2017

Dimensionality and Instrument Validation in Factor Analysis: Effect of the Number of Response Alternatives

Alexander G. Hall
University of South Carolina

DIMENSIONALITY AND INSTRUMENT VALIDATION IN FACTOR ANALYSIS: EFFECT OF THE NUMBER OF RESPONSE ALTERNATIVES

by

Alexander G. Hall

Bachelor of Arts
University of New Mexico, 2015

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Arts in

Experimental Psychology

College of Arts and Sciences

University of South Carolina

2017

Accepted by:

Amanda Fairchild, Director of Thesis

Alberto Maydeu-Olivares, Reader

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

ACKNOWLEDGEMENTS

I would first like to thank my major professor and thesis chair, Dr. Amanda Fairchild, whose dedicated feedback has time and again proven to be both reliable and valid – essential characteristics in the field of psychometrics. It has been a pleasure being able to work alongside you. I would also like to thank my thesis committee member, Dr. Alberto Maydeu-Olivares, for his continued assistance and guidance concerning both this thesis and my professional development. A great many things become possible when we understand that everything is connected, and everything is easy. Finally, I'd like to thank my parents and brother, without whom none of this would be *probable*.

ABSTRACT

Despite the great prevalence in both research and application of Factor Analysis (FA), widespread misinterpretation continues to pervade the psychological community in its application for the development and evaluation of psychometric tools. Fundamental measurement questions such as the number of response alternatives needed, and the power to detect poor model fit in non-normal or misspecified data, still remain in need of further investigation. For example, the power of the chi-square statistic used in structural equation modeling decreases as the absolute value of excess kurtosis of the observed data increases. This issue is further compounded with discrete variables, where increasing kurtosis manifests as the number of item response categories is reduced; in these cases, the fit of a confirmatory factor analysis model will improve as the number of response categories decreases, regardless of the true underlying factor structure or X^2-based fit index used to examine model fit. Such artifacts have critical implications for the assessment of model fit, as well as validation efforts. To garner additional insight into the phenomenon, a simulation study was conducted to evaluate the impact of distributional nonnormality, model misspecification and model estimator on tests of model fit when true factor structure is known. Results indicate that effects of excess kurtosis and number of scale categories are exacerbated by model misfit. We discuss results and provide substantive recommendations. We also demonstrate an empirical example of how number of response options impacts dimensionality assessment through evaluation of the Beck Hopelessness Scale (BHS).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BHS ..... Beck Hopelessness Scale

CFA ..... Confirmatory (restricted) Factor Analysis

EFA ..... Exploratory (unrestricted) Factor Analysis

FA ..... Factor Analysis

ML ..... Maximum Likelihood

MLMV ..... Maximum Likelihood with Satorra-Bentler mean and Variance χ^2

PCA ..... Principal Components Analysis

RMSEA ..... Root Mean Squared Error of Approximation

SRMR ..... Standardized Root Mean Squared Residual

ULS ..... Unweighted Least Squares

FOREWORD

"Methods of experimental design and data analysis derive their value from the contributions they make to the more general enterprise of science." – Maxwell & Delaney, 2004

"It is apparent that the common practice of factor analysis lags behind theoretical knowledge and the possible uses of it." – Richard L. Gorsuch, 1983

Statement of problem

Despite the great prevalence in both research and application of Factor Analysis (FA), widespread misuse continues to pervade the psychological community in its application for the development and evaluation of psychometric tools. Fundamental measurement questions such as the number of response alternatives needed, and the power to detect poor model fit in non-normal or misspecified data, still remain in need of further investigation. For example, the power of the chi-square statistic used in structural equation modeling decreases as the absolute value of excess kurtosis of the observed data increases. This issue is further compounded with discrete variables, where increasing kurtosis manifests as the number of item response categories is reduced; in these cases, the fit of a confirmatory factor analysis model will improve as the number of response

categories decrease, regardless of the true underlying factor structure or X^2-based fit index used to examine model fit. Such artifacts have critical implications for the assessment of model fit, as well as validation efforts. To garner additional insight into the phenomenon, a simulation study was conducted to evaluate the impact of distributional nonnormality, model misspecification and model estimator on tests of model fit when true factor structure is known. Results indicate that effects of excess kurtosis and number of scale categories are exacerbated by model misfit. We discuss results and provide substantive recommendations. We also demonstrate an empirical example of how number of response options impacts dimensionality assessment through evaluation of the Beck Hopelessness Scale (BHS).

CHAPTER 1

INTRODUCTION

In its broadest form, psychology as a subdivision of the greater scientific pursuit, seeks to study the mind and its functions via description and inference. In both forms, a cornerstone of the endeavor is measurement, which requires that we be able to aptly describe phenomena. More precisely, Bollen (1989, pg. 180) defines measurement as "the process by which a concept is linked to one or more latent variables, and these are linked to observed variables." Implicit in this, is the independence of 'concept' as a construct of the human creation. In keeping with this understanding, as well as to align with contemporary work, we will use the term 'construct' to reference any variable captured through measurement. Bollen continues that "the concept [construct] can vary from one that is highly abstract… to one that is more concrete," which for practical interpretation can be clarified by the distinction between observed [manifest] and unobserved (latent) constructs (p. 180). While the measurement of manifest constructs (e.g., height, weight, number of toes) is generally straightforward, latent constructs rarely if ever correspond in a 1:1 sense with physically measurable reality. By definition, the scope of psychology extends to unobservable constructs (e.g., intelligence, depression, personality) such that neither description nor inference would be feasible in the absence of measurement.

Bollen's commentary on the span of the concept (construct) from abstract to concrete can also be seen within the development of a single construct. Inextricably bound in the idea of measurement is validity, which refers to a measurement's ability to capture the truth of the construct (Bollen, 1989). The conceptualization of validity as applied to latent measurement has undergone dramatic transformation(s) over the last several decades, moving from a relatively concrete conceptualization involving multiple types of related but independent validity, which often had a single validity coefficient (used in much the same way as a *p*-value to reject or fail to reject a measure's validity; Bollen,1989), to more contemporary work which considers a highly context dependent, holistic account that does not make quantitative decision rules on the basis of singular coefficients (Cronbach, 1980; Cronbach & Meehl, 1955; Lissitz, 2009; Messick, 1989)[1]. There is also a divergence between the ways in which validity is discussed in an experimental context versus a psychometric context[2]- the concentration of this project is on the latter.

Despite the lack of unanimity in the field with respect to validity, key aspects critical for psychometric application remain largely constant and agreed upon. One such facet is that validity cannot be universally proven, but instead must be established on a case by case basis for a given use (Bollen, 1989; Cronbach, 1971; Lissitz, 2009; Messick, 1989). Other key constants of validity include our understanding of the functionality of content, criterion, and construct validity. Putting aside the relationship of this triad (often in contemporary work construct validity is operationalized as subsuming content and

[1] For an overview of validities development through the 1980's see Shepard (1993).
[2] For language on validity utilized in experimental psychology see Maxwell & Delaney (2004).

criterion validity) these three elements are generally considered essential aspects of validity (Crocker & Algina, 1986).

Content validity is the most qualitative of the validity dimensions and is often given special consideration in measure development as it seeks to assure that the manifest variables are consistent with the construct's conceptualization (Messick, 1989). Generally content validity relies on substantive experts. Criterion validity, as the name implies draws empirical comparisons as to the degree of correspondence between a measure and a criterion variable – usually measured by their correlation (Shepard, 1993). In cases where the criterion exists in the same temporal space as the measure being validated, it is called concurrent validity. In cases where the criterion occurs in the future (such as test scores used to predict later achievement) it is called predictive validity. Construct validity assesses the degree to which a construct's measure relates to other manifest variables in a way that is consistent with theoretically derived predictions. That is if scores on a measure are related to other similar constructs (convergent validity) but independent from or unrelated to dissimilar constructs (discriminant validity), an instrument demonstrates evidence for good construct validity. Though far from exhaustive, this overview lends insight into the most cited definition of validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions," which in turn conforms well to the understanding of a construct as a human constructed concept (Messick, 1989).

Developing a valid measure

Bollen (1989) describes the measurement process as consisting of four steps: 1) conceptualization, 2) dimensionality identification 3) measure forming, 4) structural specification. In turn, recommendations for measure development closely mirror this; 1) define the domain of interest so as to assure the manifest variables are representative of the construct and come from the corresponding universe of items applicable to the construct, 2) examine item analysis, reliability analysis, FA etc. and 3) seek to examine the measures' convergent and discriminant relationships with other established measures (Benson & Hagtvet, 1996). After identifying the construct of interest, the most important element in these processes is identifying that construct's dimensionality. As we see, this is because its dimensionality is a critical component of all subsequent steps (inclusive of measure validation). For example, when you discuss convergent and discriminant validity, hypothesized relationships are based directly off of the purported dimensionality so that an error in estimated number of dimensions will change the interpretation of those dimensions and therefore change the other measures you're examining for convergence and discrimination. In CFA, incorrectly specified dimensionality might compromise measurement further by limiting a researcher's ability to identify poorly functioning items, as factor pattern and factor structure weights of a given solution are inextricably linked to specified dimensionality, and additional sources of covariation among observed measures may not be accounted for by the specified factors and will remain classified as unexplained variation. Finally, and most globally, incorrect dimensionality assessment leads to future incorrect model specification, which necessarily compromises substantive

inferences made from measures, and has even recently been identified as critical for sound substantive inferences.

Difficulties of validating latent constructs

Even in the case of manifest constructs, validity can only be "proved" to the extent (in a given context) that its operational definition can be agreed upon. The classic example of this is temperature; that a thermometer is a valid instrument of measurement for the manifest construct of temperature is true insofar as it truly aligns with the construct of temperature. Evaluating validity evidence for latent constructs – the validation of which must necessarily still concern itself with properties of measured variables – is a more precarious process. Inherent challenges in dealing with latent constructs include the need to set a meaningful scale for the variable (Bollen, 1989), as well as incorporating latent variables into statistical analysis. Without explicitly including latent variables, one is left to assume that correlations – which are not a measure of validity – accurately reflect associations that involve latent constructs. Bollen presents this difficulty and primes his answer, FA, by rhetorically asking "What if we could estimate the relationship between a latent variable and its measure?" (p. 195).

Factor Analysis

FA refers to a subset of covariance structure analysis in which latent variables are used to formally operationalize measurement of latent constructs. Confirmatory factor analysis (CFA), also called restricted factor analysis, can be subsumed under the greater scope of structural equation modeling, while exploratory factor analysis (EFA), also called unrestricted factor analysis contributes only indirectly to measurement

operationalization. That is, in the context of EFA, the models are used primarily to help determine the number of dimensions that a latent construct comprises. More generally, FA analyzes construct dimensionality of a given instrument via an evaluation of the variation and covariation among a set of manifest variables associated with the measure (Brown & Cudeck, 1993). In short, FA presumes that a latent construct underlies the shared variation in a set of manifest variables. In this way, FA can be seen as providing a more parsimonious representation of relationships between the manifest variables.

Originally conceived by Spearman (1904) and further developed by Thurstone (1947), FA is an extension of the general linear model which states that each manifest variable consists of the variance of one or more common factor(s) and one unique factor (Brown & Moore, 2012). That it is an extension of the general linear model means that each manifest variable can be defined by a weighted additive function of the factors. FA differentiates itself from principal components analysis (PCA) in two key ways: 1) FA aims to reproduce covariance matrices while PCA only aims to maximize explained variance, and 2) FA includes an error term, implicitly acknowledging measurement error rather than presuming error-free instrumentation. Put more simply, even the best selected manifest variables will not be perfectly representative of their constructs, PCA is not measurement, it is data reduction and can be very useful when data simplification is desired (or when collected variables have high collinearity), but not when actual dimensionality assessment is the goal.

While the distinctions between PCA and FA are relatively clear, those between EFA and CFA are subtler; EFA and CFA aim to reproduce the observed relationships among a set of manifest variables with a more parsimonious and causally explanatory set of latent

variables. However, as their secondary titles of unrestricted and restricted suggest, their differences are both theoretical and practical, and manifest most clearly in the underlying assumptions made on the measurement model (Brown & Moore, 2012). In EFA the researcher generally attempts to determine the number of dimensions and evaluate manifest variables without a priori hypotheses about the underlying pattern of relationships among the variables. In CFA, the construct dimensionality is explicitly specified (on the basis of past work or strong theoretical rationale), and a corresponding pattern of manifest variables is posited. Furthermore, the evaluation of fit in CFA (explicated below) places it squarely in a SEM framework in a way traditional EFA cannot.

Evaluating model fit in CFA

Evaluation of model fit in CFA is concerned with both global fit and local fit. Generally, global model fit involves examining the difference between the covariance matrix of the sample and model-predicted covariance matrix. Much of overall model fit is done by examining the residual covariance matrix – which equals 0 under the null hypothesis, where a positive residual means that the model underpredicts the covariance between two variables, and a negative one means the predicted covariance is too high (Bollen, 1989). In evaluating global fit, the X^2 statistic is often used as a measure of absolute model goodness of fit. In this context, the X^2 measures the discrepancy between the observed sample covariance matrix (S) and the model implied covariance matrix ($\sum \hat{\theta}$). The X^2 provides a proportion-based test of the proposed, theoretical model against a saturated, just-identified model wherein variable correlations are thought to be zero, or close to zero and more arbitrary in nature (Bentler & Bonnet, 1980). Goodness of fit is

described relative to perfect fit through the minimization a discrepancy function, $F = [S, \sum \hat{\theta}]$. Fit is assessed via one of several estimation algorithms (e.g., WLS, ML, etc.) that converge to similar solutions under idealized conditions. While overall model fit is an important component of model fit evaluation, it does not necessarily reflect all the components of a model – for example, parameter estimates may not reach statistical significance, or conform with the predicted directionality. Given the large samples necessary in factor analysis to obtain accurate parameter estimates and satisfy assumptions, the generalized test of exact fit has limited utility as a stand-alone statistic (MacCallum, Browne & Sugawara, 1996).

Another issue lies in the logic of null-hypothesis statistical testing; in confirmatory factor analysis, the goal is to find support for the model as being a reliable representation of the data, such that failure to reject the null hypothesis is desired. Such a desire cannot generally be justified philosophically or practically. The logic of a hypothesis test dictates that failure to reject the null is not equivalent to confirmation of the null. Practically, in testing the hypothesis that population covariance matrix is equal to the model implied covariance matrix, χ^2 is defined as the minimum value of the fit function multiplied by (n-1) (Bentler & Bonnet, 1980).

Additional measures of absolute fit, such as the RMSEA, are often used to complement understanding of model fit in conjunction with the model X^2. Many of these measures are simple modifications of the X^2. For example, the RMSEA is: $\sqrt{\frac{\hat{\lambda}}{(n-1)(df)}}$, where $\hat{\lambda}$ is the estimated noncentrality parameter. The X^2 statistic only follows a central X^2 distribution if the proposed model is correct in the population; in the presence of

model misspecification, the test statistic follows a noncentral X^2 distribution. The noncentrality parameter informs the extent of discrepancy between S and $\sum \hat{\theta}$. Other measures of absolute fit, such as the standardized root mean squared residual (*SRMR*), do not consider the model X^2 in their calculation. Rather, the measure considers the square root of the average squared residuals on a standardized, correlation metric.

Previous literature has mentioned how distributional nonnormality can impact parameter estimates and statistical inferences derived from different model fit indices. Building on the work of Muthén and Kaplan (1985), Cudeck and Browne (1992) discussed how sample estimates of the discrepancy function were attenuated in the presence of nonnormality, such that model fit improved. Though their focus centered on how the ADF fit function is impacted by kurtosis, Olsson, Foss and Troye (2003) more generally conveyed that fit functions respond undesirably to aberrations from normality and indicated that, "a low chi-square may point not only to good fit, but also to lower power" (p. 301). Curran, West and Finch (1996) also noted decreased power to detect model misfit with increased values of kurtosis. Finally, Yuan, Bentler and Zhang (2005) described the bias that arises in goodness of fit estimators with increased skewness and kurtosis. Despite the attention given to these aspects of nonnormality and their influence on model fit estimators, previous work has not formally connected the moments to varying scale coarseness nor discussed the implications of these findings with respect to compromised validity. Moreover, previous related research did not consider the Satorra Bentler chi-square statistic in its evaluation. This is a notable difference as the statistic is intended to give estimates of standard error and goodness-of-fit which are robust to distributional non-normality.

CHAPTER 2

EFFECT OF THE NUMBER OF RESPONSE ALTERNATIVES

Previous methodological work has demonstrated that reducing the number of response alternatives on a set of items decreases the probability of rejecting an incorrect one-factor model using X^2-based fit indices (e.g., Green, Akey, Fleming, Hershberger, & Marquis, 1997; Maydeu-Olivares, Kramp, Garcia-Forero, Gallardo-Pujol, & Coffman, 2009). Maydeu-Olivares et al. (2009) conducted a repeated-measures experiment to investigate this phenomenon with real data. In the study, two questionnaires intended to measure a single construct were each administered to individuals with 2, 3, and 5 response alternatives. Maydeu-Olivares et al. observed that as they reduced the number of response alternatives in the questionnaires, the fit of a one factor model generally improved. Because it could be argued that such results were simply due to the inaccuracy of applying a common factor model to discrete responses (McDonald & Ahlawat, 1974), they also fit a one dimensional ordinal factor model to the data under the same conditions and examined results: findings held, such that fit improved as the number of response alternatives decreased.

This methodological artifact has critical implications for the validity of model fit assessment as a means to examine instrument dimensionality. This should be plain from the first section of this paper, but can be highlighted by the example of unscrupulous, or merely ill-informed researchers, who can improve the fit of their structural equation

models (*SEM*s) by reducing the number of response categories for items (e.g., converting 5-point or 7-point ratings into 3-point ratings). This issue is of particular concern as factor analysis remains the psychometric workhorse for theory construction in a number of social sciences, and it seems essential that we have confidence that our tools for model testing base support for a given theory on germane content rather than construct-irrelevant anomalies.

Consider competing frameworks for personality theory as an example. Eysenck and colleagues (Eysenck, Eysenck, & Barrett, 1985) suggest that there are three basic dimensions of human personality. In contrast the Big Five model of personality posits, as its name indicates, that five dimensions account for human personality. Looking at the instrumentation underlying these theories with the aforementioned discussion in mind begs several questions; specifically, the questionnaire typically associated with Eysenck's model consists of binary response options, whereas Big Five questionnaires generally consist of five-point item responses (e.g., Costa & McCrae, 1985; 1992). Is it possible that the different substantive conclusions across these competing theoretical frameworks are due in part to the differential number of response alternatives used to measure their respective constructs of personality? The answer is likely multi-faceted and we are not championing one theory over another in this paper. Rather we simply want to emphasize that the different number of response options these researchers employed in their instrumentation cannot be ruled out as one of the possible reasons contributing to the differential substantive conclusions of these theories.

Model fit generally improves in CFA when response categories are merged

The most straightforward way to examine the effect of reducing the number of scale categories in ratings is to collapse the extreme categories in items with an odd number of categories. For instance, merging adjacent extreme categories in 5-point item response options so that they become 3-point response items, or turning 7-point item response options into 5-point or even 3-point response items. We demonstrated the effect of merging response categories on subsequent model fit in CFA using real data from two widely used questionnaires: the NEO Five Factor Inventory (*NEO-FFI*; Costa & McCrae, 1985) and the Social Problem Solving Inventory-Revised (*SPSI-R*; D'Zurilla, Nezu, & Maydeu-Olivares, 2002). Data (*N=794*) were taken from Maydeu-Olivares et al. (2000). In both cases, the questionnaires used 5-point item response options: 0, 1, 2, 3 and 4. We fit a one factor model to each questionnaire in their original form, then again fit a one factor model after collapsing the extreme categories to turn the data into 3-point response option items (i.e., 0 & 1 = 0; 2 =1; 3 & 4=2). Both variants were examined under two conditions: (a) the common factor model where items were treated as continuous, and (b) an ordinal factor model where the items were treated as discrete. Under the common factor model, maximum likelihood (*ML*) estimation was used with a mean and variance adjusted X^2 test statistic. For the ordinal factor model, unweighted least squares (*ULS*) estimation was used, again with a mean and variance adjusted X^2 test statistic based on polychoric correlations. Results are shown in Table 2.1 We provide the mean and variance adjusted X^2, the Root Mean Squared Error of Approximation (*RMSEA*, Browne & Cudeck, 1993; Steiger, 1990) and the Standardized Root Mean Squared Residual (*SRMR*, Bentler, 1995).

We see in this table that regardless of response categories or estimator employed, there is a wide range of model misspecification when fitting a one factor model to these scales. This bolsters the notion that neither the NEO-FFI nor the SPSI-R inventory satisfy a one factor structure. For the purposes of our illustration, however, a more interesting pattern is also apparent. We see that when a common factor model is used, the X^2 statistic and all associated absolute goodness-of-fit indices improve when the 5-point NEO-FFI items are turned into 3-point items. The same findings hold true for the SPSI-R scales, with the exception of the AS scale. We obtain similar results when applying an ordinal factor analysis model, such that the X^2 and all X^2-based absolute goodness-of-fit indices improve for the scales when categories are collapsed. The SRMR, however, (i.e., the one absolute fit index employed that is not based on the X^2) only improves in 4 out of the 10 scales analyzed.

The remainder of this work demonstrates how these determinants manifest in a confirmatory factor analysis setting to provide context for the simulation study, and then we report results of the simulation study in which we examine power of both the common factor and ordinal factor models to reject a one-factor model with increasing levels of model misspecification as the number of response options (and hence skewness and kurtosis) increases. We show that when the observed data are discrete, kurtosis depends on scale coarseness such that the fewer number of response options, the more likely the items demonstrate excess kurtosis. This excess kurtosis engenders loss of power in subsequent model fitting. Moreover, we illustrate that there is a synergistic relation between model misfit and kurtosis on the power to reject incorrect models such that as model misspecification and kurtosis increase, power decreases even when using robust

estimators to accommodate non-normality (Muthén, 1993; Satorra & Bentler, 1994). Finally, we demonstrate that there is an additional impact of scale coarseness on goodness of fit indices apart from the effect of kurtosis alone.

Table 2.1 *Goodness of fit results for applying a one factor model to subscales of the NEO-FFI and SPSI-R inventories*

Model	*K*	*Fit index*	*NEO-FFI*					*SPSI-R*				
			N (*df*=54)	*E* (*df*=54)	*O* (*df*=54)	*A* (*df*=54)	*C* (*df*=54)	*NPO* (*df*=35)	*PPO* (*df*=5)	*RPS* (*df*=170)	*AS* (*df*=14)	*ICS* (*df*=35)
Common Factor Model	5	X^2	254.85	333.04	353.13	227.82	268.16	422.95	11.08	630.55	46.980	280.91
		RMSEA	.068	.081	.084	.064	.071	.120	.040	.059	.055	.095
		SRMR	.044	.061	.056	.058	.054	.066	.021	.048	.025	.061
	3	X^2	175.65	275.38	320.48	152.38	152.91	274.64	7.84	556.48	49.32	180.10
		RMSEA	.053	.072	.079	.048	.048	.094	.027	.054	.057	.073
		SRMR	.039	.057	.056	.050	.040	.057	.017	.046	.028	.051
Ordinal Factor Model	5	X^2	387.44	502.04	290.68	351.05	544.69	912.68	20.49	1163.98	102.23	490.72
		RMSEA	.088	.102	.074	.083	.107	.180	.063	.087	.090	.130
		SRMR	.051	.071	.063	.073	.072	.077	.026	.056	.025	.071
	3	X^2	219.99	278.91	206.89	189.23	189.99	515.45	8.44	593.73	68.14	269.33
		RMSEA	.062	.072	.060	.056	.056	.133	.030	.057	.071	.093
		SRMR	.053	.079	.076	.084	.069	.077	.026	.064	.033	.072

Note. *K* = number of response alternatives. *df* are unchanged by the number of response categories and model considered (i.e., common vs. ordinal factor model). Five subscales were examined for each inventory: *N*, *E*, *O*, *A* and *C* for the *NEO-FFI*; *C*, *NPO*, *PPO*, *RPS*, *AS* and *ICS* for the *SPSI-R*.

CHAPTER 3

SIMULATION STUDY

Given our findings with real data, we sought to further investigate how the number of categories impacted the behavior of X^2 goodness of fit statistics in a controlled, statistical simulation where population parameters were known. We evaluated the impact of fitting a one-factor model to generated data under several conditions, inclusive of varying degrees of departures from normality, choice of model estimator as well as degree of model misspecification. Mplus (Muthén & Muthén, 2011) was used for the simulations.

Data generation

We generated multivariate normal data with mean zero and an independent clusters, two factor model covariance structure. Population factor loadings and error variances were set to .7 and .51 across parameter combinations, and the number of items per factor was set to 5 to correspond to a 10-item questionnaire. Sample size was set to N = 500 observations for all conditions, to ensure that parameter estimates were accurately estimated but that power had not reached an asymptote so that differences in power could be observed (Forero & Maydeu-Olivares, 2009; Hu & Bentler, 1995). Observed item responses were obtained by discretizing the multivariate normal continuous data via threshold parameters. Threshold values were chosen such that the underlying population

probabilities associated with a given threshold corresponded to desired levels of item skewness and kurtosis.

Population parameters

We varied five factors in the simulation study: (a) three levels of number of item categories (K= 2, 3 and 5 response categories), (b) two levels of item kurtosis (0 and excess kurtosis; excess values of kurtosis were differentially defined corresponding to level of scale coarseness; see Table 3.1), (c) two levels of item skewness (0 and high skew; values of high skew were differentially defined corresponding to level of scale coarseness; see Table 3.1), (d) three levels of model misspecification (ρ = .8, .9, 1, where ρ = 1 is commensurate with a one-factor solution, and thus defines no model misspecification), and (e) two levels of model estimation (the common factor model, where item responses were treated as continuous data and the ordinal factor model, where item responses were treated as discrete data). Maximum likelihood estimation with robust standard errors and a mean and variance adjusted X^2 test statistic (i.e., *MLMV*; Satorra & Bentler, 1994) was used to estimate the common factor model. Unweighted least squares (*ULS*; Jöreskog, 1977) was used to estimate the ordinal factor analysis model from polychoric correlations with a mean and variance corrected X^2 goodness of fit statistic (Muthén, 1993). This estimator was chosen instead of WLSMV (i.e., the default estimator in Mplus for discrete data) as it has been shown to yield slightly better results (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009).

A partial factorial design was used as fully crossing all parameter combinations would have yielded conditions that were not viable. For example, with binary data it is

not possible to have high kurtosis and no skew, nor excess kurtosis and high skew (this case represents 12 conditions). Note that we factor both item skewness and kurtosis in the design so as to be able to disentangle the effects of item skewness from those of item kurtosis. Additionally, we included undiscretized multivariate normal data (to be estimated by factor analysis with the above 3 levels of misspecification) to provide a benchmark for the remaining conditions. As a result, the number of conditions investigated was 72 – 12 + 3 = 63. For each condition, r=1000 replications were used.

Simulation outcomes

We evaluated the power to reject incorrect factor models in CFA, as defined by the proportion of simulation replications where the model X^2 was rejected across parameter combinations. In parameter combinations where $\rho = 1$ (i.e., correct model specification), this rejection rate reflects a Type 1 error estimate. Results were evaluated against the nominal $1-\beta$=.80 and α=.05 criteria, respectively (Cohen, 1988). We report two additional absolute fit criteria: (a) the RMSEA and (b) the SRMR, to comment on their performance with respect to levels of population parameters.

Relationship between item kurtosis and scale coarseness - choice of population item skewness and kurtosis values

In designing the simulation study, probability values underlying the threshold parameters where chosen so that maximum values of skewness and kurtosis were obtained for K = 2, 3, 5 with the restriction that population probabilities were larger than .02. This restriction was imposed to ensure accurate estimation in ordinal factor analysis, given that when population probabilities are too small, sample contingency tables may

present empty cells and thus hinder estimation of polychoric correlations. The population values of item skewness and kurtosis used in the simulation are displayed in Table 3.1. We see in this table that larger values of kurtosis and skewness were specified when a coarser response scale was examined. This is due in part to the relationship between item skewness and kurtosis and the number of response options for the item, as shown below.

Let the item responses be coded as 0, 1, …, $K - 1$, where K denotes the number of responses alternatives, and $\pi_0, \pi_1, \ldots, \pi_{K-1}$ denotes the item population probabilities with the constraint that $\pi_0 = 1 - (\pi_1 + \ldots + \pi_{K-1})$, as probabilities must add up to one. Also, let

$$\mu_1 = \sum_{k=0}^{K} k\pi_k \tag{1}$$

and

$$\mu_j = \sum_{k=0}^{m-1} \left[\left(k - \mu_1 \right)^i \pi_k \right], \qquad j = 2, \ldots, 4. \tag{2}$$

The population item mean and variance are μ_1 and μ_j, respectively. The population item skewness and kurtosis are (e.g., Maydeu-Olivares, Coffman, & Hartmann, 2007)

$$\text{skewness} = \frac{\mu_3}{\mu_2^{3/2}} \tag{3}$$

and

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2}. \tag{4}$$

The kurtosis of a normal random variable is 3. For that reason, some authors use excess kurtosis instead, where excess kurtosis = 3 – kurtosis and can take negative values.

An item's mean, variance, skewness and kurtosis are not mathematically independent, however. When $K = 2$, item kurtosis can be expressed as a function of the item's variance: $\text{excess kurtosis} = \frac{1 - 6\mu_2}{\mu_2}$. When $K > 2$, the relationship also depends on the item probabilities. For instance, when $K = 3$, $\text{excess kurtosis} = \frac{12\pi_2\pi_0 + \mu_2 - 6\mu_2^2}{6\mu_2^2}$, whereas when $K = 4$, $\text{excess kurtosis} = \frac{72\pi_3\pi_0 + 12\pi_3\pi_1 + 12\pi_2\pi_0 + \mu_2 - 6\mu_2^2}{\mu_2^2}$. To illustrate these relationships, for $K = 2$ we computed the values of item kurtosis for every possible value of $\pi_1 = .1, .2, \ldots, .9$ in increments of .1. Similarly, for $K = 3$ we computed the values of item kurtosis for every admissible combination of $\pi_1 = .1, .2, \ldots, .9$, and $\pi_2 = .1, .2, \ldots, .9$. For $K = 5$, we computed item kurtosis for every possible admissible combination π_1, π_2, π_3, and π_4 in increments of .1. The resulting kurtosis values are presented graphically as a function of item variance in Figure 3.1. For these probability arrays, higher values of kurtosis are obtained the coarser the response scale. On average, kurtosis values for $K = 2$, 3, and 5 are 3.27, 2.40 and 2.03, respectively. Most importantly, the maximum values that kurtosis attains for these probability arrays are lower the finer the response scale. This demonstration informs our choice of chosen population skewness and kurtosis values for the simulation.

Simulation results

All replications converged across conditions. No improper solutions (i.e., Heywood cases) were obtained. Results are presented in Table 3.2 for both the common factor and the ordinal factor models.

Results demonstrate that when the factor structure is correctly specified (i.e., $\rho = 1$), the rejection rates of the mean and variance adjusted X^2 are generally accurate when fitting a one-factor model to the data; estimates ranged from .02 to .07 (see Table 3.2). Notably, there was not an increased Type 1 error rate associated with fitting the common factor model to the data, however. This would be expected given the data were generated according to an ordinal factor analysis model; the common factor analysis model is misspecified for item responses, as the relationship between the items and the common factors cannot be linear (McDonald, 1999). These findings demonstrate that the X^2 test statistic lacks power to detect this aspect of model misspecification (Maydeu-Olivares, Cai, & Hernández, 2011). A comparison of the rejection rates across the ordinal and common factor models illustrate that distinctions between the two solutions are nominal, with differences in rejection rates centered at zero and predominately < |.03| in magnitude.

When the model is incorrectly specified (i.e., $\rho = .8$ or $\rho = .9$), rejection rates become inaccurate in certain circumstances. Figures 3.1 through 3.3 demonstrate these results for the common factor model (note that continuous data conditions were arbitrarily assigned a value of $K = 10$ for display purposes in the figures). The left panel of Figure 3.2 illustrates rejection rates of the model X^2 as a function of number of categories, item skewness, item kurtosis and degree of model misspecification. We see in this figure that the main drivers of rejection rates are model misspecification, kurtosis and number of categories, in this order. Skew has little impact on results. Holding model misspecification constant, rejection rates are higher for low kurtosis parameter combinations. Holding model misspecification and kurtosis constant, rejection rates are

higher (and thus more powerful) as the number of response categories increases. However the relationship between kurtosis and rejection of the model X^2 is non-monotonic; higher power is observed in instances where the value of kurtosis is further from the kurtosis of a normal variable. For instance when skewness = -.2.67, kurtosis = 8.11, $\rho = .8$ and $K = 2$, the rejection rate at $\alpha=.05$ is .41; but when skewness = -2.53, kurtosis = 8.39, $\rho = .8$, and $K = 3$, the rejection rate is .66 when a common factor model is fitted. Adequate power is observed once skewness = -1.94, kurtosis = 6.18, $\rho = .8$, and $K = 5$, such that the rejection rate is .90. When skewness = 0, kurtosis ~ 3 and $\rho = .9$, the rejection rate of the model X^2 at $\alpha=.05$ fitting a common factor model is .48 for $K = 3$, .79 for $K = 5$, and .93 for continuous data; the latter demonstrates that all else equal, greater degree of model misspecification yields more accurate rejection rates.

The right panel of Figure 3.2 illustrates rejection rates of the model X^2 as a function of item standard deviation, item skewness, item kurtosis and degree of model misspecification. We see that a similar pattern of rejection rates emerges in this graph as compared to the left panel, demonstrating that the increased rejection rates of the model X^2 are not simply a function of the increased kurtosis associated with decreasing the number of scale categories. Rather, there remains a discernible effect of inaccurate rejection rates when holding the value of kurtosis constant across varying item standard deviation. This speaks to the fact that there is not a one to one relationship between item standard deviation and kurtosis. Further, in some sense the number of scale categories appears to act as a proxy for item standard deviation, such that there is an additional impact of scale coarseness on rejection rates above and beyond that associated with level of kurtosis.

We conjectured that the effects of degree of model misspecification, kurtosis and scale coarseness on the power of the statistic would carry over to any goodness of fit statistic that are a function of the X^2 test statistic, such as the Comparative Fit Index (CFI, Bentler, 1990), the Tucker-Lewis Index (TLI, Tucker & Lewis, 1973) or the Root Mean Squared Error of Approximation (RMSEA, Browne & Cudeck, 1993; Steiger & Lind, 1980). In this simulation we have only examined the effects of these drivers on the RMSEA, as the index represents an absolute (rather than comparative) measure of model misfit. Results are shown graphically in Figure 3.3 for the common factor model, where average values of the model RMSEA are illustrated as a function of number of categories, item skewness, item kurtosis and degree of model misspecification. As with reference to Figure 3.2, results are also plotted as a function of item standard deviation to bolster demonstration of the additional impact of scale coarseness relative to kurtosis alone.

Figure 3.3 demonstrates that the main drivers of RMSEA values are model misspecification, number of categories, and kurtosis, in that order. RMSEA increases as model misspecification and number of categories increase. However, the relationship between kurtosis and RMSEA is non-monotonic; lower values of the goodness of fit index are observed in instances where the value of kurtosis is further from the kurtosis of a normal variable. The value of the RMSEA is highest at kurtosis equal 3 (i.e., excess kurtosis equal to 0, the kurtosis of a normal random variable), and results generally follow the pattern of those observed in evaluating rejection rates of the model X^2.

Figure 3.4 illustrates average values of the model SRMR as a function of number of categories, item skewness, item kurtosis and degree of model misspecification. The

relationship between the SRMR and number of categories is more complex than either model X^2 rejection rates or average value of the RMSEA. First, we notice that the range of values of the average SRMR obtained in the simulation is smaller than either for the model X^2 or for the RMSEA. The largest SRMR obtained is .046, whereas the largest RMSEA is .080. Also, unlike the RMSEA, the sample SRMR is non-zero even when the model is correctly misspecified (Maydeu-Olivares, 2017). As a result, the effect of the number of categories on the SRMR is smaller than on the RMSEA generally, and the effect of using continuous data relative to 5-point items is marginal. The results for the ordinal factor analysis model are similar, but the SRMR has a larger range than in the case of the common factor model (see Table 3.2).

To examine more closely the relationship between model fit and number of response categories, we fitted a general linear model to the rejection rates, RMSEA, and SRMR obtained using as factors skewness (high, low), kurtosis (high, low), and model misspecification (.8, .9, 1). We excluded the conditions involving continuous data in these analyses. A model with main effects and all two-way interactions yielded an R^2 of 90% for rejection rates, 89% for RMSEA and 92% for SRMR. In all three cases, the skewness effects were not statistically significant at the 5% level. For rejection rates and RMSEA none of the two-way interactions was statistically significant; for the SRMR none of the interactions involving model misspecification was significant. Next, we examined the effects of including the number of categories as an additional predictor. Thus, we used skewness, kurtosis, model misspecification as factors, mean centered number of categories as covariate, and all their two-way interactions. We obtained an R^2 of 98% for rejection rates, and over 99% for RMSEA and SRMR. The number of items'

response categories predicts fit beyond what is explained by items' skewness and kurtosis. This is an unexpected finding. In predicting rejection rates, skewness was not statistically significant and the only significant interactions were kurtosis × correlation level and number of alternatives × correlation level; in predicting RMSEA only the interactions between number of alternatives × skewness and number of alternatives × kurtosis were not significant; in predicting SRMR, only the interactions between number of alternatives × skewness, number of alternatives × kurtosis, and skewness × correlation level were not significant.

We obtained similar results in the ordinal factor analysis case. A model with skewness, kurtosis, model misspecification main effects and all two-way interactions yielded R^2 of 87% and 85% for rejection rates and RMSEA. For SRMR, R^2 was only 70%. In all three cases, the main effect of skewness was not statistically significant, nor any associated interactions. When we examined the effect of including the number of categories as an additional predictor, we obtained an R^2 of 98% for rejection rates, over 99% for RMSEA and 99% for SRMR. Again, the number of items' response categories predicts fit beyond what is explained by items' skewness and kurtosis in all three cases. In predicting rejection rates, skewness and its interactions were not statistically significant; in predicting RMSEA, skewness and the interaction of skewness × number of categories were not statistically significant; and in predicting SRMR, none of the interactions involving skewness and correlation levels were statistically significant.

Why does number of categories predict fit beyond what is explained by items' skewness and kurtosis, and degree of model misspecification? Pending future work, we

conjecture that it is because number of categories acts as a proxy for items' standard deviation. Their relationship is remarkably linear and their correlation for the values in our simulation is .87. More generally, we computed the correlation between item standard deviation and number of categories (K = 2, 3, 4, 5) for every possible admissible combination of probabilities in increments of .1 (see remarks above on relationship between item kurtosis and variance), the correlation is .88.

To investigate this conjecture, we estimated general linear models as above replacing number of categories by (mean centered) item standard deviation as a covariate. For the common factor model, R^2 for rejection rates, RMSEA and SRMR were 98%, >99%, and 99%, respectively. In predicting rejection rates, the kurtosis main effect was not significant and the only significant interaction was item standard deviation × correlation level. In predicting RMSEA, the skewness and kurtosis main effects were not statistically significant nor was the item standard deviation × kurtosis interaction. Finally, in predicting SRMR the kurtosis main effect was not statistically significant, and the only significant effects were correlation level × standard deviation and skewness × kurtosis. For the ordinal factor model we obtained very similar results. R^2 for rejection rates, RMSEA and SRMR were again 98%, >99%, and 99%, respectively. In predicting rejection rates, the kurtosis and skewness main effects were not significant and the only significant interaction was item standard deviation × correlation level. In predicting RMSEA, the kurtosis main effect was not statistically significant nor were the item standard deviation × correlation level and kurtosis × skewness interactions. Finally, in predicting SRMR the kurtosis main effect was not statistically significant, and the only significant effects were correlation level × standard deviation and skewness × kurtosis.

Table 3.1 *Population probabilities and thresholds used to generate simulation data, with corresponding mean, variance, skewness and kurtosis parameters*

K	*Mean*	*Variance*	*Skewness*		*Kurtosis*		π_1	π_2	π_3	π_4	π_5	τ_1	τ_2	τ_3	τ_4
5	3.48	.89	H	-1.94	H	6.18	.02	.04	.08	.16	.70	-2.05	-1.55	-1.08	-.52
5	2.00	.60	L	.00	H	5.00	.05	.10	.70	.10	.05	-1.64	-1.04	1.04	1.64
5	3.01	1.45	H	-1.05	L	3.01	.05	.10	.12	.25	.48	-1.64	-1.04	-.61	.05
5	2.00	.88	L	.00	L	3.00	.06	.20	.48	.20	.06	-1.55	-.64	.64	1.55
3	1.80	.26	H	-2.53	H	8.39	.05	.10	.85			-1.64	-1.04		
3	1.00	.16	L	.00	H	6.25	.08	.84	.08			-1.41	1.41		
3	1.55	.55	H	-1.28	L	3.03	.15	.15	.70			-1.04	-.52		
3	1.00	.33	L	.00	L	2.99	.17	.67	.17			-.97	.97		
2	.90	.09	H	-2.67	H	8.11	.10	.90				-1.28			
2	.60	.24	L	-.41	L	1.17	.40	.60				-.25			

Note. *K* = number of response alternatives, π = probability, τ = threshold; *H* = high, *L* = low. Excess kurtosis = kurtosis – 3. Empty cells reflect conditions that were not considered in the simulation design.

Table 3.2 *Simulation Results*

Population Parameters				*Common factor model*			*Ordinal factor model*		
K	*Skewness*	*Kurtosis*	ρ	*1-β*	*RMSEA*	*SRMR*	*1-β*	*RMSEA*	*SRMR*
	.00	3.00	.80	1.00	.080	.044	--	--	--
∝	.00	3.00	.90	.93	.044	.027	--	--	--
	.00	3.00	1.0	.05	.009	.017	--	--	--
	.00	3.00	.80	1.00	.068	.026	1.00	.075	.050
	.00	3.00	.90	.79	.036	.026	.81	.038	.032
	.00	3.00	1.0	.04	.009	.027	.05	.008	.022
	.00	5.00	.80	.99	.052	.039	1.00	.058	.053
5	.00	5.00	.90	.47	.027	.028	.49	.028	.037
	.00	5.00	1.0	.05	.009	.022	.03	.007	.029
	-1.05	3.01	.80	1.00	.060	.043	1.00	.069	.051
	-1.05	3.01	.90	.64	.032	.029	.72	.035	.034
	-1.05	3.01	1.0	.05	.009	.022	.05	.009	.024
	-1.94	6.18	.80	.90	.043	.039	.98	.054	.055
	-1.94	6.18	.90	.38	.024	.033	.48	.027	.039
	-1.94	6.18	1.0	.04	.009	.027	.05	.009	.031
	.00	2.99	.80	.99	.053	.037	.99	.056	.054
	.00	2.99	.90	.48	.027	.026	.47	.027	.038
	.00	2.99	1.0	.05	.009	.021	.03	.008	.030
	.00	6.25	.80	.73	.035	.037	.65	.033	.064
3	.00	6.25	.90	.20	.018	.030	.11	.013	.050
	.00	6.25	1.0	.05	.010	.027	.02	.005	.044
	-1.28	3.03	.80	.94	.046	.042	.96	.050	.057
	-1.28	3.03	.90	.36	.023	.031	.39	.024	.041
	-1.28	3.03	1.0	.04	.009	.026	.04	.008	.034
	-2.53	8.39	.80	.66	.033	.046	.74	.035	.066
	-2.53	8.39	.90	.19	.018	.038	.19	.018	.053
	-2.53	8.39	1.0	.04	.010	.034	.04	.009	.046
	-.41	1.17	.80	.93	.044	.038	.94	.047	.059
	-.41	1.17	.90	.34	.022	.029	.36	.023	.044
2	-.41	1.17	1.0	.07	.010	.025	.05	.009	.037
	-2.67	8.11	.80	.41	.025	.045	.41	.025	.079
	-2.67	8.11	.90	.14	.014	.039	.11	.013	.067
	-2.67	8.11	1.0	.05	.009	.037	.03	.008	.061

Note. *1-β* is condition-level rejection rate of the X^2 was evaluated at α=.05. ρ = factor correlation. *RMSEA* and *SRMR* are condition-level estimates, averaged across replication.

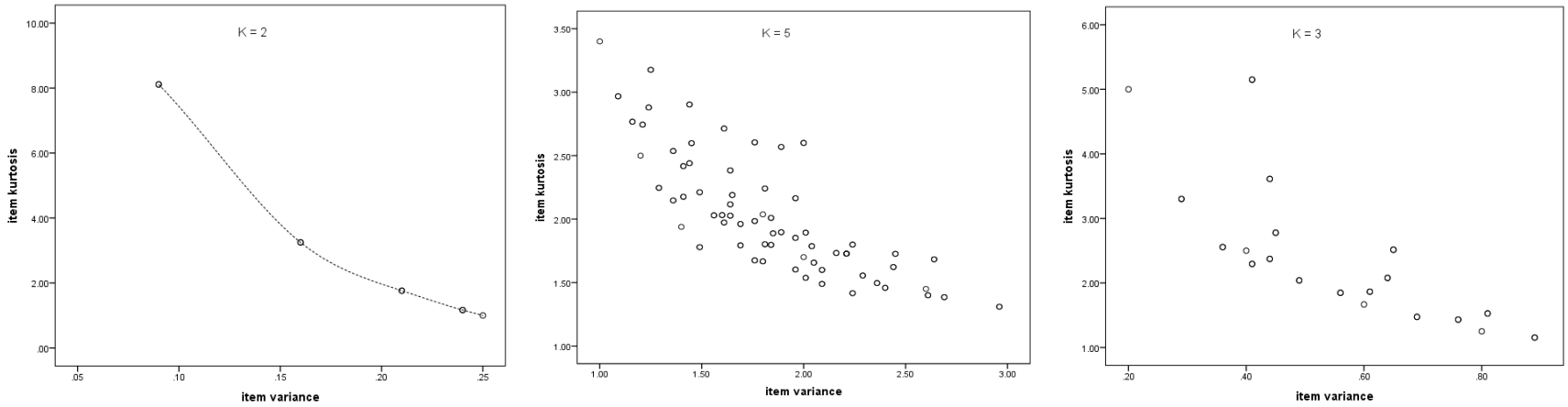

Figure 3.1 *Plot of item kurtosis as a function of item variance and number of response alternative*

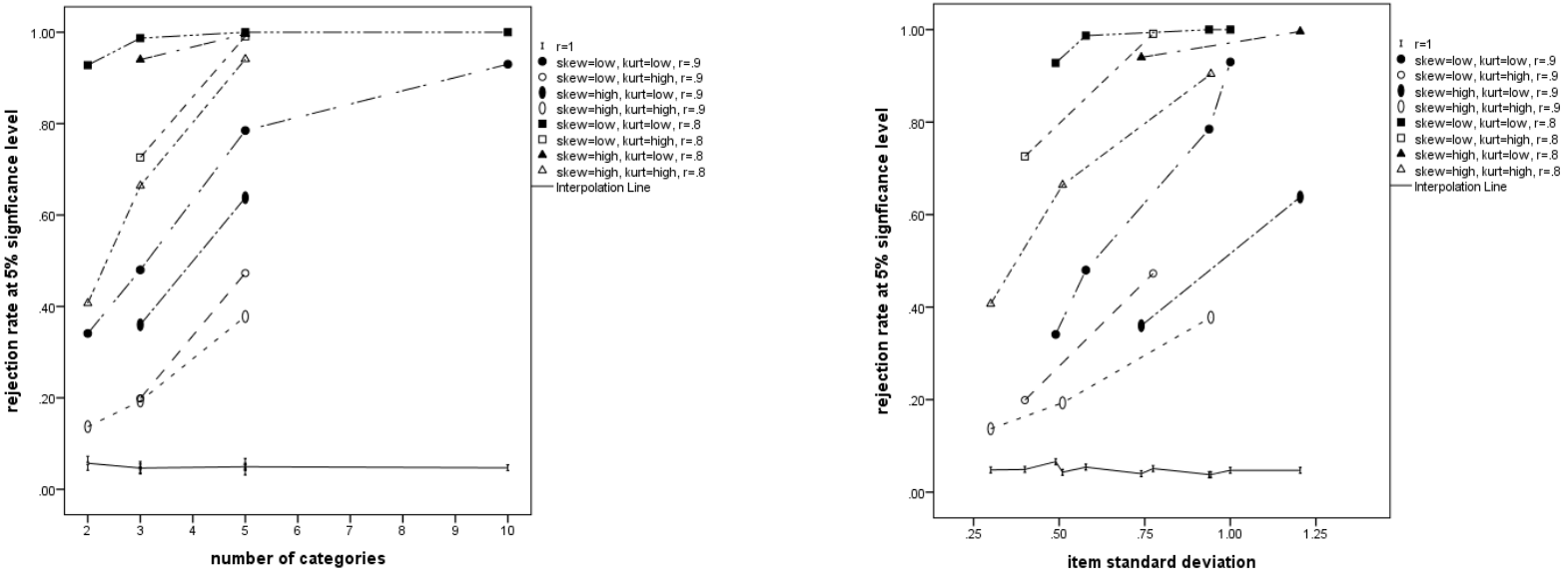

Figure 3.2 *Common factor model results: Plot of rejection rates at* $\alpha = .05$ *of the mean and variance corrected* X^2 *statistic as a function of skewness and kurtosis levels, degree of model misspecification, and number of response alternatives or item standard deviation*

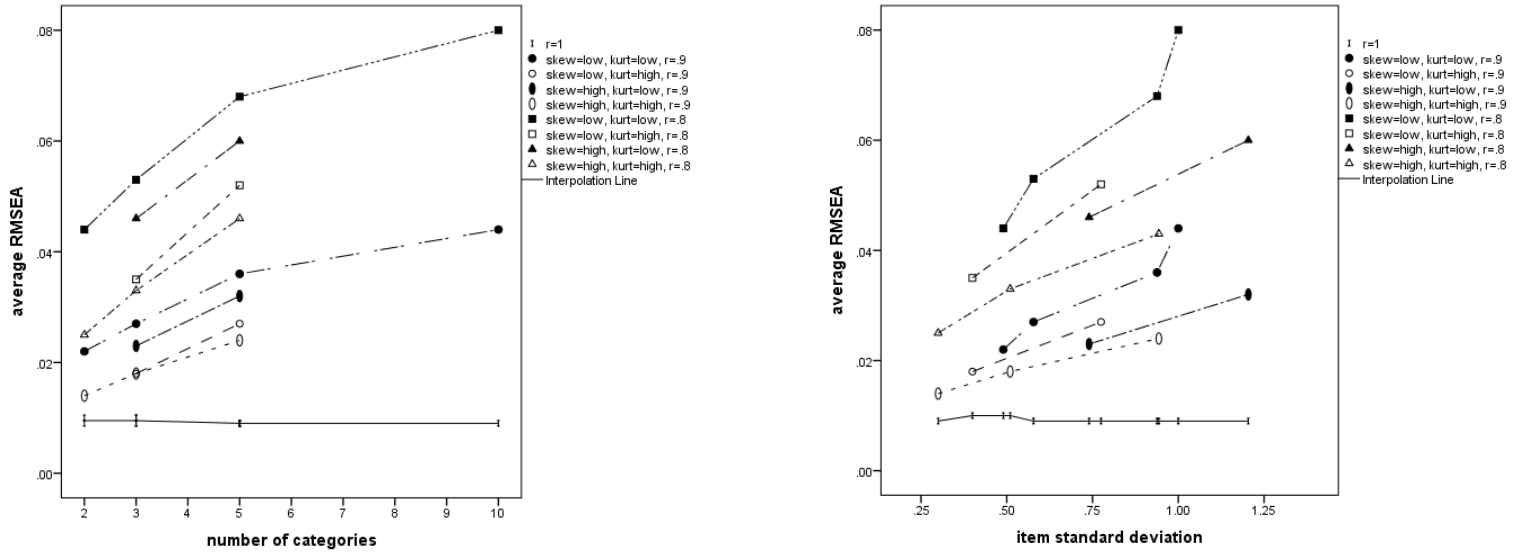

Figure 3.3 *Common factor model results: Plot of average RMSEA values as a function of skewness and kurtosis, degree of model misspecification and number of response alternatives or item standard deviation*

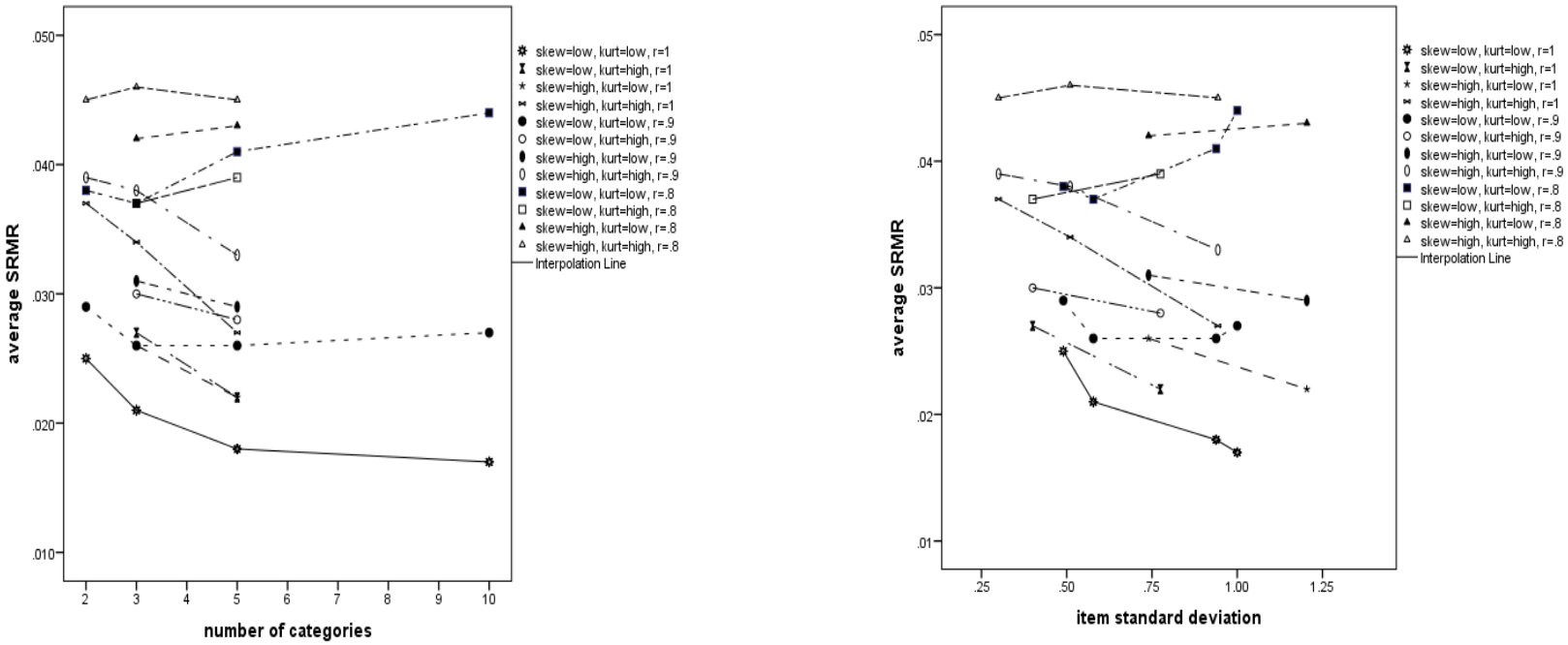

Figure 3.4 *Common factor model results: Plot of average SRMR values as a function of skewness and kurtosis, degree of model misspecification and number of response alternatives or item standard deviation*

CHAPTER 4

VARYING THE NUMBER OF RESPONSE ALTERNATIVES IN THE BECK HOPELESSNESS SCALE

Given the implications of our work extend into the validity of many measures used in applied research which have historically relied on low scale point response options, we were interested in investigating an additional substantive example. We conducted an empirical examination of the 20 item Beck Hopelessness scale (BHS) in its original two scale point form, as well as in a nine scale point form. We expected to observe empirical findings in line with simulation results, predicting fit indices indicative of better fit for models stemming from dichotomous response option data. As in the simulation, we utilized a maximum likelihood estimation with robust standard errors and a mean and variance adjusted X^2 test statistic (*MLMV:* Satorra & Bentler, 1994) for instances where we treated the data as continuous, and unweighted least squares (*ULS*; Jöreskog, 1977) with a mean and variance adjusted X^2 test statistic based on polychoric correlations, where the data were treated as discrete. Additionally, we included two popular comparative indices of model fit, the Tucker-Lewis Index (TLI: Tucker & Lewis, 1973), and the Comparative Fit Index (CFI: Bentler, 1990). This allows us to examine our early assertion where we hypothesize that the above demonstrated effects of degree of model misspecification, kurtosis and scale coarseness on the power of the statistic will carry over to any goodness of fit statistic which are a function of the X^2 test statistic.

Finally, we predicted there would be only small differences in χ^2 based fit indices when contrasting our treatment of the data as continuous versus discrete.

Selection of scale

For the purposes of this study, there was a desire to select a measure whose dimensionality has been the subject of controversy, and whose original response format was dichotomous. The BHS met both of these criteria as well as complementing past work by Maydeu-Olivares et al., (2009) which examined empirical changes in apparent fit within the domain of personality research. Their work included two, three, and five response option alternatives in relatively short measures (5-12 items). This provided additional rationale to select the BHS, as it was outside of the personality domain, and included a greater number of items. Past work also provided the incentive for utilizing a scale which included more possible response options than five, prompting our selection of the nine response option format.

Participants

The participants consisted of 952 undergraduate students at a large public university in a southeastern state who volunteered to participate in these studies. Age of our participants ranged from 18 to 25 years (M =19.75, SD = 1.57), and gender was primarily female (77%). Data were collected online using SurveyMonkey software (SurveyMonkey Inc., Palo Alto, California, USA, www.surveymonkey.com), as part of a larger instrument battery of measures unrelated to the current work. All procedural methods were approved by the University of South Carolina Institutional Review Board.

The students were divided into two samples which received the same battery of measures with the sole exception being the order in which they received the measures of interest for the present study. As indicated by Maydeu-Olivares et al., (2009) when examining scale coarseness, it is possible to use either a randomized one-way design or a repeated measure design. The latter approach (implemented here) allows us to capture intra-individual effects due to the adjustment of scale coarseness, provides us with increased precision, and therefore is a more powerful design. However, the repeated measure design must be carefully screened for testing effects. To address this, *Study A* (*n* = 503) whose age ranged from 18 to 25 years (*M* =19.94, *SD* = 1.55), and were 80% female received the 9-response option form of the BHS first, and the dichotomous form of the BHS later in the battery. *Study B* (*n* = 449) whose age ranged from 18 to 25 years (*M* =19.59, *SD* =1.57), and were 74.4% female received the dichotomous option form of the BHS first, and the 9-response form of the BHS later in the battery. In both cases, the two forms of the measure were separated by three unrelated measures. Different labels were used for the two response options (0 = *True*, 1 = *False*), and the nine response options (1 = *Strongly agree*, 5 = *Neither agree nor disagree*, 9 = *Strongly disagree*).

Screening invalid responses

We utilized a three-step process to screen and remove invalid responses from the data. First, participants who completed the entire battery (consisting of approx. 500 items) in five minutes or less were removed. This step was done prior to calculating the descriptive statistics presented above. Second, data from the nine response option was discretized by creating a new variable which recoded responses less than or equal to four, as zero, and responses greater than or equal to six, as one. These new values were then

subtracted from the participant's dichotomous response, and the square of that value was summed. This provided us with a count for the number of times each participant's nine-outcome response substantially differed from their dichotomous response. Participants whose count was greater than five were removed (this amounted to the removal of 41 participants). Finally, given our desire to use the estimator MLMV and the generally low prevalence of missingness in the remaining data, participants with one or more missing value were removed (this amounted to four participants) and left us with our final participant total to be factor analyzed (N = 907).

Evaluating potential order response bias

The final step in our preliminary evaluation of our data was to look for potential order effects that might invalidate our within-subject design. Examining the respective demographics of our two response orders (*A* and *B*) from above, we can see that the range, average, and standard deviation for age, as well as the gender breakdown are very similar. Additionally, we conducted *t*-tests for each item between *A* and *B* and after adjusting α for multiple comparisons, had only one significant finding. This significant difference (t = 4.64) was for the dichotomous form of item 12, "I don't expect to get what I really want," where participants in *A* had an estimated mean difference between the two groups of .12. This corresponds to a 32% endorsement in *A* compared to a 19% endorsement in *B.* Additionally, one unfortunate limitation of this current work is that a data collection error necessitated the complete removal of item number 10 from both forms. Despite the aforementioned limitations, we proceeded with our data analysis plan.

Beck Hopelessness scale

As it was originally conceived, the construction of the Beck hopelessness scale (BHS) was drawn from two sources to represent two theoretical dimensions of hopelessness (Beck & Weissman, Lester & Trexler, 1974). The first dimension was comprised of 11 items reflecting pessimistic statements, and includes items such as "I might as well give up because I can't make things better for myself." The second dimension had 9 items concerning optimistically framed future expectations, an example item read "I look forward to the future with hope and enthusiasm." Despite this operationalization Beck et al., (1974) eventually concluded that their measure – and subsequently the construct of hopelessness as they had operationalized it – consisted of three dimensions; "affectively toned association" which labeled feelings about the future, "loss of motivation," and "future expectations". It should be noted that this decision appears to simply split future oriented optimism into "feelings about the future" and "future expectations," and was based off of eigenvalue greater than 1 criterion in a principal component (PC) analysis, which was mistakenly identified as factor analysis.

Subsequent research into hopelessness settled into a consistent debate between a one and two dimensional form. Scheier and Carver (1985) posited that positive and negative outcome expectancies comprised two extremes of a unidimensional construct. Dember et al. (1989) on the other hand considered hopelessness as two dimensions representing positive and negative life outlook. Chang, Maydeu-Olivares, and D'Zurilla (1997) considered hopelessness through a more methodologically rigorous series of measurement work, and concluded that there was FA evidence to suggest two highly correlated but partially independent dimensions. This was further supported through an

examination of the concurrent and discriminant validity of the separate dimensions where they found that in a two factor form, pessimism but not optimism was related to depressive symptoms. Purported factor structures tested in this work can be seen in Table 4.1.

Based on past research, we expected to observe a decrease in reliability as the number of response options decreased (Maydeu-Olivares et al., 2009). However, due to the length of our test we expected these reliability gains to be mitigated compared to what has been observed in shorter measures. In the dichotomous data $\alpha = .84$ while for the nine response data $\alpha = .93$.

Results

Both Pearson's and Spearman's correlations along with 95 percent confidence intervals among the BHS scale scores using dichotomous and nine response options are displayed in Table 4.2. The correlations for Pearson's estimates are considerably higher than those from Spearman's estimates, ranging from .405 to .704 and .317 to .722 respectively. Item correlations were surprisingly low, compared to past research by Maydeu-Olivares et al., (2009) which examined 2, 3, and 5 response alternatives in the personality and affect domains which had correlations ranging from .62 to .78; given the higher magnitude of change in the number of scale points as well as the use of a measure in a different domain we did anticipate the possibility of finding correlations significantly different than 1, however we did not anticipate this degree of attenuation.

Factor Analysis

Results for evaluating χ^2 based likelihood ratio, RMSEA, CFI, and TLI fit indices in continuous as well as ordinal models with one, two, and three factors are presented in Table 4.3. Our results demonstrate that there is substantial variability in apparent model fit between the dichotomous and nine response conditions, and to a lesser but non-trivial extent between conditions that treat the data as continuous and discrete. For the dichotomous response data, χ^2 based likelihood ratio fit appears to improve slightly when treated as discrete data. In the case of the nine response data, using χ^2 based likelihood ratio, fit appears to improve considerably when treated as continuous. Likewise, when examining RMSEA between two and nine response conditions we observe that the two response format appears better in every case regardless of whether the data were treated as continuous or discrete. As concerns the continuous or discrete treatment of our response option conditions, RMSEA shows a trivial improvement when treated as ordinal in the two response condition, and a substantial improvement when treated as continuous in the nine response condition. For CFI and TLI treating the data as discrete results in greater improvement than treating the data as continuous, however the relationship between response option condition and fit reveals an interaction where the two response condition treated as continuous indicates an apparent decline in fit when compared with the nine response condition treated as continuous, but the two response condition treated as discrete shows improvement over the nine response condition when treated as discrete.

Model selection

In the preceding section, care was taken to describe the fit indices in relative terms of improvement or decline, rather than in model selection terms of "good" or "adequate" fit. This was done to highlight the subjective fit criteria used for model selection in substantive fields, where interpretation of dimensionality is often based exclusively off of global fit indices. Specifically, such decisions are generally centered around subjective cut points that have been recommended in the literature. One such recommendation is that an RMSEA < .05 indicates close fit, RMSEA < .08 indicates a fairly close fit, and RMSEA > .1 indicates poor model fit (Hu & Bentler, 1999). Relatedly, Hu and Bentler, (1999) also suggest that CFI and TLI should both be greater than .95, interpretable as saying that the specified model is at least 95% better than a model which assumes all variables are uncorrelated. As concerns the likelihood ratio test, difficulties have been discussed earlier in this work, but in the applied literature, a *p*-value greater than .01 is often said to indicate good fit. In our current work, our sample *p* was less than .001 for every model examined, therefore the χ^2 is only useful in examining the magnitude of change between models. With this noted, we can now provide model selection evaluation based primarily off of RMSEA, CFI, and TLI.

In examining the adequacy of model fit based on RMSEA, we observe that in the case of the dichotomous data, a one-factor model appears to have reasonably good fit. In the case of the nine response option, a one-factor model appears to have poor model fit, while a two factor model (if treated as continuous) appears to have relatively good fit. Interestingly, while model fit for two and three factor models appear roughly equal under the two response condition, when we examine them under the nine response option

condition, a three-factor model appears to fit considerably worse than a two factor model. In regards to our hypothesis, as predicted, differences in RMSEA between continuous and discrete treatment in the two response condition are trivial, however, in the nine response condition they are large enough where one might realistically change their substantive conclusion regarding model adequacy. This is because in the two-factor solution of the nine response condition, treating the data as continuous results in an RMSEA of .061 which may very well be considered adequate, while the discrete treatment produced an RMSEA of .098 which would likely be rejected as not indicative of adequate model fit.

More interesting still, when examining CFI and TLI as the primary basis for evaluation of model fit, a slightly different and much more pronounced pattern appears. Here, contrary to our expectation, the more important consideration is whether the data are treated is continuous or discrete; in both conditions treating the data as discrete results in dramatic improvement over continuous. None of the models for the two or nine response condition would be considered adequate where the data is treated as continuous, whereas all of the models might be said to have approximate fit if treated as ordinal. If adhering strictly to the .95 cutoff discussed above, only the two factor solution would be accepted for either the two response condition or the nine response condition. However, where treated as discrete, both the CFI and TLI are still higher for every case in the two response condition than in their nine response condition counterpart, and an applied researcher looking to confirm their one factor solution would very likely conclude that it was adequate (particularly if they were working with the two response condition.

In considering the empirical model fit of dichotomous and nine response confirmatory factor models on the Beck Hopelessness scale, we can see how a range of different conclusions about the factor structure of the BHS might be reached as a consequence of insufficient power to reject poor fitting one and three factor solutions. The purpose of the present demonstration was to examine how the number of response options on an instrument impacts the relationship between reliability and model fit in a measure outside of the personality domain, with a relatively large number of items and a large difference between the number of possible response options. Despite the increased number of items used (as compared to the simulation) in this study, substantive conclusions towards factor structure continue to be effected by increased scale coarseness. As predicted, model fit in almost every condition appeared worse for the nine response option condition than in the two response condition. Interestingly with the increased magnitude of the scaling difference observed we did notice non-trivial changes in apparent fit between data treated as ordinal and continuous. Specifically, despite nine response categories still being discrete data, treating it as continuous resulted in much improved apparent model fit with the exception of the TLI and CFI.

Other considerations

Though the focus of this applied example has been on the effect of scale coarseness on power to reject incorrectly specified factor models, we can also use our empirical example to demonstrate our simulation findings about the effects of skewness and kurtosis. In Table 4.4 we can see skewness and kurtosis for each item between the two response and nine response condition. The average skewness for the two response condition was -2.05 (1.14) as compared to -1.08 (.62) for the nine response condition.

More importantly, the average excess kurtosis (where no kurtosis is indicated by 0) for the two response condition was 3.45 (4.03) as compared to .85 (1.63) for the nine response condition. Further, as can be seen, for every item in the case of kurtosis, and all but two items in the case of skewness, the two response condition has more extreme non-normality. As described earlier in this work, this is a mathematical necessity – it's impossible for a dichotomous response variable to be normally distributed, however the degree of non-normality also reflects a fundamental difficulty in applied measures of this sort in general. Namely, clinical measures of psychological constructs are generally intended to detect extremes. In the case of the BHS, it may be argued that an undergraduate population fails to reflect the intended clinical population of interest, however, even as administered in a clinical population we would expect the overwhelming majority of individuals to have scores which reflected low to moderate levels of hopelessness. Clearly, our reliance on parametric statistics can have great consequences in terms of substantive conclusions if we are unaware of the results that stem from violating assumptions of normality. A final note on this applied example, an intentional focus has been on considering model fit exclusively through fit indices which we postulate comprises the entire evaluative process for some substantive researchers.

Table 4.1 *Purported factor structures and item loadings*

Number of Factors	Dimension Name	Associated Items
1-Factor	Hopelessness	1, 2, 3, 4, 5, 6, 7, 8, 9, 11 ,12, 13, 14, 15, 16, 17, 18, 19, 20
2-Factor	Optimism	1, 3, 5, 6, 8, 13, 15, 19
	Pessimism	2, 4, 7, 9, 11, 12, 14, 16, 17, 18, 20
3-Factor	Affect	1, 5, 6, 13, 15, 19
	Motivation	2, 3, 9, 11, 12, 16, 17, 20
	Expectations	4, 7, 8, 14, 18

Table 4.2 *Correlations Between Different Number of Response Format Options for Beck Hopelessness Scale (BHS)*

Item	Pearson ρ (95%CI)	Spearman's ρ (95%CI)
1	.545 (.465, .611)	.408 (.346, .464)
2	.405 (.296, .507)	.317 (.234, .390)
3	.482 (.410, .553)	.410 (.350, .459)
4	.658 (.618, .698)	.658 (.611, .695)
5	.646 (.602, .690)	.633 (.591, .674)
6	.428 (.350, .500)	.353 (.300, .400)
7	.686 (.620, .744)	.496 (.442, .553)
8	.714 (.679, .748)	.722 (.690, .751)
9	.522 (.451, .587)	.442 (.380, .499)
11	.608 (.513, .687)	.412 (.344, .481)
12	.579 (.521, .638)	.529 (.472, .582)
13	.433 (.367, .496)	.393 (.336, .443)
14	.645 (.596, .686)	.622 (.578, .633)
15	.596 (.538, .644)	.507 (.457, .555)
16	.571 (.489, .645)	.420 (.363, .477)
17	.482 (.389, .576)	.357 (.282, .423)
18	.704 (.666, .739)	.699 (.657, .737)
19	.404 (.324, .487)	.336 (.258, .399)
20	.559 (.464, .635)	.412 (.343, .467)
Sum	.864 (.840, .885)	.778 (.743, .813)

Table 4.3 *CFA model fit*

	Two Response			Nine Response		
	1 factor	2 factor	3 factor	1 factor	2 factor	3 factor
CFA Continuous	X^2 = 676.98 RMSEA = .062 CFI = .737 TLI = .705	X^2 = 466.64 RMSEA=.048 CFI = .842 TLI = .821	X^2=419.32 RMSEA=.049 CFI = .849 TLI = .824	X^2 =1315.50 RMSEA=.093 CFI = .769 TLI = .740	X^2 =652.63 RMSEA=.061 CFI = .900 TLI = .887	X^2 =782.31 RMSEA=.075 CFI = .866 TLI = .845
CFA Ordinal	X^2 = 670.05 RMSEA = .060 CFI = .936 TLI = .929	X^2 =431.58 RMSEA=.045 CFI = .966 TLI = .961	X^2 = 399.97 RMSEA=.047 CFI = .940 TLI = .930	X^2 =2977.39 RMSEA=.141 CFI = .902 TLI = .890	X^2 =1509.85 RMSEA=.098 CFI = .953 TLI = .947	X^2 =2274.07 RMSEA=.132 CFI = .925 TLI = .913

Note. Estimator for continuous data MLMV, estimator for all discrete data = ULS

Table 4.4 *Skewness and kurtosis*

Item	Two Response Skewness	Nine Response Skewness	Two Response Excess Kurtosis	Nine Response Excess Kurtosis
1	-3.194	-1.407	8.218	1.467
2	-3.340	-2.139	9.178	4.204
3	-2.550	-1.030	4.514	.351
4	-.071	-.098	-1.999	-1.118
5	-.782	-.337	-1.392	-.886
6	-3.218	-1.041	8.376	.612
7	-2.683	-1.720	5.210	2.324
8	-.320	-.159	-1.902	-.943
9	-1.912	-.953	1.660	-.081
11	-2.810	-1.835	5.908	2.721
12	-1.086	-.798	-.822	-.448
13	-2.209	-.723	2.888	.247
14	-.713	-.519	-1.494	-.697
15	-2.059	-.861	2.245	.166
16	-2.949	-1.654	6.712	2.380
17	-2.830	-1.771	6.023	2.673
18	-.221	-.415	-1.956	-1.003
19	-3.001	-1.169	7.021	1.047
20	-3.026	-1.840	7.175	3.081

CHAPTER 5

DISCUSSION

In considering the utility of factor analysis as a means to enhance our understanding of theoretical constructs, as well as of validity evidence for associated scales of those constructs, attention to issues that give rise to statistical artifacts in model assessment has been underemphasized. A foundational example of this concerns item scaling and the influence it imparts on making judgments regarding scale dimensionality. When designing measures, there is often an unwarranted willingness to choose and/or transform scale coarseness at will, without consideration for how these choices impact model fit. Though others have noted that increasing the number of response options generally improves the *reliability* of a scale (with gains optimized somewhere between 5-9 scale categories; Green et al., 1997; Lissitz & Green, 1975; Symonds, 1924), little discussion has been held on how scale coarseness affects structural *validity* of any kind.

This thesis has shown that choice of scale coarseness and the resultant distributional properties critically influence results of confirmatory factor analysis (a tool for assessing structural validity evidence) at different levels of measure discretization. Additionally, scale coarseness impacts the power of the X^2 test statistic beyond what is explained by the items' kurtosis alone, such that an increase in the number of categories used for an item response scale yields greater power to detect incorrect models. Additional measures of absolute model fit (i.e., the RMSEA and the SRMR) were also

affected. We conjecture that this influence is initiated by the association between an item's standard deviation and its number of categories. We further surmise that the SRMR was less influenced than either the X^2 or the RMSEA as it is a standardized statistic that does not invoke the use of a weighted mean in its computation (the X^2 and RMSEA are both weighted and unstandardized).

The contents of this research are weakly related to the old literature conducted on difficulty (i.e., spurious) factors in an exploratory factor analysis setting, which demonstrated that factor analysis of categorical data faces the problem that items with similar distributional properties tend to correlate based solely on this distributional similarity and result in spurious factors (Green et al., 1997). Item difficulty can be understood in the same way as variability in item means, such that when items differ widely in difficulty level (i.e., in item means), 'spurious' factors, in addition to 'genuine' factors of content, are obtained. McDonald and Ahlawat (1974, p.84) indicate that "These [factors] have been attributed to: a) 'attenuation' of a correlation coefficient below what it 'should' be if the difficulty levels were the same, [and] b) non-linear relationships of items on the factors of content."

Gorsuch (1974) argued that spurious factors are likely to appear because the magnitude of coefficient is inappropriately sensitive to differences in difficulty levels between items (Green, 1983). Much of the literature that has explored spurious factor extraction used the common factor model, with non-optimal estimators and eigenvalue-based methods for determining the number of factors. More recent research (Bernstein & Teng, 1989; Green, Akey, Fleming, Hershberger, & Marquis, 1997), however, has investigated this topic in confirmatory factor analysis using maximum likelihood

estimation and the likelihood ratio test statistic (without Satorra-Bentler corrections to account for non-normality).

This article differentiates itself from the difficulty factor literature in several ways. First, we streamlined our examination by considering the case where all the items within a condition have the same item mean (i.e., item difficulty in classical test theory language). Second, we examined conditions under which true multidimensionality can be hidden behind underpowered test statistics influenced by model misspecification, kurtosis, and number of categories, rather than examining when spurious factors suggest a multidimensional structure. Finally, in addition to investigating results under the common factor model, we also evaluated results using the true model that was used to generate the data (i.e., the ordinal factor model), and thus took into account the discrete nature of the data using non-linear functions between the items and latent traits as suggested by McDonald and Ahlawat (1974). We have shown that even in this case there is strong evidence suggesting that decreasing the number of response options decreases power. In other words, even when all items had the same item mean and the ordinal factor model is used, reducing the number of categories will increase the likelihood of finding spurious factors (more factors that those used to generate the data are needed to provide a good fit). We believe following Bernstein and Teng (1989) and Green et al. (1997) that the problem may be compounded by including within a condition items of different characteristics. Hence, we feel that additional research is needed in which both the average item variance/kurtosis (as in the present study) are manipulated along with the composition of items within a condition to investigate which of these two aspects is the main driver of power in item factor analysis as well as in ordinal factor analysis.

Concluding remarks

In closing, the common factor model –and its cousin, structural equation modeling with latent variables- is a cornerstone of psychometrics and as such it is often used for theory building in a number of social sciences. Yet, we have shown that the ability of these methods to reject incorrect models is seriously hampered when the number of response options used in items decreases. With the aforementioned in mind, we recommend that researchers use a large number of response options (i.e., a finer scale) when constructing items and overall instrumentation for use in evaluating constructs of interest. Not only does employing a finer response scale increase reliability as demonstrated in previous research, but it also increases the power of test statistics to reject incorrect substantive models and thus crucially contributes to developing effective validity evidence for construct measurement. Future directions for this research may be to consider these issues in an EFA context. Though often discounted by methodologists, exploratory factor analysis remains in regular use by substantive researchers to assess dimensionality of novel instrumentation. While we would expect to find similar results to those seen in a confirmatory context, such research (to our knowledge) has yet to be examined.

REFERENCES

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: the hopelessness scale. *Journal of consulting and clinical psychology*, *42*(6), 861.

Benson, J., & Hagtvet, K. A. (1996). The interplay among design, data analysis, and theory in the measurement of coping. *Handbook of coping: Theory, research, applications*, 83-106.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–46. DOI:10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588. DOI:10.1037/0033-2909.88.3.588

Bernstein, I., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, *105*, 467–477. DOI:10.1037/0033-2909.105.3.467

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen & J. s. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park: Sage. DOI:10.1177/0049124192021002005

Brown, T.A., & Moore, M.T (2012). Confirmatory factor analysis. In *Handbook of structural equation modeling* (pp. 361-379). Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Testing structural equation models*, 154, 136-162.

Chang, E. C., D'Zurilla, T. J., & Maydeu-Olivares, A. (1994). Assessing the dimensionality of optimism and pessimism using a multimeasure approach. *Cognitive therapy and research*, *18*(2), 143-160.

Chang, E. C., Maydeu-Olivares, A., & D'Zurilla, T. J. (1997). Optimism and pessimism as partially independent constructs: Relationship to positive and negative affectivity and psychological well-being. *Personality and individual Differences*, *23*(3), 433-440.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd edition. New York: Psychology Press. DOI:10.4324/9780203771587

Costa, P. T., & McCrae, R. R. (1985). *The NEO Personality Inventory*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual*. Odessa, FL: Psychological Assessment Resources.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement, 2nd ed. (pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight. *New directions for testing and measurement*, *5*(1), 99-108.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302.

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika, 57*, 357-369. DOI: 10.1007/bf02295424

Curran, P. J., West, S. G. & Finch, G. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16–29. DOI: 10.1037/1082-989x.1.1.16

Dember, W. N., Martin, S. H., Hummer, M. K., Howe, S. R., & Melton, R. S. (1989). The measurement of optimism and pessimism. *Current Psychology, 8*(2), 102-119.

D'Zurilla, T. J., Nezu, A. M., & Maydeu-Olivares, A. (2002). Manual of the Social Problem-Solving Inventory-Revised. North Tonawanda, NY: Multi-Health Systems, Inc. DOI:10.1037/10805-001

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences*, *6*, 21–29. DOI:10.1016/0191-8869(85)90026-1

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, *14*, 275–99. DOI:10.1037/a0015825

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. *Structural Equation Modeling*, *16*, 625–641. DOI:10.1080/10705510903203573

Gorsuch, R. L. (1974), Factor Analysis. Philadelphia: W. B. Saunders Company.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement*, *7*, 139-147. DOI:10.1177/014662168300700202

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling*, *4*, 108–120. DOI:10.1080/10705519709540064

Hu, L.-T. & Bentler, P.M. (1995). Evaluating model fit. In R.H. Hoyle (Ed.), *Structural Equation Modeling. Concepts, Issues and Appliations* (pp. 76-99). London: Sage. DOI:10.1080/10705519909540118

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Jöreskog, K.G. (1977). Factor analysis by least-squares and maximum-likelihood methods. In K. Enslein, A. Ralston & H.S. Wilf (Eds.), *Statistical Methods for Digital Computers* (pp. 125-153). New York: Wiley.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. IAP.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, *60*(1), 10.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130. DOI:10.1037/1082-989x.1.2.130

Maydeu-Olivares, A. (2017). Assessing the Size of Model Misfit in Structural Equation Models. *Psychometrika*, 1-26.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the Fit of Item Response Theory and Factor Analysis Models. *Structural Equation Modeling*, *18*, 333–356. DOI:10.1080/10705511.2011.581993

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*, 157–176. DOI:10.1037/1082-989X.12.4.433

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior research methods, 41,* 295-308. DOI:10.3758/BRM.41.2.295

Maydeu-Olivares, A., Rodriguez-Fornells, A., Gomez-Benito, J., D'Zurilla, T. J., Gómez-Benito, J., & D'Zurilla, T. J. (2000). Psychometric properties of the Spanish adaptation of the Social Problem-Solving Inventory-Revised (SPSI-R). *Personality and Individual Differences, 29,* 699–708. DOI:10.1016/S0191-8869(99)00226-3

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *BJMSP, 27,* 82–99. DOI: 10.1111/j.2044-8317.1974.tb00530.x

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 205–234). Newbury Park: Sage.

Muthen, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *BJMSP, 45,* 19-30. DOI: 10.1111/j.2044-8317.1992.tb00975.x

Muthén, L.K. and Muthén, B.O. (1998-2012). *Mplus* User's Guide. Seventh Edition. Los Angeles: Muthén & Muthén.

Olsson, U.H., Foss., T. & Troye, S.V. (2003). Does the ADF fit function decrease when the kurtosis increase? *BJMSP, 56,* 289-303. DOI:10.1348/000711003770480057

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. C. Clogg (Eds.), *Latent variable analysis. Applications for developmental research* (pp. 399–419). Thousand Oaks: Sage.

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: assessment and implications of generalized outcome expectancies. *Health psychology*, *4*(3), 219.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education*, *19*, 405-450.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292.

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, *25*, 173–180. DOI:10.1207/s15327906mbr2502_4

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. Paper presented at the Annual Meeting of the Psychometric Society, Iowa.

Symonds, P. M. (1924). On the Loss of Reliability in Ratings Due to Coarseness of the Scale. *Journal of Experimental Psychology*, *7*(6), 456.

Thurstone, L. L. (1947). Multiple factor analysis.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10. DOI:10.1007/bf02291170

Yuan, K.H., Bentler, P.M. & Zhang, W. (2005). The effect of skewness and kurtosis on meanand covariance structure analysis. *Sociological Methods & Research,* 34-32