Published in final edited form as:

Stat Med. 2015 October 30; 34(24): 3246–3280. doi:10.1002/sim.6541.

A Tutorial on Structural Equation Modeling for Analysis of Overlapping Symptoms in Co-occurring Conditions Using MPlus

Douglas D. Gunzler, PhD¹ and Nathan Morris, PhD²

¹Case Western Reserve University, Center for Health Care Research & Policy, MetroHealth Medical Center, Cleveland, Ohio

²Case Western Reserve University, Department of Epidemiology and Biostatistics, Cleveland, Ohio

Abstract

Structural Equation Modeling (SEM) is a very general approach to analyzing data in the presence of measurement error and complex causal relationships. In this tutorial, we describe SEM, with special attention to exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and multiple indicator multiple cause (MIMIC) modeling. The tutorial is motivated by a problem of symptom overlap routinely faced by clinicians and researchers, in which symptoms or test results are common to two or more co-occurring conditions. As a result of such overlap, diagnoses, treatment decisions and inferences about the effectiveness of treatments for these conditions can be biased. This problem is further complicated by increasing reliance on patient-reported outcomes, which introduces systematic error based on an individual's interpretation of a test questionnaire. SEM provides flexibility in handling this type of differential item functioning and disentangling the overlap. Scales and scoring approaches can be revised to be free of this overlap, leading to better care. This tutorial uses an example of depression screening in Multiple Sclerosis patients in which depressive symptoms overlap with other symptoms, such as fatigue, cognitive impairment and functional impairment. Details of how MPlus software can be used to address the symptom overlap problem, including data requirements, code and output are described in this tutorial.

Keywords

structural equation modeling; factor analysis; multiple indicator multiple cause model; differential item functioning; measurement invariance; patient reported outcomes

1. Introduction

1.1 Motivation

Symptoms or test findings that are common to two or more conditions may confound diagnosis, treatment decisions and inferences about the effectiveness of treatment. For

example, depression is the most frequent psychiatric diagnosis in multiple sclerosis (MS) patients, with lifetime risk estimated at ~50% [1]. Commonly, a patient has both MS and depression and experiences a symptom, such as fatigue, that frequently occurs in both conditions. As a result, it may be difficult to determine whether the fatigue is a symptom of MS or a symptom of depression. We use the problem of symptom overlap between MS and depression to motivate a more general discussion of structural equation modeling (SEM).

Note also, that the symptom overlap problem described above is more generalizable than it may at first seem. For example, a scale measuring satisfaction with your teacher and letter grade earned in class would be expected to have some overlap. In the general setting, two or more variables might have overlapping manifestations (are cross-loaded). Thus, it might be difficult to disentangle these variables, confounding any inference about them separately.

1.2 Relevant Introductory Texts

Kline [2] has written an influential introductory text on the topic of structural equation modeling (SEM), accessible to an applied researcher, while Bollen [3] provides a comprehensive and thorough overview of general structure equation systems, commonly known as the LISREL (linear structural relations) model. This LISREL approach was originally developed by Joreskog [4], and several similar or equivalent formulations of SEM exist [5]. Textbooks such as Bollen [3] and Kline [2] also provide an introduction to special cases of SEM such as exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and the multiple indicator multiple cause (MIMIC) model.

Brown [6] has provided an in depth introduction of CFA including details about the MIMIC model. Woods et al. [7] have performed a thorough application of the MIMIC model for differential item functioning (DIF) testing. On the subject of EFA, Tucker and MacCullum [8] have an uncompleted book on factor analysis, available at <http://www.unc.edu/~rcm/book/factornew.htm>, that nevertheless provides a completed overview of EFA. In SEM computing, Byrne [9-12] has provided resources for researchers for using SEM in MPlus, EQS, LISREL, PRELIS, SIMPLIS and AMOS. Further, the MPlus website [13] is filled with manuals and examples to lead an applied researcher conducting SEM analysis.

1.3 Tutorial aims and outline

Our aim in this tutorial is to first provide a general background on SEM. We use this general background to describe several special cases of SEM: EFA, CFA and MIMIC modeling. We then provide a blueprint on how to apply each of these techniques in succession to understand and correct for diagnostic overlap of two or more conditions [2, 3]. For each of these techniques the analyst must make a series of complex model decisions (e.g., choosing how many factors, what is the “best fitting” model). We provide details on a general sequence of models which can be used by working through the motivating example and using the MPlus software to show how to make such decisions based on theory and empirical evidence.

Modern SEM has progressed rapidly with many new developments appropriate for other settings not discussed in this tutorial i.e. longitudinal settings. Therefore, in order to give a

clear overview, we focus on an important problem in the cross-sectional setting for our general discussion.

We start, in Section 2, with background information about a motivating example. This is followed, in Section 3, with a basic introduction to the SEM framework. In Section 4, we provide an overview of factor analysis, both EFA and CFA. We next introduce the MIMIC model and DIF in Section 5. In Section 6 we provide an algorithm for adjusting a scale to isolate the latent dimension of the intended condition under study in the original scale along with practical applications of the adjusted scale for clinical use. Section 7 summarizes the paper. We provide the MPlus code for our example in the Supporting Information (available from the journal web page).

2. Background information about motivating study

2.1 Latent constructs, patient-reported outcomes, and differential item functioning

Scales representing a condition that may overlap with another condition may not be capturing the intended construct of interest. For example a depression screening scale may not accurately estimate depressed mood in a MS patient, due to overlapping symptoms of both conditions. Researchers need reliable and accurate measures of symptoms or conditions that cannot be measured directly. Such measures that are unobserved are considered *latent constructs*. Unlike directly observable measures such as height or weight, researchers may not be able to measure variables such as depression directly.

To measure such a latent construct as depression, we can capture indicators from a multiple item scale such as the PHQ-9 that represent the underlying construct. These items are directly observed and in theory if we also account for additional measurement error in our construct, accurately represent the measure that cannot be observed directly.

Patient-reported outcomes may similarly be used as a scale to help support treatment decision making through capturing an intended measure from the patient's perspective. Patient-reported outcomes have an important place in psychiatry and neurology to measure the extent of disease or condition at the individual level, because they reflect the self-reported health state of the patient directly [14]. For example, the PHQ-9 is a patient-reported outcome for depression screening [15, 16] and the MS Performance Scales© (PS) are patient-reported outcomes routinely used to describe MS symptoms [17, 18]. However, their effective integration in personalized medicine requires addressing certain conceptual and methodological challenges, since each individual patient may have a different view of how to fill out a test questionnaire, leading to a specific type of systematic error known as *differential item functioning (DIF)*.

In *Item Response Theory (IRT)* we refer to a more classic definition of DIF in which probability of response is determined by observable subgroups. This is different than our PRO definition in which we defined DIF based on some viewpoint that we don't specifically observe in our models. In the classic definition, DIF occurs when individuals from different groups (e.g., levels of MS-related fatigue) with the same latent trait (level of depression) have a different probability of giving a certain response on a questionnaire or test (e.g.,

items for sleep problems and fatigue of the PHQ-9). This classic IRT type of DIF is also very important to account for when performing analysis of overlapping symptoms of co-occurring conditions.

2.2 Prior research on depression screening scales with MS patients

Prior studies have assessed whether signs and symptoms of MS may overlap with those of depression, using other depression screening scales (Beck Depression Inventory, Beck Depression Inventory-II, Beck Depression Inventory Fast Screen, Chicago Multiscale Depression Inventory) [19-24]. The results have led to mixed recommendations of whether certain items confound the depression screening scales. For example, Mohr et al. [19, 20] recommended to omit the Beck Depression Inventory items for assessing work ability, fatigue and health concerns while Aikens et al. [21] encouraged full application of the scale for assessing depressive symptoms in patients with MS. One prior study specifically looked at the PHQ-9, in examining if PHQ-9 scores are biased in MS patients due to the MS symptoms of fatigue and poor concentration [25]. In the analyses, the authors compared the PHQ-9 items for fatigue and poor concentration in 173 MS patients to 3304 other subjects from the general population in the sample under study. They concluded that there was no evidence to exclude these items from a modified PHQ-9 score.

However, in this study we uniquely focus on a generalizable approach for determining the factor structure of a measure and then using this factor structure in a series of models to evaluate the overlap of multiple symptoms of multiple conditions simultaneously. Our approach will then allow us to form practical adjusted scales which better represent single conditions individually, i.e., free of the phenomenon of symptom overlap. Further, our motivating study extends any prior work in the area by (1) using the largest observational cohort analyzed to date of only MS patients, eliminating the potential selection bias from a noncomparable control group (2) conducting psychometric analysis of the measurement properties of depression self-rating and MS disability measures. While our motivating example focuses on the PHQ-9, it is straightforward to generalize to other scales.

2.3 Study design and measures under study

The PHQ-9 is a nine item self-reported depression screening tool to be used in connection with expert clinical judgment and/or further rating tools and not as an actual depression diagnosis [16]. Patients specify frequency in the past 2 weeks (0 = not at all to 3 = every day) of nine symptoms, yielding a total score (range: 0-27). Scores of 5, 10, 15, and 20 are validated thresholds for mild, moderate, moderately severe and severe depression. Scores on this self-reported instrument are often used to guide treatment decisions [15]. In particular, a PHQ-9 ≥ 10 has been previously established as a screening cutoff for depressive disorder [15, 26]. The PHQ-9 has been validated using multiple modes for administration, clinical populations, and diverse racial/ethnic groups [16]. The internal consistency of the PHQ-9 in two large studies from primary care and obstetrics-gynecology clinics was relatively high (Cronbach's $\alpha=0.89$ and 0.86) [15], with a similar internal consistency reported in preliminary studies with MS patients [27].

Cleveland Clinic's Knowledge Program (KP) links patient-reported PHQ-9 data to its EPIC electronic health record (EHR), yielding powerful opportunities to study and improve patient care and clinical research [28], given a researcher effectively accounts for the measurement error involved with patient reported outcomes and other objective performance disease specific measures. The Mellen Center [29] for Multiple Sclerosis manages more than 20000 visits and 1000 new patients every year for MS treatment. The KP tracks illness severity and treatment efficacy over time across the Mellen Center population.

We use a retrospective cohort study design. The inclusion criteria for our sample incorporates patients making at least one visit to the Mellen Center with measurements of PHQ-9 score and a timed 25-foot walk available. Data are available for 3507 MS patients from 2008-2011 that meet our inclusion criteria.

The sample is typical of the United States' MS population in that the Mellen Center population is mostly white (83%) and female (73%). MS is typically diagnosed in patients in their early 30's and in our sample the average age was 46 (SD = 12). These patients had their first MS diagnosis an average of 10 (SD = 9) years ago with 81% relapsing and 16% progressive with the remaining patients falling into other categories, or under evaluation for a potential MS diagnosis. Nearly 30% (n=1005) of patients had PHQ-9 ≥ 10 at their entry to the KP. The distribution of PHQ-9 scores represents a wide range of depression severity levels. See Gunzler et al. [30] for a baseline table of characteristics of the Mellen Center MS population.

The KP collects the eight single item MS Performance Scales© (PS) [17, 18] which are patient-reported disability measures. These include MS-related fatigue and cognitive domains with 6 ordinal responses, our main covariates for self-reported fatigue and cognitive decline. Reliability, criterion and construct validity have been established for these PS domains in previous studies of MS patients [17, 18].

In addition to patient-reported measures, our single item measures for functional impairment, the timed 25-foot walk and 9-hole peg test, are objective performance measures of lower (timed 25-foot walk) and upper (9-hole peg test) extremity function [31]. The timed 25-foot walk is a test of quantitative mobility and leg function performance, while the 9-hole peg test is a brief, standardized, quantitative test of arm and hand function. In our study, functional impairment is defined by these two correlated objective performance measures [32-34]. We did not consider models which adjust for only one of the two functional impairment measures to be feasible.

In summary for section 2, given our objective for our motivating example, in analyzing this EHR data with patient-reported outcomes, we must contend with numerous issues in order to provide sound statistical inference. These issues we must account for include symptom overlap, DIF, measurement error and complex relationships among a web of variables relating to co-occurring conditions. As we will show in this tutorial, SEM is a very effective technique for dealing with all these issues.

3. Structural Equation Modeling (SEM)

3.1 Basic definition

We begin the methods sections of the overview with a general background of SEM before describing the more specific techniques within the SEM framework useful for analysis of overlapping symptoms of co-occurring conditions. SEM is an extremely general modeling framework which addresses two key issues of real world importance: measurement error (i.e. latent variables) and causal networks [2, 3]. In classical approaches to SEM this is achieved purely by looking at the covariance between variables. Suppose we aim to estimate v unknown model parameters from a total of w observed and w^* unobserved (latent) variables. Then classic approaches to SEM model the relationship between the covariance and the parameters. That is, $\Sigma = \Sigma(\theta)$ for a vector of unknown parameters θ of dimension $v \times 1$ and the variance-covariance matrix of our observed variables $\Sigma(\theta)$ of dimension $w \times w$. In more modern approaches to SEM, we often have to look outside of the covariance between variables. For example, with categorical variables and multilevel models the covariance between variables alone is not a sufficient statistic for determining the likelihood. In such cases we may require information from the fourth-moment or individual level data instead of the covariance matrix. SEM is a generalization of numerous statistical techniques, such as ANOVA, linear regression and factor analysis. It can handle many more specialized techniques such as modeling feedback loops, latent constructs and path analytic models. SEM approaches are well-suited for many studies due to their facility in dealing with latent (unobserved) variables and the assessment of complex mediating relationships in causal analysis [2, 3, 35, 36]. Similarly, for making causal inference and assessing model fit, SEM techniques are also advantageous for performing analyses with observed variables only.

More specific to the nomenclature of SEM, we can view SEM as a general technique for using a conceptual model, path diagram and system of linked regression-style equations to capture complex and dynamic relationships among a web of observed and unobserved variables. The conceptual model is a general idea of the relationships under study. For example, a researcher could hypothesize that smoking leads to lung cancer. We explain the concepts of the path diagram and the general LISREL form of structural equations in 3.3 and 3.4.

Conducting SEM analyses involves four steps (1) specifying the model (2) assessing model fit (3) making any model modifications (4) testing hypotheses of interest. David A. Kenny defines model specification as the “translation of theory, previous research, design, and common sense into a structural model” [37]. In this process, a researcher indicates causal paths and directionality between variables (latent or observed) under study. While testing hypotheses of interest is typical of most statistical modeling approaches, even this is unique in SEM. Specifically, the SEM framework allows us to simultaneously estimate parameters and thus simplifies testing to a single analyses and conduct tests that are adjusted for all other model relationships and covariates. We discuss model fit and model modifications in sections 3.5 and 3.6.

3.2 MPlus

Significant advances have been made over the past few decades in the theory and applications as well as software development for fitting SEM models. For example, in addition to specialized packages such as LISREL [4], MPlus [13], EQS [5], and Amos [38], procedures for fitting SEM are also available from general-purposes statistical packages such as R, SAS, STATA and Statistica. Many of these software packages are extraordinarily versatile modeling packages which extend the classical LISREL SEM formulation to include new types of observed and unobserved traits such as categorical traits and time to event traits.

In our analyses we use MPlus Version 7.0 [13]. MPlus is more generally a program for latent variable modeling of which classical SEM is a special case [39]. Examples of other latent variable modeling techniques include exploratory factor analysis, Bayesian networks and item response theory.

MPlus Version 7.0 and above allows a user to perform analyses in the SEM framework (i.e. EFA, MIMIC, multigroup) either by inputting code, using a language generator or drawing a path diagram directly, with a variety of different choices for model estimators.

Data must reside in an external ASCII file (i.e. file type .dat or .txt) containing no more than 500 variables and with a maximum record length of 5,000 characters [12]. The data must be numeric except for certain missing value flags and should not include a header for variable names. As an example, to perform EFA on the nine PHQ-9 items, a text file called 'MS' is saved on the C drive with the data with the total summed PHQ-9 score along with the nine items and the MS performance scales subscale for fatigue (11 variables). Here are 10 lines of data from the input text file 'MS.txt' on the C drive, where a missing value is coded as '-999':

```
6  1  1  0  1  0  0  3  0  0  1
13 3  1  1  3  2  1  1  1  0  4
0  0  0  0  0  0  0  0  0  0  2
6  0  0  2  1  2  0  1  0  0  2
2  0  0  0  1  0  1  0  0  0  2
1  0  0  0  1  0  0  0  0  0  1
1  0  0  0  1  0  0  0  0  0  0
8  0  1  1  3  1  1  1  0  0 -999
9  1  0  2  3  0  0  2  1  0  4
```

MPlus code format for running a model on an existing external ASCII file begins with a title, identifying the data sources, naming the variables, and specifying which variables will be used in analyses. Other options here include identifying the numeric code for missing data and specifying categorical or count outcomes. All variable names must begin with a letter and must be eight characters or less. An example of code for performing EFA on the nine items of the PHQ-9 (see section 4.2), using the data file 'MS' on the C drive:

```
Title: Exploratory Factor Analysis PHQ9;
```

```
Data:
```



```

FILE= C:\ MS.txt;
Variable:
Names are
PHQ9
PHQ9_A1 PHQ9_A2 PHQ9_A3 PHQ9_A4
PHQ9_A5 PHQ9_A6 PHQ9_A7 PHQ9_A8
PHQ9_A9 Fatigue
;
usevar=
PHQ9_A1 PHQ9_A2 PHQ9_A3 PHQ9_A4
PHQ9_A5 PHQ9_A6 PHQ9_A7
PHQ9_A8 PHQ9_A9;
missing are all
(-999);
ANALYSIS:
estimator=MLR;
Factor Analysis PHQ9;
SUMMARY OF ANALYSIS
Number of groups                                1
Number of observations                          3505
Number of dependent variables                   9
Number of independent variables                 0
Number of continuous latent variables           0
Observed dependent variables
Continuous
PHQ9_A1 PHQ9_A2 PHQ9_A3 PHQ9_A4 PHQ9_A5 PHQ9_A6
PHQ9_A7 PHQ9_A8 PHQ9_A9
Estimator                                        MLR
Rotation                                        GEOMIN
Row standardization                            CORRELATION
Type of rotation                              OBLIQUE
Epsilon value                                  Varies
Information matrix                             OBSERVED
Maximum number of iterations                    1000
Convergence criterion                          0.500D-04
Maximum number of steepest descent iterations   20
Maximum number of iterations for H1             2000
Convergence criterion for H1                    0.100D-03
Optimization Specifications for the Exploratory Factor Analysis Rotation Algorithm
Number of random starts                        30
Maximum number of iterations                    10000
Derivative convergence criterion                0.100D-04

```

Due to sample size and since each of the PHQ-9 items has at least four ordinal categories, the measures can be viewed as approximating an interval scale [40-43]. Treating scales such as the PHQ-9 items as categorical in EFA also introduces threshold parameters (3 for every

item for the PHQ-9) which won't be as straightforward to interpret. Thus, we treat the PHQ-9 items as continuous, as was commonly done in prior studies involving the PHQ-9 [40-41]. Note that MPlus is not case sensitive and writing 'missing' or 'MISSING' are interpreted in the same manner.

Other major headings for MPlus coding include: ANALYSIS, MODEL, OUTPUT and PLOT. In the ANALYSIS section an estimator can be specified. For instance if a robust maximum likelihood is used (standard maximum likelihood, or ML, is often the default setting for continuous variables), we specify the MLR estimator in the coding for this section. MLR is a commonly used estimator in the SEM framework for interval data that is slightly skewed or where parametric assumptions may be violated [13]. Note that to examine the potential for bias based on variable distributions we verified results using a mean and variance adjusted weighted least squares estimator (WLSMV option in MPlus) treating our outcomes (PHQ-9 items) as ordinal categorical measures.

The MODEL section is where both the measurement and structural components of the SEM model are specified. We reserve this discussion until after introducing more concepts regarding the SEM framework. OUTPUT allows the user to request more technical details in the output. For example, here standardized estimates or modification indices can be requested along with factor scores. Finally, with PLOT, the user can request various types of plots corresponding to the particular analyses. The MPLUS User's Guide (also available on the statmodel.com website) provides example codes and applications for many different models within the SEM framework, and it is a great resource for learning how to program with MPlus [13]. Books have also been written on the topic of using MPlus for SEM analyses, such as *Structural Equation Modeling with Mplus* by Barbara Byrne [12]. We have provided a brief overview of MPlus coding for a basic cross-sectional model, but depending on the model complexities there are more MPlus coding headings (i.e. multilevel latent growth model, monte carlo simulation).

Once code is entered, users can either choose to 'Run MPlus' on the drop down menu under the heading 'Mplus' or hit Alt+R on the keyboard to run the program. Mplus will then provide technical output for the analyses, which we will discuss in later sections of this tutorial. Included in the output will be a summary of the analyses. For example, for the exploratory factor analysis, MPlus will provide the sample size, variable information and information about the estimator and convergence:

3.3 Path Diagram

Geneticist Sewell Wright, the originator of path analysis, implied that if we make a causal assumption, and assume the direction of causality, then we can measure "the importance of a given path of influence from cause to effect [44]." A path diagram is a graphical representation of the conceptual model and consists of nodes representing the variables and arrows showing relations among the variables. In a path diagram, unobserved variables, including latent constructs and error terms, are represented by ellipses or circles. Observed variables are represented by rectangles. When a variable is measured with error, we add an error term with an arrow from it to the variable. Arrows are generally used to represent relationships among the variables, giving a graphical language to causal relationships. A

single straight arrow indicates a causal relation from the base of the arrow to the head of the arrow. A curved two-headed arrow indicates there may be some association between the two variables. Error terms not connected in a path diagram, indicate stochastic independence across the error terms. However, error terms may be correlated. For example, the only way to correlate two endogenous variables is through their error terms. If we suspect two variables are highly correlated, they should share similar errors.

In general path diagrams can be understood as implying certain conditional independence relations among variables. Such conditional independence relations can be read off the diagram using the “d- separation” rule. D-separation is a criterion for deciding, from a given causal graph, whether a set X of variables is independent of another set Y , given a third set Z [45]. The particular variables x and y are d-connected by z if they are not d-separated by z (i.e. z does not block the causal path between them). See Bollen [3] and Pearl [45] for more a more complete explanation of this rule and for details about modeling complex relationships involving latent constructs using path diagrams and SEM. Also, d-connected is another way of describing *mediation*, in which a middle variable, or mediator, helps explain the mechanism by which an independent variable influences an outcome [46].

As an example of a path diagram, figure 1 represents the path diagram for the causal path from baseline *time since diagnosis* in MS patients to *depression*. Here, also indirectly, *time since diagnosis* effects *depression* through *cognitive decline*. Since this path between *time since diagnosis* and *depression* is not blocked by *cognitive decline*, these two measures are d-connected by *cognitive decline*. *Depression* is a latent variable while the other variables are observed. The PHQ-9 items are observed indicators of *depression*. Figure 1 is also an example of a hypothesized mediation.

3.4 The LISREL formulation of SEM

In SEM, variables may be divided into two classes: *exogenous* and *endogenous*. Endogenous variables act as an outcome in at least one of the structural equations, while exogenous variables are always independent variables. In the path diagram, endogenous variables are those nodes with no arrows pointing into them, and, thus, their variation is not explained by other factors in the model. From figure 1, *depression* and *cognitive decline* are endogenous variables, while *time since diagnosis* is an exogenous variable.

In the LISREL formulation for SEM there are also two sets of equations which can be written in a general matrix form: the structural equations and the measurement equations [3]. The *structural* equations show potential causal links between *endogenous* and *exogenous* variables, and the *measurement* equations explicitly model measurement error and latent variables [2, 3].

First, consider the matrix form for the measurement equations:

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu}_y + \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \\ \mathbf{x} &= \boldsymbol{\mu}_x + \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \end{aligned} \quad (1)$$

Here, ξ represents a vector of r unobserved latent *exogenous* variables which are measured by the q observed variables $\mathbf{x}$. Similarly, η is a vector of m unobserved latent *endogenous* variables which are measured by the p observed variables $\mathbf{y}$. The equations for $\mathbf{y}$ and $\mathbf{x}$ include vectors of intercepts, μ_y and μ_x , matrices of slopes, Λ_y and Λ_x , respectively, and vectors of corresponding random error terms, ε of dimension $p \times 1$ and δ of dimension $q \times 1$, respectively. μ_y is of dimension $p \times 1$ and μ_x is of dimension $q \times 1$, and Λ_y is of dimension $p \times m$ and Λ_x is of dimension $q \times r$. Λ_y and Λ_x are often referred to as loading matrices.

Next, the structural model, which relates the unobserved latent variables to each other, can be expressed in the following form:

$$\eta = \mu_\eta + \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta \quad (2)$$

μ_η is a $m \times 1$ matrix of intercepts for the unobserved endogenous latent variables, $\mathbf{B}$ is a $m \times m$ matrix of slopes relating the unobserved endogenous latent variables to each other, $\mathbf{\Gamma}$ is $m \times n$ matrix of slopes for the unobserved exogenous latent variables and ζ is a $m \times 1$ vector of random error terms for the unobserved endogenous latent variables.

In the special case of no latent variables, such as in a path analytic model, there is no measurement model because all variables are measured without error (i.e. $\mathbf{y} = \eta$ and $\mathbf{x} = \xi$). Thus the form of the structural model in (2) can be simplified to observed variables only:

$$\mathbf{y} = \mu_y + \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \zeta \quad (3)$$

Here, μ_y is a $p \times 1$ matrix of intercepts, $\mathbf{B}$ is a $p \times p$ matrix of slopes for the observed endogenous variables, $\mathbf{\Gamma}$ is $p \times q$ matrix of slopes for the observed exogenous variables and ζ is $p \times 1$ vector of random error terms for the observed endogenous variables.

Given a little algebra, under the assumption that $\mathbf{I} - \mathbf{B}$ is invertible, from (2) all endogenous variables can be moved onto the left side of the equation, while all exogenous variables remain on the right side:

$$\begin{aligned} \eta &= \mu_\eta + \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta \\ (\mathbf{I} - \mathbf{B})\eta &= \mu_\eta + \mathbf{\Gamma}\xi + \zeta \\ \eta &= (\mathbf{I} - \mathbf{B})^{-1}\mu_\eta + (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma}\xi + (\mathbf{I} - \mathbf{B})^{-1}\zeta \\ \eta &= (\mathbf{I} - \mathbf{B})^{-1}(\mu_\eta + \mathbf{\Gamma}\xi) + \zeta^* \end{aligned} \quad (4)$$

where $\mathbf{I}$ is the identity matrix of dimension $m \times m$ and $\zeta^* = (\mathbf{I} - \mathbf{B})^{-1}\zeta$. As follows, (3) can be rewritten in a similar form to (4) given no latent variables:

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}(\mu_y + \mathbf{\Gamma}\mathbf{x}) + \zeta^* \quad (5)$$

Note, however, in many cases, such as the mediation process we describe in section 3.3, the model is not linear (i.e., it is curvilinear) since $(\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Gamma}$ from (4) and (5) is not linear in terms of the parameters [36]. For an example of this, assume a simple three variable mediation process where y , z and x are all observed variables. The structural equations for the mediation process are then:

$$\begin{aligned} y &= \mu_y + B_z z + \gamma_{xy} x + \zeta_y, \\ z &= \mu_z + \gamma_{xz} x + \zeta_z \end{aligned} \quad (6)$$

We can express these equations in the form of (3):

$$\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix} + \begin{bmatrix} 0 & B_z \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} + \begin{bmatrix} \gamma_{xy} \\ \gamma_{xz} \end{bmatrix} x + \begin{bmatrix} \zeta_y \\ \zeta_z \end{bmatrix} \quad (7)$$

Likewise, we can express them in a form similar to (5):

$$\begin{aligned} \begin{bmatrix} y \\ z \end{bmatrix} &= \begin{bmatrix} 1 & -B_z \\ 0 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix} + \begin{bmatrix} \gamma_{xy} \\ \gamma_{xz} \end{bmatrix} x \right) + \begin{bmatrix} 1 & -B_z \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \zeta_y \\ \zeta_z \end{bmatrix} \\ &= \begin{bmatrix} \mu_y + B_z \mu_z + (\gamma_{xy} + B_z \gamma_{xz}) x + \zeta_y^* \\ \mu_z + \gamma_{xz} x + \zeta_z^* \end{bmatrix} \end{aligned} \quad (8)$$

The above SEM is clearly not linear in the parameters because of the terms $B_z \gamma_{xz}$ and $B_z \mu_z$ in the first row of the matrix in (8).

3.5 Estimating parameter and standard errors

Let $\mathbf{v}_i$ be a vector of all endogenous and exogenous variables, observed and latent. Then let $\mathbf{\Sigma}(\boldsymbol{\theta})$ be the theoretical variance-covariance matrix of $\mathbf{v}_i$ for the given parameters $\boldsymbol{\theta}$. In order to estimate parameters $\boldsymbol{\theta}$ and standard errors and to test hypotheses of interest in the SEM framework, we must decide upon an appropriate estimation method. *The key idea used in classic SEM estimation methods is to try to choose the model parameters ($\boldsymbol{\theta}$) which yields an implied/theoretical covariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta})$ as “close” as possible to the observed/sample covariance matrix $\mathbf{S}$.* However, it is crucial to define precisely what is meant by “close.” In classical SEM, the distance between $\mathbf{\Sigma}(\boldsymbol{\theta})$ and $\mathbf{S}$ must be given by some discrepancy function $F(\mathbf{\Sigma}(\boldsymbol{\theta}), \mathbf{S})$, then numerical optimization is used to find the $\boldsymbol{\theta}$ which minimizes this function [47, 48]. In some more modern versions of SEM, the discrepancy function is modified to include the intercepts/means and covariate coefficients $F(\mathbf{\Sigma}(\boldsymbol{\theta}), \mathbf{B}(\boldsymbol{\theta}), \mathbf{S}, \mathbf{C})$ where $\mathbf{B}(\boldsymbol{\theta})$ represents the coefficients and intercepts/means as a function of the model parameters and $\mathbf{C}$ is an estimate of the corresponding coefficients and intercepts/means [49, 50].

For an example, for performing estimation using ML, the negative of the log-likelihood may also be considered as a discrepancy function, because minimizing the negative of the log-likelihood is the same as maximizing the log-likelihood. The MLR approach uses ML to estimate the parameters, but uses a robust sandwich type estimator (Huber-White sandwich estimator) to calculate standard errors that are robust to model assumptions such as

multivariate normality [51]. Bootstrapping is a similar but more computationally intensive approach to creating robust standard errors [52].

A discrepancy function known as the generalized least squares (GLS) function can be applied when the variances of the observations are unequal or when there is a certain degree of correlation between the observations. Interestingly, it may be shown that if the model is correctly specified *and the observed variables are multivariate normally distributed*, then GLS is as asymptotically efficient as the maximum likelihood estimator [53].

Both ML and MLR provide a method for dealing with missing data under the missing at random (MAR) assumption, where MPLUS, for example, uses a slight modification of ML, full information ML (FIML) [13]. In this approach, all parameters and standard errors are derived from the joint distribution of the endogenous and exogenous variables, given assumptions such as multivariate normality and conditional independence, thus providing a single analyses framework for all hypothesis testing based on the model. Under these assumptions, the marginal likelihood after integrating out the missing values can be maximized. In doing so, the paradigm of defining a discrepancy function for the covariance matrix (i.e. $F(\Sigma(\theta), \mathbf{S})$ or $F(\Sigma(\theta), \mathbf{B}(\theta), \mathbf{S}, \mathbf{C})$) no longer works because $\mathbf{S}$ (and $\mathbf{C}$) is not a sufficient statistic for the data. Thus, individual level data is needed.

Numerous approaches exist for using SEM when dealing with non-normal data. In some cases, a simple transformation may be sufficient. However, within the classical SEM paradigm, a robust discrepancy function [49] is available, based on adjusting the mean or both the mean and variance in multiple stages. This weighted least squares estimator uses an asymptotically distribution free (ADF) discrepancy function using direct estimation of the fourth-order moments of the residuals. Similarly, Muthén originally developed a three-stage weighted least squares estimator for handling categorical indicators of a latent construct under the assumption that the underlying latent construct is normally distributed [49].

In more modern versions of SEM, binary, ordinal or nominal categorical outcomes are handled using ML. For example, SEM with a logit or probit interpretation may be achieved using numerical approximations such as Monte Carlo integration and adaptive Gaussian quadrature [54]. Note that the probit model may also be interpreted as assuming a latent construct exists which leads to different categorical outcomes if the construct exceeds some threshold. Similarly, an estimator is available for count outcomes with a log interpretation [13]. Note that once again, for these extensions of SEM, $\mathbf{S}$ (and $\mathbf{C}$) is not a sufficient statistic for the data, and thus individual level data is needed.

In small data sets, results may be biased using frequentist SEM approaches Bayesian analyses has been shown to perform well on such data [55]. Recent updates in latent variable software, such as MPlus, allow for Bayesian estimation using a Markov chain Monte Carlo algorithm.

3.6 Assessing Model Fit

A reasonable expectation of model fit in SEM is that a researcher won't be able to prove a model, but will be able to show that discrepancy between model and data doesn't surpass the

limits of chance. There is also no one “gold standard” of model fit in SEM, and thus we provide a sketch of some of the statistics and indices that will be useful for analyses.

Under certain regularity conditions, plugging $\hat{\theta}$, our estimate of θ which minimizes the distance between $\Sigma(\theta)$ and S and $B(\theta)$ and C in our discrepancy function $F(\Sigma(\theta), B(\theta), S, C)$ (which can be simplified when using the more classic definition of the discrepancy function) we obtain our minimized discrepancy function $F(\hat{\Sigma}(\hat{\theta}), \hat{B}(\hat{\theta}), S, C)$. If the model is correct and fitted to S and C , $F(\hat{\Sigma}(\hat{\theta}), \hat{B}(\hat{\theta}), S, C)$ will directly yield a test statistic with an asymptotic chi-squared distribution:

$$T = (n - 1) \hat{F}(\hat{\Sigma}(\hat{\theta}), \hat{B}(\hat{\theta}), S, C) \sim \chi_{df_M}^2 \quad (9)$$

where $\chi_{df_M}^2$ is the central chi-squared distribution with degrees of freedom (df_M). The degrees of freedom for the T statistic ($df_M = p' - k$) will depend on the number of freely estimated parameters, k (see Appendix A for definition of free parameters), in our vector of

parameters θ , and the number of observed variables ($g = p + q$) in our model, $p' = \frac{g(g-1)}{2}$ [56]. T in (9) can similarly be obtained using the ML estimator, under the assumption of multivariate normality, using likelihood ratio theory in comparing a model to a saturated model. In this case, the likelihood ratio test (LRT) statistic is $-2\ln(\hat{\theta}; y) + 2\ln(\theta_{sat}; y)$, where $\ln(\theta_{sat}; y)$ is the maximum log likelihood of the saturated model [3].

Given that MLR is considered by many as a preferable estimator to ML in modern SEM, we briefly describe the correction of the formula (9) for MLR, which is asymptotically equivalent to the Yuan-Bentler T2* test statistic [13, 57]. Under MLR, T in (9) is adjusted to

$\frac{T}{c}$ for an estimated scaling factor c based on the robust asymptotic covariance matrix [57, 58]. If data are truly multivariate normally distributed, then the estimated scaling factor $c = 1.00$ and MLR will lead to the same solution as ML. However, the scaling factor c is corrected for too-fat distribution tails (> 1.00) or too-thin distribution tails (< 1.00).

Different scaling factors c in $\frac{T}{c}$ have been proposed for different estimators that are robust to non-normality, such as the Satorra-Bentler scaled chi-square test statistic for the maximum likelihood estimator with standard errors and a mean-adjusted chi-square test statistic robust to non-normality (MLM) [58].

This asymptotically chi-squared distributed test statistic in (9) (or the scale corrected tests statistics) will provide a basis for assessing model fit, and in itself tests overall model fit. The null hypothesis is that there is no difference between the proposed model and the data structure, while the alternative hypothesis is that there is a difference between the proposed model and the data structure. *Thus, a large chi-squared test with a corresponding small p-value indicates that the model does **not** fit the data.*

Commonly, studies will reject the null as the chi-squared statistic in (9) is affected by non-normality, correlation size, low power, and sample size (both too small or too large). To help

account for non-normality, the alternative scale corrected statistics as mentioned above are commonly used [13, 57, 58]. However, there may be stability issues regarding the Satorra-Bentler scaled chi-square test statistic in smaller samples [59]. Thus, the Bollen-Stine bootstrapping method [59] may be a more appropriate method in these samples for assessing normality followed by bootstrapping to assess parameter estimates. While these adjustments protect against the problems associated with nonnormality, they do not protect the model from being rejected because there is some seemingly small misspecification in it such as a slight nonlinearity in the data.

A commonly used index, Root Mean Square Error of Approximation (RMSEA) [60], is a point estimate that builds on this chi-squared statistic T but is parsimony and sample size corrected:

$$RMSEA = \sqrt{\frac{T - df}{df(n - 1)}}. \quad (10)$$

Confidence intervals can be constructed around the point estimate, because the RMSEA asymptotically follows a rescaled noncentral χ^2 distribution for a given sample size, degrees of freedom, and noncentrality parameter $\lambda = T - df$, estimated in the sample as λ reflecting the degree of misfit in the proposed model [61]. A close fit hypothesis can be tested for the model using RMSEA. There are several limitations to the fit index, namely, RMSEA may not exactly follow a non-central chi-square distribution, may be sensitive to nonnormality, and may favor larger models.

Another commonly reported fit index, the Comparative Fit Index (CFI) [62], is an incremental fit measure comparing the fit of the model to a baseline model. Let M be the substantive model as discussed above and I be the independence model made up of the diagonal elements of $\Sigma = \Sigma(\theta)$, $\sigma_{11}, \sigma_{22}, \dots, \sigma_{ww}$. Given normal asymptotic theory holds, CFI can be derived using $\lambda = \chi_M^2 - df_M$ and from the independence model I $\lambda_I = \chi_I^2 - df_I$.

$$CFI = \frac{\lambda_I - \lambda}{\lambda_I} \quad (11)$$

If the CFI index is greater than one, it is set at one and if less than zero, it is set to zero. The closer the CFI index is to one, the better the model fit. The formulas given in (10) and (11)

can be corrected for MLR based on using the statistic $\frac{T}{c}$ in place of T .

The Tucker-Lewis Index (TLI) [63] is another commonly reported incremental fit measure with a higher penalty for adding parameters than CFI, and without the zero to one range restriction. A commonly used absolute fit index, based on standardized difference between the observed correlation and the predicted correlation, is the Standardized Root Mean Square Residual (SRMR) [64].

Some general rule of thumb guidelines in SEM literature are that RMSEA ≤ 0.05 indicates an excellent fit while < 0.08 is acceptable; CFI and TLI > 0.90 are acceptable and > 0.95 are excellent fit. In addition, all three indices should reach acceptable (preferably excellent) levels before designating a model as good fitting. SRMR value ≤ 0.08 represent a good fit with the model.

We recommend, with no existing gold standard test for assessing model fit, using a hybrid approach combining both prior theory and multiple data-driven tests to address research hypotheses while best representing the data structure. This will be shown in our motivating example for determining the number of factors in sections 4.2 and 4.3.

3.7 Making Model Modifications

We can compare between different nested models (see Appendix B for definitions of nested and non-nested models) using model fit statistics such as the Chi-Square Test of Difference. Calculating the Chi-Square Test of Difference using a more robust estimator such as MLR requires additional steps involving the correction factors for the different nested models involved in the comparison [13, 57, 58]. We can also compare between models that are either nested or non-nested using Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and Browne-Cudeck Criterion (BCC) [2]. Among a series of models, the lower the value of AIC, BIC, and BCC the better the model fit. R^2 is used for model comparison between models that are either nested or non-nested but typically isn't as well-defined as it is in more traditional approaches such as linear regression. These tests of model comparison aim to minimize specification errors between an initial model and the unknown, true model characterizing the population and measures under study.

Due to computational technology, adaptive algorithms for specification searches have been introduced to the SEM literature. Specification searches are algorithms for modification of an initial model with the goal of finding the best model for the given data structure. Specification search procedures already built into software can be of great use in cutting labor time by running many model comparisons and tests at once. Specification searches have been adapted by Marcoulides and Drezner based on genetic algorithms (2001) and ant colony optimization (2003) [65, 66] and in the TETRAD program [67].

In AMOS 5.0, an exploratory specification search for the best theoretical model given an initial model uses fit function criteria from chi-square, chi-square-df, AIC, BCC, BIC, chi-square divided by the degrees of freedom and significance level [68]. Using an initial model, the algorithm will examine all possible models under examination. For example, in a path model with 10 causal paths, 5 of the paths may be “known”, while 5 of the paths may be “unknown”. These “known” paths are not optional for testing of the model; all models investigated by the algorithm will include these paths as specified. Such paths are determined according to logic, theory, and prior empirical evidence. “Unknown” paths are optional for testing of the model and thus may need to be evaluated for inclusion or exclusion in the model [50]. Potentially $2^5 = 32$ models (or some subset of the 32 models) may now need to be evaluated for a “best-fitting” model based on different potential combinations of these 5 “unknown” paths. AMOS provides a short list of only the “best-fitting” models for each combination of paths (“best-fitting” model with 2 paths, “best-

fitting” model with 3 paths, ..., “best-fitting” model with 5 paths). Based on the fit function criteria, all or some may suggest the “best-fitting” model. Even if one or more statistical criteria suggest a better model than the initial model, ultimately the researcher is left to decide whether a model with improved fit advances prior theory and is consistent with prior evidence.

A piecewise approach could be used for a model with a large number of optional paths using the exploratory specification search. The model could be broken down into smaller, more manageable pieces with fewer “unknown” paths and tested. Eventually, the researcher can test additional “unknown” paths in steps after previously optional paths have become designated as “known” paths to the model. The process is completed when a final model is identified after all optional paths have been tested by the researcher.

MacCallum et al. has demonstrated the shortcomings of specification searches when results are data-driven through simulation and real studies using goodness-of-fit indices and modification indices [69, 70]. Unless sample size is sufficiently large there can be difficulties in converging to a generalizable finding. When results are data-driven, based at least in part from fitting the model to a particular sample there is a non-trivial possibility that characteristics of the sample may influence the particular modifications that are performed. *Thus a conservative approach making few modifications with clear interpretability is certainly warranted* [69, 70].

Alternative methods for difference testing in order to determine the value in adding or dropping a causal path or multiple causal paths to a model can be done with a single specified initial model. We fix the intercepts or thresholds and parameter estimates for the causal paths we are considering including in the model, and perform invariance testing, through examining *modification indices*. Modification indices are calculated individually for every path that is fixed to zero, through use of a chi-square test statistic with one df. The higher the value of the modification index for a causal path, the better the predicted improvement in overall model fit if that path were added to the model. A modification index of four or above is statistically significant at the $\alpha = 0.05$ level. Jöreskog suggested that a modification index should be at least five before the researcher considers adding the causal path and modifying the hypothesized model [4].

Modification indices can be extended for multiple model constraints while estimating the individual modification index for each constraint, conditioned on the rest of the model, using multivariate calculus [71]. This approach can be used to test multiple intercepts, thresholds or parameters at once, where the degrees of freedom for the chi-squared test statistic will reflect the number of parameters under constraint.

While modification indices and Lagrange multipliers can be evaluated with empirical criteria as mentioned above, if the sample size is not large or the magnitude of the estimate is small, using such criteria may automatically include or exclude a path. This may not be consistent with the conceptual and theoretical grounds in the study. Therefore it is preferred to examine the actual estimate in relation to other such estimates in the model instead [2].

A manual approach to specification search would involve constraining each parameter for an “unknown” path one at a time, checking the value of the modification index. If the value is high as discussed previously, then we free the parameter. Similarly, we can repeat this process for multiple “unknown” parameters at the same time using Lagrange multipliers.

Saris et al. [72] proposed using both the modification index and the expected parameter change (EPC) simultaneously for performing model modifications. EPC represents how much the given parameter would be expected to change if it were freely estimated [6, 13]. Completely standardized EPCs allow the researcher to make relative comparisons across parameters. Parameters with large modification indices and large EPCs should be freed first for model modification provided it makes theoretical sense. Parameters with large modification indices and small EPCs or small modification indices and large EPCs might still be considered for modification if it makes theoretical sense. However, this inconsistency between the two measures may indicate a sample size issue. Finally, small modification indices together with small EPC values indicate, at least empirically, that a parameter should not be freed.

3.7 Sample Size and Power

Various rule of thumb guidelines have been derived for sample size in the SEM framework. Since SEM analysis may involve many different causal paths and estimating free parameters in a covariance matrix of large dimensions, a very large sample is typically required. For example, Kline [2] refers to a rule of thumb sample size of 200, given that a model is not too complex and that normal assumptions hold, as acceptable for a study (referred to as the “typical” SEM sample size). Further, Kline [2] discusses a separate sample size to free parameter rule, referencing Jackson [73], that the ratio of sample to parameters should be ideally 20:1, with 10:1 acceptable but less than ideal. For a more data driven approach, Monte Carlo simulations have been used for both sample size calculations and power [74]. This also involves context, as, for example for performing factor analysis we can calculate power based on how large a sample would be necessary to detect two factors rather than one, if indeed the factor structure in our sample was two factors. For mediation analysis, the sample size or power might be derived under simulation generated under full mediation, where the direct effect is equal to zero, vs. a suitable alternative effect size to be considered for the direct effect.

MacCallum introduced an alternative type of power analyses for covariance structure models using RMSEA [75]. A web page has been set up by Preacher et al. that generates R code that can perform power analysis using RMSEA [76].

Satorra first introduced a method for evaluating the power of the likelihood ratio test for SEM [77]. However, he later introduced a more straightforward method for power analyses using modification indices [78]. Modification indices can be compared to a table of the noncentral chi-square distribution with one degree-of-freedom in order to assess power. The use of modification indices for power analysis has been extended for use in tandem with EPCs by Kaplan [79] and Saris et al. [72].

Multivariate methods of model modification (for two or more parameters, simultaneously) have been developed for power analysis using a multivariate Wald test [80]. While multivariate methods minimize Type I or Type II errors incurred by model modification relative to univariate approaches, one does so at the expense of monitoring individual changes that might detract from the interpretation of the model. Kaplan and George [81] evaluated power in the multiple group confirmatory factor analysis setting using the Wald test.

3.8 Identifiability

SEM estimation proceeds by finding the parameter (θ) which best fits the observed data. Unfortunately, for some models, multiple possible values of θ will fit the data equally well. This is caused by a condition known as non-identifiability where multiple possible values of θ lead to the exact same distribution of observed variables. For example, consider the mediation model in figure 1. If the error terms for *cognitive decline* (ϵ_z) and *depression* (ϵ_y) are correlated, where $\text{corr}(\epsilon_z, \epsilon_y) > 0$, or if a reciprocal relationship exists between *cognitive decline* and *depression*, such that *cognitive decline* $\rightarrow$ *depression* and *depression* $\rightarrow$ *cognitive decline*, then the structural equation model corresponding to figure 1 would not be identifiable. In this case, there would be too many unknown parameters to estimate, and not enough known information to estimate these parameters. One way to make such a model identifiable would be with the use of instrumental variables, as discussed in the econometrics literature [82] which also meets the assumption of sequential ignorability as discussed in the causal inference literature [45, 83]. There are a number of rules which can be used to determine the identifiability of a model. Typically, for a non-identifiable model, the second derivative of the fitting function with respect to the parameters $d^2F/d\theta^2$ is numerically not of full rank leading the SEM software to issue a warning or error. Such warnings are provided in the MPlus software that may help to diagnose identifiability problems.

Consider a structural equation model with k unknown parameters and $p' = \frac{g(g-1)}{2}$ is the number of entries in the covariance matrix of the observed variables. The model is *underidentified* if $k > p'$. As a result, the model has the problem of infinite best fitting solutions and cannot be identifiable. On the other hand, when $k = p'$, the model, although not guaranteed, could be identifiable. An identifiable model with $k = p'$ is called just identified or *saturated*. Since saturated models will fit the data exactly, tests of model fit as in section 3.6 no longer can yield useful results. The preference in using SEM is for an *overidentified* model, where $k < p'$. Overidentified structural equation models have positive degrees of freedom, thus providing some meaning for tests of model fit in order to determine if the model fits the data structure well.

SEM researchers can explicitly, if theoretically appropriate, make an underidentified or just identified model overidentified by using information to make more parameters known. For example, without loss of generality if a mean is known to be zero, and the researcher is estimating that mean in the model, then that parameter can be constrained to zero, increasing model degrees of freedom. This will also be shown in our motivating example analyses in

which we constrain certain DIF paths to zero, in order to find a more parsimonious, best fitting solution.

A problem related to non-identifiable models is the result of multiple different models being *observationally equivalent*. *Observationally equivalent models* yield the same predicted correlation structure and identical model fit statistics and indices [2], but a very different interpretation of the data. The space of all possible models may be divided up into sets of observationally equivalent models. SEM may be used to decide which set of equivalent models fits the data best, but it cannot help to distinguish which of the equivalent models fits the data best. Thus, logic, theory and prior evidence must be used to make a decision about which model to use. For example with two variables X and Y, it is impossible to distinguish using the data alone whether X causes Y, Y causes X or if some third variable causes them both. The decision to choose one model above another must be based on theoretical reasoning regarding the subject matter. In the context of a particular study, we may hypothesize drinking intensity causes depression, since our prior research in a population of alcohol dependent individuals has shown that depression is a potential mediator of the relationship between drinking intensity and suicidal ideation [35]. The existence of such observationally equivalent models has damaging consequences for the success of purely algorithmic model specification searches (see section 3.7).

3.9 Multigroup analyses

Researchers are often interested if groups differ across populations. In SEM nomenclature, *group invariance* denotes that results are the same across groups [2, 6]. Showing group invariance may be important for proving model stability and is a reason for performing multigroup analyses. Similarly, we may use multigroup analyses to test for *moderation* (i.e. interaction). In terms of multigroup analyses in SEM, moderation would imply that different subgroups influence the strength of the relationships in the model.

In performing multigroup analyses, our objective would be to compare between nested models to determine if we have group invariance to either prove model stability or to test for moderation. Given the complexities of SEM analyses, there may be different parts of the model of interest for multigroup analyses and testing for group invariance. A researcher may be interested in factor loadings and intercepts only, or specific parameter estimates or correlation parameters for group invariance, depending on the specific research question of interest. In our motivating example, interest is in subgroup measurement differences by age, sex (male or female), race, MS type (relapsing or progressive) and baseline time since diagnosis. Since in our example we are testing such subgroup measurement differences one at a time, we use a Bonferroni correction to control the family-wise error rate. Groups are compared using the model fit statistics and indices (i.e. χ^2 , RMSEA, CFI, TLI and SRMR) for each subgrouping between two models (i.e. MALE model and FEMALE model). First an unconstrained model is specified freely estimating all parameters across groups. Thus, for example, we would have unique model estimates for all parameters for the subset of the data that is MALE and unique model estimates for all parameters for the subset of the data that is FEMALE. Then, we would specify a constrained model in which we would be constraining the estimates for all factor loadings and item intercepts in the two measurement models to be

equal across groups. Therefore, if there is a subgroup difference, the model fit would be significantly worse under the constraint of group equality.

3.10 Limitations of SEM

The specified model in an application of SEM must be plausible to obtain meaningful results. There is no model magic that allows a researcher to automatically assume causality when using SEM to perform causal inference. Causal assumptions should be based on strong scientific theory and prior evidence. For example, temporal order of variables must be known. Application of SEM also should include initially a thoughtful determination of which model fitting method would be appropriate for the data as discussed in Section 3.5. For example it is very common in SEM to use an estimator that is robust to non-normality (i.e. MLR). SEM requires a large sample size. However, it is difficult to determine exactly what a large sample is in any given situation. As previously discussed in Section 3.7, the number of parameters vs. sample size is an important consideration.

As discussed previously in Section 3.8, model identifiability is an issue that often arises when performing SEM analyses. Similarly, there may be multiple equivalent models that fit the data equally well. Again, there is no statistical inference, only scientific theory and prior evidence, that will allow a researcher to choose between these equivalent models. Further, there are some similar limitations with SEM as with traditional methods, as covariance and correlation matrices analyzed may be influenced by missing data and outliers. A researcher needs to take such issues into account when performing any statistical inference. A realistic notion of what can be concluded from model fit statistics and indices, since there is no “gold standard”, is that we won't be able to prove a model, but that discrepancy between model and data doesn't surpass limits of chance. However, if the discrepancy does exceed what would be expected by chance, SEM provides methods for estimating the size of the discrepancy.

4. Factor Analysis

4.1 Common Factor Model

Since we have provided a more general overview of the SEM framework in section 3, we now begin our discussion of the specific techniques within the SEM and more general latent variable framework for analysis of overlapping symptoms in co-occurring conditions. Factor analysis is a technique for reducing the number of measures under study to a smaller number of factors and detecting a structure in the relationships between measures [2, 3, 6, 8, 13]. For example, the nine items of the PHQ-9 have been hypothesized to show an underlying trait of depression [40, 41]. It also has been hypothesized that these nine items break up into two factors representing somatic and affective domains of depression [84].

We express the common factor model:

$$\mathbf{Y} = \boldsymbol{\mu}_F + \boldsymbol{\Lambda}_F \mathbf{F} + \boldsymbol{\delta} \quad (12)$$

Here $\mathbf{Y}$ is a vector of dimension $p \times 1$. $\boldsymbol{\mu}_F$ is a $p \times 1$ vector of common factor intercepts, $\boldsymbol{\Lambda}_F$ is a $p \times m$ matrix of common factor loadings, while $\mathbf{F}$ is a $m \times 1$ vector of common factors. The vector of unique factor terms (i.e. disturbance terms) $\boldsymbol{\delta}$ is of dimension $p \times 1$ and is assumed to be uncorrelated with $\mathbf{F}$. The covariance matrix of factor models has a very simple form. Let $\text{Var}(\boldsymbol{\delta}) = \boldsymbol{\Delta}$. Then:

$$\text{Var}(\mathbf{Y}) = \boldsymbol{\Lambda}_F \text{Var}(\mathbf{F}) \boldsymbol{\Lambda}_F^T + \boldsymbol{\Delta} \quad (13)$$

We typically assume $\boldsymbol{\Delta}$ is a diagonal matrix with $\gamma_1, \dots, \gamma_p$ in the diagonal.

The model is not identifiable without additional constraints, as discussed in section 3.8. Thus, there are an infinite number of parameters which fit the data equally well. To make the model identifiable we first need to fix the “scale” of the latent factors $\mathbf{F}$. There are two common approaches to doing this. First, we can fix $E(F_i)=0$ and $\text{Var}(F_i)=1$, in which case the latent factors are measured on a scale with mean 0 and variance 1. Second, we can fix $E(F_i)=0$ and fix one of the factor loadings as a reference loading to one, in which case the factor is measured on the same scale as the reference variable. Factor loadings can be interpreted, for each item, as leading to the shared variance between that item and each underlying latent factor. Standardizing factor loadings gives a more straightforward indication of the magnitude of strength of the shared variance, on the same scale as a correlation coefficient. After standardization, both methods to identify the model then lead to the same standardized factor loadings. In a single factor model with at least three indicators, fixing the scale makes the model identifiable. However, fixing the scale may still not be enough to obtain identifiable multiple factor models. See Bollen [3] for more detailed rules for determining identifiability of factor analytic models.

There are broadly speaking two different types of factory analytic approaches which handle the identifiability problem. *Confirmatory factor analysis (CFA)* [6] is a technique within the SEM framework that handles the identifiability problem by fixing some of the loadings to 0. Typically using CFA, the observed variables are assumed to be indicators of only one latent factor, so each variable will load on only that factor. On the other hand, when using *exploratory factor analysis (EFA)* [8, 85] there are an infinite number of different model parameters which fit the data equally well. EFA is also not typically an SEM procedure, but is a more general latent variable procedure. Therefore, using EFA a researcher aims to choose the set of factor loadings which leads to the most meaningful substantive interpretation of the underlying factors. The techniques for choosing the most meaningful parameters are known as factor rotation techniques.

Rotation has a geometric interpretation in that a rotation can be viewed as equivalent to a change of the positioning of the axes of the space spanned by the latent factors. While all rotations fit the data equally well (i.e. all rotations have the same covariance structure), some rotations are more interpretable than others. Rotation techniques start with an initial solution and then change the direction of the initial factors so as to optimize a particular function that reflects distance to what is referred to as a ‘simple structure’. This ‘simple structure’ aims for a more meaningful substantive interpretation [85]. For example, the varimax rotation

optimizes a function that leads to some factor loadings being large and other factor loadings being small for more meaningful interpretation of individual factors as loading more heavily on some items than others. Varimax is an example of an *orthogonal rotation* technique. In orthogonal rotations the original factor space is rotated while there perpendicularity is preserved. For an orthogonal rotation, we assume factors are uncorrelated. Fixing $\text{Var}(\mathbf{F}) = \mathbf{I}$, equation (13) simplifies to $\text{Var}(\mathbf{Y}) = \mathbf{\Lambda}_F \mathbf{\Lambda}_F^T + \mathbf{\Psi}$.

In oblique rotation the new axes are allowed to cross at an angle different than 90 degrees, allowing an interrelationship between factors. In other words, factors can be correlated. In a lot of research, and commonly in the social sciences, the assumption of correlated factors is more reasonable than independent factors. For example, we would not assume *somatic* and *affective* domains of the PHQ-9 to be completely independent, since they both capture aspects of *depression*. However, a very high correlation between factors ($\rho = 0.80$) may be problematic in that it implies poor discriminant validity and suggests that a more parsimonious factor solution is preferred [6, 86]. Thus, the two factors could be combined into one more parsimonious factor. Alternatively, we could use a second order CFA model. In this case, the two *somatic* and *affective* factors would be the second order factors that could be used to form one first order factor for *depression*.

4.2 Determining the Number of Factors using EFA

A common application for EFA is to deduce how many factors are in the data. For EFA, we assume $\mathbf{\Gamma}$ in (13) is the diagonal matrix with $\gamma_1, \dots, \gamma_p$ in the diagonal. While we outline objective rules for EFA, it is important to note that, factor analysis in general, involves both theory and data analyses. While some items may clearly load onto a factor, commonly items either cross-load onto multiple factors, or do not have a strong relationship with any factor (standardized factor loading < 0.40). Also, interpretation for two or more factors is dependent on theory (i.e. somatic and affective domains of depression). The magnitude and clarity of the factor loadings along with the interpretation often helps facilitate, beyond any statistical inference, in deciding upon the factor structure of a measure of interest.

In EFA we look for the effective number of common factors (minimum value of m) [85]. Some general modern rules of thumb that could be used in the process of determining m are as follows:

- All primary factor loadings should have a standardized estimate > 0.40 (with no secondary factor loading > 0.30). If these conditions are not met, remove problematic variables one at a time and re-run the series of models.
- The residual correlation matrix should have a mean of all residual correlations < 0.05 . The residual correlation matrix is the difference between the implied correlation matrix and the observed correlation matrix.
- Conservatively, if the correlations among any two factors is > 0.80 , consider a more parsimonious solution (i.e. combining factors).
- Review the correlation matrix between items that make up a factor and evaluate for patterns of high correlation among the items.

- m is determined by the number of factors above the elbow in a scree plot.
- The latent factor should make sense in theory (i.e. *depression* for the PHQ-9 items).
- In the SEM framework, we also look for this solution of m through examining model fit indices.

In order to determine the number of latent factors in the PHQ-9, we examine the eigenvalues, which represent the variance accounted for by each underlying factor, in a scree plot along with model fit criteria. The number of eigenvalues > 1 represent unique latent factors according to Kaiser's rule [87]. However, sampling variability may produce eigenvalues > 1 even if all eigenvalues of a correlation matrix are exactly one and no large components exist [88].

We therefore also use parallel analysis [88, 89] for further analysis based on the eigenvalues. In parallel analysis, eigenvalues are computed from a random dataset with the same numbers of observations and variables as the original dataset, which in our case is the Knowledge Program data. When the eigenvalues from the random data are larger than the reported eigenvalues for the factor analysis, then the factors are mostly random noise. In MPlus, the PARALLEL= option can be added to the ANALYSIS section, where in this case specifying PARALLEL=50 means simulating 50 random datasets for our parallel analysis, and recording the average eigenvalues over those 50 datasets. Eigenvalues are relatively robust from dataset to dataset, and therefore we do not require a large number of random datasets for sound inference.

Exploratory factor analysis (EFA) of the nine items of the PHQ-9 is used to determine how many latent factors represent aspects of *depression*. We use an oblique (geomin) rotation.

The scree plot (Figure 3) was output from MPlus and showed one latent construct (highest eigenvalue = 4.82, no other eigenvalues > 1) for the PHQ-9 within this MS population. As shown in Figure 3, the number of factors above the elbow in this scree plot is one. The parallel analysis on the same plot also indicates one factor, as only one of the random eigenvalues from the parallel analysis is plotted below an eigenvalue from the sample under study in Figure 3. The analysis based on the eigenvalues suggests one latent factor for *depression* from the PHQ-9.

However, the model fit statistics and indices tell a different story regarding the factor structure in this analyses. The two factor model showed improvement over the one factor model (table 1) providing a mixed message of whether the one or two factor model was a better fit. No other factor structure, i.e. three or more factors, had intelligible factor loadings on the items. Thus we could rule out these solutions. For example, in the three factor solution, despite further model fit improvement, there did not seem to be a clear theoretical interpretation of how the items loaded onto the three factors. Also, item 2 contained a factor loading greater than one (1.328) on one of the factors. A factor loading greater than one may signify an estimated error variance (residual variance) that is negative, which would mean that too many factors have been extracted. However, in this case, there were no negative residual variances. The factor loading greater than one can also occur with correlated factors

[90], which then leads to standardized factor loadings that do not have the interpretation as correlation coefficients.

Another rule of thumb guideline as mentioned, is that a primary standardized factor loading should be ≥ 0.40 , meaning that the latent factor explains at least $(100 \times .4^2)\% = 16\%$ of the variance in the item, in order to attribute an item to a factor. Therefore, in Table 2, both the 1-factor and 2-factor models have strong factor structures according to this criteria. We look to use CFA to help decide upon the number of factors.

4.3 Confirmatory Factor Analysis

As mentioned earlier, in *confirmatory factor analysis* (CFA) we pre-specify the factor structure to make the model identifiable. There may be correlation among error terms when using CFA, which will help us determine a factor structure given our mixed evidence as discussed in section 4.2. If we specify this non-zero residual correlation, then the matrix in (13) will have off-diagonal terms (i.e. corresponding to the error term correlation between PHQ-9 items for sleep and fatigue).

In examining the two potential factors in a CFA model, items for anhedonia, feel depressed, feelings of failure and self-harm load on one factor while items for sleep problems, fatigue, appetite change, poor concentration and psychomotor symptoms load on a second factor. This is close to the cognitive and somatic interpretation of the factor structure for the PHQ-9. However, the correlation between the two factors is very large ≈ 0.86 , which implies poor discriminant validity and suggests that a more parsimonious single factor solution is preferred [6, 86].

We further use prior theory and examine the completely standardized EPC values for each modification index for the covariance between error terms for each pair of PHQ-9 items. Our prior theory suggests that there may be strong pairwise relationships between related fatigue items (sleep problems and fatigue) and cognitive and functional impairment items (poor concentration and psychomotor symptoms) in line with our model for overlapping symptoms. The completely standardized EPC and modification index (MI) were also reasonably high for these residual correlations (sleep problems and fatigue EPC = 0.331, MI = 207.32 and poor concentration and psychomotor symptoms EPC = 0.213, MI = 88.16). Another very high completely standardized EPC also had a theoretical basis. Feel depressed and feelings of failure (standardized EPC = 0.323, MI = 136.15) are both affective symptoms of depression. Items for anhedonia and feel depressed also had a fairly high completely standardized EPC = 0.280 (MI = 94.18) and correlations between them could arguably be included as well. However, to keep our model parsimonious and to avoid correlation structures that are inconsistent (i.e. feel depressed is correlated with both feelings of failure and anhedonia, but anhedonia and feelings of failure are not correlated) we did not include this residual correlation.

The model fit statistics and indices are very similar for the one factor CFA model specifying these three pairwise error term covariances as the two factor CFA model (see Table 1). We now use these results to express the CFA measurement model for *Depression* in matrix notation, as expressed in the form of the common factor model in (12):

$$\mathbf{Items} = \boldsymbol{\mu}_D + \boldsymbol{\Lambda}_D \text{Depression} + \boldsymbol{\delta} \quad (14)$$

$$\mathbf{Items} = \begin{bmatrix} \text{Anhedonia} \\ \text{Depressed} \\ \text{Sleep} \\ \text{Fatigue} \\ \text{Appetite} \\ \text{Guilt} \\ \text{Concentration} \\ \text{Psychomotor} \\ \text{Self - Harm} \end{bmatrix}, \boldsymbol{\mu}_D = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_9 \end{bmatrix}, \boldsymbol{\Lambda}_D = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_9 \end{bmatrix}, \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_9 \end{bmatrix}.$$

$\delta_1 \dots \delta_9$ are the disturbance terms. **Items** is a vector of the 9 observed items for the PHQ-9 instrument. *Depression* is a unidimensional unobserved latent variable formed from the observed endogenous variables **Items**. The equations for Items includes a vector of factor intercepts, $\boldsymbol{\mu}_D$, and factor loadings, $\boldsymbol{\Lambda}_D$, both of dimension 9×1 .

Further, assuming $\text{Var}(\delta_i) = \gamma_i$ and $\text{Cov}(\delta_i, \delta_j) = \gamma_{ij}$ and $\text{Var}(\text{Depression}) = 1$, then

$$\text{Var}(\mathbf{Items}) = \boldsymbol{\Lambda}_D \boldsymbol{\Lambda}_D^T + \text{Var}(\boldsymbol{\delta}) \quad (15)$$

where,

$$\text{Var}(\boldsymbol{\delta}) = \begin{bmatrix} \gamma_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma_2 & 0 & 0 & 0 & \gamma_{26} & 0 & 0 & 0 \\ 0 & 0 & \gamma_3 & \gamma_{34} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma_{34} & \gamma_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma_5 & 0 & 0 & 0 & 0 \\ 0 & \gamma_{26} & 0 & 0 & 0 & \gamma_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma_7 & \gamma_{78} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma_{78} & \gamma_8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \gamma_9 \end{bmatrix} \quad (16)$$

Using above notation from (13), $\text{Var}(\mathbf{Items}) = \boldsymbol{\Lambda}_D \boldsymbol{\Lambda}_D^T + \text{Var}(\boldsymbol{\delta})$.

We note that (17-19) is equivalent to a multiple factor solution with equal factor loadings. However, this solution is more parsimonious than a two-factor solution and maintains the unidimensional interpretation of a one factor solution.

4.4 Cross-loaded modeling

The most general SEM-based model for analyses of overlapping symptoms involving latent constructs is the cross-loaded model. This type of model can be categorized as a multi-factor CFA model with cross loaded items.

Without loss of generality, we discuss the simple cross-loaded model from figure 4 with three observed indicators of latent factor A , representing symptoms of Condition A, and three observed indicators of latent factor B , representing symptoms of Condition B. The potential overlapping symptoms of A and B are modeled through cross-loaded items (Item 3 and Item 1*). For an example, A can represent items in a depression screening scale for an underlying latent factor of depression and B can represent a multi-item MS-related fatigue scale for an underlying latent factor of MS-related fatigue with cross loading on potential items of overlap between the two scales. Figure 4 can be extended for more items, specified correlations between items as in our motivating example, latent factors, and cross-loaded items. The iterative process is to use EFA to determine the factor structure for the different measures under study (or in conjunction with CFA as in our motivating example), specify a cross loaded model based on hypothesis of theoretical relevance and perform analyses of the overlapping symptoms. A special case of this cross-loaded model with only one latent factor for Condition A (depression) and with observed covariates representing symptoms of Condition B (MS) is shown in our motivating example.

5 Multiple Indicator Multiple Cause (MIMIC) model

5.1 MIMIC Model

The multiple indicator multiple cause (MIMIC) model is another important special case of SEM. It is a measurement model (i.e. factor model) with observed covariates [2, 6, 7]. The observed covariates explain the latent construct in the MIMIC Model. The general form of a MIMIC model:

$$\begin{aligned} y &= \mu_y + \Lambda_y \eta + \varepsilon, \\ \eta &= \mu_\eta + \mathbf{B}\eta + \Gamma \mathbf{x} + \zeta \end{aligned} \quad (17)$$

Using matrix algebra:

$$y = \mu_y + \Lambda_y (\mu_\eta + \mathbf{B}\eta + \Gamma \mathbf{x} + \zeta) + \varepsilon = \mu_y + \Lambda_y \mu_\eta + \Lambda_y \mathbf{B}\eta + \Lambda_y \Gamma \mathbf{x} + \Lambda_y \zeta + \varepsilon = \mu_\eta^* + \mathbf{B}^* \eta + \Gamma^* \mathbf{x} + \varepsilon^* \quad (18)$$

where $\mu_\eta^* = \mu_y + \Lambda_y \mu_\eta$, $\mathbf{B}^* = \Lambda_y \mathbf{B}$, $\Gamma^* = \Lambda_y \Gamma$, and $\varepsilon^* = \Lambda_y \zeta + \varepsilon$. It follows that, if we assume multivariate normality for $\zeta \sim N(\mathbf{0}, \Psi)$ and $\varepsilon \sim N(\mathbf{0}, \tau)$, and ζ is independent of ε :

$$\begin{aligned} E(\varepsilon^*) &= \Lambda_y E(\zeta) + E(\varepsilon) = 0, \\ \text{Var}(\varepsilon^*) &= \Lambda_y \text{Var}(\zeta) \Lambda_y^T + \text{Var}(\varepsilon) + \Lambda_y \text{Cov}(\zeta, \varepsilon) = \Lambda_y \Psi \Lambda_y^T + \tau \end{aligned} \quad (19)$$

In a research question, we might be interested in only one overlapping symptom from a single item scale (i.e. performance scales fatigue domain).

However, in Figure 5 we show the path diagram representing the MIMIC Model for our motivating example before accounting for any overlapping symptoms between depression and MS, where multiple observed covariates describing MS symptoms are associated with the latent construct for *depression*. These observed covariates are also correlated with each

other. While a traditional MIMIC Model assumes the observed covariate causes the latent construct, given strong prior theory, a correlation should be used if no causal ordering is clear. Thus, our model is a modification of the traditional MIMIC model.

5.2 Conceptual model of overlap of depression with other symptoms of multiple sclerosis

In performing analysis of overlapping symptoms in co-occurring conditions, a prior hypothesis based on strong theory is essential in determining the conceptual model. Here we explain in our motivating example how to form our conceptual model based on a priori information. This information will then be used to extend the corresponding MIMIC model for figure 5 to account for overlapping symptoms between depression and MS.

PHQ-9 items for sleep problems and fatigue address fatigue. Attempting to separate fatigue as a depressive symptom from MS-related disability in the PHQ-9 is likely problematic for a patient due to the multiple dimensions, such as physical, mental and emotional, that may be involved in describing fatigue symptoms [91]. The PHQ-9 addresses concentration and speed of movement and speech. Concentration problems that are perceived as memory problems stemming from cognitive impairment in MS can potentially be impaired concentration due to comorbid depression [92]. Further, functional impairment could potentially be associated with concentration problems. Cognitive impairment, functional impairment and depression in MS can potentially lead to difficulties in movement and speech. Thus, potentially items for poor concentration and psychomotor symptoms on the PHQ-9 overlap with MS-related cognitive impairment and functional impairment.

Therefore based on prior theory, we hypothesize that PHQ-9 items for sleep problems, fatigue, poor concentration and psychomotor problems have the potential to overlap with symptoms described by our MS disability scales. We draw a path diagram for the specified model, to provide a visual imagery of the hypothesized relationships. Based on the factor analytic results as described in sections 4.2 and 4.3, we use a unidimensional depression screening scale with item correlations.

5.3 Differential Item Functioning (DIF) in a MIMIC model

An important feature of the MIMIC model given our analyses of overlapping symptoms, is that it permits detection and adjustment for differential item functioning (DIF) [7]. Here we refer to DIF in the classic, IRT sense where people from different groups (e.g., levels of MS-related fatigue) with the same latent trait (level of depression) have a different probability of giving a certain response on a questionnaire or test (e.g., items for sleep problems and fatigue of the PHQ-9). Thus, we extend the MIMIC model in figure 5 to the model in figure 6 accounting for overlapping symptoms between depression and MS by including the DIF paths. In fact, the model corresponding to figure 5 is simply the model in figure 6 with all the DIF paths constrained to zero (or removed). We will refer to the model corresponding to figure 6 in this analyses as the MIMIC model with DIF paths. Further, the MIMIC model with DIF paths is a special case of the multi-factor CFA model with cross loaded items, since we have only one factor and then observed covariates instead of multi-factors.

We developed the factor structure for *Depression* with residual correlations in the CFA model before adjusting for DIF paths with MS disability measures. In our analyses the residual correlations are statistically significant in the MIMIC model with DIF paths. Thus, this is evidence that the same residual correlation structure is appropriate for the MIMIC model after adjustment for the DIF paths.

While Item Response Theory-based Approaches are also used for problems of DIF, there are unique advantages to the MIMIC modeling approach within the context of our motivating problem. We can express *depression* as a latent variable with measurement error, which is likely more accurate for this MS population than the summed PHQ-9. While an IRT-based approach can do this, within *depression*, multiple factors or correlated items are easily modeled using a MIMIC modeling based approach. A more robust estimator, MLR, is readily available in MPlus for valid inference with the skewed PHQ-9 items within this analyses. Covariates, such as the patient reported outcomes and other MS disability measures in this study, may be continuous. Model fit statistics and indices and modification indices are readily available to provide evidence for modifying our model. Very importantly for practical clinical use, the MIMIC modeling approach can be used to create an adjusted scale, free of symptom overlap based on factor scores, as will be shown in section 6.

A limitation to using the MIMIC model is that they test for uniform, but not nonuniform DIF [7]. Uniform DIF is constant across *depression*, while nonuniform DIF varies across *depression*. We address this limitation through a multigroup analyses for subgroup measurement differences of the MIMIC model within the multigroup CFA framework which we discuss in Section 5.4.

Extending the 1-factor CFA model to the MIMIC model with DIF paths in our study, we can express the corresponding model from figure 6 in matrix notation to model the overlapping symptoms between depression and MS via DIF paths A, B, C, D, E and F. Following the general expression for the MIMIC model in (18) the MIMIC model with DIF paths and correlation (rather than causation) between the covariates and the latent outcome simplifies to:

$$\text{Items} = \mu_D^* + \mathbf{B}^* \text{Depression} + \mathbf{\Gamma}^* \mathbf{MS} + \epsilon^* \quad (20)$$

$$\begin{aligned}
\mathbf{Items} &= \begin{bmatrix} Anhedonia \\ Depressed \\ Sleep \\ Fatigue \\ Appetite \\ Guilt \\ Concentration \\ Psychomotor \\ Self - Harm \end{bmatrix}, \boldsymbol{\mu}_D^* \\
&= \begin{bmatrix} \mu_1^* \\ \vdots \\ \mu_9^* \end{bmatrix}, \mathbf{B}^* \\
&= \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_9 \end{bmatrix}, \boldsymbol{\Gamma}^* = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ A & 0 & 0 & 0 \\ B & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & C & E(TW) & E(PT) \\ 0 & D & F(TW) & F(PT) \\ 0 & 0 & 0 & 0 \end{bmatrix}, \mathbf{MS} \\
&= \begin{bmatrix} MS \text{ Fatigue} \\ MS \text{ Cognitive} \\ Timed Walk \\ Peg Test \end{bmatrix}, \boldsymbol{\varepsilon}^* \\
&= \begin{bmatrix} \varepsilon_1^* \\ \vdots \\ \varepsilon_9^* \end{bmatrix}
\end{aligned}$$

A through F in $\boldsymbol{\Gamma}^*$ are the parameter estimates for the DIF paths A through F in Figure 1 (TW = Timed Walk and PT = Peg Test) and $\varepsilon_1^*, \dots, \varepsilon_9^*$ are the disturbance terms. **Items** is a vector of the 9 observed items for the PHQ-9 instrument. *Depression* is a unidimensional unobserved latent variable formed from the observed endogenous variables **Items**. The equations for **Items** includes a vector of factor intercepts, $\boldsymbol{\mu}_D^*$, and factor loadings, $\mathbf{B}^*$, both of dimension 9×1 .

Further, assuming $\text{Var}(\varepsilon_i^*) = \gamma_i$ and $\text{Cov}(\varepsilon_i^*, \varepsilon_j^*) = \gamma_{ij}$ and $\text{Var}(\text{Depression}) = 1$, then

$$\text{Var}(\mathbf{Items}) = \mathbf{B}^* \mathbf{B}^{*T} + \boldsymbol{\Gamma}^* \text{Var}(\mathbf{MS}) \boldsymbol{\Gamma}^{*T} + \mathbf{B}^* \text{Cov}(\text{Depression}, \mathbf{MS}) \boldsymbol{\Gamma}^{*T} + \text{Var}(\boldsymbol{\varepsilon}^*) \quad (21)$$

where,

$$\text{Var}(\epsilon^*) = \begin{bmatrix} \gamma_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma_2 & 0 & 0 & 0 & \gamma_{26} & 0 & 0 & 0 \\ 0 & 0 & \gamma_3 & \gamma_{34} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma_{34} & \gamma_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma_5 & 0 & 0 & 0 & 0 \\ 0 & \gamma_{26} & 0 & 0 & 0 & \gamma_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma_7 & \gamma_{78} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma_{78} & \gamma_8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \gamma_9 \end{bmatrix} \quad (22)$$

and the variance-covariance matrix for **MS** is as follows:

$$\text{Var}(\mathbf{MS}) = \text{Var} \begin{bmatrix} MS \text{ Fatigue} \\ MS \text{ Cognitive} \\ Timed \text{ Walk} \\ Peg \text{ Test} \end{bmatrix} = \begin{bmatrix} \kappa_{11} & \dots & \kappa_{14} \\ \vdots & \ddots & \vdots \\ \kappa_{14} & \dots & \kappa_{44} \end{bmatrix}. \quad (23)$$

Also,

$$\text{Cov}(\mathbf{Depression}, \mathbf{MS}) = [\varpi_1 \dots \varpi_4]. \quad (24)$$

Now, using robust maximum likelihood methods as described in section 3.5, we can estimate the parameters in this MIMIC model with DIF paths. The MIMIC model from figure 5, in which paths A through F are constrained to zero, is a more parsimonious model than the MIMIC model with DIF paths.

5.4 Establishing overall DIF in a MIMIC model

In performing DIF analyses for overlapping symptoms of co-occurring conditions, the first step is to establish the existence of an overall DIF effect. This can be done in our study by comparing the MIMIC model with DIF paths to the MIMIC model without DIF paths using model fit statistics and indices as discussed in section 5.3. In a more general cross-loaded model, similarly, this would be accomplished by comparing a model with DIF paths for all cross-loaded items, to a model constraining all such DIF paths to zero.

As noted by Woods et al. [7] the comparison between these two models is a bit unrealistic as it is very unlikely to have no DIF at all. However, such an analyses is important when considering overlapping symptoms, since any DIF is leading to bias in our measures. Thus, when correcting measures through adjusted scales, the amount of DIF will be reflected in how much adjustment is made to the original scale. If the amount of DIF is so small that it can be attributed to measurement error or noise, then we might not make any changes to the original scale.

In Table 3, the fit criteria (as marked by an increase in CFI and TLI, and decrease in χ^2 , RMSEA and SRMR) showed that the MIMIC model with the DIF paths (figure 6) is a better fitting model, and overall a relatively good fit for the data structure using the criteria as

described in section 3.6. These results were verified by the χ^2 Difference Test comparing the two models ($p < 0.001$).

Estimates for the standardized factor loadings for *depression* are lower in the MIMIC model with DIF paths compared to the model without DIF paths for the items for sleep problems, fatigue, poor concentration and psychomotor symptoms (Table 4). The decrease in these estimates signifies less shared variance between each of these four items and *depression*.

The standardized estimates of the DIF paths with the exception of E are statistically significant. The magnitude of the estimates is large for the overlap of the item for fatigue with MS-related fatigue ($B=0.51$), medium for the item for sleep problems with MS-related fatigue and the items for poor concentration and psychomotor symptoms with MS-related cognitive impairment ($A=0.21$, $C=0.38$ and $D=0.23$) and small for the item for psychomotor symptoms with the 9-hole peg test aspect of MS-related functional impairment (F for peg test= 0.11).

As a result of the statistical evidence in this section, we find overlap of *depression* with other symptoms for MS patients that may be confounding a patient's PHQ-9 score. Simultaneously, the MIMIC model with DIF paths is estimating the overlap via DIF paths and then correcting *depression* by adjusting the factor loadings.

5.5 Model modification in a MIMIC model with DIF

Once overall DIF has been established, there are two ways to proceed, forward or backwards. These approaches will determine how each individual DIF relationship (i.e. in our example A through F) influences model fit in isolation, in order to determine which specific overlapping symptoms in a hypothesized model of overlap are significant. The more standard approach would be to proceed forward, to free the DIF path with the largest modification index and then check if model fit improves over the MIMIC model without DIF paths. If model fit improves, then repeat this procedure now freeing the DIF path with the next largest modification index and then repeat until the best fitting model has been established.

Our manner of model modification for this study takes the backwards approach of looking for evidence to drop paths, which may be more fitting for a problem of overlapping symptoms or test findings. Our hypothesized model in figure 6 has included particular DIF paths of theorized overlapping symptoms between depressive symptoms and other symptoms for MS patients, regardless of data structure. Therefore, we are looking for evidence to drop one of these paths, not to build up to this model.

We constrain to zero (or remove) one DIF path at a time in examining the separate models with A, B, C, D, E and $F=0$ for improvement over our MIMIC model with all DIF paths. From Table 3, evaluating model fit indices and statistics, we observe improvement over the MIMIC model with DIF paths only for $E=0$ (where both $E(TW)$ and $E(PT)$ are constrained to zero at the same time). To illustrate the difference of $E=0$ from (20), we set $E=0$

$$\mathbf{\Gamma}^* = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ A & 0 & 0 & 0 \\ B & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & C & 0 & 0 \\ 0 & D & F(TW) & F(PT) \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (25)$$

The chi-square test of difference is not significant, thus indicating that the more parsimonious model E=0 is a better fit.

We examine the two modification indices simultaneously for E and F. For path E, both our estimates of the modification indices are less than five. Bollen and Long, as applied by Sudhahar et al., state that the appropriate procedure is to remove all the DIF paths that are statistically not significant where changes in model fit are not observed [93, 94]. In table 3 we see similar model fit for E=0 and the full MIMIC model and in table 4 we see that the DIF paths for E in the full MIMIC model are close to zero and not significant ($p > 0.05$).

5.6 Multigroup analyses to test for subgroup measurement differences

To address a limitation of the MIMIC modeling approach and test for nonuniform DIF across subgroups, we perform multigroup analyses for the MIMIC model in the multigroup CFA framework.

For the general MIMIC model the multigroup confirmatory factor analytic model for m groups:

$$\begin{aligned} \mathbf{y}^{(g)} &= \boldsymbol{\mu}_y^{(g)} + \boldsymbol{\Lambda}_y^{(g)} \boldsymbol{\eta}^{(g)} + \boldsymbol{\varepsilon}^{(g)}, \\ \boldsymbol{\eta}^{(g)} &= \boldsymbol{\mu}_\eta^{(g)} + \mathbf{B}^{(g)} \boldsymbol{\eta}^{(g)} + \boldsymbol{\Gamma}^{(g)} \mathbf{x}^{(g)} + \boldsymbol{\zeta}^{(g)} \end{aligned} \quad (26)$$

where $g = 1, \dots, m$.

In our example, the multigroup model for our best fitting MIMIC model with E=0 will reduce to:

$$\mathbf{Items}^{(g)} = \boldsymbol{\mu}_D^{*(g)} + \mathbf{B}^{*(g)} \text{Depression} + \boldsymbol{\Gamma}^{*(g)} \text{MS} + \boldsymbol{\varepsilon}^{*(g)} \quad (27)$$

where $g=1,2$. In this model, we test for potential subgroup differences one at a time, by age, sex (male or female), race, MS type (relapsing or progressive) and baseline time since diagnosis, using a Bonferroni correction to control the family-wise error rate, by extending our best fitting model in the multigroup CFA framework [6]. In our SEM-based approach we form dichotomous variables for age (threshold is mean $\approx$ median = 46), race (white or other) and baseline time since diagnosis (less or greater than 10). Figure 7 illustrates the multigroup model for sex.

We compare the model fit statistics and indices, for each subgrouping between two models: a) freely estimating all parameters across groups (i.e. males and females) such as in (27) and b) constraining the factor loadings and item intercepts to be equal across groups (assumption of measurement invariance). For b), (27) reduces to

$$\text{Items}^{(g)} = \mu_D^* + \mathbf{B}^* \text{Depression} + \mathbf{\Gamma}^{*(g)} \text{MS} + \epsilon^{*(g)} \quad (28)$$

If there is a subgroup difference (group noninvariance), the model fit would be significantly worse under the constraint of group equality.

First, we apply an omnibus test of invariance using a chi-square test of difference to compare the models freely estimating parameters across each subgroup to the models constraining factor loadings and item intercepts to be equal across subgroups [6]. In this case, we do not reject the null hypothesis of measurement invariance for any of the subgroupings (i.e. no statistically significant p-values). Further, we do not observe any differences in CFI > .01 between the two sets of models for each subgrouping. We observe similar model test statistics and other indices in the multigroup CFA framework between the two sets of models for each subgrouping.

6. Correcting a scale for symptom overlap

Using the MIMIC modeling approach described above will lead a researcher to a “best fitting” model. In our case, the model with E=0 was the “best fitting” model. For practical clinical use, clinicians would benefit from a revised scale free of the overlapping symptoms to isolate the latent dimension of intended use of the scale. In our case we are adjusting the PHQ-9. The latent construct for *depression* isolates the depression latent dimension within the PHQ-9 for a more pure estimate of how the MS patient's mood may influence their other self-report responses and objective performance measures within the intended population.

We describe our techniques to develop an adjusted scale in section 6.1 as well as a method for performing a transformation to the same empirical distribution as the standard scoring for clinical use without having to significantly revise psychometric properties such as thresholds in section 6.2. The described approach could be applied to a broad range of models (e.g. a cross-loaded model, multigroup model).

6.1 Factor scores

A factor score is an individualized score on a latent construct representing a scale of interest [6]. Factor scores are of particular clinical interest, as they may represent meaningful measurements of patient characteristics. For example knowing that a patient is highly depressed (where *depression* is the unobserved factor) is clinically actionable.

Note that there are a fixed number of factor loadings regardless of the sample size, while there is a factor score for each individual. Thus the number of factor scores increases with the sample size. Factor scores are unobserved quantities, and cannot be known with certainty regardless of the sample size. As such factor scores can be predicted, but not estimated. A naïve method would be to just use the factor loadings as regression coefficients and

calculate factor scores similar to a predicted value for a fitted linear regression. However, a more optimal approach, is to apply Bayes' theorem to calculate the posterior distribution of the factor scores given the observed indicators. That is, we wish to calculate the distribution $P_{\theta}(f_i | y_i) \propto P_{\theta}(y_i | f_i)P_{\theta}(f_i)$ using the assumption of multivariate normality where the θ is replaced by some estimate of the parameters. The standard approach to predicting f_i is to choose the value of f_i which maximizes $P_{\theta}(f_i | y_i)$. This approach is known as the maximum a posteriori (MAP) method.

Using MPLUS, we calculate the factor scores within our sample adjusting for symptom overlap of depression and MS using the MAP method [13]. Factor scores can be output through MPlus into another data file (such as in SPSS) using a few lines of code in the OUTPUT section of the MPlus program. For example, saving factor scores into an SPSS file on the C drive called FS can be done using the coding:

```
OUTPUT:
  SAVEDATA: FILE IS C:\FS.sav;
  SAVE IS fscores;
```

Rigorous psychometric evaluation and interpretation is necessary for validating adjusted scale based on factor score cutoffs for clinical use. However, the proposed score adjustment algorithms could be useful in further analyses in studying a condition (i.e. depression) within the sample, accounting for the overlap of symptoms between co-occurring conditions (i.e. as an outcome in a regression model). A researcher could always use the same MIMIC model with DIF paths specified in this paper where the latent construct is the outcome of a set of predictors or path analysis in further SEM analyses.

6.2 Probability Integral Transformation of Factor Scores to Empirical Distribution of Scale Under Study

These factor scores can be linearly transformed to fit any potentially useful range (i.e. 0-27, 0-100, etc.) In our example, a probability integral transformation is used to transform these factors scores into the PHQ-9 scores within this population to approximately maintain the interpretation and thresholds of the PHQ-9 [95]. This is a one-to-one transformation, so the distribution of PHQ-9 scores in this population will remain the same, just that certain individuals will be matched to different PHQ-9 scores. This type of transformation relies on large sample theory in that you would need an appropriate range of different PHQ-9 scores to make this transformation valid. This transformation is built on continuous cumulative distribution functions and thus there are no special steps required for tied adjusted scores between two or more individuals (unlikely for factor scores). To perform the transformation in our sample [30] (see Appendix C for more technical details of the transformation):

1. Use a kernel density estimator to estimate the density of the PHQ-9 scores in this population.
2. Obtain the cumulative distribution function by numerical integration.
3. Numerically invert the cumulative distribution function for the PHQ-9.
4. Repeat steps 1) and 2) for the factor scores.

5. Transform the factor scores of each individual into a PHQ-9 score for this population by transforming the cumulative distribution function for the factor scores using the inverse of the cumulative distribution function for the PHQ-9.

The resulting factor scores will be distributed the same as the PHQ-9 within this population. For ease of use in a clinical setting, the recommendation is to round these transformed scores off to the nearest whole number. Note, within an EHR-based database, every time a new or series of new scores is entered, we can compute the factor scores. Then, either all subjects can be updated or we can keep prior records the same, and just use the functions to transform the new scores only.

The assumptions of the above described transformation approach is that within the population under study, we have correctly identified the empirical distribution (mean, standard deviation, range and cutoffs) based on the standard score, but have misclassified the particular scores of individuals. This further assumes that latent construct under study (i.e. *depression*) is different in patients with the co-occurring condition compared to the normal population, with some symptoms worse (i.e. anhedonia and feel depressed in our motivating example) while some symptoms overlap between the co-occurring conditions and inflate standard scales and diagnoses. Theoretical evidence that these assumptions are valid is warranted before using such a transformed scale in the clinical setting. For example, with our motivating study, previous studies have discussed that the etiology of depression is different in MS patients compared to the general population [24], making this part of the assumption reasonable. Given our large sample under study, as well, with a range of demographic characteristics and depressive and MS symptoms, also makes it reasonable to assume that we have correctly identified the empirical distribution of PHQ-9 scores internally within the Mellen Center population.

6.3 Using subscales without DIF

One way around this problem of overlapping symptoms is to use or form new scales and test findings that show no evidence of DIF within the population under study. While it is unlikely to find a scale or diagnoses completely devoid of DIF for patients with co-occurring conditions where symptoms overlap, certain scales will ultimately prove useful in helping facilitate the use of other scales. For example, the PHQ-2 comprises the first two items of the PHQ-9 and has been used as a depression screening tool [96]. Thus, using this shortened scale will bypass evaluating items for sleep problems, fatigue, poor concentration and psychomotor symptoms. A PHQ-2 score of three has been used as a threshold for depression for screening purposes [96].).

The PHQ-2 is highly correlated with the PHQ-9 in this MS study population (pearson $\rho \approx 0.87$). Further, a receiver operating characteristic (ROC) analysis of PHQ-9 vs. PHQ-2 in this study population, showed high specificity=95.7% and a large positive predictive value (PPV) = 91.2%, and area under the curve (AUC) = 0.852, though at the expense of the test sensitivity = 63.8. In general, the PHQ-2 has shown wide variability in sensitivity in previous validation studies and more research is needed to see if its diagnostic properties approach those of the PHQ-9 [97].

6.4 Application of the adjusted scales on two real patients

In a clinical setting, validating and assessing the reliability of a new scale is essential before usage with real patients with co-occurring conditions. However, given the assumptions we stated in 6.3, transformation of an adjusted scale to the empirical distribution of the original scale provides clinicians with a tool to use immediately. Using such a transformed scale in conjunction with tools showing no evidence of DIF can provide clinicians with a much clearer picture of symptoms of one of the co-occurring conditions. For example, in our study, we suggest adding two more entries in the Knowledge Program data base for the Mellen Center population, one for the PHQ-2 and one for the adjusted scoring algorithm. The simplest and most efficient approach for clinicians is almost certainly to first use the PHQ-2 for pre-screening for depression in MS patients, before using the modified algorithm as a screening tool. A positive PHQ-2 screen combined with a positive transformed adjusted algorithm screen will better identify patients in need of depression treatment as opposed to patients with MS-related symptoms not indicative of depression. This approach will give clinicians information for improved use of the PHQ-9, regarding better diagnosis, treatment decisions and inferences about effectiveness of treatment free of the overlap between depressive symptoms with other symptoms for MS patients.

Using these methods, to get an idea of how different the transformed adjusted score population is from the standard score population, we define a point estimate

$$\hat{\theta}_D = \text{transformed adjusted score} - \text{standard score} \quad (29)$$

Within our sample, $\text{mean}(\theta_D) \approx 0$ and $\text{var}(\theta_D) = 1.30$. Only 44% of subjects maintain the same transformed adjusted score as the original PHQ-9 score ($\theta_D = 0$), with 16% of subjects $|\theta_D| \geq 2$ and three subjects $\theta_D = 6$. Further, 74 subjects had a PHQ-9 score ≥ 10 and now have a transformed adjusted score < 10 , while 65 subjects had a PHQ-9 score < 10 and now have a transformed adjusted score ≥ 10 . Thus, our adjusted scale can lead to differing conclusions in screening for depression for a MS patient.

For a patient-level example, using these methods, patient A in our application has a PHQ-2 of zero and did not screen positive for depression via the transformed adjusted scale score. However, the standard PHQ-9 score for patient A is above the threshold of 10. Meanwhile, patient B screened positive for depression via the PHQ-2 and through the transformed adjusted scale score, but did not via the standard PHQ-9 score. Patient A highly endorsed the four items showing DIF, while patient B had a low endorsement of these items. Even for a clinician unwilling to make use of these new scales before further validation, at the very least, we have established that the use of the PHQ-9 without the PHQ-2 for MS patients is ill advised and to pay special attention to how a patient endorses items showing DIF.

7. Discussion

In this tutorial, we first provided an introduction to SEM as background information before providing an overview of factor analysis and MIMIC modeling within the context of analysis of overlapping symptoms in co-occurring conditions. These methods provide

flexibility for disentangling symptom overlap, allowing researchers to model measurement error, adjusting for confounding and making causal inference closely inline to an *a priori* conceptual model. Scales and scoring approaches can be revised, free of this overlap, using the factor scores derived directly from the model, leading to better diagnosis, treatment decisions and inferences about effectiveness of treatment. We have discussed the assumptions, strengths and limitations of these approaches for the problem, along with extensions to the most general setting for the problem.

To highlight the use of these SEM approaches and to show how to make complex modeling decisions for this type of analyses, we discussed in detail a motivating example, in which depressive symptoms overlap with other symptoms for MS patients. After showing significant overlap of depression symptoms with MS-related fatigue and cognitive and functional impairment using a MIMIC modeling approach, an adjusted depression screening scale was derived from the PHQ-9 using factor scores. This adjusted depression screening scale can be used in conjunction with the PHQ-2, to isolate the depression latent dimension of the PHQ-9, for a more pure estimate of an MS patient's mood. In practice, such adjusted scales may help clinicians prevent over or under prescribing anti-depressants and fatigue and MS medication, and provide better tailored care. When using these methods on a particular population such as the Mellen Center, the results may have limited external validity.

In future work, we plan to extend these methods to longitudinal data analyses using a multilevel mediation model in the latent growth curve modeling framework [98] while accounting for overlapping symptoms of co-occurring conditions to support treatment decision making. We will also provide methodology to explore the predictive value of the adjusted depression screening scales by linking the response to particular anti-depressants, and subgroups of patients identified via growth mixture modeling (GMM) [98, 99].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support for this study was provided by a grant from NIH/NCRR CTSA KL2TR000440 and by a grant from Novartis. The funding agreement ensured the authors' independence in designing analyses, interpreting the data, writing, and publishing the report. We appreciate the contributions from Drs. Randall Cebul, Thomas Love, Adam Perzynski, Irene Katzan, Neal Dawson, Center for Health Care Research and Policy, Drs. Robert Bermel and Deborah Miller, Mellen Center, Dr. Jarrod Dalton, Cleveland Clinic, and Dr. Martha Sajatovic, Departments of Psychiatry and Neurology at Case Western Reserve University School of Medicine.

Appendix

A. Definition of free parameters

While we may have k^* total parameters among our vector of parameters θ , we will have k free parameters to estimate, while the other $k^* - k$ parameters may be fixed to a value given a prior assumption. A free parameter is one that can be adjusted by the optimization routine to make the model fit the data. For example, when performing mediation we often fix the covariance of the error terms ζ_y, ζ_z from (6) to zero under the hypothesis of full mediation

for model identifiability. Thus, using ML estimation, in (6) we would have seven free parameters to estimate ($\mu_y, B_z, \gamma_{xy}, \mu_z, \gamma_{zy}, \sigma_y, \sigma_z$), where σ_y, σ_z are the variances of the respective error terms ζ_y, ζ_z , but eight total parameters.

B. Nested vs. non-nested models

It is necessary to make the distinction between *nested* and *non-nested* models when discussing model comparisons. In Figure 1, let model A include all causal paths, and model B includes all the paths with solid lines (all paths except for the direct effect from *time since diagnosis* to *depression*). Then, we have two *nested models*, in that model B can be derived from putting a parameter restriction on model A, mainly constraining a causal path to zero. Both models have the same observed measures.

We can derive other nested models from both model A and model B. Fixing the parameter for a causal path to a constant, such as from *time since diagnosis* to *cognitive decline*, from model A leads to a model nested within model A. The fixed parameters in model B are a subset of the fixed parameters in model A. In *nonhierarchical* models, our observed measures in the two models of comparison are not hierarchically related. Also referred to as *non-nested models*, by definition, none of the individual models can be obtained from each other by restricting parameters or causal paths or through a limiting process. but eight total parameters.

C. Technical Details of Probability Integral Transformation of Factor Scores to Empirical Distribution of PHQ-9 Scale

1. Use a kernel density estimator to estimate the density of the PHQ-9 scores in this population (to avoid confusion we label as *PHQ9*):

$$\hat{f}(PHQ9) = \frac{1}{nh} \sum_i K\left(\frac{PHQ9 - PHQ9_i}{h}\right)$$

$$K(u) = \begin{cases} \frac{1}{27} & \text{If } 0 \leq u \leq 27, \\ 0, & \text{otherwise.} \end{cases} \quad (30)$$

where *PHQ9* is a random variable from a continuous range of random variables between 0 and 27, $\{PHQ9_1, \dots, PHQ9_n\}$ represent the sample scores of sample size $n=3507$ from the empirical density f and h is the bin width.

2. Obtain the cumulative distribution function $F(PHQ9)$ by numerical integration of $f(PHQ9)$.
3. Numerically invert the cumulative distribution function for the *PHQ9* to obtain $F^{-1}(PHQ9)$.
4. Repeat steps 1) and 2) for the factor scores fs to obtain $f(fs)$ and $F(fs)$.
5. Transform the factor scores of each individual into a score for this population by way of $F^{-1}(PHQ9) = fs \Leftrightarrow F(fs) = PHQ9$

References

1. Panel GC. The Goldman Consensus statement on depression in multiple sclerosis. *Mult Scler.* 2005; 11:328–337. [PubMed: 15957516]
2. Kline, RB. Principles and practice of structural equation modeling. Guilford press; 2011.
3. Bollen, K. Structural equations with latent variables. Wiley; New York, NY: 1989.
4. Joreskog, K.; Sorbom, D. LISREL 8 user's reference guide. Scientific Software; Chicago: 1996.
5. Bentler, PM. EQS Structural Equations Program Manual Author: Peter M Bentler. BMDP Statistical Software; 1989. p. 254
6. Brown, TA. Confirmatory factor analysis for applied research. Guilford Press; 2012.
7. Woods CM, Oltmanns TF, Turkheimer E. Illustration of MIMIC-model DIF testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of psychopathology and behavioral assessment.* 2009; 31:320–330. [PubMed: 20442793]
8. Tucker, LR.; MacCallum, RC. Unpublished manuscript. Ohio State University Columbus: 1997. Exploratory factor analysis.
9. Byrne, BM. Structural equation modeling with AMOS: Basic concepts, applications, and programming. Routledge; 2013.
10. Byrne, BM. Structural equation modeling with EQS: Basic concepts, applications, and programming. Routledge; 2013.
11. Byrne, BM. Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Psychology Press; 2013.
12. Byrne, BM. Structural equation modeling with Mplus: Basic concepts, applications, and programming. Routledge; 2013.
13. Muthén LK, Muthén BO. Mplus. The comprehensive modelling program for applied researchers: user's guide. 2012; 5
14. Alemayehu D. Conceptual and Analytical Considerations toward the Use of Patient-Reported Outcomes in Personalized Medicine. *American Health & Drug Benefits.* 2012; 5:310–317. [PubMed: 24991329]
15. Kroenke K, Spitzer RL, Williams JB. The Phq-9. *Journal of general internal medicine.* 2001; 6:606–613. [PubMed: 11556941]
16. Blacker, D. Psychiatric rating scales Kaplan and Sadock's Comprehensive Textbook of Psychiatry. 8th. Sadock, BJ.; Sadock, VA., editors. Philadelphia: Lippincott Williams & Wilkins; 2005. p. 929-955.
17. Schwartz CE, Vollmer T, Lee H. Reliability and validity of two self-report measures of impairment and disability for MS. *Neurology.* 1999; 52:63–63. [PubMed: 9921850]
18. Marrie RA, Goldman M. Validity of performance scales for disability assessment in multiple sclerosis. *Multiple sclerosis.* 2007; 13:1176–1182. [PubMed: 17623733]
19. Mohr DC, Goodkin DE, Likosky W, Beutler L, Gatto N, Langan MK. Identification of Beck Depression Inventory items related to multiple sclerosis. *Journal of behavioral medicine.* 1997; 20:407–414. [PubMed: 9298438]
20. Mohr DC, Hart SL, Goldberg A. Effects of treatment for depression on fatigue in multiple sclerosis. *Psychosomatic medicine.* 2003; 65:542–547. [PubMed: 12883103]
21. Aikens J, Reinecke M, Pliskin N, Fischer J, Wiebe J, McCracken L, Taylor J. Assessing depressive symptoms in multiple sclerosis: is it necessary to omit items from the original Beck Depression Inventory? *Journal of behavioral medicine.* 1999; 22:127–142. [PubMed: 10374139]
22. Crawford P, Webster NJ. Assessment of Depression in Multiple Sclerosis: Validity of Including Somatic Items on the Beck Depression Inventory-II. *International Journal of MS Care.* 2009; 11:167–173.
23. Benedict RH, Fishman I, McClellan M, Bakshi R, Weinstock-Guttman B. Validity of the beck depression inventory-fast screen in multiple sclerosis. *Multiple sclerosis.* 2003; 9:393–396. [PubMed: 12926845]

24. Chang C, Nyenhuis D, Cella D, Luchetta T, Dineen K, Reder A. Psychometric evaluation of the Chicago Multiscale Depression Inventory in multiple sclerosis patients. *Multiple sclerosis*. 2003; 9:160–170. [PubMed: 12708812]
25. Sjonnesen K, Berzins S, Fiest KM, Bulloch AG, Metz LM, Thombs BD, Patten SB. Evaluation of the 9-item Patient Health Questionnaire (PHQ-9) as an assessment instrument for symptoms of depression in patients with multiple sclerosis. *Postgraduate medicine*. 2012; 124:69–77. [PubMed: 23095427]
26. Ferrando SJ, Samton J, Mor N, Nicora S, Findler M, Apatoff B. Patient health questionnaire-9 to screen for depression in outpatients with multiple sclerosis. *International Journal of MS Care*. 2007; 9:99–103.
27. Kim J, Amtmann D, Cook KF, Johnson KL, Bamer AM, Chung H, Askew RL. Psychometric properties of three depression scales in people with multiple sclerosis. *International Journal of MS Care*. 2012; 14:85–86.
28. Katzan I, Speck M, Dopler C, Urchek J, Bielawski K, Dunphy C, Jehi L, Bae C, Parchman A. The Knowledge Program: an innovative, comprehensive electronic data capture system and warehouse. *AMIA Annual Symposium Proceedings*. 2011:683–692. [PubMed: 22195124]
29. Mellen Center for Multiple Sclerosis Treatment and Research CC, Neurological Institute. 2013. Retrieved from http://my.clevelandclinic.org/neurological_institute/mellen-center-multiple-sclerosis/default.aspx
30. Gunzler D, Perzynski A, Morris N, Bermel R, Lewis S, Miller D. Disentangling Multiple Sclerosis & Depression: An Adjusted Depression Screening Score for Patient-Centered Care. *Journal of behavioral medicine*. 2015; 38:237–250. [PubMed: 24880636]
31. Polman CH, Rudick RA. The Multiple Sclerosis Functional Composite A clinically meaningful measure of disability. *Neurology*. 2010; 74:S8–S15. [PubMed: 20421572]
32. Whitaker J, McFarland H, Rudge P, Reingold S. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. *Multiple sclerosis*. 1995; 1:37–47. [PubMed: 9345468]
33. Rudick R, Fischer J, Antel J, Confavreux C, Cutter G, Ellison G, Lublin F, Miller A, Petkau J, Rao S. Clinical outcomes assessment in multiple sclerosis. *Annals of neurology*. 1996; 40:469–479. [PubMed: 8797541]
34. Fischer J, Rudick R, Cutter G, Reingold S. The Multiple Sclerosis Functional Composite measure (MSFC): an integrated approach to MS clinical outcome assessment. *Multiple sclerosis*. 1999; 5:244–250. [PubMed: 10467383]
35. Conner KR, Gunzler D, Tang W, Tu XM, Maisto SA. Test of a clinical model of drinking and suicidal risk. *Alcoholism: Clinical and Experimental Research*. 2011; 35:60–68.
36. Kowalski, J.; Tu, XM. *Modern applied U-statistics*. John Wiley & Sons; 2008.
37. Kenny DA. *Terminology and Basics of SEM*. 2011
38. Arbuckle, J. *Amos 6.0 user's guide*. Marketing Department, SPSS Incorporated; 2005.
39. Muthén BO. Beyond SEM: General latent variable modeling. *Behaviormetrika*. 2002; 29:81–118.
40. Hansson M, Chotai J, Nordstöm A, Bodlund O. Comparison of two self-rating scales to detect depression: HADS and PHQ-9. *British Journal of General Practice*. 2009; 59:e283–e288. [PubMed: 19761655]
41. Huang FY, Chung H, Kroenke K, Delucchi KL, Spitzer RL. Using the Patient Health Questionnaire-9 to Measure Depression among Racially and Ethnically Diverse Primary Care Patients. *Journal of general internal medicine*. 2006; 21:547–552. [PubMed: 16808734]
42. Johnson DR, Creech JC. Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*. 1983:398–407.
43. Zumbo BD, Zimmerman DW. Is the selection of statistical methods governed by level of measurement? *Canadian Psychology/Psychologie canadienne*. 1993; 34:390.
44. Wright S. Correlation and causation. *Journal of agricultural research*. 1921; 20:557–585.
45. Pearl, J. *Causality: models, reasoning and inference*. 2nd. Cambridge Univ Press; 2009.
46. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*. 1986; 51:1173. [PubMed: 3806354]

47. Browne MW. Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*. 1984; 37:62–83. [PubMed: 6733054]
48. Browne, MW. Covariance structures. In: Hawkins, DM., editor. *Covariance structures*. Cambridge University; Cambridge, UK: 1982.
49. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*. 1984; 49:115–132.
50. Morris NJ, Elston RC, Stein CM. A framework for structural equation models in general pedigrees. *Human heredity*. 2009; 70:278–286. [PubMed: 21212683]
51. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967; 1:221–233.
52. Nevitt J, Hancock GR. Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural equation modeling*. 2001; 8:353–377.
53. Olsson UH, Foss T, Troye SV, Howell RD. The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural equation modeling*. 2000; 7:557–595.
54. Skrondal, A.; Rabe-Hesketh, S. *Structural equation modeling: categorical variables*. Wiley Online Library; 2005.
55. Lee, SY.; Song, XY. *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. John Wiley & Sons; 2012.
56. Bollen KA, Maydeu-Olivares A. A polychoric instrumental variable (PIV) estimator for structural equation models with categorical variables. *Psychometrika*. 2007; 72:309–326.
57. Yuan KH, Bentler PM. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological methodology*. 2000; 30:165–200.
58. Satorra A, Bentler PM. Corrections to test statistics and standard errors in covariance structure analysis. 1994
59. Bollen KA, Stine RA. Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*. 1992; 21:205–229.
60. Browne MW, Cudeck R, Bollen KA, Long JS. Alternative ways of assessing model fit. *Sage Focus Editions*. 1993; 154:136–136.
61. Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*. 2008; 36:462–494. [PubMed: 19756246]
62. Bentler PM. Comparative fit indexes in structural models. *Psychological bulletin*. 1990; 107:238. [PubMed: 2320703]
63. Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*. 1973; 38:1–10.
64. Hu, Lt; Bentler, PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*. 1998; 3:424.
65. Marcoulides GA, Drezner Z. Specification searches in structural equation modeling with a genetic algorithm. *New developments and techniques in structural equation modeling*. 2001:247–268.
66. Marcoulides GA, Drezner Z. Model specification searches using ant colony optimization algorithms. *Structural equation modeling*. 2003; 10:154–164.
67. Scheines, R.; Spirtes, P.; Glymour, C.; Meek, C. *Tetrad ii: Tools for discovery*. Lawrence Erlbaum Associates; Hillsdale, NJ: City: 1994. *Tetrad ii: Tools for discovery*.
68. Schumacker RE. Teacher's Corner: Conducting Specification Searches With Amos. *Structural equation modeling*. 2006; 13:118–129.
69. MacCallum R. Specification searches in covariance structure modeling. *Psychological bulletin*. 1986; 100:107.
70. MacCallum RC, Roznowski M, Necowitz LB. Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological bulletin*. 1992; 111:490. [PubMed: 16250105]

71. Arfken G. Lagrange multipliers. *Mathematical methods for physicists*. 1985:945–950.
72. Saris WE, Satorra A, Sörbom D. The detection and correction of specification errors in structural equation models. *Sociological methodology*. 1987; 17:105–129.
73. Jackson DL. Revisiting sample size and number of parameter estimates: Some support for the N: q hypothesis. *Structural equation modeling*. 2003; 10:128–141.
74. Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. *Structural equation modeling*. 2002; 9:599–620.
75. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*. 1996; 1:130.
76. Preacher, KJ.; Coffman, DL. Computing power and minimum sample size for RMSEA [Computer software]. 2006. Available from <http://quantpsy.org/>
77. Satorra A, Saris WE. Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*. 1985; 50:83–90.
78. Satorra A. Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*. 1989; 54:131–151.
79. Kaplan D. Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate behavioral research*. 1989; 24:285–305.
80. Kaplan D, Wenger RN. Asymptomatic Independence and Separability in Covariance Structure Models: Implications for Specification Error, Power, and Model Modification. *Multivariate behavioral research*. 1993; 28:467–482.
81. Kaplan D, George R. A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal*. 1995; 2:101–118.
82. Heckman JJ. Econometric causality. *International Statistical Review*. 2008; 76:1–27.
83. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychological methods*. 2010; 15:309. [PubMed: 20954780]
84. Kalpakjian CZ, Toussaint LL, Albright KJ, Bombardier CH, Krause JK, Tate DG. Patient Health Questionnaire-9 in spinal cord injury: an examination of factor structure as related to gender. *The journal of spinal cord medicine*. 2009; 32:147. [PubMed: 19569462]
85. Raykov, T.; Marcoulides, GA. *Introduction to psychometric theory*. Taylor & Francis; 2010.
86. Marsh HW, Hau KT, Wen Z. In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural equation modeling*. 2004; 11:320–341.
87. Costello A, Osborne J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Pract Assess Res Eval*. 2005; 10(7) 10.pareonline.net/getvn.asp 2011.
88. Franklin SB, Gibson DJ, Robertson PA, Pohlmann JT, Fralish JS. Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*. 1995; 6:99–106.
89. Horn JL. A rationale and test for the number of factors in factor analysis. *Psychometrika*. 1965; 30:179–185. [PubMed: 14306381]
90. Jöreskog KG. How large can a standardized coefficient be. The help-file of the LISREL program. 1999
91. Krupp, LB. *Fatigue in multiple sclerosis: a guide to diagnosis and management*. Demos Medical Publishing; 2004.
92. Siegert R, Abernethy D. Depression in multiple sclerosis: a review. *Journal of Neurology, Neurosurgery & Psychiatry*. 2005; 76:469–475.
93. Bollen, KA.; Long, JS. *Testing structural equation models*. Sage; 1993.
94. Clement Sudhahar J, Israel D, Selvam M. Banking Service Loyalty Determination Through SEM Technique. *Journal of Applied Sciences*. 2006; 6:1472–1480.
95. Casella, G.; Berger, RL. *Statistical inference*. Duxbury Press; Belmont, CA: 1990.
96. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical care*. 2003; 41:1284–1292. [PubMed: 14583691]

97. Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *Journal of general internal medicine*. 2007; 22:1596–1602. [PubMed: 17874169]
98. Duncan, TE.; Duncan, SC.; Strycker, LA. An introduction to latent variable growth curve modeling: Concepts, issues, and application. Routledge Academic; 2013.
99. Muthén B, Muthén LK. Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*. 2000; 24:882–891.

Figure 1.

Figure 2.

Figure 3.

Figure 4.

Figure 5.

Figure 6.

Figure 7.

Table 1
Model fit statistics and indices from the exploratory factor analysis and confirmatory factor analysis models

N = 3505					
	1-factor EFA*	2-factor EFA	3-factor EFA	1-factor CFA with item correlations**	2-factor CFA
χ^2	754.356 (27)	266.470 (19)	109.505 (12)	386.57 (24)	401.83 (26)
CFI	0.922	0.973	0.989	0.961	0.960
TLI	0.896	0.949	0.968	0.941	0.944
RMSEA (95% CI)	0.088 (0.082, 0.093)	0.061 (0.055, 0.068)	0.048 (0.040, 0.057)	0.066 (0.060, 0.071)	0.064 (0.059, 0.070)
SRMR	0.042	0.021	0.013	0.032	0.031

p < 0.001 for all χ^2 tests of model fit.

* Same model fit values for the 1-factor CFA model *without* item correlations, with the trivial exception that the χ^2 statistic = 754.344, as the 1-factor EFA.

** item correlations are shown in Fig. 1.

Table 2
Standardized factor loadings from the exploratory factor analysis

N = 3505			
Depression	1-factor	2-factor	
		factor 1	factor 2
1. anhedonia	0.80	0.54	0.31
2. feel depressed	0.82	0.89	0.00
3. sleep problems	0.64	-0.02	0.74
4. fatigue	0.69	0.00	0.77
5. appetite change	0.66	0.18	0.53
6. feelings of failure	0.77	0.70	0.11
7. poor concentration	0.70	0.25	0.50
8. psychomotor symptoms	0.59	0.15	0.48
9. self-harm	0.51	0.62	-0.12

Table 3
Model fit statistics and indices from models to estimate the overlap of depressive symptoms with MS-related fatigue, cognitive impairment and functional impairment using a MIMIC modeling approach

N = 3507									
	No DIF	DIF	A=0	B=0	C=0	D=0	E=0	F=0	
X²	1592.55 (56)	475.45 (48)	573.02 (49)	1207.70 (49)	913.31 (49)	603.12 (49)	479.13 (50)	491.08 (50)	
CFI	0.877	0.966	0.958	0.907	0.931	0.956	0.966	0.965	
TLI	0.842	0.949	0.938	0.864	0.899	0.935	0.951	0.949	
RMSEA (95% CI)	0.088 (0.085, 0.092)	0.050 (0.046, 0.055)	0.055 (0.051, 0.059)	0.082 (0.078, 0.086)	0.071 (0.067, 0.075)	0.057 (0.053, 0.061)	0.049 (0.045, 0.054)	0.050 (0.046, 0.054)	
SRMR	0.046	0.028	0.031	0.039	0.036	0.032	0.029	0.030	
X² Diff p-value*	<0.001	reference	<0.001	<0.001	<0.001	<0.001	0.159	<0.001	
Mod Indices**	--	--	98.27	733.00	430.66	120.79	4.12, 4.26	22.55, 6.07	

p < 0.001 for all X² tests of model fit.

In MIMIC model with no DIF paths A through F = 0 while in the MIMIC model with DIF paths A through F are freely estimated. In the other MIMIC models (A through F = 0), only the indicated DIF path is constrained to zero, while all other paths are freely estimated.

* X² Difference Test (scale corrected for MLR estimator) for each separate model in comparison to the MIMIC model with DIF paths.

** For E=0 and F=0 first modification index is for the 9-hole peg test, second modification index is for the timed 25-foot walk.

Table 4
Standardized factor loadings and estimates (r) of DIF paths for MIMIC models with no DIF, DIF and E=0 (best fitting) models

Model (N = 3507)			
	No DIF	DIF	E=0
<i>depression</i>			
Factor Loadings			
1. anhedonia	0.79	0.82	0.82
2. feel depressed	0.77	0.82	0.82
3. sleep problems	0.65	0.49	0.49
4. fatigue	0.73	0.34	0.35
5. appetite change	0.67	0.66	0.66
6. feelings of failure	0.72	0.75	0.75
7. poor concentration	0.72	0.48	0.48
8. psychomotor symptoms	0.61	0.43	0.43
9. self-harm	0.47	0.51	0.51
Estimates (r)			
DIF paths			
A	0	0.21*	0.21*
B	0	0.52*	0.52*
C	0	0.38*	0.38*
D	0	0.23*	0.23*
E (peg test)	0	-0.02	0
E (timed walk)	0	-0.02	0
F (peg test)	0	0.11*	0.11*
F (timed walk)	0	-0.01	0.00

*
 $p < 0.001$

Table 5
Application of Adjusted Depression Screening Scale on Two Real Patients

PHQ-9 Item # and Answer Values										Comparison of Scoring Approaches		
Patient	1	2	3	4	5	6	7	8	9	Standard Scoring	PHQ-2	Transformed Factor Scores
A	0	0	3	3	2	0	3	1	0	12	0	8
B	3	2	1	1	2	0	0	0	0	9	5	13

Transformed Factor Scores are rounded to the nearest whole number.
For Patient A factor score = 0.159, Patient B factor score=0.701
PHQ-9 10 has been previously established as a screening cutoff for depressive disorder
PHQ-2 3 has been used as a threshold for depression for screening purposes