PATH-RELATED FIT INDICES IN STRUCTURAL EQUATION MODELING: INTRODUCTION OF TWO GENERAL FRAMEWORKS AND EVALUATION OF PERFORMANCE OF FIT INDICES DERIVED FROM THEM

by

STEFANIE SUSANNE BECK

(Under the Direction of Charles E. Lance)

ABSTRACT

Causal analysis of covariance structures through structural equation modeling represents an indispensable tool for theory development and theory testing. In order to decide on whether to accept or reject a theoretical model, researchers rely on goodness-of-fit indices. However, the majority of currently used fit indices are global, that is, they are a function of the fit of both the measurement and the structural model. As has been shown by McDonald and Ho (2002) and O'Boyle and Williams (2011), this often leads researchers to erroneously accept misspecified models because good fit of the measurement model masks bad fit of the structural model. This study aims to provide alternative, more accurate fit indices. Two general frameworks for fit indices that rely on fit of the structural model only were developed, testing James' et al. (1982) condition nine and ten. Path-related fit indices were derived from the two frameworks and their performance under several different cutoff values was tested in a simulation of six population models. Their performance was compared to the performance of four of the most popular and widely used global fit indices *CFI, RMSEA, TLI,* and *SRMR.* Results show that all newly developed path-related fit indices are considerably more accurate in rejecting even slightly misspecified models than any of the global fit indices. Recommendations and implications for theory and practice are discussed.

INDEX WORDS: Structural Equation Modeling, Fit Index, Goodness of Fit, Structural Model, *RMSEA-P*, Model Misfit, Monte Carlo Simulation

PATH-RELATED FIT INDICES IN STRUCTURAL EQUATION MODELING: INTRODUCTION OF TWO GENERAL FRAMEWORKS AND EVALUATION OF PERFORMANCE OF FIT INDICES DERIVED FROM THEM

by

STEFANIE SUSANNE BECK

Prediploma, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, 2008

M.Sc., The University of Georgia, 2012

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

PATH-RELATED FIT INDICES IN STRUCTURAL EQUATION MODELING:

INTRODUCTION OF TWO GENERAL FRAMEWORKS AND EVALUATION OF

PERFORMANCE OF FIT INDICES DERIVED FROM THEM

by

STEFANIE SUSANNE BECK

Major Professor: Charles E. Lance

Committee: Nathan T. Carter
Robert P. Mahan

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2013

DEDICATION

This dissertation is dedicated to my family, especially to my parents Susanne and Michael Beck who supported me in every way imaginable throughout my studies. There is no doubt that without their continuous support and countless sacrifices, I would not have achieved what I have today. I would also like to thank David for his love and encouragement throughout this process, and for making everything easy.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my major professors Dr. Chuck Lance and Dr. Robert Mahan. Thank you to Dr. Chuck Lance for your support throughout the past years. Your trust in me as well as your hands-off mentoring approach allowed me to work independently, yet at the same time, I could always rely on your guidance and support. Thank you for giving me the opportunity to complete the dissertation from 5,000 miles away, and for many Skype dates. You're a very real statistical urban legend and I am proud to have been your student.

Thank you to Dr. Rob Mahan for being a wonderful mentor and providing invaluable support and advice along the way.

I would also like to thank Dr. Nathan Carter for his invaluable help with this dissertation, including teaching me how to conduct Monte Carlo simulations, troubleshooting, providing much encouragement along the way and figuring out the last missing PSI-value.

Thank you to Dr. Brian Hoffman for being an extraordinary professor and providing guidance, many interesting conversations, and research and applied opportunities throughout the last four years.

I could not have done it without my friends at UGA who have become my home away from home. Julia Sauer was my support system and the greatest friend in both good and difficult times. I am also very thankful for the friendship of Dr. Carlos Faraco, Allison Siminovsky and Stephanie Downey and I am confident that it will last despite the geographical distance.

Finally, I would like to thank Dr. Larry Williams, Dr. Stanley Mulaik, Dr. Gordon Cheung and Russell Ecob for their useful advice on Monte Carlo simulations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Over the past three decades, structural equation modeling (SEM) has become an essential tool for model testing in a variety of scientific disciplines. SEM allows researchers to investigate causal interrelationships between a set of variables. Using SEM, one can determine whether the propositions stated in a theoretical model hold when evaluated against empirical data (Moshagen, 2012). This makes SEM an indispensable technique for developing, evaluating and modifying theories (Anderson & Gerbing, 1988). The popularity of SEM stems from the many advantages it provides compared to other techniques such as regression: Using SEM, one can assess complex models with multiple dependent variables. Furthermore, one can estimate the full model at once, which allows one to test various hypotheses in one step instead of conducting separate analyses as in path analysis. One of the greatest advantages of SEM is that it allows researchers to model measurement error contained in a model's variables, which makes it superior to regular regression techniques with composite measures (Kline, 2005), which assumes outcome and predictor variables are measured without error. Thus, problems with measures may be identified so that they are not wrongfully attributed to theoretical, substantive causal relationships. Moreover, SEM provides researchers with a means to simultaneously assess the interrelationships between manifest indicators and their underlying latent factors as well as the interrelationships and causal structures among latent factors (Anderson & Gerbing, 1988). In addition to theory testing, there are other useful applications of SEM such as scale development, construct validation, and measurement invariance (MacCallum & Austin, 2000).

Researchers have used SEM to evaluate theoretical models in such diverse fields as psychology (Aryee, Chen, Sun, & Debra, 2007), economics (Küpper & Burkhart, 2009), marketing (Kemp & Bui, 2010), accounting (Stone, Bryant, & Wier, 2010), educational sciences (Velathuyam & Aldridge, 2013), and even bioinformatics (Wu et al., 2013). For example, SEM has been employed to study consumer behavior (Kemp & Bui, 2011), employee outcomes of abusive supervision (Aryee et al., 2007), antecedents of childrens' motivation to learn (Velathuyam & Aldridge, 2013), and determinants of patients' adherence to medication (Boyer et al., 2012). By evaluating causal models, researchers test scientific theories, thereby enhancing knowledge about construct interrelationships and building a foundation for subsequent research. Furthermore, establishing causal relationships between variables may serve as a basis for researchers and practitioners to design interventions, which oftentimes may be costly and time-consuming. Due to its theoretical and practical implications, it is crucially important that researchers can accurately evaluate proposed theoretical models using SEM. The goal of the current study is to improve one of the tools researchers use to interpret the results of SEM to help them draw the right conclusions from their data and advance theory-building and practice in their respective fields. Specifically, several new measures to more accurately evaluate causal models will be developed and their performance will be tested in a simulation study.

Improving these tools is necessary because the measures researchers currently use to evaluate their models lead in many cases to a misinterpretation of their results. Before outlining the reasons for this, some introductory comments about SEM will provide the context for the ensuing discussion.

Latent variable SEMs consist of two distinct sub-models: (a) the measurement model and (b) the structural model. In the measurement portion of the model, the relationships between observed variables and the latent factors representing them are examined. In the structural portion of the model, the hypothesized causal relationships between latent variables are evaluated. As such, different hypotheses are assessed in the measurement model and structural models. In order to test these hypotheses, the overall composite model is assessed against empirical data. If the pattern of covariances in the sample data is similar to an estimated covariance matrix based on the hypothesized relationships in the model, it is said to have good fit to the data. Misfit in the measurement model indicates that the observed variables might not measure what they were designed to measure. If misfit occurs in the structural portion of the model, the hypothesized structural relations do not hold in the sample.

In order to quantify the degree to which a model fits the sample data and to decide whether to accept or reject a model, researchers have developed a multitude of different goodness-of-fit indices (e.g. Bentler, 1990; Bentler & Bonett, 1980; Mulaik, James, van Alstine, Bennett, Lind, Stilwell, 1989; Tucker & Lewis, 1973). Fit indices provide researchers with a single numerical value that informs researchers about how well their theoretical model fits the sample data. Practically all of these fit indices can be considered global, which means that they provide a measure of how well the overall composite model fits the data. One confounding problem with using global fit indices is that when evaluating the source of misfit, they do not provide information on the location of the misspecification. In other words, if a model does not fit the data well, a global fit index does not indicate whether the misspecification occurred in the measurement model, the structural model, or both. Global fit indices alone do not tell the researcher if there was an issue with their measures, which might be fixed by modifying the measurement model, or if the hypothesized causal relationships were not confirmed in their data.

Even more problematic, a global fit index might indicate good fit of the composite model, however, this does not rule out the possibility that there is misfit in either portion of the model.

As such, there is the danger that researchers obtain favorable goodness-of-fit indices despite misspecifications in the structural model, consider their hypotheses as supported and publish their research, which then might become the foundation for subsequent research.

Soon after goodness-of-fit indices gained popularity among researchers, various authors recognized this problem and started questioning the appropriateness of evaluating global model fit to draw conclusions about causal relations (e.g. Marsh, 1987; Marsh & Hocevar, 1985; Mulaik et al., 1989; Sobel & Bohrnstedt, 1985). Despite their warnings, global fit indices continue to be widely used and researchers generally take fit index values above recommended cut-off values as evidence of support for their hypotheses (O'Boyle & Williams, 2011). A recent re-investigation of published studies show how pervasive the problem is: In more than two thirds of studies examined, authors stated that their hypotheses were supported when overall fit of their model was good. However, a separate examination of the measurement and structural models yielded that the causal relationships did not hold in the sample data (O'Boyle & Williams, 2011).

A handful of researchers have devoted attention to the issues related to global fit indices and have suggested alternative ways of assessing model fit that focus more on the structural portion of the model (Marsh, 1987; Marsh & Hocevar, 1985; McDonald & Ho, 2002; Mulaik et al., 1989; O'Boyle & Williams, 2011; Sobel & Bohrnstedt, 1985; Williams & O'Boyle, 2011). In their recent study, Williams and O'Boyle (2011) discussed two fit indices that focus exclusively on the structural portion of models and do not take into account any properties of the measurement model. They conducted a re-analysis of simulation data published earlier by Williams and Holahan (1994) and examined the performance of their two proposed fit indices. Williams and O'Boyle (2011) were able to show that they were much more accurate than global fit indices in accepting correctly specified models and rejecting misspecified models. Therefore, they recommended that these path-related fit indices should from thereon be used in addition to global fit indices to assess model fit.

While Williams and O'Boyle's (2011) work is an important step towards more accurate model testing, they failed to include several key points: First, they presented only two specific path-related fit indices. However, these fit indices can be shown to be special cases of a general framework for the evaluation of the structural portion of a model that is able to incorporate many stand-alone fit index researchers may choose. Such general frameworks for path-related fit indices have not been described to date.

A second issue with Williams and O'Boyle's article is that they did not provide cutoff-points for the two path-related fit indices whose performance they assessed. This is problematic because path-related and global fit indices do not reflect the same properties of a model: As mentioned above, the values of path-related fit indices are a function of fit in the structural portion of the model, whereas global fit indices are a function of fit in the composite model. Therefore, cutoff points recommended for global fit indices will most likely not apply to path-related fit indices. Without appropriate cutoff values, there is no useful information to be gained from path-related fit indices. As such, it is necessary to empirically derive cutoff values specifically for path-related fit indices that allow researchers to decide whether to accept or reject a model based on the values obtained.

The current study will address these important points and provide several novel contributions. Two general frameworks will be outlined from which researchers can derive fit indices that focus only on the structural portion of a model. Fit indices based on these frameworks will provide a test of two complementary hypotheses: Fit indices based on the first framework will allow researchers to assess structural model fit in regards to the hypothesis that causal relationships not estimated in a model are in fact zero. Three exemplary path-related fit indices that incorporate different stand-alone fit indices will be developed based on this framework. Fit indices derived from the second framework test the hypothesis that the estimated paths in a model are are indeed non-zero. Using simulation data, the ability of these fit indices to identify correctly specified and misspecified models will be tested. Furthermore, recommendations for appropriate cutoff values will be given. This will facilitate researchers' decision whether to accept or reject a model.

By using fit indices derived from the two frameworks presented in this study, future researchers will be able to subject their models to a more stringent test than by using global fit indices alone and can thus be more confident in their models if they are found to provide acceptable fit to the data.

The remainder of this chapter is organized as follows: first, some technical details of SEM will be outlined to provide a foundation for the discussion of fit indices. Next, issues with model testing that motivated the development of fit indices will be discussed. A typology of existing global fit indices will be provided and their respective properties will be discussed briefly. Shortcomings of these global fit indices will then be outlined and an overview over different researchers' suggestions to alleviate these issues will be given. Finally, the general frameworks for path-related fit indices will be developed and several different fit path-related fit indices will be derived from these frameworks.

SEM – Definition and Process

Structural equation models can be defined as a set of complex statistical hypotheses (McDonald & Ho, 2002). These hypotheses involve both the measurement model and the structural portion of a model. As mentioned above, in the measurement model, a researcher examines whether the hypothesized relationships between latent variables and the corresponding manifest variables chosen to represent the latent variables are supported in the sample data

(Lance & Vandenberg, 2002). More specifically, one examines whether on the one hand, observed variables have a non-zero relationship with the latent variables they were designed to measure, and on the other hand, are not directly related to any other latent variables in the model.

In the structural portion of the model, the relationships between the latent variables are assessed in respect to two equally important hypotheses. These two hypotheses were presented in James' et al. (1982) seminal book on causal analysis and arguably represent the foundation of hypothesis testing in SEM. According to James et al. (1982), ten conditions must be fulfilled to make causal assumptions, of which conditions 9 and 10 are central to this study. Condition 9 states that, in order to claim that a causal model is consistent with the data, the parameters estimated in the model must be significantly different from zero. Condition 9 can be tested either by examining each estimated parameter individually for significance or by conducting nested model comparisons, which will be explained in more detail below. Condition 10, on the other hand, focuses on the parameters in a model that are not being estimated. Not estimating specific parameters is tantamount to hypothesizing that they are equal to zero, which represents an important hypothesis in and of itself, a fact that is very often overlooked by researchers (Williams & O'Boyle, 2011). In other words, the paths not included in a model are of equal importance as the estimated paths.

McDonald and Ho (2002) therefore suggested that in developing their model, researchers should not only provide rationale on why they hypothesize certain relationships, but also state why they did not estimate certain parameters in their model. In summary, assessing an SEM informs the researcher about whether the measurement model is correctly specified, and whether in the structural model, estimated paths are non-zero while the paths excluded from the model are in fact zero.

After the pattern of parameters hypothesized to be zero/non-zero has been specified, the model is estimated. The estimation procedure involves an iterative process, whereby the discrepancy of a covariance matrix $\Sigma(\hat{\Theta})$ implied by the interrelationships of variables in the hypothesized model and an observed covariance matrix *S* is minimized through a fit function (Lance & Vandenberg, 2002). If the hypothesized model fits the sample data well, the residuals between the model-implied covariance matrix $\Sigma(\hat{\Theta})$ and the sample covariance matrix *S* are expected to converge asymptotically to zero. While there are several different fit functions used in SEM (e.g. maximum likelihood, weighted least squares, generalized least squares), the maximum likelihood fit function F_{ML} is by far the most widely employed (Lance & Vandenberg, 2002; Moshagen, 2012). It is defined as

$$F(\theta) = \ln|\Sigma(\theta)| - \ln|S| + tr\left[S\Sigma(\theta)^{-1}\right] - p \quad (1)$$

where $\ln|\Sigma(\hat{\Theta})|$ is the natural logarithm of the determinant of the model-implied covariance matrix, $\ln|S|$ is the natural logarithm of the determinant of the sample covariance matrix, $tr\left[(S\Sigma(\hat{\Theta})^{-1}\right]$ denotes the trace of *S* post-multiplied by the inverse of $\Sigma(\hat{\Theta})$,and *p* refers to the number of manifest variables in the model. In the limit if $\Sigma(\hat{\Theta})$ and *S* are identical, the difference between $\ln|\Sigma(\hat{\Theta})|$ and $\ln|S|$ will be equal to zero, the inverse of $\Sigma(\hat{\Theta})$ will be equal to *S*, too, so that their product will yield a *p* x *p* identity matrix *I*. As such, F_{ML} minimizes the discrepancies between covariances ($\ln|\Sigma(\hat{\Theta})|$ - $\ln|S|$) and variances ($tr\left[(S\Sigma(\hat{\Theta})^{-1}\right]$ - *p*) of $\Sigma(\hat{\Theta})$ and *S* (James et al., 1982; Lance & Vandenberg, 2002).

Several assumptions underlie the maximum likelihood fit function (Marsh, Balla & McDonald, 1988): The observed variables must have a multivariate normal distribution and be linearly related to the latent constructs. The analysis must be based on a sample covariance matrix, as opposed to a correlation matrix (Jöreskog & Sörbom, 1981). Furthermore, the model must be identified and sample size *N* must be large (Marsh et al., 1988). If these assumptions are met, then *N*-1 times the value of the maximum likelihood discrepancy function F_{ML} is asymptotically distributed as the χ^2-statistic. This property allows for statistical hypothesis testing. Smaller χ^2-values indicate better model fit, so that the χ^2-statistic can be considered a badness-of-fit test. Good model fit (or a lack of badness of fit) is indicated by a non-significant χ^2-value, so that a model is rejected when its χ^2-value is significantly different from zero. The null hypothesis states that model fit is perfect except for sampling error and that the model-implied covariance matrix does not differ from the sample covariance matrix. It is rejected when χ^2 is significantly different from *df*, the expected value under H_0 (Bentler & Bonett, 1980).

However, there are serious issues with using the χ^2-statistic as a determinant of model fit. $(N\text{-}1)^*F_{ML}$ is only asymptotically distributed as χ^2, which means that only with large samples, the fit function follows the χ^2-distribution that enables statistical significance testing. Therefore, the χ^2-statistic is directly sample-size dependent and increases with larger samples. This means that, as sample size increases, so does the danger of rejecting a model that in reality fits the data well, thereby committing a Type I error. Even trivially small discrepancies between $\Sigma(\hat{\Theta})$ and *S* will lead to rejection of the null hypothesis and thus the rejection of one's model. At the same time, one runs the danger of committing a Type II error when testing one's model with data derived from small samples, as with smaller samples, models are more likely to not be rejected even if in reality, they may not fit the data well (Bentler & Bonett, 1980). As such, the χ^2-statistic is not always clearly interpretable (Bentler, 1990).

This is problematic for various reasons: In order to support their hypotheses, researchers might be tempted to test their models purposefully on small samples or use only subsamples of their sample. This, however, runs counter to the very purpose of theory development and testing, namely, uncovering relationships between variables that can be generalized to the population or large parts thereof (Marsh et al., 1988). Moreover, a statistically significant χ^2 may not necessarily mean that the fixed parameters were poorly specified (Jöreskog, 1969; Mulaik et al., 1989). Rather, there could be other problems with the model due to a violation of any one of James' et al. (1982) conditions of causality such as, for example, lack of self-containment.

Therefore, Mulaik et al. (1989) claimed that researchers should not be interested as much in absolute fit as in close fit of their model to the data. Consequently, there might not be much value in assessing model fit in terms of a strictly dichotomous accept-reject decision strategy. In accordance with Mulaik et al. (1989), researchers have argued that theoretical models are designed with the goal of reflecting approximations instead of perfect representations of reality. Therefore, the null hypothesis of perfect model fit may be unrealistically stringent, irrelevant and of little practical value, and should thus not be used as the sole basis for the decision to accept or reject a theoretical model (McDonald, 1989; Steiger, 2007). Instead, a shift from classical hypothesis testing approaches towards a comparative model testing approach was recommended (Bentler & Bonett, 1980; Mulaik et al.; 1989; Tanaka, Panter, Winborne, & Huba, 1990). This motivated the development of alternative means to assess model fit. These alternatives – hierarchical model testing and goodness-of-fit indices- will be described on the following pages.

Hierarchical Model Testing

In light of the limitations of the χ^2-statistic, several authors (e.g. Bentler & Bonett, 1980, James et al., 1982) have recommended using hierarchical model testing. Thereby, a sequence of nested models is estimated and one's hypothesized model is compared to both more and less restrictive nested models. The least restrictive model in this sequence is a saturated structural model (SS) in which, once the causal order among variables is established, all possible relationships between latent variables are freely estimated. This model always has the best possible fit and the number of its degrees of freedom is equal to the degrees of freedom of the measurement model. On the other end of the continuum lies a null model, which, in its most rigid form, has all paths set to zero. As such, the null model represents the hypothesis of no relationships between any measured variables. This model displays the worst possible fit and the highest number of degrees of freedom. In between the saturated structural model and the null model lies one's hypothesized, or target model (T). Furthermore, there are more restrictive models (T-x) than the target model, where one or more paths that are estimated in the target model are set to zero, and less restrictive models (T+x), in which one or more parameters that are fixed to zero in the target model are freely estimated.

The difference between the target model's and an alternative model's χ^2-statistics then provides useful information for hypothesis testing: If the difference in χ^2 between a target model and a more restrictive model (T-x) is significant, one can conclude that the additional paths estimated in the target model are relevant and improve model fit. If, on the other hand, the χ^2-value of the target model is not significantly smaller than the χ^2-value of a more restrictive model, it tells the researcher that the additional paths included in the target model do not improve model fit. Since parsimonious models with fewer estimated parameters are preferable (Mulaik et al., 1989), the target model should be abandoned in favor of a more parsimonious model in that case.

Likewise, a target model whose χ^2 is significantly larger than the χ^2-value of a less restrictive model (T+x) may be abandoned, as a larger χ^2 in this case indicates that the model's restrictions lead to worse model fit. If, on the other hand, the target model's χ^2 is not significantly larger than the χ^2 of a less restrictive model, the target model fulfills James' et al.

(1982) condition 10 requirement, that is, no important paths have been wrongfully omitted from the model.

In summary, if one's target model fits the data significantly better than a more restrictive model, and at the same time, it does not display significantly worse fit than a model with additional parameters estimated, one may conclude that the hypothesized model is consistent with the data[1].

Hierarchical model testing provides important information about the fit of one's target model in relation to other relevant models and helps circumvent the issues associated with the χ^2-statistic. However, Bentler and Bonett (1980) noted that there would be much value in additionally obtaining a single index of the increase in goodness-of-fit in the comparison of two models. Therefore, a multitude of goodness-of-fit indices have been developed to provide information about model fit beyond the χ^2-statistic. In the following, a typology of existing fit indices will be provided.

Goodness-of-Fit Indices

Ever since Bentler and Bonett's (1980) seminal article, goodness-of-fit indices have played a central role in the evaluation of structural equation models (Kenny & McCoach, 2003). Goodness-of-fit indices provide a single numerical measure of whether a hypothesized model does not have any paths inappropriately constrained to zero and as such, meets James' et al. (1982) condition 10 requirement. Since the number of fit indices has grown large over the years, several researchers have provided different conceptual frameworks to organize them (e.g. Tanaka, 1993). In the context of the present study, the widely recognized (e.g. Bollen, 1989; Hu & Bentler, 1998; Marsh et al., 1988; Meade, Johnson & Braddy, 2008; Tanaka, 1993) distinction between absolute and incremental fit indices is most relevant and will therefore be used. This

[1] Although it represents only one out of an infinite number of theoretically possible models.

distinction refers to the reference point against which model fit is evaluated: the reference point in absolute fit indices is perfect model fit, whereas in incremental fit indices, the target model is being compared to another reference model (Meade et al., 2008).

Absolute Fit Indices

Absolute (alternatively: stand-alone) fit indices provide a direct measure of how well a covariance matrix derived from a hypothesized model reproduces the sample data, whereby no explicit reference to another model is made (Hu & Bentler, 1998). As such, they may be considered as the multivariate equivalent to the coefficient of determination (R^2) in regression analysis (Fan, Thompson & Wang, 2007, Tanaka & Huba, 1989). They are functions of the discrepancy of $\Sigma(\hat{\Theta})$ and *S*, sample size, and degrees of freedom (McDonald & Ho, 2002).

Researchers have combined these elements in various ways to build different indicators of absolute model fit. Technically, the χ^2-statistic may also be considered an absolute fit index. As described above, however, the assumption that the χ^2-statistic is asymptotically distributed as χ^2 represents a test of the unrealistic hypothesis of perfect model fit, which is most likely rejected with large sample sizes. Researchers have therefore suggested that in cases of small model misspecifications, where the degree of discrepancy is limited, the distribution of the χ^2-statistic may be better approximated by the noncentral χ^2-distribution. The noncentral χ^2-distribution has been described as a more realistic and thus useful reference distribution (Rigdon, 1996). The degree of noncentrality is indicated by the parameter λ, which is asymptotically equal to χ^2-df (McDonald, 1989). The size of the noncentrality parameter (NCP) λ can be considered an indicator of the degree of model misspecification, whereby a small NCP is associated with a small degree of model misspecification (Bentler, 1990).

Therefore, the NCP may be considered another absolute fit index, and a number of researchers have chosen to incorporate the NCP into their absolute and incremental fit indices.

Another absolute fit index is χ^2/df (Jöreskog, 1969), which has the same underlying rationale as χ^2. By dividing a model's χ^2 by its degrees of freedom, however, χ^2/df has an additional penalty function for lack of parsimony built in, since the number of degrees of freedom decreases with the number of parameters estimated Marsh et al., 1988). The χ^2/df – ratio is expected to approximate 1.0 when the model is consistent with the data, as $E(\chi^2) = df$ under the central χ^2-distribution. When model fit is not perfect, the expected χ^2 –value equals the sum of *df* and the NCP, so that $E(\chi^2) = df + \text{NCP}$. As such, with increasing model misfit, the χ^2/df – ratio gets larger than 1.0 due to χ^2 being a function of both *df* and the NCP, which increases with model misfit as well.

Among the group of absolute fit indices, Hu and Bentler (1998) recommended using the root-mean-squared error of approximation (*RMSEA*; Steiger & Lind, 1980), and the standardized root-mean-square residual (*SRMR*; Jöreskog & Sörbom, 1981). These indices were shown to be sensitive to model misspecification, while being less sensitive to sample size and estimation method than other fit indices. Other absolute fit indices include Akaike's Information Criterion (*AIC*; Akaike, 1987), the goodness-of-fit index (*GFI*; Jöreskog & Sörbom, 1984), the adjusted goodness-of-fit index (*AGFI*; Jöreskog & Sörbom, 1984), and the critical *N* (Hoelter, 1983). However, due to various issues with these indices such as low sensitivity for model misspecification, and high sensitivity to distribution, estimation method, and sample size, Hu and Bentler (1998) did not recommend their use.

Incremental Fit Indices

Incremental fit indices measure the fit of a hypothesized model by comparing it to a reference model, which is usually either a baseline model or a fully parameterized model (Meade et al., 2008). As such, incremental fit indices provide a measure of how well a model fits the data compared to either the best- or the worst-fitting model in a series of nested models. Incremental fit indices incorporate stand-alone indices and can be further categorized into three groups. Type I fit indices are normed and only use information from the fit function. Bentler and Bonett (1980) were among the first researchers to develop a type I incremental fit index. Their normed fit index (*NFI*) is defined as follows

$$NFI = \frac{F_{(B)} - F_{(T)}}{F_{(B)}} \tag{2}$$

where F_B and F_T refer to the test statistic for a baseline model and the hypothesized target model, respectively. The difference in fit between the target model and the worst-fitting baseline model is assessed in reference to the worst possible fit. In other words, obtaining a high *NFI* means that the target model provides a meaningful improvement in fit compared to a model where no relationships between any manifest variables are hypothesized.

Type II indices additionally use information from the expected value of the fit function under the central χ^2-distribution and are considered non-normed, since their value can fall outside the range of 0-1. One early example is the Tucker-Lewis Index (Tucker & Lewis, 1973) that incorporates the χ^2/df stand-alone index:

$$TLI = \frac{\dfrac{\chi^2_{(B)}}{df_{(B)}} - \dfrac{\chi^2_{(T)}}{df_{(T)}}}{\dfrac{\chi^2_{(B)}}{df_{(B)}} - 1} \tag{3}$$

whereby $E(\chi^2) = df$. If the target model is consistent with the data, the expected value of χ^2/df is 1.0 and the expressions in the numerator and denominator become identical, so that $TLI = 1.0$ with perfect model fit. Similar to the *NFI*, the reference model is a baseline model that specifies no relationships at all between any of the variables.

Type III indices are similar to type II indices in that they incorporate expected values of the fit function, however, instead of the central χ^2-distribution, a noncentral χ^2-distribution is assumed for their expected values (Meade et al., 2008). Bentler's (1990) comparative fit index (*CFI*), which is one of the most popular and widely used fit indices (Williams & O'Boyle, 2011), belongs to the group of type III indices and is defined as follows:

$$CFI = 1 - \frac{NCP_T}{NCP_B} \quad (4)$$

where the NCP of the target model is being compared to the NCP for a baseline model. The *CFI*, too, incorporates the most rigid form of baseline models, specifying no relationships at all. As the NCP of a true target model approximates zero, the expected value of the *CFI* is 1.0 when model fit is perfect.

Apart from these examples, researchers have developed numerous other type I, type II, and type III fit indices. Their properties are comprehensively documented in Marsh et al.,'s (1988) and Hu and Bentler's (1998, 1999) and other authors' (e.g. Rigdon, 1996; Tanaka, 1993) work.

Issues with Existing Fit Indices

While fit indices were designed to circumvent the issues related to hypothesis testing with the χ^2-statistic, they themselves present challenges that limit their ability to identify correctly specified and misspecified models. Ideally, the size of a fit index is influenced only by the degree of model misspecification, so that misspecified models obtain a lower value than correctly specified models (alternatively, in case of badness-of-fit indices such as *RMSEA* and *SRMR*, misspecified models should obtain a higher value). However, many fit indices have shown to be sensitive to sample size, distribution, estimation method, number of variables in the model (Kenny & McCoach, 2003; Moshagen, 2012) and effects of violation of normality and independence (Hu & Bentler, 1998, 1999). At the same time, fit indices are not always sensitive to model misspecification (Hu & Bentler, 1998, 1999).

Furthermore, there still do not exist reliable cut-off points for what constitutes acceptable model fit. Bentler and Bonett (1980) suggested that models with *TLI*- and *NFI*-values lower than .90 might be severely misspecified. This led many researchers to erroneously reason that fit index values above .90 indicate good model fit (Lance, Butts, & Michels; 2006). Hu and Bentler (1998,1999) conducted extensive simulation studies and recommended a cutoff value of .95 for many fit indices such as the *CFI* and the *TLI*, which has been subsequently adopted by most researchers. Recent simulation studies have shown, however, that these cutoff values may still be too lenient and not help researchers in choosing the right model (Williams & O'Boyle, 2011). Arguably one of the most problematic aspects about all incremental and some standalone fit indices, however, concerns the type of baseline model incorporated in them.

Baseline Models in Fit Indices

There are several different baseline models that may be used in incremental fit indices, and which one of them is the appropriate one has been subject to debate (e.g. Mulaik et al., 1989; Sobel & Bohrnstedt, 1985; Williams & O'Boyle, 2011). Williams and Holahan (1994) discussed three distinct types of baseline models. First, an absolute null model (AN) is characterized by the absence of a measurement model, that is, there is no link between latent variables and their manifest indicators. Furthermore, there are no correlations among exogenous latent variables and no structural parameters linking the latent variables. The variance/covariance matrix of an absolute null model is a diagonal matrix, with all off-diagonal elements being zero (i.e., $\Sigma(\hat{\Theta}) = \sigma_i^2 I$). As described above, when built into incremental fit indices, the comparison to an absolute null model represents the hypothesis that the target model fits the data better than a model specifying no relationships between the variables at all.

Second, a less stringent baseline model is an uncorrelated factors model (UF), which specifies a measurement model linking latent variables and their indicators, but neither correlations among exogenous latent variables nor structural parameters relating latent variables. Third, in a structural null model (SN), both the measurement model and correlations among exogenous latent variables are estimated, but all directed paths are set to zero. As such, out of these three baseline models, the structural null model resembles most closely the target model in that the only difference lies in the directional paths estimated in the target model that are set to zero in the structural null model.

Bentler and Bonett (1980) argued that one should use the "most restrictive, theoretically defensible model" (p.600) as a basis for comparison. Counter to their recommendation, most researchers incorporated the absolute null model in their fit indices, such as the *NFI*, the *TLI*, and the *CFI*. This approach has been criticized by a number of researchers: Sobel and Bohrnstedt (1985) were the first authors to voice strong concerns about this practice. According to them, the choice of a baseline model should be guided by the current state of knowledge about the relationships to be tested. They argued that if there existed some prior knowledge about relationships among variables, that is, if the research model were in part confirmatory, it would not be appropriate to compare the target model to an absolute null model. By choosing an absolute null model as baseline model, researchers would ignore what is already known about relationships among at least some of the variables in the model. Furthermore, the only information gained from a fit index based on an absolute null model is that estimating some relationships provides better fit to the data than hypothesizing no relationships between variables at all. However, it does not tell the researcher if one's model provided meaningful and valid information over what is already known.

Sobel and Bohrnstedt (1985) further argued that the absolute null model should only be employed in purely exploratory cases where there is no existing theory to guide model development. McDonald and Ho (2002), too, pointed out that there is no compelling rationale for using the absolute null model as a baseline model.

Mulaik et al. (1989) agreed with Sobel and Bohrnstedt (1985) in that the use of the absolute null model in incremental fit indices may mask small, yet important differences between two nested structural models and argued for the use of less restrictive null models. They noted that the overall fit of a structural equation model may oftentimes be heavily influenced by the fit of the measurement model and much less by the fit of the structural model. In the measurement model, a substantial number of parameters that are irrelevant for the structural relations in the model are estimated and thus influence overall goodness-of-fit. As researchers add manifest indicators to their model to decrease the risk of obtaining improper solutions (Ding, Velicer & Harlow, 1995), more parameters in the measurement model need to be estimated. With more parameters in the measurement model, the influence of the measurement model on global fit indices increases, so that it may drive fit to a larger degree than the structural portion of the model. As such, it may be well possible to obtain a high goodness-of-fit index/low badness-of-fit index when only the measurement model is correctly specified.

At the same time, the structural model, that researchers are primarily interested in because it represents the causal relationships in the model, may be misspecified. Likewise, when model fit is found to be not acceptable, global fit indices do not provide any information about whether the misspecification occurred in the measurement portion or the structural portion of the model, or both (McDonald & Ho, 2002).

Marsh and Hocevar (1985) noted that the problem of confounding sources of misfit extends to higher-order factor models. In this special case, a global fit index might indicate good model fit when the relationship between indicators and first-order factors is correctly specified, however, the relations between first- and second-order factors might be misspecified. Likewise, second-order factors might provide a good representation of the covariances among first-order factors, but if the relationships between indicators and first-order factors are misspecified global fit indices might still indicate poor model fit.

Solutions to the Baseline Problem

In order to alleviate the problem of confounding sources of misspecification, researchers have suggested different solutions: Sobel and Bohrnstedt (1985) recommended that researchers individually tailor the baseline model to the respective research question they wish to examine. That way, the baseline model would reflect all current knowledge about the relationships among variables in the model, whereas the target model should additionally incorporate as of yet unknown relationships among variables. However, tailoring the baseline model of a fit index individually to a specific research context might necessarily lead to highly idiosyncratic choices of nontrivial null models, so that the comparability of goodness of fit of different models estimated in different studies would be severely limited (Marsh et al ., 1988).

Marsh (1987) introduced the target coefficient (*TC2*), a fit index for higher-order factor models that allows one to test whether the relationships between first- and second-order factors are correctly specified. The *TC2* is defined as

$$TC2 = \frac{(F_U - T)}{(F_U - F)} \tag{5}$$

where F_U denotes an uncorrelated factors model, T denotes the target model and F denotes an oblique factors model. The uncorrelated factors model represents the baseline with the worst possible model fit whereas the oblique model represents the best possible model fit. As such, the improvement in fit of the target model over the baseline model is evaluated against the total amount of fit contributed by the relationship between first-and second-order factors. This way, the confounding influence of the relationships between indicators and first-order factors is excluded from the analysis. However, by using an uncorrelated factors model as the baseline model, one still includes correlations between first-order factors in the analysis even though they are not relevant to the model comparison. Therefore, one confounding element remains, as model fit might to some degree be driven by the correlations among factors.

Similarly, Mulaik et al. (1989) developed the relative normed fit index (*RNFI*), which incorporates the uncorrelated factors model instead of the absolute null model as a baseline model.

It is defined as

$$RNFI = \frac{(F_U - F_J)}{\left[(F_U - F_M) - (d_J - d_M)\right]} \tag{6}$$

where F_U is a measure of fit of the uncorrelated factors model, F_J is a measure of fit of the structural model, and F_M is a measure of fit of the measurement model. D_J and d_M denote the degrees of freedom of the structural and the measurement model. The difference in fit between the baseline model F_U and the structural model F_J is compared to the difference in fit between the baseline model and the measurement model. The degrees of freedom in the denominator provide a correction for sample-size bias following Marsh's et al. (1988) recommendations. The *RNFI* allows one to determine the increase in model fit related to the structural model relative to model fit due to the measurement model. Similar to Marsh's TC2, one weakness of the RNFI lies in its use of the uncorrelated factors model as a baseline model: there are still parameter estimates involved in the model comparison that are not directly relevant for causal relationships in the structural model. In general, one should therefore not use the uncorrelated factors model as a baseline model if the focus is on the structural relationships of the model.

In their seminal paper, Anderson and Gerbing (1988) recommended using a two-step approach in model testing. In step one, the measurement model should be estimated separately, and only after the measurement model has been respecified until it fits the data well, one should assess overall model fit. However, as McDonald and Ho (2002) noted, Anderson and Gerbing (1988) still recommended that once the measurement model has been adjusted overall model fit should be assessed with global fit indices that rely on absolute null model comparisons.

Regardless of the criticisms on the use of the absolute null model and the alternatives some authors have provided, to this day researchers continue to use fit indices based on it. Rigdon (1996) noted that, despite being aware of the issue underlying the use of the absolute null model in fit indices, researchers may be hesitant to develop better fit indices that focus more on fit of the structural model. He stated that "… certainly, without population distributions for the resulting indexes, researchers who adopt alternative baseline models will be forced to develop and defend criteria for evaluating the index values that result. Consequently, it is unlikely that there will be a movement toward an alternate baseline model anytime soon" (p.377).

McDonald and Ho (2002) were the first to systematically examine the impact of the type of baseline model. They assessed the difference between global fit indices and a fit index that focuses only on path model relationships in regards to their ability to identify correctly specified and misspecified models. The authors suggested that fit of the measurement model and the structural model should be evaluated separately in order to disentangle model fit of the measurement model and the structural model. As the asymptotic χ^2-distribution of a composite model's discrepancy function consists of two independent additive noncentral χ^2-values, one can obtain a χ^2-value for the measurement model and one for the structural model. Likewise, the corresponding degrees of freedom for the measurement model and the structural model are additive. This important property of the asymptotic χ^2-distribution allows researchers to decompose overall model fit into separate measures of model fit contributed by the measurement model and by the structural model, respectively.

In order to demonstrate the impact of the measurement model on overall model fit, McDonald and Ho (2002) re-examined models from 14 published studies. In these studies, χ^2-values and degrees of freedom for the composite model as well as separate χ^2-values and degrees of freedom for the measurement model and/or the structural model were provided. The authors analyzed χ^2-values and degrees of freedom of the measurement model and the structural model. Additionally, they calculated separate *RMSEA*-values for both the measurement and the structural model. The *RMSEA* based on the structural portion of the model was later termed *RMSEA-P* (O'Boyle & Williams, 2011) and represents one of only two known fit indices that focus exclusively on causal relations of a model. It is defined as

$$RMSEA-P = \sqrt{\frac{(\chi_p^2 - df_p)}{(df_p * (n-1)}} \qquad (7)$$

where $\chi_p{}^2$ refers to the χ^2-statistic of the structural model, df_p are the degrees of freedom of the structural model, and *n* denotes the sample size. As such, the *RMSEA-P* is a measure of the degree of misspecification in the structural model per degree of freedom, adjusted for sample size.

McDonald and Ho (2002) found that the number of degrees of freedom for the structural model was generally (much) smaller than for the measurement model, which indicates that model fit was more strongly influenced by the measurement portion of the model. A troubling finding, however, was that in many cases in McDonald and Ho's (2002) analysis, model fit of the composite model appeared satisfactory when the structural portion of the model did not fit the data well and yielded high χ^2- and *RMSEA-P* values. Here, the good fit of the measurement model masked the bad fit of the structural model.

These findings are alarming, because they imply that a substantial part of existing research involving structural equation models may have arrived at faulty conclusions based on favorable values of global fit indices. Although McDonald and Ho's (2002) findings clearly support earlier critics' concerns about global fit indices, researchers continue to use and manage to publish their studies in top-tier journals. However, it is highly questionable whether the relationships between latent variables reported in their studies would have been supported when examined in isolation, separated from the measurement model (Williams and O'Boyle, 2011).

O'Boyle and Williams (2011) further advanced the discussion about fit indices that focus on the structural portion of a model. Extending McDonald and Ho's (2002) work, they re-examined 43 studies published in top-tier journals between 2001 and 2008 that contained structural equation models. In these models, they decomposed overall model fit into fit of the measurement model and fit of the structural model. O'Boyle and Williams (2011) conducted χ^2-difference tests between the composite model and the measurement model to determine whether the difference in fit between these models was large enough to assume that the structural portion contributed to overall fit on top of the measurement model. Additionally, they calculated the *RMSEA-P* as proposed by McDonald and Ho (2002).

Similar to McDonald and Ho (2002), they found that in the majority (70%) of the studies examined, fit indices for the composite model indicated good model fit but when examined in isolation, the structural portion of the model fit the data poorly and yielded fit indices that did not meet conventional cutoff criteria. This means that in the majority of the models examined, researchers have omitted potentially relevant paths linking latent variables, failing to meet James' et al. (1982) condition 10 requirement. O'Boyle and Williams (2011) thus recommended that researchers rely more on fit indices that focus on the fit of the structural model, such as the *RMSEA-P* instead of inferring from global fit indices that their model fits the data well.

To provide an additional path-related fit index, Williams and O'Boyle (2011) subsequently presented the noncentrality structural covariance index (*NSCI-P,* whereby the "P" stands for "path"). The *NSCI-P* was originally developed by Williams and Holahan (1994), however, it remained overlooked by the literature until Williams and O'Boyle recently re-introduced it. It involves a comparison of the NCPs of the structural null model, the target model and the saturated structural model and is defined as

$$NSCI\text{-}P = \frac{(\chi^2_{SN} - \chi^2_{T}) - (df^2_{SN} - df^2_{T})}{(\chi^2_{SN} - \chi^2_{SS}) - (df^2_{SN} - df^2_{SS})} \qquad (8)$$

In the *NSCI-P,* the difference in fit of the structural null model and the target model is evaluated against the total amount of fit available in the structural model, that is, the difference between the structural null model and the saturated structural model.

Using simulation data from Williams and Holahan (1994), Williams and O'Boyle (2011) demonstrated that both the *NSCI-P* and the *RMSEA-P* are considerably more accurate in identifying correct models than global fit indices such as the *RMSEA* and *CFI*. Based on these results, they recommended that researchers should always include path-related fit indices in model evaluation.

As mentioned above, the two path-related fit indices *RMSEA-P* and *NSCI-P* are to date the only known fit indices that incorporate a structural null model as a baseline model. They provide researchers with an important tool to evaluate their hypotheses more accurately. However, it is not clear why Williams and O'Boyle (2011) limited their treatment of path-related fit indices to these two specific indices. Instead, analogous to better-known global fit indices, it would be helpful to provide a general framework for path-related fit indices. This would allow for the development of different path-related fit indices by incorporating various stand-alone indices. As described above, a number of researchers have called for the development and application of path-related fit indices, however, nobody has specifically outlined what a general framework for path-related fit indices might look like. Therefore, one goal in this study is to outline such a framework and test whether fit indices based on this framework are better able to identify correctly specified and misspecified models than global fit indices.

General Framework and Exemplary Path-Related Fit Indices for Condition 10

As with incremental global fit indices, a general framework for path-related fit indices involves nested model comparisons. Contrary to global fit indices, these model comparisons must be limited to the latent variable level for path-related fit indices. As such, baseline models that include parts of the measurement model cannot be incorporated in path-related fit indices. This leaves the uncorrelated factors model and the structural null model as possible baseline models for path-related fit indices. The uncorrelated factors model may still contain potentially confounding information, as estimating the correlations between exogenous latent variables does not provide direct information on causal relationships. Therefore, the structural null model is the most appropriate baseline model for path-related fit indices. As such, the range of nested model comparisons contains a structural null model, T-x models, the target model, T+x models, and a saturated structural model. In order to facilitate the discussion of the frameworks for path-related fit indices, it is helpful to express total fit contained in the structural portion of the model with the following formula:

$$\text{Total Fit (Structural Model)} = \frac{(F_{SN}-F_T)+(F_T-F_{SS})}{(F_{SN}-F_{SS})} \tag{9}$$

where F_{SN} represents fit of the structural null model, F_T represents fit of the target model and F_{SS} represent fit of the saturated structural model. This formula shows that total fit in the structural model can be expressed as the sum of two elements: The first element is the discrepancy in fit between the structural null model and the target model (F_{SN} - F_T). As such, it is a measure of the improvement in fit obtained by estimating the hypothesized paths in a model over a model with no structural relationships. This is equivalent to James' et al. (1982) condition 9 test. The second element denotes the discrepancy between the fit of the target model and the fit of the best possible model with all unidirectional paths estimated (F_{SN} - F_{SS}). As such, this expression is a measure of James' et al. (1982) condition 10, in that it measures the changes in model fit obtained by setting specific paths to zero. Overall, this formula shows that perfect model fit is a function of both estimating significant structural paths and setting non-significant structural paths to zero. From this formula, path -related fit indices for condition 9 and condition 10 tests can be derived.

As has been demonstrated above, a path-related fit index would include a nested model comparison between the target model and a saturated structural model in order to provide an omnibus test of James' et al. (1982) condition 10, This gives a measure of the degree to which setting a number of paths in the target model to zero reduces model fit as compared to a model with perfect fit in the structural portion of the model. The resulting difference should be evaluated against the maximum amount of fit available only in the structural portion of the model. By doing so, one excludes the potentially misleading influence of the measurement model. This number is obtained by comparing the structural null model, representing the worst possible fit the structural portion of a model can have, to the saturated structural model. This differentiates path-related fit indices from global fit indices where the total amount of fit is defined as the difference between the saturated structural model and the absolute null model. As such, a general path-related fit index testing condition 10 (from hereon termed *C10*) would take the following form:

$$C10 = \frac{F_{SN} - F_T}{F_{SN} - F_{SS}} \qquad (10)$$

A researcher may pick any stand-alone fit index *F* and create path-related fit indices by incorporating them into the above framework. In order to demonstrate this, the performance of three path-related *C10* fit indices in identifying correctly specified and misspecified models will be examined in this study. The first index that will be examined is a *C10*-index as proposed

above incorporating the χ^2-statistic:

$$\frac{\chi^2_T - \chi^2_{SS}}{\chi^2_{SN} - \chi^2_{SS}} \tag{11}$$

Alternatively, one may additionally build in a penalty function for lack of parsimony and design a path-related *C10* fit index based on Jöreskog's (1969) stand-alone index χ^2/df:

$$\frac{\frac{\chi^2_T}{df_T} - \frac{\chi^2_{SS}}{df_{SS}}}{\frac{\chi^2_{SN}}{df_{SN}} - \frac{\chi^2_{SS}}{df_{SS}}} \tag{12}$$

Finally, a path-related fit index based on NCPs takes the following form:

$$\frac{(\chi^2_T - \chi^2_{SS}) - (df_T - df_{SS})}{(\chi^2_{SN} - \chi^2_{SS}) - (df_{SN} - df_{SS})} \tag{13}$$

These three fit indices are examples of path-related fit indices based on the framework introduced above, however, researchers may incorporate any stand-alone fit index of their choice, such as for example the *GFI*, the *AGFI* or the *SRMR*.

General Framework and Exemplary Path-Related Fit Indices for Condition 9

While researchers have devoted much attention to developing fit indices that provide a test of James' et al. (1982) condition 10, the *NSCI-P* and the *TC2* are currently the only omnibus goodness-of-fit indices that provide a test of condition 9. As mentioned above, researchers have relied on two approaches recommended by James et al. (1982) to test whether a model fulfills condition 9. First, one may examine each parameter separately by testing it for significance or creating confidence intervals. However, James et al. (1982) pointed out that significance tests of individual parameters might not be independent and have an unknown bias. Second, one may test each parameter by comparing the model in which the respective parameter is estimated to a nested model where the same parameter is fixed to zero and conduct a single df chi-square difference test to determine if fit worsens if the parameter is set to zero.

Given the large number of omnibus tests for condition 10, it is not entirely clear why there are only two corresponding omnibus tests for condition 9 to date. Above all, these two indices have remained largely overlooked by the literature until recently. A potential reason may lie in the great influence of James' et al. (1982) work on other researchers, where the authors explicitly stated that goodness-of-fit tests only relate to condition 10. However, their argument was based on the use of the χ^2–statistic to assess overall model fit, which indeed represents a condition 10 test. In their quest to develop fit indices as an alternative to the χ^2–test, researchers perhaps focused too much on condition 10 tests and did therefore not consider developing fit indices for condition 9. In the case of Marsh's (1987) *TC2*, it is likely that it didn't become popular because Marsh only briefly mentions it in the appendix of an article on higher-order factor analysis, but does not discuss it at all in his later work on goodness-of-fit indices.

Analogous to condition 10 omnibus fit indices, a fit index testing condition 9 can be developed by combining relevant model comparisons into one index. More specifically, it involves two model comparisons. First, the improvement of fit of a target model over a structural null model should be assessed. This number should then be evaluated against the total amount of fit contributed by the structural portion of the model. As described above, the total amount of fit contained in a structural model can be obtained by comparing the structural null model to the saturated structural model. As such, the general structure of a measure of overall goodness-of-fit (in the following termed *C9*) testing condition 9 can be defined as follows:

$$C9 = \frac{F_{SN} - F_T}{F_{SN} - F_{SS}} \qquad (14)$$

Similar to *C10* fit indices, researchers can select any stand-alone fit index of their choice to create a *C9* index. In order to demonstrate this, two *C9* indices will be designed by employing the general framework presented above, and their ability to identify correct and misspecified models will be tested on simulation data. A *C9* index in its simplest form incorporates the χ^2–statistic:

$$\frac{\chi^2_{SN} - \chi^2_{T}}{\chi^2_{SN} - \chi^2_{SS}} \tag{15}$$

Analogous to the *C10* indices presented above, one may alternatively build a *C9* index based on the $\chi 2/df$ stand-alone index, if model parsimony is supposed to be taken into account:

$$\frac{\frac{\chi^2_{SN}}{df_{SN}} - \frac{\chi^2_{T}}{df_{T}}}{\frac{\chi^2_{SN}}{df_{SN}} - \frac{\chi^2_{SS}}{df_{SS}}} \tag{16}$$

As mentioned above, the *NSCI-P* is another path-related fit index containing the structure of the general *C9* framework that incorporates NCPs as standalone-indices. As with *C10* fit indices, these *C9* indices are exemplary for fit indices built from the general *C9* framework, however, any other stand-alone fit index may be incorporated into the framework to design a path-related fit index. An overview over the general frameworks and the fit indices derived from them is shown in Table 1.

Cutoff Values of Path-Related Fit Indices

Cutoff values provide researchers with guidance on which fit index values represent acceptable model fit. While there exist a number of simulation studies on cutoff values for global fit indices (e.g. Hu & Bentler, 1998; 1999), only one study to date (Williams & O'Boyle, 2011) briefly mentions potential cutoff values for path-related fit indices. As path-related fit indices do not focus on the same elements of a model as global fit indices, they are expected to be different from cutoff indices for global goodness-of-fit indices. As such, traditional cutoff values might not be applicable to path-related fit indices. Therefore, the main goal of this study is to systematically examine cutoff values specifically for path-related fit indices and determine which ones are most effective in accepting correct models and rejecting misspecified models.

The performance of fit indices under specific cutoff values is generally determined through hierarchical model comparisons on simulated data. In simulated data, the underlying population model is known, as such, the "true" target model is also known. A series of nested models is fit to the data that are created on basis of the population model. This series includes a null model, models with paths incorrectly set to zero, the target model, models with non-significant paths estimated, and a saturated structural model. Different cutoff values can then be evaluated according to two criteria. The first criterion concerns the frequency of Type-II error under a certain cutoff value, which is the percentage of cases in which the correct target model is incorrectly rejected in favor of more or less restrictive misspecified models. The second criterion refers to the power of a cutoff value, that is, the percentage of cases in which misspecified models are rejected under a certain cutoff value.

Potential Cutoff Values for C10 Indices

Fit indices based on the *C10* framework presented above involve a comparison of the target model to the best-fitting model in a series of nested models. If a target model is correctly specified and no paths are inappropriately fixed to zero, it is expected that estimating additional structural paths will not lead to a substantial improvement in model fit. As such, it is expected that the difference in fit between the correctly specified model and the saturated structural model is approximately zero. In the *C10* framework, this quantity is divided by the amount of total fit available in the structural portion of the model, so that the overall expression approximates zero as well. As such, *C10* indices may work as "light switches", in that they yield a value of approximately zero for correctly specified models. A pilot study conducted prior to this study, where the performance of the three *C10* indices presented above was assessed using simulated data, lends tentative support for a cutoff value of zero.

In this study, the performance of the three proposed *C10* fit indices in regards to the criteria described above will be evaluated using cutoff values of 0, 0.01, 0.025 and 0.05. Furthermore, their performance will be compared to four of the most popular global fit indices. More specifically, Type II error rate and power for the *RMSEA*, *CFI*, *TLI*, and the root mean squared residual *SRMR* under their respective cutoff values (.95 for *CFI* and *TLI*, .06 for *RMSEA* and .08 for *SRMR*; Hu & Bentler, 1998, 1999) will be compared to Type II error rate and power of the *C10* fit indices under the four cutoff values mentioned above.

Potential C9 – Cutoff Values

As described above, *C9* fit indices are based on a comparison of the fit of the structural null model to the fit of the target model. This quantity is then divided by a term that expresses the difference in fit of the structural null model and the saturated structural model. If the parameters estimated in a target model are significant as hypothesized, target model fit will approximate the fit of the saturated structural model. Consequently, the values in the numerator and in the denominator are expected to become approximately equal when the paths estimated in the target model are significant. As such, it is expected that *C9* indices become approximately 1.0 when a model is correctly specified. A previously conducted pilot study provided first support for this assumption. Analogous to *C10*, four different cutoff values for *C9* indices will be examined through a simulation study, and their Type II error rate and power will be determined. The cutoff values for *C9* indices examined in this study are .95, .975, .99 and 1.0.

In summary, this study will contribute to the existing literature on fit indices by providing two general frameworks for path-related fit indices that test James' et al. (1982) condition 9 and 10. Based on these frameworks, exemplary path-related fit indices will be created using various stand-alone indices and will be tested on simulated data. Several potential cutoff values will be assessed and their performance in identifying correctly specified and misspecified models will be evaluated. Additionally, the performance of *C10* fit indices will be compared to the performance of popular global fit indices. Overall, the findings from this study might provide future researchers with a means to more accurately assess model fit.

Table 1.
Overview over general frameworks and fit indices built from them using stand-alone fit indices.

Stand-alone indices	C9	C10
General structure	$\frac{F_{SN} - F_T}{F_{SN} - F_{SS}}$	$\frac{F_T - F_{SS}}{F_{SN} - F_{SS}}$
χ^2	$\frac{\chi^2_{SN} - \chi^2_T}{\chi^2_{SN} - \chi^2_{SS}}$	$\frac{\chi^2_T - \chi^2_{SS}}{\chi^2_{SN} - \chi^2_{SS}}$
χ^2/df	$\frac{\frac{\chi^2_{SN}}{df_{SN}} - \frac{\chi^2_T}{df_T}}{\frac{\chi^2_{SN}}{df_{SN}} - \frac{\chi^2_{SS}}{df_{SS}}}$	$\frac{\frac{\chi^2_T}{df_T} - \frac{\chi^2_{SS}}{df_{SS}}}{\frac{\chi^2_{SN}}{df_{SN}} - \frac{\chi^2_{SS}}{df_{SS}}}$
$\chi^2 - df$ *(NCP)*	$\frac{(\chi^2_{SN} - \chi^2_T) - (df^2_{SN} - df^2_T)}{(\chi^2_{SN} - \chi^2_{SS}) - (df^2_{SN} - df^2_{SS})}$	$\frac{(\chi^2_T - \chi^2_{SS}) - (df^2_T - df^2_{SS})}{(\chi^2_{SN} - \chi^2_{SS}) - (df^2_{SN} - df^2_{SS})}$

Note. C9 = condition 9 fit index; *C10* = condition 10 fit index; *SN* = structural null model; *T* = target model; *SS* = saturated structural model; *NCP* = noncentrality parameter

CHAPTER 2

METHOD

Population Models

To test the hypotheses, artificial data based on six different population models were created. These six models were previously simulated by Williams and Holahan (1994) and the data gained from their simulations were also the basis for Williams and O'Boyle's (2011) recent study. They were chosen for this study to allow for comparability of this study's results with results from Williams and O'Boyle's (2011) work. Furthermore, simulating meaningful theoretical models may increase the external validity of the results obtained in the simulation (Gerbing & Anderson, 1993).

Model 1a was originally presented by MacCallum (1986). As shown in Figure 1, it contains three exogenous and two endogenous latent variables, whereby each latent variable is represented by two manifest indicators. The model contains six structural parameters with values of .4 and .6. The remaining structural parameters are fixed to zero. In order to determine whether the number of indicators influences the fit index values, this model was also simulated with four indicators per latent variable (model 1b). Model 2a was taken from Mulaik et al. (1989) and slightly modified by Williams and Holahan (1994). It is presented in Figure 2 and contains four exogenous and three endogenous latent variables, whereby the exogenous latent variables are correlated at .3. Nine structural paths have values of either .4 or .6, and six additional paths are set to zero. Analogous to model 1b, model 2b is identical to model 2a with the exception that instead of two indicators, each latent variable is represented by four indicators.

Model 3 stems from Duncan, Haller and Portes (1971) and is presented in Figure 3. It contains six perfectly measured exogenous latent variables and two endogenous latent variables with two indicators each. The values of the ten estimated structural parameters range from .08 to .42, and four additional paths are fixed to zero. Finally, model 4 is a longitudinal model originally presented by Ecob (1987). As shown in Figure 4, one latent variable is perfectly measured, whereas the other latent variable is represented by three indicators. Both variables are measured at three occasions. Eight structural parameters range in value from -.14 to .78, the remaining four paths were set to zero.

The differences between models in regards to various characteristics represent one strength of the six population models examined in this study: The models contain a varying number of indicators per factor (between one and four), a varying number of latent variables (between five and eight), and a different number of omitted paths in the target model (Williams & Holahan, 1994). Given that the size of the measurement model relative to the size of the structural model may influence global fit indices' sensitivity to model misspecification, it is beneficial to examine population models with different measurement models. Furthermore, the examined models are representative of a number of variable interrelationships that are of interest to researchers, as they include a longitudinal model, a model with non-recursive variable relationships, and mediational mechanisms. This allows for generalizability of this study's results to a broad range of theoretical models.

Study Design

Nested models were fit to the data generated based on the six population models at four different sample sizes. A literature review conducted prior to the simulation indicated that the most commonly used sample sizes in simulation studies on fit indices are 100, 200, 500 and 1000; therefore, these sample sizes were used in this study as well.

Where possible, seven nested models were fit to the data generated from each of the six population models described in the previous paragraph. The first model is a structural null model (SN), where only a measurement model and the correlations among latent variables are specified, whereas all latent variable relations are set to zero. The second model (T-3) is a model with major misspecifications. In the T-3 model, a measurement model and latent variables are specified, however, three paths that are estimated in the target model are fixed to zero in this model. The third model is a model with minor misspecification (T-1) where one path estimated in the target model is set to zero. Fourth, the target model (T) is identical to the population model based on which the simulation data were generated. In the fifth model (T+1), one path whose true value in the population model equals zero is added to the target model. The sixth model (T+3) contains three additional parameters not estimated in the target model. Finally, the last model is a saturated structural model (SS) in which all structural paths are estimated.

As such, the order of models tested is SN → T-3 → T-1 → T → T+1 → T+3 → SS, going from the most restrictive model SN to the least restrictive model SS. For each model tested, 1000 replications were run. This sequence of models was tested for the population models taken from Mulaik et al. (1989) and Duncan et al. (1971). In the case of the population model of MacCallum (1986), adding a path to the target model yielded a saturated structural model. As such, models T+1 and T+3 were not tested with this population model. Similarly, in Ecob's (1987) population model, there was only one path to be added between the target model and the saturated structural model. Therefore, model T+3 was not estimated in the population model from Ecob (1987). In total, fit indices were derived from 4*7*1000 = 28,000 replications for each of the population models taken from Mulaik et al. (1989) and Duncan et al. (1971). This number is composed of four different sample sizes, seven different nested models and 1000 replications per estimation. For the model from MacCallum (1986), 4*5*1000 = 20,000 replications were run. Finally, for the Ecob (1987) model, fit indices were derived from 4*6*1000 = 24,000 replications.

Procedure

First, simulated data were generated based on variable relationships in the six population models described above. In their studies, Duncan et al. (1971), Ecob (1987) and Mulaik et al. (1989) only provided path coefficients and intercorrelations between variables, however, they did not provide covariance matrices. Therefore, a population covariance matrix for each of these three models had to be calculated first. Generally, covariance matrices can be obtained by using the following formula (Jöreskog & Sörbom, p.5):

$$\Sigma = \begin{bmatrix} \Lambda_y (I-B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I-B)^{-1}\Lambda_y' + \Theta_\varepsilon & \Lambda_y (I-B)^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x \Theta\Gamma'(I-B')^{-1}\Lambda_y' & \Lambda_x \Phi\Lambda_x' + \Theta_\delta \end{bmatrix} \tag{17}$$

Software programs such as SAS (Sas Institute, 2011) allow for a calculation of covariance matrices based on specified parameters. Once a population covariance matrix for a particular model has been calculated, the population model specifying the variable interrelationships based on which the covariance matrix was created (the target model) was fit to the covariance matrix. This provided a test of whether the model had been correctly specified and whether the parameter estimates were correctly reflected in the covariance matrix. If a χ^2–value of 0 was obtained and the parameter estimates were identical to the parameter estimates of the population model, the population covariance matrix was subjected to Cholesky decomposition. The resulting transformed matrix was then entered into PRELIS (Jöreskog & Sörbom, 1996), and for every sample size, 1000 data sets that add random error components to the population covariance matrix were created. This resulted in a total of 24,000 data sets (six population models times four different sample sizes times 1000 replications).

In the next step, the nested models described above were fit to the data sets generated in the previous step using LISREL (Jöreskog & Sörbom, 1996). The resulting fit indices for each model were saved in fit files separate from the regular LISREL output. A syntax written in SAS was then used to “harvest” a number of different fit indices and provide their means and standard deviations across the 1000 data sets for each model. These mean values then served as input for the calculation of the path-related fit indices *C9* and *C10*.

Once the *C10* and *C9* fit indices had been calculated, their values were examined. For each path-related fit index, Type II error and power were calculated under the different proposed cutoff values, (0, 0.01, 0.025, 0.05 for *C10* indices; .95, .975, .99, 1.00 for *C9* indices). That way, it was determined which cutoff value performed best in identifying correctly specified and misspecified models. Furthermore, the different path-related fit indices were compared to each other in regards to their performance, and the best-performing *C10* and *C9* indices were determined. Finally, Type II error rates and power were calculated for the three global fit indices *RMSEA*, *CFI*, *TLI,* and *SRMR* under their respective recommended cutoff values (.95 for *CFI* and *TLI*, .06 for *RMSEA,* .08 for *SRMR;* Hu & Bentler, 1998,1999). Their Type II error rates and power were then compared to Type II error and power of the path-related fit indices to determine which group of fit indices performed better in identifying correct and misspecified models.

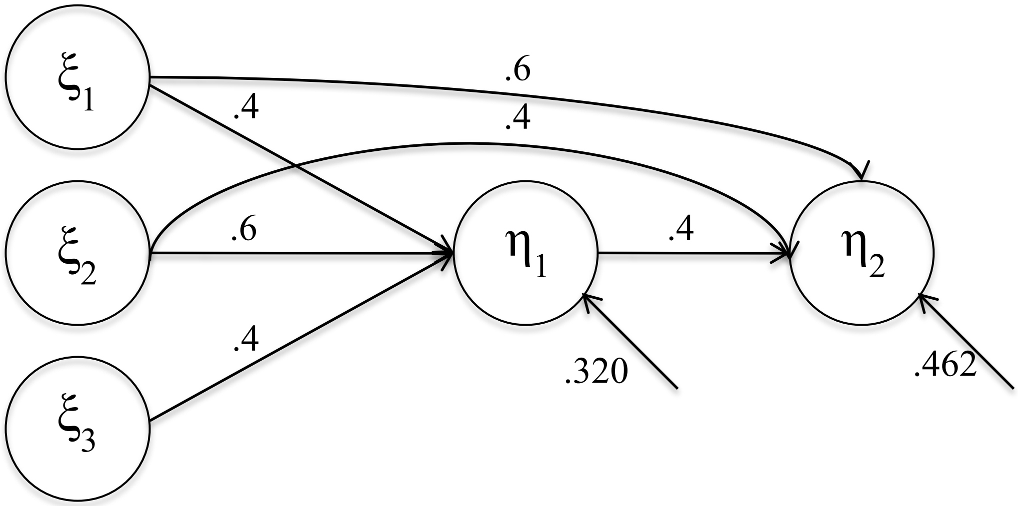

Figure 1. Population model originally presented by MacCallum (1986). Simulated both with two and four indicators per latent variable in this study.

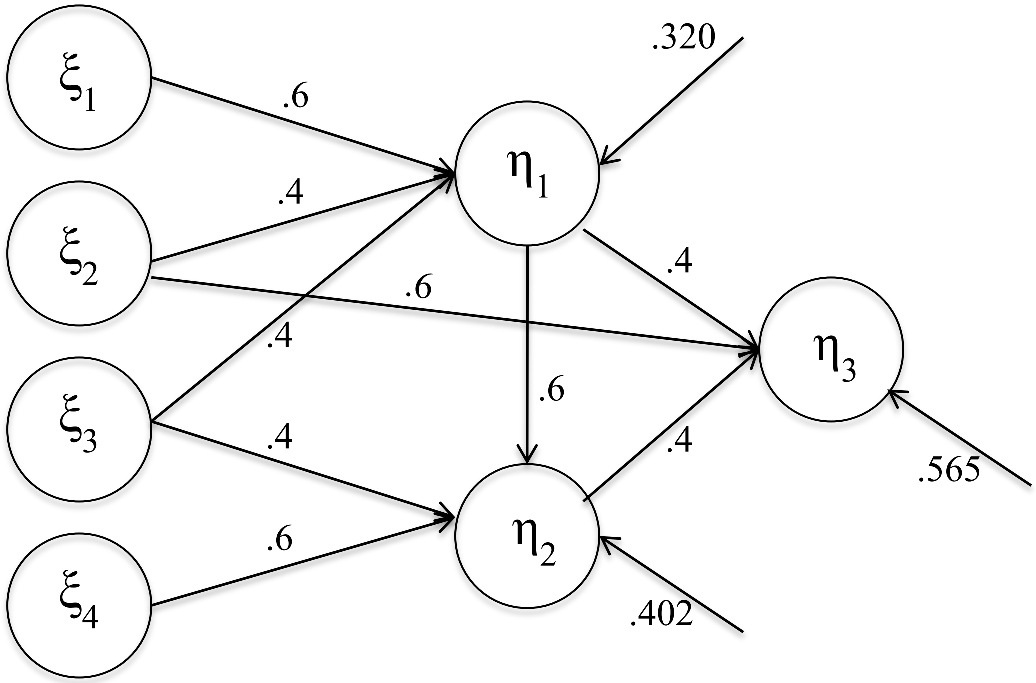

Figure 2. Population model originally presented by Mulaik et al. (1989). Simulated both with two and four indicators per latent factor in this study.

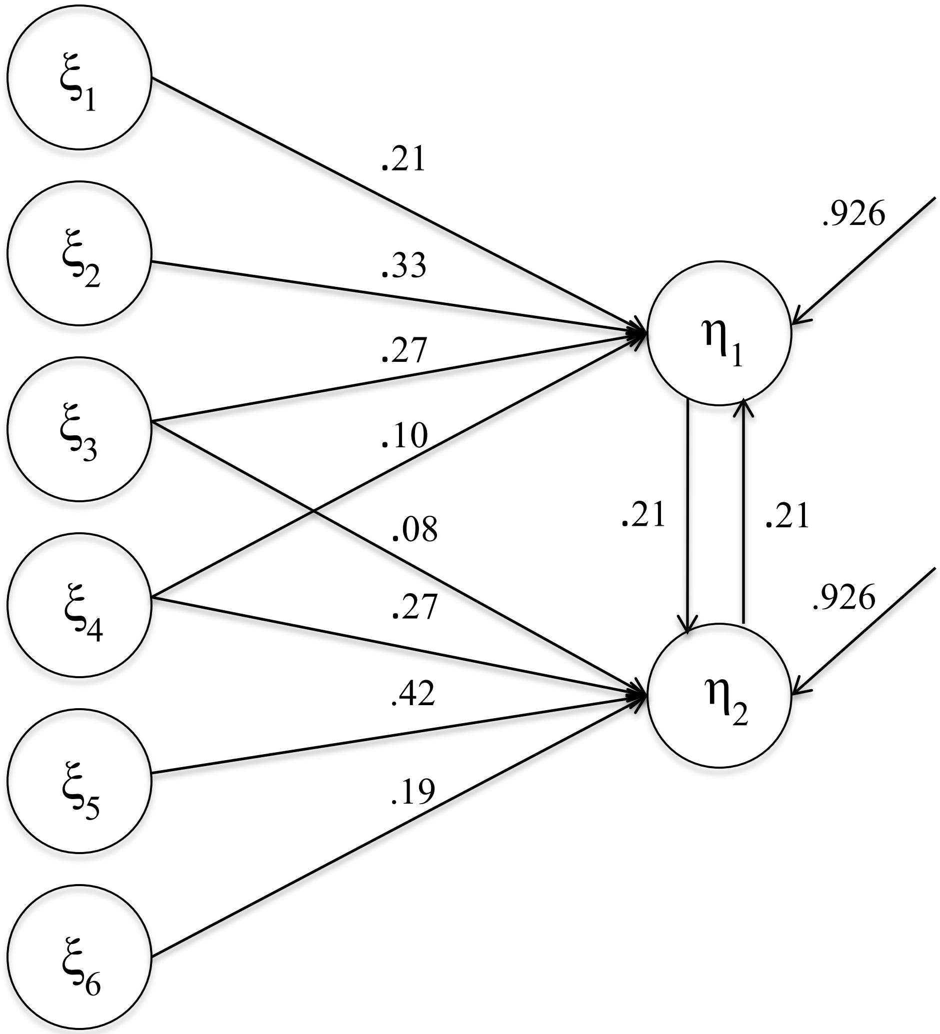

Figure 3. Population model originally presented by Duncan et al. (1971). Exogenous latent variables are assumed to be perfectly measured, each endogenous latent variable is represented by two indicators.

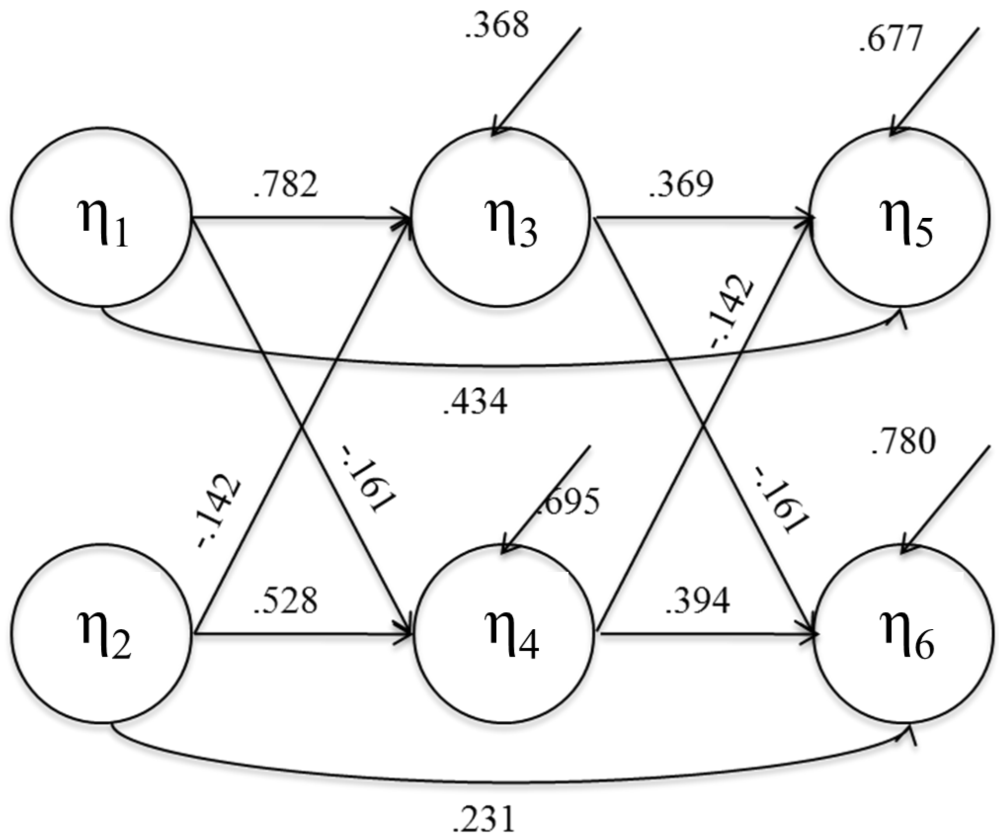

Figure 4. Population model taken from Ecob (1987). ξ_1, η_1 and η_3 are measured perfectly, ξ_2, η_2 and η_4 are each represented by three indicators.

CHAPTER 3

RESULTS

Within each of the six population models, fit indices were derived for every nested model at sample sizes 100, 200, 500, and 1000. *C10* and *C9* fit indices were calculated from each model's respective χ^2–value and degrees of freedom using the formulas displayed in Table 1. Since each model was run with 1000 replications, the displayed fit index values for a particular model represent the mean of the single fit index values yielded in these replications. For some of the models estimated, not all 1000 replications produced admissible and convergent solutions. This happened particularly frequently with severely misspecified models, such as SN or SS models as well as at low sample sizes. Fit indices from non-admissible and non-convergent model estimations were excluded from the calculation of mean fit index values for any particular model. As such, for some of the nested models that were estimated, the resulting fit indices do not represent the averaged fit indices of 1000, but a smaller number of replications. Tables 2-7 display the resulting fit indices for each population model and indicate the number of admissible and convergent solutions that were yielded by estimating the individual models.

Similar to results from a previous pilot study, *C10* fit index values were 1.000 for the worst-fitting SN model and approached zero as model fit became perfect. In the population model from Duncan et al. (1971), χ^2/df and $\chi^2–df$ fit indices for the T-1, T, T+1 and T+3 models became slightly negative at some sample sizes. In Ecob's (1987) model, χ^2/df and $\chi^2–df$ fit indices took on slightly negative values for the T and T+1 models. In Mulaik's et al. (1989) model with four indicators per factor, one χ^2/df fit index value became slightly negative for the target model at sample size 100.

C9 fit index values were zero for the SN model and approached 1.000 with improving model fit. In Duncan's et al. (1971) population model, fit indices based on χ^2/df and $\chi^2–df$ reached values slightly greater than 1.000 in the T-1, T, T+1 and T+3 models. Similarly, values slightly greater than one were obtained in Ecob's (1987) population model for fit indices based on χ^2/df and $\chi^2–df$ in the T and T+1 models. In Mulaik's et al. (1989) population model in the T model at sample size 100, a value slightly greater than one was obtained for the fit index based on χ^2/df.

The main goal of this study was to determine the most effective fit index and the cutoff value under which it performs best. Therefore, the performance of the *C10* and *C9* fit indices under various different cutoff values was determined and compared to the performance of *RMSEA*, *CFI*, *TLI*, and *SRMR* under their respective recommended cutoff values. Two criteria were used to evaluate the performance of each fit index. The first criterion was Type II error. For each fit index, it was assessed by counting the number of cases across the six population models in which a target model was rejected under a particular cutoff value.

The second criterion was power. It was further divided into two subcategories: First, P(T-1) was defined as the power of a fit index to reject slightly misspecified models with only one significant path removed (T-1 models) under a particular cutoff value. Second, P(T-3) refers to the power of a fit index to reject models with large misspecifications, that is, models with three significant paths taken out of the model (T-3 models) under a particular cutoff value. For each fit index, both subcategories of power were evaluated by determining the percentage of cases across all population models in which a misspecified model was correctly rejected under a particular cutoff value.

It must be noted that when specifying the T-3 and T-1 models, the paths with the lowest parameter estimates were removed from the target model. For example, for each T-1 model, the smallest parameter estimate out of all parameter estimates contained in the population model was selected and removed from the target model. Likewise, for each T-3 model, the three smallest parameter estimates were removed from the target model. This was done intentionally to allow for a more stringent test of the various fit indices' sensitivity to misspecification, since removing paths with small parameter estimates from the target model does not change the fit of the model as drastically as removing paths with large parameter estimates.

Table 8 displays Type II error, P(T-1), and P(T-3) for the *C10* fit indices and the global fit indices *RMSEA*, *CFI*, *TLI* and *SRMR.* Results are aggregated across all six population models. Furthermore, results are displayed both for each sample size individually and averaged across all sample sizes. In the following, the averaged results across all sample sizes will be discussed.

All four global fit indices performed well in regards to Type II error: In none of the six nested model sequences, the target model was incorrectly rejected, which equals a Type II error rate of 0%. However, the favorable Type II error rate came at the expense of power: Regardless of sample size, all T-1 models yielded fit index values above the cutoff of .95 for *CFI* and *TLI*, equal to a power rate of 0%. The other two global fit indices *SRMR* and *RMSEA* performed better, yielding a power rate of 8.25% and 25%, respectively. As such, when using *CFI* or *TLI,* in an empirical study with no knowledge of the "true" population model, one would have incorrectly selected the misspecified T-1 model as the best-fitting model. When using *SRMR* or *RMSEA*, one would have incorrectly selected T-1 in 91.75% and 75% of cases, respectively. The severely misspecified T-3 model was correctly rejected in 16.67% of cases for *CFI*, 33.33% of cases for *TLI,* 50% of cases for *RMSEA*, and 62% of cases for *SRMR*. This means that, depending on the fit index, severely misspecified models are accepted in 38% to 83.33% of cases when evaluating model fit with global fit indices.

Among the *C10* fit indices, the only fit index that performed as badly as the global fit indices was *C10-*χ^2 with a cutoff value of zero. Here, both the target model and all misspecified models were rejected in 100% of the cases, as such, both Type II error and the two power rates P (T-1) and P (T-3) equaled 100%. All other combinations of *C10* fit indices and different cutoff values performed better than the global fit indices in regards to power. Two distinct combinations of *C10* indices and particular cutoff values performed best: The fit indices based on χ^2/df and χ^2-*df* with a cutoff value of .01 both showed a Type I error rate of 0%. For both fit indices, the T-1 model was correctly rejected in 83.33% of all cases, whereas the T-3 model was correctly rejected in 100% of the cases. As such, these two combinations of *C10* indices and different cutoff values performed best among the path-related fit indices, which in turn worked much better than any of the global fit indices *SRMR*, *RMSEA*, *CFI*, and *TLI*.

In summary, by using either *C10-*χ^2/df or *C10-*χ^2*-df* with a cutoff value of .01, one is able to considerably increase the likelihood of rejecting misspecified models compared to global fit indices.

Type II error rates, P (T-1), and P (T-3) for *C9* fit indices are displayed in Table 9. The following combinations of fit indices and cutoff values performed best in identifying the target model and misspecified models: Both *C9-*χ^2/df and *C9-*χ^2*-df* (*NSCI-P*; Williams & Holahan, 1994) with a cutoff value of .01 yielded a Type II error of 0%, a P (T-1) rate of 83%, and a P (T-3) rate of 100%. As such, these fit indices represent a reliable omnibus test of James' et al. (1982) condition 9, that is, whether the structural paths specified in a model are significant.

Table 2.

Results from MacCallum (1986), model with two indicators per factor.

				C10 fit indices			C9 fit indices			Global fit indices				
Model	*N*	*χ2*	*df*	*χ2*	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
SN	100	323.170	32	1.000	1.000	1.000	.000	.000	.000	.398	.252	.763	.667	981
	200	612.442	32	1.000	1.000	1.000	.000	.000	.000	.396	.255	.768	.673	1000
	500	1483.440	32	1.000	1.000	1.000	.000	.000	.000	.396	.257	.770	.677	1000
	1000	2937.300	32	1.000	1.000	1.000	.000	.000	.000	.395	.258	.771	.678	1000
T-3	100	111.988	29	.290	.312	.283	.710	.688	.717	.128	.150	.931	.893	870
	200	197.091	29	.292	.319	.289	.708	.681	.711	.124	.155	.932	.895	926
	500	450.362	29	.292	.320	.290	.708	.680	.710	.120	.158	.933	.896	989
	1000	873.243	29	.291	.321	.291	.709	.679	.709	.120	.158	.933	.897	1000
T-1	100	61.870	27	.121	.139	.117	.879	.861	.883	.070	.099	.971	.952	996
	200	96.501	27	.121	.141	.119	.879	.859	.881	.067	.103	.972	.953	1000
	500	200.746	27	.121	.142	.120	.879	.858	.880	.065	.105	.972	.954	1000
	1000	375.837	27	.120	.142	.120	.880	.858	.880	.064	.105	.972	.954	1000

(table continued)

Table 2 continued

Model	*N*	*χ2*	*df*	*C10 fit indices*			*C9* fit indices			Global fit indices				Rep.
				χ2	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	
T	100	26.947	26	.004	.000	.000	.996	1.000	1.000	.035	.020	.997	.999	1000
	200	26.378	26	.002	.000	.000	.998	1.000	1.000	.024	.014	.999	1.000	1000
	500	25.877	26	.001	.000	.000	.999	1.000	1.000	.015	.009	1.000	1.000	1000
	1000	25.987	26	.000	.000	.000	1.000	1.000	1.000	.011	.006	1.000	1.000	1000
SS	100	25.887	25	.000	.000	.000	1.000	1.000	1.000	.034	.020	.997	.999	999
	200	25.389	25	.000	.000	.000	1.000	1.000	1.000	.024	.014	.999	1.000	1000
	500	24.908	25	.000	.000	.000	1.000	1.000	1.000	.015	.009	1.000	1.000	1000
	1000	25.054	25	.000	.000	.000	1.000	1.000	1.000	.011	.007	1.000	1.000	1000

Note. N = sample size; df = degrees of freedom; *SRMR* = standardized root mean squared residual; *RMSEA* = root mean squared error of approximation; *CFI* = comparative fit index; *TLI* = Tucker-Lewis-Index; Rep. = number of replications that produced admissible and convergent solutions (out of 1000)

Table 3.
Results from MacCallum (1986), model with four indicators per factor.

				C10 fit indices			C9 fit indices			Global fit indices				
Model	*N*	*χ2*	*df*	*χ2*	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
SN	100	471.122	167	1.000	1.000	1.000	.000	.000	.000	.373	.107	.934	.925	1000
	200	753.511	167	1.000	1.000	1.000	.000	.000	.000	.371	.110	.938	.930	1000
	500	1620.810	167	1.000	1.000	1.000	.000	.000	.000	.370	.112	.939	.931	1000
	1000	3077.380	167	1.000	1.000	1.000	.000	.000	.000	.370	.113	.939	.931	1000
T-3	100	267.618	164	.314	.313	.308	.686	.687	.692	.144	.062	.978	.974	1000
	200	349.506	164	.312	.314	.309	.688	.686	.691	.140	.064	.980	.977	1000
	500	612.941	164	.309	.314	.308	.691	.686	.692	.138	.065	.981	.978	1000
	1000	1055.790	164	.307	.312	.306	.693	.688	.694	.136	.065	.981	.979	1000
T-1	100	212.187	162	.127	.127	.124	.873	.873	.876	.080	.039	.989	.987	1000
	200	240.375	162	.126	.128	.124	.874	.872	.876	.072	.042	.992	.990	1000
	500	343.718	162	.125	.128	.124	.875	.872	.876	.067	.043	.992	.991	1000
	1000	520.115	162	.123	.127	.123	.877	.873	.877	.065	.043	.993	.991	1000

(table continued)

Table 3 continued

Model	*N*	*χ2*	*df*	*C10 fit indices* *χ2*	*χ2/df*	*χ2-df*	*C9* fit indices *χ2*	*χ2/df*	*χ2-df*	Global fit indices *SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
T	100	175.532	161	.004	.000	.000	.996	1.000	1.000	.049	.013	.996	.996	1000
	200	167.688	161	.002	.000	.000	.998	1.000	1.000	.035	.009	.999	.999	1000
	500	162.667	161	.001	.000	.000	.999	1.000	1.000	.022	.006	1.000	1.000	1000
	1000	161.901	161	.000	.000	.000	1.000	1.000	1.000	.015	.004	1.000	1.000	1000
SS	100	174.404	160	.000	.000	.000	1.000	1.000	1.000	.049	.013	.996	.996	972
	200	166.567	160	.000	.000	.000	1.000	1.000	1.000	.034	.009	.999	.999	986
	500	161.735	160	.000	.000	.000	1.000	1.000	1.000	.021	.006	1.000	1.000	991
	1000	160.917	160	.000	.000	.000	1.000	1.000	1.000	.015	.004	1.000	1.000	997

Note. *N* = sample size; df = degrees of freedom; *SRMR* = standardized root mean squared residual; *RMSEA* = root mean squared error of approximation; *CFI* = comparative fit index; *TLI* = Tucker-Lewis-Index; Rep. = number of replications that produced admissible and convergent solutions (out of 1000)

Table 4.
Results from Mulaik et al. (1989), model with two indicators per factor

				C10 fit indices			C9 fit indices			Global fit indices				
Model	*N*	*χ2*	*df*	*χ2*	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
SN	100	489.737	71	1.000	1.000	1.000	.000	.000	.000	.418	.224	.824	.775	974
	200	918.458	71	1.000	1.000	1.000	.000	.000	.000	.417	.229	.825	.776	985
	500	2214.920	71	1.000	1.000	1.000	.000	.000	.000	.417	.233	.825	.776	963
	1000	4372.020	71	1.000	1.000	1.000	.000	.000	.000	.417	.234	.825	.776	956
T-3	100	171.670	65	.281	.297	.270	.719	.703	.730	.166	.106	.955	.937	949
	200	287.650	65	.277	.298	.271	.723	.702	.729	.164	.113	.954	.935	986
	500	642.952	65	.276	.299	.273	.724	.701	.727	.164	.117	.953	.934	998
	1000	1234.150	65	.275	.299	.274	.725	.701	.726	.164	.118	.952	.933	1000
T-1	100	88.764	63	.094	.094	.081	.906	.906	.919	.072	.047	.989	.984	1000
	200	121.288	63	.086	.091	.080	.914	.909	.920	.067	.060	.988	.983	1000
	500	225.172	63	.083	.091	.080	.917	.909	.920	.064	.066	.987	.981	1000
	1000	397.087	63	.081	.091	.080	.919	.909	.920	.063	.068	.986	.980	1000

(table continued)

Table 4 continued

Model	*N*	*χ2*	*df*	*C10 fit indices* *χ2*	*χ2/df*	*χ2-df*	*C9* fit indices *χ2*	*χ2/df*	*χ2-df*	Global fit indices *SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
T	100	53.646	62	.015	.004	.001	.985	.996	.999	.038	.004	.999	1.005	1000
	200	52.015	62	.007	.002	.000	.993	.998	1.000	.027	.003	1.000	1.003	1000
	500	51.122	62	.003	.001	.000	.997	.999	1.000	.017	.002	1.000	1.001	1000
	1000	50.792	62	.001	.000	.000	.999	1.000	1.000	.012	.001	1.000	1.001	1000
T+1	100	52.608	61	.012	.004	.001	.988	.996	.999	.037	.004	.999	1.005	1000
	200	51.044	61	.006	.002	.000	.994	.998	1.000	.026	.003	1.000	1.003	1000
	500	50.171	61	.002	.001	.000	.998	.999	1.000	.016	.002	1.000	1.001	1000
	1000	49.840	61	.001	.000	.000	.999	1.000	1.000	.011	.001	1.000	1.001	1000
T+3	100	50.435	59	.008	.002	.001	.992	.998	.999	.036	.003	.999	1.006	1000
	200	48.950	59	.004	.001	.000	.996	.999	1.000	.025	.002	1.000	1.003	1000
	500	48.054	59	.001	.000	.000	.999	1.000	1.000	.016	.002	1.000	1.001	1000
	1000	48.054	59	.001	.000	.000	.999	1.000	1.000	.016	.002	1.000	1.001	1000
SS	100	47.076	56	.000	.000	.000	1.000	1.000	1.000	.034	.003	.999	1.006	1000
	200	45.821	56	.000	.000	.000	1.000	1.000	1.000	.024	.002	1.000	1.003	1000
	500	45.154	56	.000	.000	.000	1.000	1.000	1.000	.015	.002	1.000	1.001	1000
	1000	44.869	56	.000	.000	.000	1.000	1.000	1.000	.011	.001	1.000	1.001	1000

Note. *N* = sample size; df = degrees of freedom; *SRMR* = standardized root mean squared residual; *RMSEA* = root mean squared error of approximation; *CFI* = comparative fit index; *TLI* = Tucker-Lewis-Index; Rep. = number of replications that produced admissible and convergent solutions (out of 1000)

Table 5.

Results from Mulaik et al. (1989), model with four indicators per factor

				C10 fit indices			C9 fit indices			Global fit indices				
Model	*N*	*χ2*	*df*	*χ2*	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
SN	100	853.305	344	1.000	1.000	1.000	.000	.000	.000	.388	.103	.941	.935	1000
	200	1301.250	344	1.000	1.000	1.000	.000	.000	.000	.386	.106	.946	.940	1000
	500	2700.360	344	1.000	1.000	1.000	.000	.000	.000	.385	.108	.947	.942	1000
	1000	5049.930	344	1.000	1.000	1.000	.000	.000	.000	.384	.108	.947	.942	1000
T-3	100	451.035	338	.164	.150	.150	.836	.850	.850	.093	.040	.987	.985	1000
	200	497.988	338	.156	.151	.149	.844	.849	.851	.085	.042	.991	.990	1000
	500	694.369	338	.151	.151	.148	.849	.849	.852	.080	.042	.992	.991	1000
	1000	1040.810	338	.150	.151	.149	.850	.849	.851	.079	.042	.992	.991	1000
T-1	100	415.406	336	.090	.078	.077	.910	.922	.923	.081	.026	.991	.990	1000
	200	427.741	336	.082	.078	.076	.918	.922	.924	.071	.029	.995	.994	1000
	500	522.826	336	.079	.078	.076	.921	.922	.924	.065	.030	.996	.995	1000
	1000	698.511	336	.078	.078	.076	.922	.922	.924	.063	.030	.996	.995	1000
T	100	378.809	335	.014	-.001	.001	.986	1.001	.999	.052	.012	.995	.994	1000
	200	355.461	335	.006	.000	.000	.994	1.000	1.000	.037	.008	.999	.999	1000
	500	343.194	335	.003	.000	.000	.997	1.000	1.000	.023	.005	1.000	1.000	1000
	1000	338.875	335	.001	.000	.000	.999	1.000	1.000	.016	.004	1.000	1.000	1000

(table continued)

Table 5 continued

Model	*N*	*χ2*	*df*	*C10 fit indices* *χ2*	*χ2/df*	*χ2-df*	*C9* fit indices *χ2*	*χ2/df*	*χ2-df*	Global fit indices *SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
T+1	100	377.758	334	.011	.000	.001	.989	1.000	.999	.052	.012	.995	.994	1000
	200	354.463	334	.005	.000	.000	.995	1.000	1.000	.036	.008	.999	.999	1000
	500	342.147	334	.002	.000	.000	.998	1.000	1.000	.023	.005	1.000	1.000	1000
	1000	337.849	334	.001	.000	.000	.999	1.000	1.000	.016	.004	1.000	1.000	1000
T+3	100	375.540	332	.007	.000	.000	.993	1.000	1.000	.051	.012	.995	.994	1000
	200	352.320	332	.003	.000	.000	.997	1.000	1.000	.036	.008	.999	.999	1000
	500	340.131	332	.001	.000	.000	.999	1.000	1.000	.022	.005	1.000	1.000	1000
	1000	335.831	332	.001	.000	.000	.999	1.000	1.000	.016	.004	1.000	1.000	1000
SS	100	372.313	329	.000	.000	.000	1.000	1.000	1.000	.050	.012	.995	.994	1000
	200	349.305	329	.000	.000	.000	1.000	1.000	1.000	.035	.008	.999	.999	1000
	500	337.158	329	.000	.000	.000	1.000	1.000	1.000	.022	.005	1.000	1.000	1000
	1000	332.872	329	.000	.000	.000	1.000	1.000	1.000	.016	.004	1.000	1.000	1000

Note. *N* = sample size; df = degrees of freedom; *SRMR* = standardized root mean squared residual; *RMSEA* = root mean squared error of approximation; *CFI* = comparative fit index; *TLI* = Tucker-Lewis-Index; Rep. = number of replications that produced admissible and convergent solutions (out of 1000)

Table 6.

Results from Duncan et al. (1971)

				C10 fit indices			C9 fit indices			Global fit indices				
Model	*N*	*χ2*	*df*	*χ2*	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
SN	100	125.928	27	1.000	1.000	1.000	.000	.000	.000	.233	.179	.786	.644	985
	200	222.983	27	1.000	1.000	1.000	.000	.000	.000	.229	.183	.793	.655	988
	500	515.220	27	1.000	1.000	1.000	.000	.000	.000	.228	.185	.796	.661	992
	1000	1002.620	27	1.000	1.000	1.000	.000	.000	.000	.227	.185	.797	.662	1000
T-3	100	26.466	20	.105	.054	.049	.895	.946	.951	.052	.044	.984	.968	1000
	200	31.361	20	.082	.065	.052	.918	.935	.948	.044	.046	.988	.973	1000
	500	46.675	20	.066	.070	.053	.934	.930	.947	.037	.049	.989	.975	1000
	1000	73.174	20	.060	.072	.054	.940	.928	.946	.035	.050	.989	.975	1000
T-1	100	19.474	18	.043	-.015	-.003	.957	1.015	1.003	.037	.025	.993	.992	1000
	200	19.536	18	.025	-.002	.001	.975	1.002	.999	.027	.019	.996	.996	1000
	500	20.426	18	.013	.004	.003	.987	.996	.997	.018	.015	.998	.997	1000
	1000	22.958	18	.009	.006	.004	.991	.994	.996	.014	.014	.999	.997	1000
T	100	17.971	17	.029	-.022	-.008	.971	1.022	1.008	.033	.024	.993	.994	1000
	200	17.611	17	.016	-.009	-.004	.984	1.009	1.004	.024	.017	.997	.998	1000
	500	16.965	17	.006	-.004	-.002	.994	1.004	1.002	.015	.010	.999	1.000	1000
	1000	17.078	17	.003	-.002	-.001	.997	1.002	1.001	.010	.007	.999	1.000	1000

(table continued)

Table 6 continued

Model	*N*	*χ2*	*df*	*C10 fit indices* *χ2*	*χ2/df*	*χ2-df*	*C9* fit indices *χ2*	*χ2/df*	*χ2-df*	Global fit indices *SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
T+1	100	17.035	16	.021	-.020	-.007	.979	1.020	1.007	.031	.024	.993	.994	1000
	200	16.592	16	.011	-.009	-.004	.989	1.009	1.004	.022	.017	.997	.998	853
	500	15.982	16	.004	-.004	-.002	.996	1.004	1.002	.013	.010	.999	1.000	1000
	1000	16.057	16	.002	-.002	-.001	.998	1.002	1.001	.010	.007	.999	1.000	1000
T+3	100	14.738	14	.000	-.023	-.010	1.000	1.023	1.010	.024	.024	.994	.995	1000
	200	14.296	14	.000	-.011	-.005	1.000	1.011	1.005	.017	.017	.998	.999	754
	500	13.846	14	.000	-.004	-.002	1.000	1.004	1.002	.011	.010	.999	1.000	1000
	1000	14.008	14	.000	-.002	-.001	1.000	1.002	1.001	.008	.007	1.000	1.000	1000
SS	100	14.745	13	.000	.000	.000	1.000	1.000	1.000	.024	.029	.993	.987	959
	200	14.312	13	.000	.000	.000	1.000	1.000	1.000	.017	.021	.997	.995	971
	500	13.827	13	.000	.000	.000	1.000	1.000	1.000	.011	.013	.999	.999	972
	1000	13.995	13	.000	.000	.000	1.000	1.000	1.000	.007	.009	.999	.999	983

Note. N = sample size; df = degrees of freedom; *SRMR* = standardized root mean squared residual; *RMSEA* = root mean squared error of approximation; *CFI* = comparative fit index; *TLI* = Tucker-Lewis-Index; Rep. = number of replications that produced admissible and convergent solutions (out of 1000)

Table 7.

Results from Ecob (1987).

Model	*N*	*χ2*	*df*	C10 fit indices			C9 fit indices			Global fit indices				Rep.
				χ2	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	
SN	100	176.080	55	1.000	1.000	1.000	.000	.000	.000	.220	.152	.849	.819	379
	200	293.783	55	1.000	1.000	1.000	.000	.000	.000	.214	.154	.853	.823	388
	500	674.592	55	1.000	1.000	1.000	.000	.000	.000	.218	.159	.855	.826	691
	1000	1312.890	55	1.000	1.000	1.000	.000	.000	.000	.220	.161	.855	.826	892
T-3	100	55.920	48	.082	.052	.048	.918	.948	.952	.070	.027	.989	.987	631
	200	63.232	48	.072	.062	.055	.928	.938	.945	.053	.034	.991	.993	258
	500	86.796	48	.068	.070	.062	.932	.930	.938	.045	.038	.991	.994	265
	1000	123.322	48	.063	.068	.059	.937	.932	.941	.045	.005	.002	.993	226
T-1	100	50.024	46	.036	.016	.015	.964	.984	.985	.063	.021	.992	.993	526
	200	51.905	46	.027	.018	.016	.973	.982	.984	.048	.020	.995	.995	658
	500	56.048	46	.020	.018	.015	.980	.982	.985	.037	.018	.997	.997	768
	1000	65.866	46	.017	.018	.015	.983	.982	.985	.032	.019	.998	.997	853
T	100	47.303	45	.016	-.001	.000	.984	1.001	1.000	.054	.017	.994	.996	717
	200	46.826	45	.006	-.003	-.002	.994	1.003	1.002	.038	.014	.997	.998	850
	500	45.625	45	.003	.000	.000	.997	1.000	1.000	.024	.008	.999	1.000	966
	1000	45.737	45	.002	.000	.000	.998	1.000	1.000	.017	.006	1.000	1.000	992

(table continued)

Table 7 continued

				C10 fit indices			*C9* fit indices			Global fit indices				
Model	*N*	*χ2*	*df*	*χ2*	*χ2/df*	*χ2-df*	*χ2*	*χ2/df*	*χ2-df*	*SRMR*	*RMSEA*	*CFI*	*TLI*	Rep.
T+1	100	46.522	44	.010	.002	.002	.990	.998	.998	.053	.018	.994	.995	664
	200	45.996	44	.003	-.001	-.001	.997	1.001	1.001	.037	.014	.997	.998	784
	500	44.727	44	.002	.000	.000	.998	1.000	1.000	.023	.008	.999	1.000	935
	1000	44.709	44	.001	.000	.000	.999	1.000	1.000	.016	.006	1.000	1.000	987
SS	100	45.249	43	.000	.000	.000	1.000	1.000	1.000	.052	.012	.994	.996	608
	200	45.214	43	.000	.000	.000	1.000	1.000	1.000	.037	.015	.997	.998	755
	500	43.704	43	.000	.000	.000	1.000	1.000	1.000	.023	.008	.999	1.000	893
	1000	43.691	43	.000	.000	.000	1.000	1.000	1.000	.016	.006	1.000	1.000	951

Note. N = sample size; df = degrees of freedom; *SRMR* = standardized root mean squared residual; *RMSEA* = root mean squared error of approximation; *CFI* = comparative fit index; *TLI* = Tucker-Lewis-Index; Rep. = number of replications that produced admissible and convergent solutions (out of 1000)

Table 8.

Performance of C10 fit indices under different cutoff values, SRMR, RMSEA, CFI, and TLI averaged across all six population models.

		Cutoff values															
		$\chi 2$				$\chi 2/df$				$\chi 2$-*df*				*SRMR*	*RMSEA*	*CFI*	*TLI*
N	Crit.	0	.01	.025	.05	0	.01	.025	.05	0	.01	.025	.05	.08	.06	.95	.95
	β	100	66.67	16.67	0	50	0	0	0	83.33	0	0	0	0	0	0	0
100	*P(T-1)*	100	100	100	66.67	83.33	83.33	66.67	66.67	83.33	83.33	66.67	66.67	33.33	16.67	0	0
	P(T-3)	100	100	100	100	100	100	100	100	100	100	100	66.67	66.67	50	16.67	33.33
	β	100	16.67	0	0	33.33	0	0	0	50	0	0	0	0	0	0	0
200	*P(T-1)*	100	100	83.33	66.67	83.33	83.33	66.67	66.67	100	83.33	66.67	66.67	0	16.67	0	0
	P(T-3)	100	100	100	100	100	100	100	100	100	100	100	100	66.67	50	16.67	33.33
	β	100	0	0	0	16.67	0	0	0	16.67	0	0	0	0	0	0	0
500	*P(T-1)*	100	100	66.67	66.67	100	83.33	66.67	66.67	100	83.33	66.67	66.67	0	33.33	0	0
	P(T-3)	100	100	100	100	100	100	100	100	100	100	100	100	66.67	50	16.67	33.33
	β	100	0	0	0	33.33	0	0	0	33.33	0	0	0	0	0	0	0
1000	*P(T-1)*	100	83.33	66.67	66.67	100	83	67	67	100	83	67	67	0	33.33	0	0
	P(T-3)	100	100	100	100	100	100	100	100	100	100	100	100	50	50	16.67	33.33
	β	100	20.83	4.17	0	33.33	0	0	0	45.83	0	0	0	0	0	0	0
Avg.	*P(T-1)*	100	95.83	79.17	66.67	91.67	83.33	66.67	66.67	95.83	83.33	66.67	66.67	8.25	25	0	0
	P(T-3)	100	100	100	100	100	100	100	100	100	100	100	91.67	62.50	50	16.67	33.33

Note. *N* = Sample size; Avg. = Average; Crit. = Criterion, by which fit indices are evaluated; β = Type II error, i.e. the percentage of cases in which a correct model is incorrectly rejected in favor of a more or less restrictive model under a particular cutoff value; *P(T-1)* = Power of a fit index to detect small misspecifications, i.e. percentage of cases in which a model with small misspecifications (one path missing) is correctly rejected under a particular cutoff value; *P(T-3)* = Power of a fit index to detect large misspecifications, i.e.

percentage of cases in which a model with large misspecifications (three paths missing) is correctly rejected; Average = average of β, *P (T-1), P (T-3)* across all sample sizes; *SRMR* = standardized root mean squared residual; *RMSEA* = root mean squared error of approximation; *CFI* = comparative fit index; *TLI*= Tucker-Lewis- Index; underlined = best –performing fit indices

Table 9.

Performance of C9 fit indices under different cutoff values averaged across all six population models.

		Cutoff values											
		χ2				*χ2/df*				*χ2-df*			
N	Crit.	.95	.975	.99	1.00	.95	.975	.99	1.00	.95	.975	.99	1.00
100	β	0	16.67	66.67	100	0	0	0	50	0	0	0	83.33
	P (T-1)	66.67	100	100	100	66.67	66.67	83.33	83.33	66.67	66.67	83.33	83.33
	P (T-3)	100	100	100	100	100	100	100	100	66.67	100	100	100
200	β	0	0	16.67	100	0	0	0	33.33	0	0	0	50
	P (T-1)	66.67	83.33	100	100	66.67	66.67	83.33	83.33	66.67	66.67	83.33	100
	P (T-3)	100	100	100	100	100	100	100	100	100	100	100	100
500	β	0	0	0	100	0	0	0	16.67	0	0	0	16.67
	P (T-1)	66.67	66.67	100	100	66.67	66.67	83.33	100.00	66.67	66.67	83.33	100
	P (T-3)	100	100	100	100	100	100	100	100	100	100	100	100
1000	β	0	0	0	100	0	0	0	33.33	0	0	0	33.33
	P (T-1)	66.67	66.67	83.33	100	66.67	66.67	83.33	100	66.67	66.67	83.33	100
	P (T-3)	100	100	100	100	100	100	100	100	100	100	100	100
Avg.	β	0	4.17	20.83	100	0	0	<u>0</u>	33.33	0	0	<u>0</u>	45.83
	P (T-1)	66.67	79.17	95.83	100	66.67	66.67	<u>83.33</u>	91.67	66.67	66.67	<u>83.33</u>	95.83
	P (T-3)	100	100	100	100	100	100	<u>100</u>	100	91.67	100	<u>100</u>	100

Note. N = Sample Size; Avg. = Average across sample sizes; Crit. = Criterion, by which fit indices are evaluated; β = Type II error, i.e. the percentage of cases in which a correct model is incorrectly rejected in favor of a more or less restrictive model under a particular cutoff value; *P (T-1)* = Power of a fit index to detect small misspecifications, i.e. percentage of cases in which a model with small misspecifications (one path missing) is correctly rejected under a particular cutoff value; *P (T-3)* = Power of a fit index to detect large misspecifications, i.e. percentage of cases in which a model with large misspecifications (three paths missing) is correctly rejected a; Average = Average of β, *P (T-1), P (T-3)* across all sample sizes ; underlined = best –performing fit indices

CHAPTER 4

DISCUSSION

This study had three primary goals: The first goal was to develop two general frameworks for path-related fit indices that evaluate James' et al. (1982) condition 9 and condition 10. The second goal was to create exemplary fit indices based on these two frameworks by incorporating three common standalone fit indices and to test their performance under various different cutoff values, so that the best-performing combination of each fit index and a particular cutoff value could be determined. Finally, the third goal was to compare the performance of the path-related fit indices to popular global fit indices.

The results from the simulation study showed that all *C10* and *C9* fit indices performed very well in accepting the correctly specified target model and rejecting models with both small and severe misspecifications. The optimal cutoff value differed for fit indices based on χ^2 on the one hand and on χ^2/df and χ^2-df on the other hand. The fit indices based on χ^2/df and χ^2-df displayed the highest power rates for the slightly misspecified T-1 models and showed therefore the best performance. As described above, it is worth mentioning that only the paths with the smallest parameter estimates were deleted from the target model to create T-1- and T-3- models. For example, in Ecob's model, a path with an estimate of -.142 was removed from the target model to create the T-1 model, while there were path estimates as large as .782 in the model. The fact that power was still high despite the small differences between the target and the misspecified models is proof of the accuracy of path-related fit indices.

Based on the results of this study, *C10* indices reliably indicate whether any significant paths have been erroneously set to zero in a model, whereas *C9* indices reliably indicate whether the specified paths in a model are significant.

As predicted, the comparison between path-related and global fit indices showed that all path-related fit indices were considerably more accurate than the global fit indices *SRMR*, *RMSEA*, *CFI*, and *TLI*. Using *CFI* and *TLI* and their recommended cutoff values, none of the six T-1 models with small misspecifications at any of the four sample sizes would have been rejected, and only a small number of misspecified models were identified using *RMSEA* and *SRMR*. In only one of the six population models, *CFI* was able to detect the severely misspecified T-3 model, whereas the other five T-3 models would have been erroneously accepted. Using *TLI*, the T-3 model was correctly rejected in two out of the six population models. *RMSEA* identified the misspecified T-3 model in three of the population models. Finally, across all sample sizes, *SRMR* performed best among the global fit indices, since four of the six T-3 models were correctly rejected. As such, the global fit indices were shown to be too lenient in regards to model misspecification. Given that at equal fit, the more parsimonious model is preferable (Mulaik et al., 1989), one would be likely to select a model with important significant paths left out over a correctly specified model when evaluating theoretical models against empirical data.

The findings of this study imply that in the majority of cases, the perfect fit of the measurement model masked the misspecifications in the structural model. This is particularly troubling because the models evaluated in this study had rather small measurement models with relatively few parameters estimated: Duncan's et al. (1971) model had only 1.25 indicators per factor. Ecob's (1987) model had two indicators per factor, and MacCallum's (1986) and Mulaik's et al. (1989) models were each specified with two and four indicators per factor. Yet despite the small number of paths estimated in the measurement model, it still masked the bad fit of the misspecified models.

Increasing the number of manifest variables in one's models has been shown to reduce the number of non-converged and improper solutions, and to yield more accurate and stable parameter estimates as well as more reliable factors (Marsh, Hau, Balla, & Grayson, 1998). As such, researchers may aim to use models with more indicators than in the current simulation study. The measurement models may therefore be larger in empirical research than in the current simulation study. This means that the problem of measurement model fit masking structural model fit may be even more pronounced in models assessed against empirical data.

Overall, these findings are in line with the concerns many researchers have previously voiced about evaluating structural relations with global fit indices (e.g. McDonald & Ho, 2002, Mulaik et al., 1989; O'Boyle and Williams, 2011; Rigdon, 1996; Sobel & Bohrnstedt, 1985). They demonstrate that global fit indices may not represent adequate tools for helping researchers decide whether to accept or reject a theoretical model. It could also be shown that path-related fit indices based on the frameworks introduced in this study provide a viable alternative to global fit indices and are much more accurate in detecting model misspecifications.

Limitations and Directions for Future Research

While this study provides researchers with an alternative to evaluating model fit with global fit indices, there were some limitations to this research. As mentioned above, four out of the six population models examined in this study had measurement models with only two or fewer indicators. Since models tested against empirical data usually incorporate more indicators, future research should simulate models with a higher number of indicators. That way, it could be determined whether the differences between path-related and global fit indices in regards to power are even greater than in the models used for this study. Furthermore, in this study only three different standalone fit indices were incorporated into the *C10* and *C9* frameworks. Since the ideal cutoff values differed for two of the six *C10* and *C9* fit indices, it appears that there does not exist one single best-performing cutoff value that can be applied to all path-related fit indices. Future research should therefore evaluate fit indices created from the general frameworks but using additional standalone indices such as the *SRMR* or the A*GFI* and determine their respective cutoff values as well as their performance.

Another limitation of this study is that the models were simulated under ideal circumstances, that is, all variables were assumed to be perfectly normally distributed and no missing data was incorporated into the simulation. Future studies should introduce perturbations in order to provide for more realistic scenarios that may be better applicable to empirical data.

A relatively minor limitation of this study is that in the evaluation of the global fit indices, the formula to calculate *CFI* used by LISREL 8.70 and newer versions might lead to heavily inflated *CFI* values (G. Cheung, personal conversation, August 23, 2013). In this study, LISREL 8.80 was used, which might have caused inflated *CFI* values, so that *CFI* displayed the worst power to reject misspecified models among all fit indices examined. When interested in *CFI* values, future researchers might therefore consider estimating models with other software programs that use different baseline models, such as for example MPlus (Muthén & Muthén, 2011). Finally, for the simulations in this study, the variables were specified as normally distributed. Future research should examine the performance of the path-related fit indices proposed in this study when variables do not follow a normal distribution.

Implications

The findings of this study have important implications for researchers, editors and practitioners. First, they clearly support McDonald and Ho's (2002) and O'Boyle and William's (2011) recommendation for researchers to not rely only on global fit indices to assess structural relationships, but to use path-related fit indices that assess fit of the structural model. Both McDonald and Ho's (2002) and O'Boyle and Williams' (2011) analyses of published SEM models showed that a large majority of models might be misspecified because the authors assessed model fit with global fit indices. Researchers should therefore use caution when building theories based on previously published research where SEM was used, since there is a realistic chance that published theoretical models may be flawed. As such, researchers might consider re-examining and calculating structural model fit of published models if they aim to derive hypotheses and build their own research on such models. If χ^2-values and *df* are provided for the structural model, path-related fit indices can be quickly and easily calculated.

Editors of scientific journals should be aware of the issues around assessing model fit with global fit indices. They should require every author to assess fit of the structural model and accept papers only if authors provide sufficient proof that the hypothesized structural relationships are significant, and that the relationships that are not estimated are in fact zero. At the minimum, authors should be urged to provide χ^2-values and *df* for the measurement model and the structural model separately, so that other researchers can calculate path-related fit indices on their own.

Finally, practitioners should also closely examine structural model fit if they aim to design interventions based on variable relationships established through SEM. As a fictitious example, an I/O- psychologist might have found in the literature that there is a strong relationship between leader behaviors and employee turnover intentions. Having an issue with employee turnover in his organization, he might decide to conduct an expensive large-scale leadership training in his organization. However, if global fit indices were used in the studies the I/O psychologist bases his decision on the "true" relationship between those variables might be weaker or not exist at all. In this case, a lot of resources would be wasted. This example applies to all other sciences in which SEM is being used.

Conclusion

Due to its many advantages, SEM has become one of the most widely used tools for theory development and theory evaluation in many scientific disciplines. The findings researchers obtain by testing their theories using SEM become the cornerstone for other researchers' as well as practitioners' work. Misinterpretations of the variable interrelationships in one's data might lead to the perpetuation of flawed theories and impede the advancement of scientific knowledge building. Therefore, it is essential that the tools researchers use to evaluate their models are accurate. This study replicated findings that show that current tools are not sensitive enough to reject misspecified models. In addition, researchers are provided with newly developed fit indices that circumvent the issues related to global fit indices. It was shown that by using the new path-related fit indices to evaluate their data, causal models are subjected to a more stringent test and misspecified models can be better identified. As such, this research contributes to the improvement of tools for model testing, so that researchers can have greater confidence in the verity of their theoretical models.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332. doi:10.1007/BF02294359

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103(3), 411-423. doi:10.1037/0033-2909.103.3.411

Aryee, S., Chen, Z., Sun, L., & Debrah, Y. A. (2007). Antecedents and outcomes of abusive supervision: Test of a trickle-down model. *Journal Of Applied Psychology*, *92*, 191-201. doi:10.1037/0021-9010.92.1.191

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246. 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*, 588-606. doi:10.1037/0033-2909.88.3.588

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*, 303-316. doi: 10.1177/0049124189017003004

Boyer, L., Cermolacce, M., Dassa, D., Fernandez, J., Boucekine, M., Richieri, R., Vaillant, F., Dumas, R., Auquier, P., & Lancon, C. (2012). Neurocognition, Insight and Medication Nonadherence in Schizophrenia: A Structural Equation Modeling Approach. *Plos ONE*, *7*, 1-8. doi:10.1371/journal.pone.0047655

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*, 119-143. doi:10.1080/10705519509540000

Duncan, O., Haller, A., & Portes, A. (1971). Peer influence on aspirations: A reinterpretation. In H.M. Blalock (Ed.), *Causal models in the social sciences* (pp. 219-244). Chicago: Aldine-Atherton.

Ecob, R. (1987). Applications of structural equation modeling to longitudinal educational data. In P. Cuttance & R. Ecob (Eds.), *Structural modeling by example* (pp. 138-159). New York: Cambridge University Press.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In Bollen, K. A., & Long, J. S. (Eds.), *Testing structural equation models* (pp.40-65). Newbury Park, CA: Sage.

Hoelter, J. W. (1983). The analysis of covariance structures goodness-of-fit indices. *Sociological Methods & Research*, *11*, 325-344. doi: 10.1177/0049124183011003003.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424. doi:10.1037/1082-989X.3.4.424.

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*, 1-55. doi:10.1080/10705519909540118.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage Publications.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183-202. doi: 10.1007/BF02289343.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Chicago: International Education Services.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Kemp, E., & Bui, M. (2011). Healthy brands: establishing brand credibility, commitment and connection among consumers. *Journal Of Consumer Marketing*, *28*, 429-437. doi: 10.1108/07363761111165949.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, *10*, 333-351. doi: 10.1207/S15328007SEM1003_1.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York: Guilford.

Küpper, D. & Burkhart, T. (2009). Effectuation in the context of R&D projects: Characteristics and impact on project performance. Academy Of Management Annual Meeting Proceedings, 1-6. doi:10.5465/AMBPP.2009.44251806.

Lance, C. E., & Vandenberg, R. J. (2002). Confirmatory factor analysis. *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis*, 221-254.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria What Did They Really Say?. *Organizational Research Methods*, *9*, 202-220. doi: 10.1177/1094428105284919.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107. doi:10.1037/0033-2909.100.1.107.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. Annual review of psychology, 51, 201-226. doi:10.1146/annurev.psych.51.1.201.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391. doi:10.1037/0033-2909.103.3.391.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181-220.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*, 97-103. doi: 10.1007/BF01908590.

McDonald, R. P., & Ho, M. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64-82. doi:10.1037/1082-989X.7.1.64.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568. doi:10.1037/0021-9010.93.3.568.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal, 19,* 86-98. doi:10.1080/10705511.2012.634724.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*, 430-445. doi:10.1037/0033-2909.105.3.430.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

O'Boyle, E. R., & Williams, L. J. (2011). Decomposing model fit: Measurement vs. theory in organizational research using latent variables. *Journal Of Applied Psychology*, *96*, 1-12. doi:10.1037/a0020539.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*, 369-379. doi:10.1080/10705519609540052.

Sas Institute. (2011). *SAS/STAT 9.3 user's guide*. SAS Institute.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 152-78). San Francisco: Jossey-Bass.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*, 893-898. doi: 10.1016/j.paid.2006.09.017.

Steiger, J.H., & Lind, J.C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Stone, D. N., Bryant, S. M., & Wier, B. (2010). Why are financial incentive effects unreliable? An extension of self-determination theory. Behavioral Research in Accounting, 22(2), 105-132.doi:10.2308/bria.2010.22.2.105.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models (pp.136-162).* Newbury Park, CA: Sage.

Tanaka, J. S., Panter, A. T., Winborne, W. C., & Huba, G. J. (1990). Theory testing in personality and social psychology with structural equation models: A primer in 20 questions. In C. Hendrick, M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 217-242). Thousand Oaks, CA US: Sage Publications, Inc.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38,* 1-10.doi:10.1007/BF02291170.

Velayutham, S., & Aldridge, J. M. (2013). Influence of Psychosocial Classroom Environment on Students' Motivation and Self-Regulation in Science Learning: A Structural Equation Modeling Approach. *Research In Science Education*, *43*, 507-527. doi: 10.1007/s11165-011-9273-y.

Williams, L. J., & Holahan, P. J. (1994). Parsimony-based fit indices for multiple-indicator models: Do they work? *Structural Equation Modeling: A Multidisciplinary Journal*, *1*, 161-189. doi:10.1080/10705519409539970.

Williams, L. J., & O'Boyle, E. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods, 14,* 350-369. doi: 10.1177/1094428110391472.

Wu, X., Berkow, K., Frank, D. N., Li, E., Gulati, A. S., & Zhu, W. (2013). Comparative analysis of microbiome measurement platforms using latent variable structural equation modeling. BMC bioinformatics, 14(1), 79. doi:10.1186/1471-2105-14-79