
65 or older	4	1.0
Race		
White American	279	71.2
African American	48	12.2
Hispanic American	21	5.4
Asian American	28	7.1
Native American	7	1.8
Multiracial	6	1.5
Other	3	0.8
Highest level of education or degree completed		
Less than high school	1	0.3
High school graduate (diploma or GED)	36	9.2
Some college, but did not receive a degree	136	34.8
College degree	161	41.2
Graduate or professional degree	57	14.6
Current employment status		
Full time for wages	222	56.8
Part time for wages	51	13.0
Self-employed	46	11.8
Out of work/ looking for work	8	2.0
Out of work/ not looking for work	1	0.3
A student	13	3.3
A homemaker	44	11.3
In the military	-	-
Retired	1	0.3
Unable to work	5	1.3

Annual family income		
Less than \$20,000	42	10.7
\$20,000-\$39,999	94	24.0
\$40,000-\$69,999	134	34.2
\$70,000-\$99,999	83	21.2
\$100,000-\$149,999	28	7.1
\$150,000-\$199,999	6	1.5
\$200,000 or more	5	1.3
Marital Status		
Unmarried and not living with your intimate partner	70	17.9
Unmarried and cohabitating with your intimate partner	94	24.0
Married but not living with your intimate partner	8	2.0
Married and living with your intimate partner	215	54.8
Length of time for intimacy		
1-3 years	107	27.3
3.1 years to 5 years	79	20.2
5.1 years to 10 years	91	23.2
10.1 years to 15 years	49	12.5
15.1 years to 20 years	34	8.7
More than 20 years	32	8.2
Residency		
Northeast	69	17.6
Southeast	130	33.2
Midwest	81	20.7
Southwest	36	9.2
West	75	19.2
Type of area living in		
Urban area	114	29.2
Suburb	186	47.6
Rural area	91	23.3

Appendix E

Psychometric information about the Severity of Violence Against Women Scale
 Table E1
*Frequency and Descriptive Statistics of the SVAWS Used to Measure Intimate Partner
 Violence*

	Never	Once	A few times	Many times	A few times and Many times Combined
SVAWS_1	296	60	33	3	36
SVAWS_2	299	61	25	7	32
SVAWS_3	343	33	10	6	16
SVAWS_4	321	44	21	3	24
SVAWS_5	235	64	81	11	92
SVAWS_6	286	43	51	11	62
SVAWS_7	322	32	30	4	34
SVAWS_8	301	47	37	4	41
SVAWS_9	349	30	9	3	12
SVAWS_10	346	22	17	6	23
SVAWS_11	344	24	20	4	24
SVAWS_12	360	21	7	4	11
SVAWS_13	356	22	12	2	14
SVAWS_14	363	12	13	3	16
SVAWS_15	371	8	9	3	12
SVAWS_16	370	12	9	1	10
SVAWS_17	367	13	7	4	11
SVAWS_18	367	7	10	6	16
SVAWS_19	375	9	5	3	8
SVAWS_20	360	19	11	1	12
SVAWS_21	329	36	22	1	23
SVAWS_22	344	28	17	3	20
SVAWS_23	350	26	10	4	14
SVAWS_24	347	20	20	2	22
SVAWS_25	357	21	10	3	13
SVAWS_26	360	19	12	-	12
SVAWS_27	343	25	18	4	22
SVAWS_28	351	19	14	7	21
SVAWS_29	347	30	12	2	14
SVAWS_30	362	13	16	-	16
SVAWS_31	349	25	15	3	18
SVAWS_32	354	18	12	6	18
SVAWS_33	351	23	15	1	16
SVAWS_34	354	16	14	4	18
SVAWS_35	361	14	14	-	14
SVAWS_36	364	13	9	2	11
SVAWS_37	371	10	9	2	11
SVAWS_38	368	10	10	4	14
SVAWS_39	358	15	9	6	15
SVAWS_40	371	4	16	-	16
SVAWS_41	352	24	13	2	15
SVAWS_42	363	14	12	2	14
SVAWS_43	363	14	9	5	14
SVAWS_44	367	10	13	2	15
SVAWS_45	366	15	8	1	9
SVAWS_46	358	15	15	4	19

Table E2

Correlation Among the Psychological Abuse, Physical Abuse, and the Sexual Abuse

Factors Using the Multidimensional Correlational Model (N = 392)

Variables	1	2
1. Psychological Abuse		
2. Physical Abuse	.97	
3. Sexual Abuse	.95	.99

Table E3

Goodness of Fit Statistics for All Tested Measurement Models (N = 392)

Model	χ^2	<i>df</i>	<i>p</i>	RMSEA	90% CI of RMSEA	CFI	TLI
Unidimensional	1139.975	989	< .001	.020	[.014, .025]	.998	.998
Multidimensional	1097.566	986	.007	.017	[.009, .023]	.998	.998
Bifactor	1012.804	943	.057	.014	[.000, .020]	.999	.999

Note. Given the response categories were ordinal in nature, mean and variance adjusted

weight least squares (WLSMV) estimator was used in the above models. RMSEA =

Root Mean Square Error of Approximation. CI = confidence interval. CFI = comparative

fit index. TLI = Tucker-Lewis Index. The bifactor factor was the champion among the

three models.

Table E4

Confirmatory Factor Analysis Standardized Loadings, Relative Parameter Bias, and Individual Explained Common Variance

Item No.	Unidimensional	Bifactor				IECV
		General Factor	Psychological Abuse	Physical Abuse	Sexual Abuse	
SVAWS_1	.771	.753	.229			.915
SVAWS_2	.776	.763	.194			.939
SVAWS_3	.845	.855	-.050			.997
SVAWS_4	.918	.919	.067			.995
SVAWS_5	.568	.521	.390			.641
SVAWS_6	.730	.668	.549			.597
SVAWS_7	.820	.802	.262			.904
SVAWS_8	.686	.638	.432			.686
SVAWS_9	.875	.850	.287			.898
SVAWS_10	.923	.889	.331			.878
SVAWS_11	.936	.897	.361			.861
SVAWS_12	.951	.947	.127			.982
SVAWS_13	.901	.884	.251			.925
SVAWS_14	.858	.817	.370			.830
SVAWS_15	.953	.922	.329			.887
SVAWS_16	.948	.934	.215			.950
SVAWS_17	.965	.953	.215			.952
SVAWS_18	.976	.975	.084			.993
SVAWS_19	.967	.954	.199			.958
SVAWS_20	.940	.944		.013		1.000
SVAWS_21	.839	.825		.292		.889
SVAWS_22	.912	.902		.252		.928
SVAWS_23	.916	.906		.246		.931
SVAWS_24	.921	.915		.188		.959
SVAWS_25	.958	.957		.106		.988
SVAWS_26	.958	.961		.034		.999
SVAWS_27	.779	.789		-.064		.993
SVAWS_28	.939	.936		.107		.987
SVAWS_29	.926	.915		.242		.935
SVAWS_30	.965	.967		.047		.998
SVAWS_31	.945	.938		.177		.966
SVAWS_32	.938	.934		.140		.978
SVAWS_33	.959	.956		.113		.986
SVAWS_34	.982	.979		.110		.988
SVAWS_35	.978	.981		-.021		1.000
SVAWS_36	.931	.934		.026		.999
SVAWS_37	.969	.974		-.011		1.000
SVAWS_38	.977	.977		.087		.992
SVAWS_39	.966	.966		.062		.996
SVAWS_40	.989	.997		-.131		.983
SVAWS_41	.924	.918			.239	.937
SVAWS_42	.953	.953			.142	.978
SVAWS_43	.959	.951			.319	.899
SVAWS_44	.972	.972			.095	.991
SVAWS_45	.970	.971			.012	1.000
SVAWS_46	.947	.951			.003	1.000

Table E5

Factor-Level Bifactor Indices

	ECV	Omega/OmegaS	OmegaH
Intimate Partner Violence	.944	.996	.980
Psychological Abuse	.040	.987	.083
Physical Abuse	.011	.995	.010
Sexual Abuse	.005	.989	.019

Note. ECV = Explained Common Variance. Omega = Model-based estimate of internal reliability of the general factor in the multidimensional composite. OmegaS = Model-based estimate of the internal reliability of the group-specific factor in the multidimensional composite. OmegaH = the percent of the meaningful variance in raw total scores that can be attributed to the individual differences on the general factor.

Appendix F

Frequency and Descriptive Statistics of the Religious Commitment Inventory – 10

	Not at all true	A little true	Somewhat true	Quite true	Very true	<i>M</i>	<i>SD</i>	Skewness	Kurtosis
Intra_1	94	95	94	73	36	1.65	1.28	0.25	-1.03
Intra_2	36	78	105	101	69	2.23	1.22	-0.16	-0.93
Intra_3	35	60	83	113	99	2.46	1.27	-0.44	-0.87
Intra_4	43	56	100	107	85	2.35	1.27	-0.35	-0.87
Intra_5	50	74	107	99	61	2.12	1.25	-0.14	-0.96
Intra_6	41	68	89	101	93	2.35	1.30	-0.30	-1.02
Inter_1	100	89	82	63	58	1.72	1.39	0.27	-1.18
Inter_2	32	81	94	101	82	2.31	1.24	-0.20	-1.01
Inter_3	73	88	90	83	58	1.91	1.33	0.07	-1.15
Inter_4	106	78	87	74	45	1.68	1.36	0.22	-1.78

Appendix G

Local Dependency Result of the RCI-10

Unidimensional GR Model LD Statistics										
Item	χ^2	1	2	3	4	5	6	7	8	9
1	0.7									
2	1.3	2.8p								
3	1.5	4.7n	2.0n							
4	1.8	2.1n	1.6n	4.3p						
5	2.4	2.3n	0.0n	1.3n	5.4p					
6	2.6	4.1n	3.6p	1.6p	0.9n	0.8p				
7	0.6	3.5p	2.6p	4.7n	1.8n	2.7n	6.7n			
8	2.0	1.5n	0.7n	5.2n	1.7n	2.3n	3.6n	4.9n		
9	0.9	2.2p	3.6n	7.2n	3.8n	2.8n	9.9n	3.6p	4.5p	
10	1.2	0.9p	3.2n	4.7n	7.1n	7.9n	13.7n	4.3p	4.1n	11.8p
Two-factor GR Model LD Statistics										
Item	χ^2	1	2	3	4	5	6	7	8	9
1	0.6									
2	1.1	2.9p								
3	1.5	4.7n	2.4n							
4	1.9	2.1n	2.4n	3.8p						
5	2.4	2.3n	0.1n	1.4n	4.2p					
6	2.3	4.1n	3.0p	0.9p	1.3n	0.8n				
7	0.5	4.3p	3.1p	3.9n	1.5n	1.9n	5.2n			
8	2.0	1.5p	3.0p	6.1p	2.9p	3.9p	2.8p	5.7n		
9	1.7	3.3p	3.4n	7.7n	4.2n	2.8n	11.1n	2.2n	3.5n	
10	1.8	1.9p	3.4n	4.5n	7.3n	9.1n	14.4n	2.7n	5.4n	5.2p
Bifactor GR Model LD Statistics										
Item	χ^2	1	2	3	4	5	6	7	8	9
1	0.6									
2	1.6	1.7n								
3	1.2	3.3n	2.2n							
4	1.4	0.9n	1.6n	3.3n						
5	2.1	2.8n	0.0n	2.0n	2.9n					
6	2.1	4.2n	3.4p	0.9p	0.0n	0.7n				
7	0.6	3.0p	2.9p	4.1n	1.5n	2.2n	6.1n			
8	2.3	1.9n	0.9n	5.5p	2.0p	2.7n	2.8n	5.4n		
9	2.2	2.8p	4.1n	7.4n	4.0n	1.9n	9.5n	2.2n	3.7n	
10	1.9	1.7p	3.6n	3.5n	6.2n	7.0n	12.0n	2.8p	4.7n	5.6n

References

- Agresti, A. (1990). *Categorical Data Analysis*. New York, NY: Wiley.
- Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, NJ: Educational Testing Service.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.
- Retrieved from: <http://www.jstor.org/stable/2346101>
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3–5. doi:10.1177/1745691610393980
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed). New York, NY: Springer.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
- doi: 10.1007/BF02293801
- Browne, M. W., & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Newbury Park, CA: SAGE Publications.
- Cai, L. (2015). flexMIRT[®]: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

- Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical & Statistical Psychology*, 66(2), 245-276. doi:10.1111/j.2044-8317.2012.02050.x
- Cai, L., & Monroe, S. (2014). A New Statistic for Evaluating Item Response Theory Models for Ordinal Data. *CRESST Report 839. National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.
- Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.
- Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C. -W., & Oguzhan, O. (2018). MIRT: Multidimensional item response theory. In *R package, version 1.28*.
Retrieved from: <https://cran.r-project.org/web/packages/mirt/index.html>
- Chen, W. -H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289. doi:10.3102/10769986022003265.
- Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40, 5-32. doi:10.1007/bf02291477
- Cunradi, C. B., Caetano, R., & Schafer, J. (2002). Socioeconomic predictors of intimate partner violence among white, black, and Hispanic couples in the United States. *Journal of Family Violence*, 17(4), 377-389. doi:10.1023/A:1020374617328
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

- DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, 13(4), 354-378. doi:10.1080/15305058.2013.799067
- DeWall, C. N. (2010, October). *God give me self-control strength: Unlocking the mystery between religiosity and self-control*. Paper presented at the Centre for Research on Self and Identity, University of Southampton, Southampton, England.
- Diener, E., Emmons, R. A., Larsen, R. J. & Griffin, S., (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71-75.
doi: 10.1207/s15327752jpa4901_13
- Ellison, C. G. & Anderson, K. L. (2001). Religious involvement and domestic violence among U.S. couples. *Journal for the Scientific Study of Religion*, 40(2), 269-282.
doi:10.1111/0021-8294.00055
- Ellison, C. G., Trinitapoli, J. A., Anderson, K. L., & Johnson, B. R. (2007). Race/ethnicity, religious involvement, and domestic violence. *Violence Against Women*, 13(11), 1094-1112. doi:10.1177/1077801207308259
- Follingstad, D. R., Coker, A. L., Chahal, J. K., Brancato, C. J., & Bush, H. M. (2016). Do guns in the home predict gender and relationship attitudes? An exploratory study. *Journal of Aggression, Maltreatment & Trauma*, 25(10), 1097-1116.
doi:10.1080/10926771.2016.1225144
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14, 275-299.
doi:10.1037/a0015825

- Fossati, A., Widiger, T. A., Borroni, S., Maffei, C., & Somma, A. (2015). Item response theory modeling and categorical regression analyses of the five-factor model rating form. *Assessment*, 1-17. doi:10.1177/1073191115621789
- Frazier, P., Greer, C., Gabrielsen, S., Tennen, H., Park, C., & Tomich, P. (2013). The relation between trauma exposure and prosocial behavior. *Psychological Trauma: Theory, Research, Practice, and Policy*, 5(3), 286-294. doi:10.1037/a0027255
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4), 347-360.
- Retrieved from: <http://www.jstor.org/stable/1434586>
- Hansen, M., Cai, L., Monroe, S., & Li, Z. (2014). Limited-information goodness-of-fit testing of diagnostic classification item response theory models. *CRESST Report 840*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.
- Hanson, R. A., Templin, J.L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Healy, K. (2005). *Social work theories in context: Creating frameworks for practice*. Houndmills, UK: Palgrave Macmillan.
- IRTPRO user guide 4.2. (2018). Retrieved from <http://www.ssicentral.com/irt/>
- Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, 7(109), 1-10. doi:10.3389/fpsyg.2016.00109

- Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, 75(3), 393-419. doi:10.1007/s11336-010-9165-5
- Johnson, A. J., & Stephens, R. L. (2015). Concluding thoughts. In A. J. Johnson (Ed.), *Religion and men's violence against women* (pp. 453-469). New York, NY: Springer. doi:10.1007/978-1-4939-2266-6_29
- Jurich, D. P. (2014). *Assessing model fit of multidimensional item response theory and diagnostic classification models using limited-information statistics*. James Madison University.
- Kass, R. E., & Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Koehler, K. J. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, 75(370), 336-344. doi:10.1080/01621459.1980.10477473
- Kreidler, M. C. (1995). Victims of family violence: The need for spiritual healing. *Journal of Holistic Nursing*, 13(1), 30-36. doi:10.1177/089801019501300105
- Maulana, Helms-Lorenz, and van de Grift (2015). Pupils' perceptions of teaching behaviour: Evaluation of an instrument and importance for academic motivation in Indonesian secondary education. *International Journal of Educational Research*, 69, 98-112.
- Marshall, L. L. (1992). Development of the severity of violence against women scales. *Journal of Family Violence*, 7(2), 103-121. doi:10.1007/BF00978700

- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71-101.
doi:10.1080/15366367.2013.831680
- Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2ⁿ contingency tables: A unified framework. *Journal of the American Statistical Association, 100*(471), 1009-1020.
doi:10.1198/016214504000002069
- Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713-732.
doi:10.1007/s11336-005-1295-9
- Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305-328.
doi:10.1080/00273171.2014.911075
- Maydeu-Olivares, A., & Montaña, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika, 78*(1), 116-133. doi:10.1007/s11336-012-9293-1
- McCullough, M. E. & Willoughby, B. L. B. (2009). Religion, self-regulation, and self-control: Associations, explanations, and implications. *Psychological Bulletin, 135*(1), 69-93. doi:10.1037/a0014213
- McCullough, M. E. & Worthington, E. L., Jr. (1995). Promoting forgiveness: A comparison of two brief psychoeducational group interventions with a waiting-list control. *Counseling and Values, 40*(1), 55-68.
doi:10.1002/j.2161-007X.1995.tb00387.x

- McCullough, M. E., Worthington, E. L., Jr., Maxey, J., & Rachal, K. C. (1997). Gender in the context of supportive and challenging religious counseling interventions. *Journal of Counseling Psychology, 44*(1), 80-88.
doi:10.1037/0022-0167.44.1.80
- Miller, W. A., Shepperd, J. A., & McCullough, M. E. (2013). Evaluating the Religious Commitment Inventory for adolescents. *Psychology of Religion and Spirituality, 5*(4), 242-251. doi:10.1037/a0031694
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*(4), 551-560. doi:10.1007/bf02293813
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*(4), 599-620.
doi:/10.1207/S15328007SEM0904_8
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide (7th ed.)*. Los Angeles, CA.
- Orlando, M., & Thissen, D. (2003). Further examination of the performance of S-X2, an item fit index for dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298. doi:10.1177/0146621603027004004
- Pavot, W. & Diener, E. (1993). Review of the satisfaction with life scale. *Psychological Assessment, 5*(2). 164-172. doi:10.1037/1040-3590.5.2.164
- Pavot, W. G., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment, 57*(1), 149-161.
doi:10.1207/s15327752jpa5701_17

- Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157-175. doi:10.1080/14786440009463897
- Pew Research Center. (2018). *June 4-September 30, 2014 - Pew research center 2014 religious landscape study (RLS-II)* [survey]. Retrieved from <http://www.pewresearch.org>
- Polak, J. & Grabowski, D. (2017). Preliminary psychometric characteristics of the Polish version of the Religious Commitment Inventory–10 (RCI-10 PL) by Everett Worthington and colleagues. *Roczniki Psychologiczne*, 20, 213-234. doi:10.18290/rpsych.2017.20.1-6en
- Presley, C. A., Meilman, P. W., & Leichliter, J. S. (2002). College factors that influence drinking. *Journal of Studies on Alcohol Supplement*, 14, 82-90. doi:10.15288/jsas.2002.s14.82
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696. doi:10.1080/00273171.2012.715555
- Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, 61(3), 509-528. doi:10.1007/bf02294552
- Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. *Sociological Methodology*, 29, 81-111. doi:10.1111/0081-1750.00061

- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544-559.
doi:10.1080/00223891.2010.496477
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31. doi:10.1007/s11136-007-9183-7
- Renzetti, C. M., DeWall, C. N., Messer, A., & Pond, R. (2015). By the grace of god: Religiosity, religious self-regulation, and perpetration of intimate partner violence. *Journal of Family Issues, 38*(14), 1974-1997.
doi:10.1177/0192513X15576964
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. doi:10.1037/met0000045
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Saltzman, L. E., Fanslow, J. L., McMahon, P. M., & Shelley, G. A. (1999). *Intimate partner violence surveillance: Uniform definitions and recommended data elements (Version 1.0)*. Atlanta, GA: Centers for Disease Control and Prevention.
Retrieved from: <https://stacks.cdc.gov/view/cdc/7537>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17(Part 2), 34.
Retrieved from: <http://www.psychometrika.org/journal/online/MN17.pdf>

- Stahre, M., Naimi, T., Brewer, R., & Holt, J. (2006). Measuring average alcohol consumption: the impact of including binge drinks in quantity–frequency calculations. *Addiction, 101*(12), 1711-1718.
doi:10.1111/j.1360-0443.2006.01615.x
- Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the Psychometrika Society meeting, Iowa City, IA.
- Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modelling. *Personality and Individual Differences, 42*, 893–898.
- Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*(1), 58-75. doi:10.1111/j.1745-3984.2000.tb01076.x
- Stone, C. A., Ankenmann, R.D., Lane, S., & Liu, M. (1993). *Scaling QUASAR's performance assessment*. Paper presented at the 1993 annual meeting of the American Educational Research Association, Atlanta
- Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183– 206). New York, NY: Routledge/Taylor & Francis Group.
- Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement, 37*, 41-57. doi:10.1177/0146621612462759

- Symonds, M. R. E., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference, and model averaging in behavioral ecology using the Akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*, 13-21.
- Ten Berge, J.M.F. & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613-625.
doi:10.1007/BF02289858
- Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–66). New York, NY: Springer-Verlag.
- Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*, 120–151. doi:10.1177/0272431613511332.
- Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, *60*, 41-63. doi:10.1016/j.jsp.2016.11.001
- Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical & Statistical Psychology*, *56*, 271-288.
doi:10.1348/000711003770480048
- Tsang, J., McCullough, M. E., & Hoyt, W. T. (2005). Psychometric and rationalization accounts of the religion-forgiveness discrepancy. *Journal of Social Issues*, *61*(4), 785-805. doi:10.1111/j.1540-4560.2005.00432.x

- Wade, N. G., Bailey, D. C., & Shaffer, P. (2005). Helping clients heal: Does forgiveness make a difference? *Professional Psychology: Research and Practice*, 36(6), 634-641. doi:10.1037/0735-7028.36.6.634
- Wade, N. G., Worthington, E. L., Jr., & Vogel, D. L. (2007). Effectiveness of religiously tailored interventions in Christian therapy. *Psychotherapy Research*, 17(1), 91-105. doi:10.1080/10503300500497388
- Wang, M., Horne, S. G., & Levitt, H. M. (2009). Christian women in IPV relationships: An exploratory study of religious factors. *Journal of Psychology & Christianity*, 28, 224-235.
- Wighting, M. J. & Liu, J. (2009). Relationships between sense of school community and sense of religious commitment among Christian high school students. *Journal of Research on Christian Education*, 18(1), 56-68. doi:10.1080/10656210902751834
- Wilks, S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60-62. doi:10.1214/aoms/1177732360
- Worthington, E. L., Jr. (1988). Understanding the values of religious clients: A model and its application to counseling. *Journal of Counseling Psychology*, 35(2), 166-174. doi:10.1037/0022-0167.35.2.166
- Worthington, E. L., Jr., Hsu, K., Gowda, K., & Bleach, E. (1988, November). *Preliminary tests of Worthington's (1988) theory of important values in religious counseling*. Paper presentation to the International Congress on Christian Counseling, Atlanta, GA.

- Worthington, E. L., Jr., Wade, N. G., Hight, T. L., Ripley, J. S., McCullough, M. E., Berry, J. W., Schmitt, M. M., Berry, J. T., Bursley, K. H., & O'Connor, L. (2003). The Religious Commitment Inventory--10: Development, refinement, and validation of a brief scale for research and counseling. *Journal of Counseling Psychology, 50*(1), 84-96. doi:10.1037/0022-0167.50.1.84
- Xu, J., Paek, I., & Xia, Y. (2017). Investigating the behaviors of M_2 and $RMSEA_2$ in fitting a unidimensional model to multidimensional data. *Applied Psychological Measurement, 41*(8), 632-644. doi:10.1177/0146621617710464

Vita

Caihong Rosina Li

Education:

May 2017 Graduate Certificate in Applied Statistics, University of Kentucky
June 2015 Master of Science, Educational Psychology, University of
Kentucky
June 2009 Bachelor of Science, Educational Technology, Tianjin Foreign
Studies University, Tianjin, China

Awards:

Fall 2016 – Spring 2017 Helen Thacker Graduate Fellowship (\$4,000)
Fall 2015 – Summer 2016 Graduate School Academic Year Fellowship (\$15,000)
Fall 2014 – Spring 2015 Graduate Student Fellowship, Ashland Inc. (\$7,400)
Fall 2014 – Spring 2015 International Student Tuition Scholarship (\$2,000)

Publication:

Usher, E. L., **Li, C. R.**, Butz, A. R., & Rojas, J. P. (2018). Grit and self-efficacy: Are both essential for children's academic success? *Journal of Educational Psychology*. [SEM]

Usher, E. L., Ford, C. J., **Li, C. R.**, & Weidner, B. L. (2018). Math and Science Self-Efficacy Development in Rural Appalachia: Converging Mixed Methods. *Contemporary Educational Psychology*. [CFA, SEM, mixed method]

Fedewa, A. L., Toland, M. D., Usher, E. L., & **Li, C. R.** (2016). Relationships Between Children's Physical and Psychological Characteristics. *The International Electric Journal of Elementary Education*, 9, 151-166. [correlational]

Mamaril, N. A., Usher, E. L., **Li, C. R.**, Economy, D. R., & Kennedy, M. S. (2016). Measuring undergraduate students' engineering self-efficacy: A scale validation. *Journal of Engineering Education*, 105, 366-395.
doi: 10.1002/jee.20121. [EFA, CFA, reliability, regression]

Usher, E. L., Mamaril, N. A., **Li, C. R.**, Economy, D. R., & Kennedy, M. S. (2015). Sources of self-efficacy in undergraduate engineering. *Proceedings of the 2015 ASEE Annual Conference and Exposition*, Seattle, Washington. [Qualitative]

Kennedy, M. S., Usher, E. L., Mamaril, N. A., Economy, D. R., **Li, C. R.**, & Sharp, J. (2015). Undergraduate students' materials science and engineering self-efficacy: Assessment and implications. *Proceedings of the 2015 ASEE Annual Conference and Exposition*, Seattle, Washington. [CFA]