

## Adventures in Path Analysis and Preparatory Analysis

Brandy R. Sinco, MS, University of Michigan, Ann Arbor, MI  
Phillip L. Chapman, PhD, Colorado State University, Fort Collins, CO

### ABSTRACT

This paper will focus on the basics of path analysis, how to run path models in Proc CALIS, and how to use SAS to test for multivariate normality. Two estimation methods for path analysis: ML (Maximum Likelihood) and FIML (Full Information Maximum Likelihood) will be explained and compared. The results from Proc CALIS will also be compared with the SPSS AMOS module and the Mplus software.

Further, SAS can be used to check data for the MAR (Missing At Random) assumption and to estimate a path model after imputing missing data. Power calculations for structural equation models can be done with Proc IML. The concepts behind SEM power calculation will be explained and an IML program to perform SEM power calculations will be presented.

The data used for the analysis in this presentation is from the REACH-Detroit project, a culturally tailored Diabetes intervention for African American and Latino/a persons in inner city Detroit.

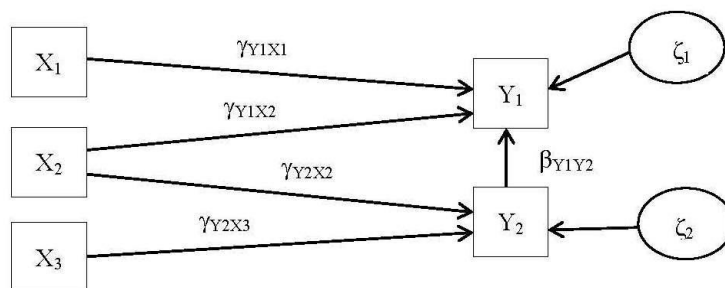
### Outline

Introduction to Path Analysis.	Pages 1-2
Evaluating the Goodness of Fit of a SEM Model.	Page 3
Degrees of Freedom and Power.	Pages 3-4
Description of the REACH Detroit Project and Hypothesized Path Model.	Pages 4-5
Multivariate Normality.	Pages 6-8
Missing Data.	Pages 8-10
Results of Fitting a Path Model to the REACH Detroit Data.	Pages 10-14
Comparison Between SAS Proc CALIS, SPSS/AMOS, and Mplus.	Page 15
Conclusions	Page 15
References	Page 16-17
Appendix: SAS Code for Multivariate Normality, Missing Data Analysis, Path Model for REACH	Pages 18-19

### Introduction to Path Analysis<sup>1-3</sup>

A Path Analysis model is a multivariate linear model based on a diagram that specifies the relationships between the variables, such as Figure 1 below. Path Analysis is the sub-model of SEM (Structural Equation Model), in which all variables (except error terms) are manifest, meaning observable. Whereas, SEM allows for both manifest and latent variables. Let  $n$  be the size of a random sample, where  $X_1$  to  $X_p$  are the independent variables and  $Y_1$  to  $Y_q$  are the dependent variables. In general,  $p$  and  $q$  do not have to be equal. Each of  $X_1$  to  $X_p$  and  $Y_1$  to  $Y_q$  are  $n \times 1$  vectors.

Figure 1: Example of a Path Diagram With Equations,  $p = 3$ ,  $q = 2$ .



- $Y_1 = X_1\gamma_{Y1X1} + X_2\gamma_{Y1X2} + Y_2\beta_{Y1Y2} + \zeta_1$ .
- $Y_2 = X_2\gamma_{Y2X2} + X_3\gamma_{Y2X3} + \zeta_2$ .

## SAS Code for Path Diagram in Figure 1

```
ods html newfile=proc path="c:\tempfiles";    ods graphics on;
Proc Calis Covariance Residual Modification Kurtosis Method=ML plots=residuals
Data=Test;
LINEQS Y1 = P_Y1X1 X1 + P_Y1X2 X2 + P_Y1Y2 Y2 + Err_Y1,
        Y2 = P_Y2X2 X2 + P_Y2X3 X3 + Err_Y2;
VARIANCE Err_Y1 = Var_EY1, Err_Y2 = Var_EY2,
        X1 = Var_X1, X2 = Var_X2, X3 = Var_X3;
VAR X1 X2 X3 Y1 Y2;
Run;    ods graphics off; ods html close;
```

## Endogenous and Exogenous Variables

An exogenous variable is an input variable with a mean and variance determined by variables outside of the model diagram, such as  $X_1$  to  $X_3$  in Figure 1 above. An endogenous variable is an output variable with a mean and variance that are related to other variables in the model, such as  $Y_1$  and  $Y_2$  in Figure 1. Exogenous variables are analogous to independent, input, or X variables in multiple regression, while endogenous variables correspond to dependent, output, or Y variables. Path analysis differs from multivariate regression by allowing a variable to be a covariate in one equation and an outcome in another equation, such as  $Y_2$  in Figure 1.

Covariances must be estimated for each pair of exogenous variables. Variances must be estimated for each exogenous variable and the error terms of the equations for endogenous variables. These parameters are automatically set by the CALIS procedure. Estimating covariances between the error terms of endogenous variable equations is optional and depends on whether the researcher hypothesizes significant covariance between endogenous error terms. Hence, error covariances are set to zero by default in the CALIS procedure.

## Design Equations and Matrices<sup>1,5</sup>

Assume a simple random sample of size  $n$  from an infinite population. Let

- $p$  = Number of manifest exogenous variables;  $q$  = Number of manifest endogenous variables.
- $r = p + q$  = Number of manifest exogenous + endogenous variables.
- $X$  = Matrix of manifest exogenous variables, with dimension  $n \times p$ .
- $Y$  = Matrix of manifest endogenous manifest variables, with dimension  $n \times q$ .
- $Z = [X \ Y]$  = Matrix of manifest endogenous and exogenous variables, dimension  $n \times r$ .
- $\mu$  = Column vector of means of manifest variables based on SEM model, dimension  $r \times 1$ .
- $\Sigma$  = Covariance matrix of manifest variables based on EM model, dimension  $r \times r$ .
- Both  $\mu$  and  $\Sigma$  are functions of the SEM model parameters.<sup>5</sup>
- $\bar{z}$  = Column vector of sample means of manifest variables, dimension  $r \times 1$ .
- $S$  = Sample covariance matrix, containing all sample variances and covariances of the columns of  $z$ , with  $(n - 1)$  in the denominator, dimension  $r \times r$ .
- $|S|$  = Generalized Variance of  $S$  = determinant( $S$ ).
- $\hat{\Sigma}$  = Estimated covariance matrix of  $Z$  based on the SEM model.
- $|\hat{\Sigma}|$  = Generalized Variance of  $\hat{\Sigma}$  = determinant( $\hat{\Sigma}$ ).

The goals of SEM are to estimate the conditional means and covariances of the endogenous variables.  $\Sigma$  must be a positive definite matrix, meaning that the determinant must be positive. SEM models are often fit by the method of Maximum Likelihood. The maximum likelihood process minimizes the discrepancy function,  $F_{ML}$ , where

$$F_{ML} = \ln|\hat{\Sigma}| - \ln|S| + tr(S\hat{\Sigma}^{-1}) + (\bar{z} - \mu)^T \hat{\Sigma}^{-1} (\bar{z} - \mu) - r \text{ and } F_{ML} = -2\log(\text{likelihood})/n.^{6,7}$$

Model  $\chi^2$ :  $\chi^2_{ML} = (n - 1)F_{ML}.$ <sup>7</sup> If the SEM model fits perfectly, the model  $\chi^2_{ML} = 0$ , because  $S = \hat{\Sigma}$ .

The LISREL (Linear Structural Relationships) equation for path analysis is  $Y = \alpha + YB + X\Gamma + \zeta$ , where  $E(\zeta) = 0$  and  $\zeta$  is uncorrelated with  $X_1$  to  $X_p$ . Further,  $B$  has zeros on the diagonal and  $(I - B)$  is non-singular. The reason for this is that  $(I - B)^{-1}$  must exist, in order to find a solution to the equation. If  $B=0$ , the path model is equivalent to multivariate regression.

## Lagrange Multiplier And Wald Statistics<sup>7</sup>

To display the Lagrange multiplier and Wald statistics, include the “Modification” option on the Proc CALIS. The Lagrange Multiplier estimates the reduction in  $\chi^2_{ML}$  by adding variables to SEM model. Whereas, the Wald Statistic estimates the increase in  $\chi^2_{ML}$  by removing variables from a SEM model. The SAS Proc CALIS documentation states that Lagrange multiplier and Wald statistics are approximations of what how the model  $\chi^2$  would change. More accurate results would be obtained by actually running the models and comparing the actual  $\chi^2$  values with a Likelihood Ratio Test.

## Evaluating the Goodness of Fit of a SEM Model<sup>1,2</sup>

Goodness of fit indices are best understood by dividing them into four categories: absolute fit, incremental fit, parsimony, and prediction ability of the model.

Let  $df_{ML}$  = degrees of freedom for the model under consideration,

$df_B$  = degrees of freedom for the null model with no covariates.

The model  $\chi^2$  is given by  $\chi^2_{ML} = (n - 1)F_{ML}$  and null model's  $\chi^2$  is  $\chi^2_B = (n - 1)F_B$

**Absolute Fit Indices** are analogous to  $R^2$  in linear regression and estimate the proportion of the sample covariance that is explained by the model. The GFI (Joreskog-Sorbom Goodness of Fit Index) is analogous to  $R^2$  in linear regression<sup>2, 6</sup>.  $GFI > .9$  indicates a good absolute fit.

**Incremental Fit Indices** compare the hypothesized model to the null model with no predictors ( $Y_1 = \varepsilon_1, \dots, Y_q = \varepsilon_q$ ). Kline recommends the CFI (Bentler's Comparative Fit Index)<sup>2, 14</sup>. A value of CFI from .90 - .95 is considered acceptable, while above .95 indicates a better incremental fit.  $CFI = 1 - (\chi^2_{ML} - df_{ML}) / (\chi^2_B - df_B)$

**Parsimony Adjusted Indices** include penalty terms in their formulas for more complex models. When two models with similar fit to the data are compared with parsimony adjusted indices, the indices will favor the less complex model. The RMSEA (Steiger-Lind Root Mean Square Error of Approximation) with a 90% confidence interval<sup>2, 15</sup>.  $RMSEA < .05$  is considered ideal, .05 to .08 indicates acceptable parsimony, .08 to .10 is considered mediocre, and above .10 signals a poor fit.<sup>1</sup> Also, “Probability of Close” fit is the p-value for the null hypothesis that  $RMSEA \leq .05$ .

**Predictive Fit Indices** estimate model fit in samples of the same size and estimate the model's ability to make predictions for the population. The SRMR (Standardized Root Mean Square Residual, Hu and Bentler, 1999) is related to the correlation residuals<sup>2, 16</sup>.  $SRMR < .10$  is the goal, with values under  $< .08$  indicating better predictive ability of the model.

Because SEM is based on maximum likelihood estimation, AIC (Akaike's Information Criteria) can be used to compare any two models and the LRT (likelihood ratio test statistic) can be used to compare two nested models. In SEM, the LRT is computed by subtracting the chi-squares between the reference model,  $\chi^2_M$ , and the nested model,  $\chi^2_0$ .  $LRT = \chi^2_M - \chi^2_0$ . The degrees of freedom will be the difference in degrees of freedom between the reference model and the nested model.

The squared multiple correlation coefficient for each endogenous variable has the same interpretation as  $R^2$  in linear regression, the proportion of variance in the dependent variable that is explained by the regression equation.

## Degrees of Freedom and Power

Since the goal of SEM is to choose a model that nearly replicates the covariance matrix, a saturated model would contain  $r(r + 1)/2$  terms, because a covariance matrix contains  $r(r + 1)/2$  unique entries.

- Let  $t$  = number of parameters estimated in a SEM =  $t$  = # path coefficients + # variances + # covariances.
- $df$  = degrees of freedom =  $r(r + 1)/2 - t$ .<sup>1, 17</sup>

So,  $df$  does not change with  $n$  as in a linear regression model. Neither increasing nor decreasing the sample size will change the degrees of freedom, but will change the power.

MacCullum (1996) proposed estimating the power in a SEM by comparing the values of the RMSEA statistic between models with good and poor fits. For a well-fitting model, the  $\chi^2$  statistic will have a noncentrality parameter,  $\lambda$ , near zero<sup>18</sup>. For a poorly fitting model, the  $\chi^2$  statistic will have the same  $df$ , but with a larger non-centrality parameter. The difference between the non-centrality parameters can be used to estimate power because  $E(\chi^2 - df) = \lambda$ .

Let  $RMSEA_0$  = RMSEA for a well-fitting model, typically .05.

$RMSEA_a$  = RMSEA for a poor-fitting model, typically .08.

df = Degrees of freedom.  
n = Sample size for complete data.  
 $\alpha$  = Size of type I error; commonly .05.

$\lambda_0$  = Non-centrality parameter assuming  $\Sigma$  is such that  $E(\chi^2) - df = (n-1)*df*RMSEA_0^2$ .  
 $\lambda_a$  = Non-centrality parameter assuming  $\Sigma$  is such that  $E(\chi^2) - df = (n-1)*df*RMSEA_a^2$ .

The null hypothesis is  $H_0: 0 \leq \lambda \leq \lambda_0$ . The alternative hypothesis is  $H_A: \lambda \geq \lambda_a$ .  
 $cval = 100(1 - \alpha)$  percentile of  $\chi^2(df, \lambda_0)$ .  
Power =  $Pr[\chi^2(df, \lambda_a) > cval]$ .

For the REACH-Detroit data,  $n=188$  complete observations, 58 df, and the lower bound for power is 0.79. This is a lower bound because the entire dataset contains 326 observations and the model was estimated with FIML (Full Information Maximum Likelihood), which includes all observations, even if the data is incomplete.

### SAS Code for SEM Power Calculation

```
Proc IML;
alpha=.05; /* significance level */
rmsea0=.05; /* null hyp value */
rmseaa=.08; /* alt hyp value */
d=58; /* degrees of freedom */
n=188; /* sample size */
ncp0=(n-1)*d*(rmsea0**2);
ncpa=(n-1)*d*(rmseaa**2);
cval = cinv(1-alpha, d, ncp0);
power = 1 - probchi(cval, d, ncpa);
Print rmsea0 rmseaa alpha d n cval[Format=10.3] power[Format=5.3] ncp0 ncpa;
Quit;
```

### Description of the REACH Detroit Project

REACH stands for Racial and Ethnic Approaches to Community Health. In the year 2000, the Centers for Disease Control and Prevention (CDC) funded 24 different REACH projects in cities across the country. In the United States, communities of color are more likely than whites to suffer from health problems such as diabetes, high blood pressure, breast and cervical cancer, hiv/aids, and infant mortality. For more information, visit [www.reachdetroit.org](http://www.reachdetroit.org) and [www.cdc.gov/reach](http://www.cdc.gov/reach).

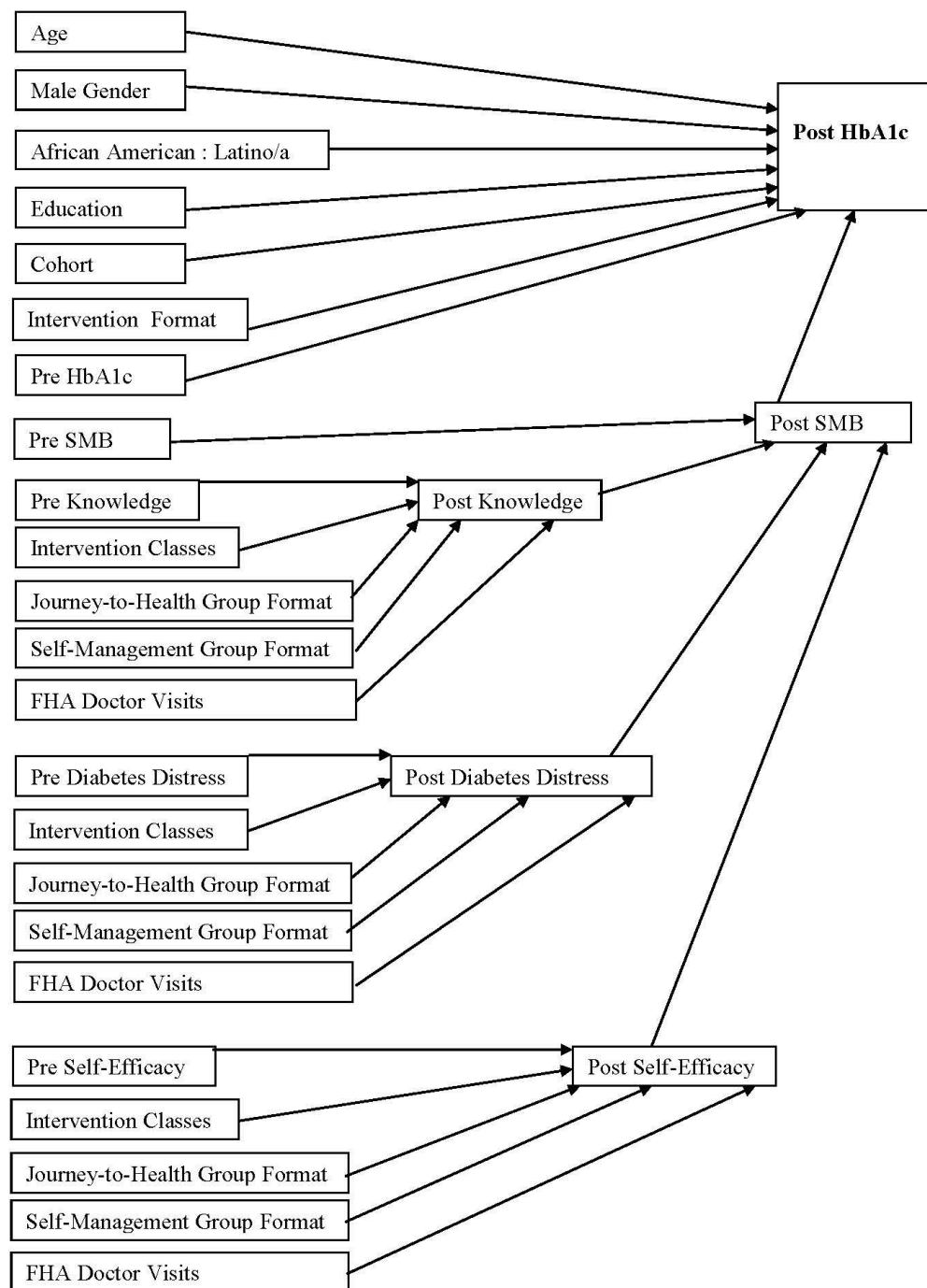
The REACH-Detroit intervention was a culturally tailored Diabetes curriculum over 11 sessions taught by PEER health educators, known as (FHAs) "Family Health Advocates". Part 1 of the curriculum was called "Journey to Health" and part 2 was called "Self-Management". Having the FHAs accompany clients to at least one doctor visit was another key part of the intervention.

This analysis project utilized data from 326 REACH clients in two cohorts. The measurements were taken from interviews and lab measurements at baseline, 6, 12 months. In cohort 1 ( $n = 180$ ), there was no control group because a randomized / control study was not acceptable to community partners in Detroit when the project began. After gaining trust in the community for a randomized control intervention, cohort 2 was designed with two groups: group 1 ( $n = 72$ ) intervention after baseline interview; group 2 ( $n = 74$ ) intervention after 6 month interview. Further, pre-Intervention was defined as baseline if cohort 1 or cohort 2 immediate arm and 6 months if cohort 2 delayed arm. Post-intervention was defined as 6 months if cohort 1 or cohort 2 immediate arm and 12 months for the cohort 2 delayed arm. The primary outcome was HbA1c (Hemoglobin A1c), the state of art method of measurement of blood sugar.

The hypothesized path to improvement in HbA1c is shown in figure 2. In English, figure 2 says that improvement in diabetes self-management behavior would lead to reduction in HbA1c. Improvements in self-management behavior would be achieved through greater knowledge and self-efficacy, along with lower diabetes distress. Participation in the intervention would lead to improved knowledge and self-efficacy, and would reduce diabetes distress. Participation measured by the number of intervention classes, attendance in group versus one-on-one format, and being accompanied to at least one doctor appointment by a FHA.

Before beginning the path analysis, all demographic, participation, and outcome measures were compared by treatment group and by cohort. Designated groups were cohort 1, cohort 2 delayed, and cohort 2 immediate. Differences in Binary measures were evaluated with a Pearson chi-square test. Ordinal measures were compared by one-way anova if the distributions were symmetric and with the Kruskal-Wallis test for asymmetrical distributions.

**Figure 2: Hypothesized Paths to HbA1c Improvement**



## Assessment of Multivariate Normality (SAS code in appendix)

### Measures and Assessment by Mahalanobis Distance, Histogram, and QQ Plot

Let  $Z_j$  be a single random variable with mean  $\mu$  and variance  $\sigma^2$ . Then, the normality of  $Z_j$  can be assessed by computing  $(Z_j - \mu)/\sigma$  and comparing its histogram and qq-plot to a standard normal variable with zero mean and unit variance. If the histogram and qq-plot of  $Z_j$  are close to a  $\text{Normal}(0, 1)$ , then assuming that  $Z_j$  is approximately normal is reasonable. Because  $(Z_j - \mu)^2/\sigma^2$  would have a chi-square distribution with one degree of freedom,  $\chi^2(1)$ , comparing the histogram and qq-plots of  $(Z_j - \mu)^2/\sigma^2$  to a  $\chi^2(1)$  is an alternative method for assessing normality.

For a  $r$ -variate random sample of size  $n$ , the Mahalanobis distance is analogous to  $(Z_j - \mu)^2/\sigma^2$  and is given by the formula below.

- Let  $Z$  be a  $n \times r$  matrix of random variables.
- $Z_{(i)}$  =  $i$ th row of  $Z$  with dimension  $1 \times r$ .
- $\mu$  = column vector of means of the columns of  $Z$ , dimension  $r \times 1$ .
- $\Sigma$  = covariance matrix of a row of  $Z_{(i)}$ , dimension  $r \times r$ .
- $d_{(i)}^2$  = Mahalanobis distance of  $Z_{(i)} = (Z_{(i)} - \mu^T) \Sigma^{-1} (Z_{(i)} - \mu^T)^T$ .

If  $Z$  is multivariate normal, then the histogram and qq-plots of the Mahalanobis distance,  $d_{(i)}^2$ , should have an approximate  $\chi^2(r)$  distribution. In practice,  $\mu$  and  $\Sigma$  are usually unknown and are estimated by the sample mean and covariance.

The Mahalanobis Distance,  $d_{(i)}^2$ , can also be computed from the eigenvalues and eigenvectors of the covariance matrix of  $Z$ . Because the covariance matrix is symmetric and non-negative definite, it can be expressed in terms of its eigenvalues and eigenvectors<sup>3</sup>. There will be  $r$  eigenvalue, eigenvector pairs for  $Z$ .

- Let  $\lambda_j$  = eigenvalue of the covariance matrix of a row of  $Z$ .
- $\Lambda$  = diagonal matrix with  $\lambda_1, \lambda_2, \dots, \lambda_r$  on its diagonal, dimension  $r \times r$ .
- $e_j$  = eigenvector corresponding to  $\lambda_j$ , dimension  $n \times 1$ .
- $e$  =  $r \times r$  matrix containing eigenvectors  $e_1$  to  $e_r$  in each column.

Then,  $d_{(i)}^2$  = Mahalanobis distance of  $Z_{(i)} = (Z_{(i)} - \mu^T) e^T \Lambda^{-1} e (Z_{(i)} - \mu^T)^T$ .

### Mardia's Statistics for Multivariate Normality

One of the assumptions in the classical SEM is that all variables, endogenous and exogenous, are multivariate normal. Because SEM tests a hypothesized covariance structure with a chi-square statistic that is sensitive to large kurtosis, multivariate kurtosis is considered a key diagnostic for data in a SEM model<sup>5</sup>.

For any univariate random variable,  $Z_j$ , with mean  $\mu$ , the skewness and kurtosis are defined by the following equations:  $\mu_3$  = skewness =  $E(Z_j - \mu)^3$  and  $\mu_4$  = kurtosis =  $E(Z_j - \mu)^4$ . For any random variable with a symmetric distribution, including a normal random variable, the skewness will be zero. For any normal random variable with variance  $\sigma^2$ , the kurtosis will be  $3\sigma^2$ . For a standard normal distribution, the kurtosis will be 3.

Geometrically, skewness is a measure of symmetry. Positive skewness indicates that a distribution has a longer tail to the right and negative skewness signals a larger tail to the left. Kurtosis measures peakedness and tail density relative to a standard normal distribution. A kurtosis above 3 indicates a higher, sharper peak and longer tails than a  $\text{Normal}(0,1)$ . Lower values indicate a lower, less distinct peak and shorter tails than a  $\text{Normal}(0,1)$ . Balanda and MacGillivray wrote that increasing kurtosis was associated with the movement of probability mass from the shoulders of a distribution into its center and tails<sup>26</sup>.

Mardia generalized the formulas for univariate skewness and kurtosis to a  $r$ -variate distribution<sup>27</sup>. He proved that the kurtosis for a  $r$ -variate standard normal variable would be  $r(r + 2)$ . So, if  $r = 1$ , the kurtosis would be 3.

The population multivariate skewness and kurtosis are given by the following equations<sup>27, 28</sup>:

- Multivariate Skewness =  $\beta_{1r} = E\{(Z_{(i)} - \mu^T) \Sigma^{-1} (Z_{(i)} - \mu^T)^T\}^3$ .
- Multivariate Kurtosis =  $\beta_{2r} = E\{(Z_{(i)} - \mu^T) \Sigma^{-1} (Z_{(i)} - \mu^T)^T\}^2$ .

For a single variable,  $Z_j$ , the multivariate kurtosis will correspond to univariate formula,  $\beta_{2r} = E\{(Z_{j(i)} - \mu^T) \sigma^{-2} (Z_{j(i)} - \mu^T)^T\}^2$ . However, the multivariate formula for skewness is the square of the univariate formula;  $\beta_{1r} = E\{(Z_{j(i)} - \mu^T) \sigma^{-2} (Z_{j(i)} - \mu^T)^T\}^3$ .

Mardia proposed the following formulas for sample skewness and kurtosis.

- $\bar{Z}$  = column vector of the means of columns of  $Z$ , dimension  $r \times 1$ .

- Multivariate Skewness:  $b_{1r} = \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \left[ \left( Z_{(i)} - \bar{Z}^T \right) S^{-1} \left( Z_{(k)} - \bar{Z}^T \right)^T \right]^3$ .
- Multivariate Kurtosis:  $b_{2r} = \frac{1}{n} \sum_{i=1}^n \left[ \left( Z_{(i)} - \bar{Z}^T \right) S^{-1} \left( Z_{(i)} - \bar{Z}^T \right)^T \right]^2$ .

If  $Z$  is  $r$ -variate normal, then  $nb_{1r}/6 \sim \chi^2[r(r+1)(r+2)/6]$  and  $b_{2r} \sim \text{Normal}[r(r+2), 8r(r+2)/n]^{27}$ . Therefore, the multivariate normality of  $Z$  can be tested with a chi-square test for  $nb_{1r}/6$  and the multivariate kurtosis by comparing

$\left( \frac{b_{2r} - r(r+2)}{\sqrt{8r(r+2)/n}} \right)$  to the percentiles of the standard normal distribution.

### Test Results for Multivariate Normality in the REACH Detroit Dataset

For the REACH data, multivariate normality was examined with histograms and qq-plots for all variables in model (Figure 3a) and only the endogenous variables (Figure 3b). When the joint normality is assessed for all variables in Figure 3a, the histogram and qq-plots more closely fit the chi-square distribution, than in Figures 3b.

For figure 3.3.1,  $r = 20$  and  $\left( \frac{b_{2r} - r(r+2)}{\sqrt{8r(r+2)/n}} \right) = -0.03$ .

The value is  $\ll 1.96$  and again indicates that multivariate normality is a reasonable assumption. The mystery is that some of the variables are binary. So, how can all variables together produce better diagnostics for multivariate normality than only the endogenous variables? The key to this mystery is that binary variables can have kurtoses that are smaller than normal variables and the contribution of binary variables can lower the multivariate kurtosis.

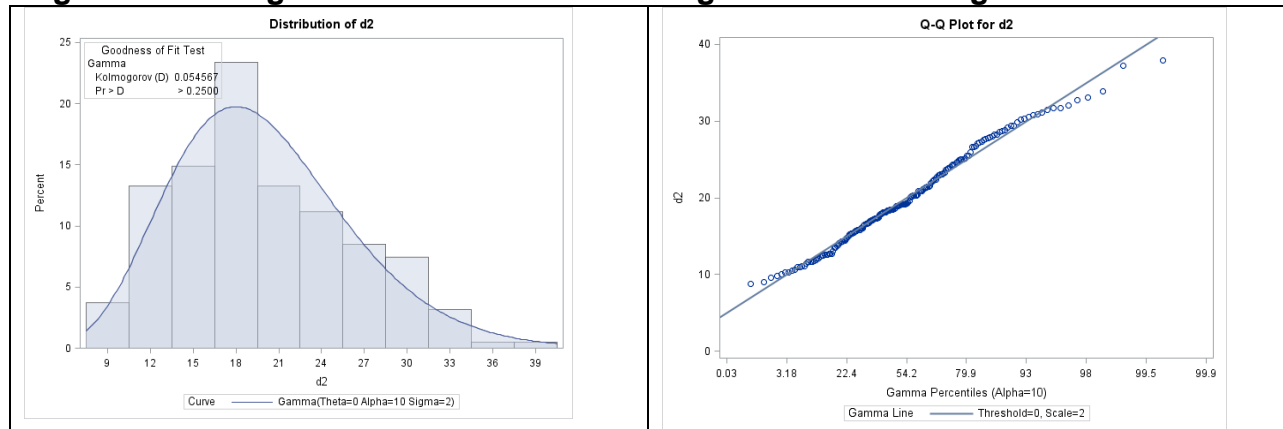
Standardized skewness is the skewness divided by the  $(\text{variance})^{3/2}$ . Similarly, standardized kurtosis refers to  $(\text{kurtosis}/(\text{variance})^2 - 3)$ . The standardized skewness of a Bernoulli random variable, where  $\pi$  is the probability of

the event, is  $\frac{(1-2\pi)}{\sqrt{\pi(1-\pi)}}$  and standardized kurtosis is  $\frac{1-6\pi(1-\pi)}{\pi(1-\pi)}$ . Therefore, the normalized skewness and

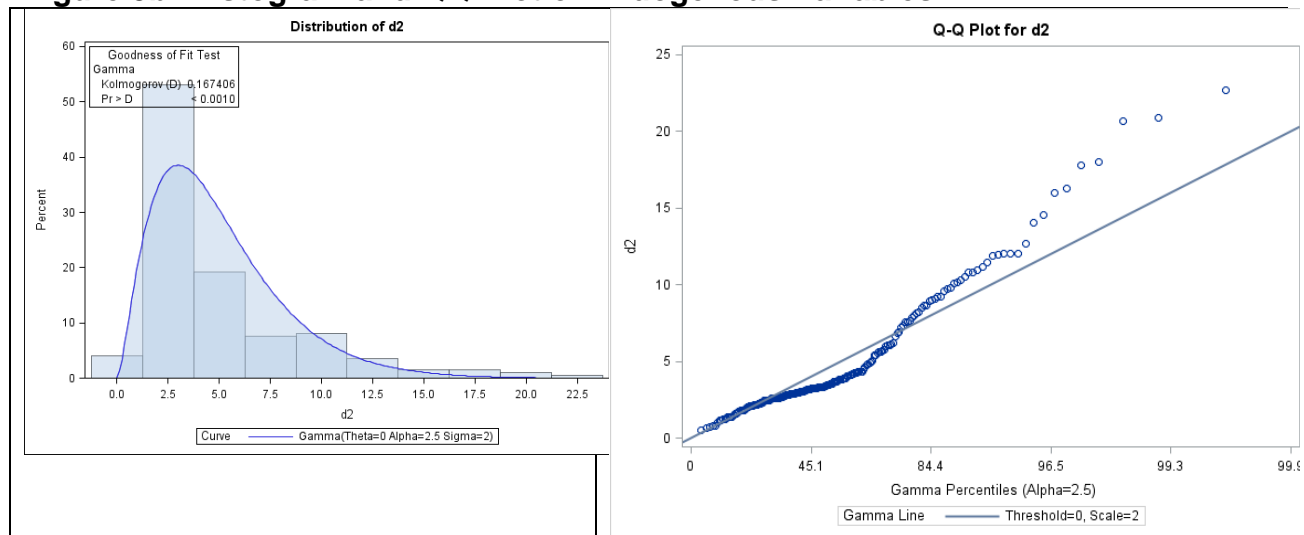
kurtosis of a binary random variable can have values lower than a normal random variable. The overall multivariate kurtosis can be lowered by including binary random variables with other random variables that have kurtosis values higher than normal random variables. The table below displays the standardized skewness and kurtosis values for binary variables with  $\pi$  from 0.25 to 0.75.

$\pi$	Skewness	Kurtosis
0.25	1.15	-0.67
0.50	0.00	-2.00
0.75	-1.15	-0.67

**Figure 3a: Histogram and QQ-Plot of All Exogenous And Endogenous Variables**



**Figure 3b: Histogram and QQ-Plot of Endogenous Variables**



Although the model  $\chi^2$  derivation is based on the assumption that all variables in a SEM are multivariate normal<sup>5</sup>, the exogenous variables do not have to be normally distributed<sup>5, 32, 33</sup>. An adequate condition is that the endogenous variables, conditional on the exogenous variables, be multivariate normal<sup>5, 33</sup>. Bentler and Chou provided examples of exogenous variables, such as gender and race/ethnicity, that are clearly non-normal. In conclusion, the normality diagnostics indicate that the multivariate skewness and kurtosis in the REACH data are sufficiently small for SEM analysis.

## Missing Data

In the REACH-Detroit dataset, missing data is an issue that needs to be addressed. Out of 326 observations, 188 are complete.

Let Y be the outcomes, or dependent variables, for a model. (Hemoglobin A1c, Knowledge, Self- Management, Diabetes Distress, Self-Efficacy). X are the predictors, or independent variables, for a model. (Participation, Demographics, Treatment Group).

M is the indicator for missing data and  $f(M)$  be the probability density function for M, where

M = 1 denotes missing data and M = 0 denotes complete data on an observation.

In the missing data literature, there are three types of missing data mechanisms<sup>34-36</sup>:

- **MCAR (Missing Completely at Random).** Missingness does not depend on the values of variables in the data set. I.E., missingness does not depend on Y (outcome) or X (covariates).  $f(M | X, Y) = f(M)$ .
- **MAR (Missing at Random).** Missing Y may depend on covariates, X, but not on Y.  $f(M | X, Y) = f(M | X)$ .



- **MNAR (Missing Not at Random).** Missingness is related to unobserved data; also called “Non-Ignorable Missing”.

**In my opinion, Missing at Random (MAR) is a reasonable assumption for the REACH-Detroit data.**

- ✓ From Table 1, the pre-intervention means for all 5 endogenous variables (HbA1c, Knowledge, Diabetes Distress, Self-Efficacy, Self-Management) do not differ significantly by whether the post-intervention values are missing. In addition, the post-intervention HbA1c means do not differ by whether pre HbA1c is missing. Due to the small numbers of missing values in the other pre-intervention variables, the only post – intervention variable that was compared by whether the pre-intervention values were missing was HbA1c.
- ✓ The only variables that differ by withdrawal are participation measures, class attendance and FHA accompanied doctor visits. This makes perfect sense because people who withdrew weren’t available to participate. None of the outcome nor demographic variables differ significantly by withdrawal.
- ✓ Withdrawal was the reason for over 70% of the missing post-intervention data.
- ✓ Last, although none of the exogenous variables are significantly associated with withdrawal, I do not think that the data is MCAR (Missing Completely At Random) because a previous analysis indicated that a higher percentage of African Americans withdrew in cohort 2.

**Table 1: Means (95% CI) of Endogenous Variables, By Whether Data at the Other Time Point Is Missing**

Measure	Other Timepoint Missing	OtherTimepoint Present	p-Value
Pre-Intervention HbA1c	8.4 (7.7, 9.0)	8.5 (8.2, 8.8)	0.6405
Post-Intervention HbA1c	7.7 (5.9, 9.6)	7.8 (7.5, 8.0)	0.8776
Pre-Intervention Knowledge	3.4 (3.2, 3.6)	3.3 (3.2, 3.5)	0.4969
Pre-Intervention Diabetes Distress	22.3 (17.2, 27.4)	24.4 (21.7, 27.1)	0.1733
Pre-Intervention Self-Efficacy	74.8 (70.8, 78.7)	70.4 (67.8, 73.0)	0.0705
Pre-Intervention Self-Management	2.9 (2.7, 3.1)	3.0 (2.9, 3.1)	0.3714

### **Multiple Imputation (MI)<sup>35, 37</sup> (SAS code in appendix )**

Multiple Imputation uses Markov Chain Monte Carlo simulation to estimate missing values in the data set. A key assumption for multiple imputation is that the missing data mechanism is at least MAR (Missing At Random). Depending on the amount of data that is missing, m imputations will be performed, with the result being m datasets. The higher the percentage of missing data, the higher the number of imputations.

After imputing m datasets, the originally intended statistical analysis procedure, such as regression or structural equation modeling, will be performed, but will be repeated on the m imputed datasets. The resulting parameter estimates from the m analyses will then be averaged from the m runs and the standard errors will be the sum of between imputation and within-imputation variances by following equations. In addition, the imputation model can contain additional variables that are either associated with missingness or with the value of the outcome.

- Let  $\theta_i$  be the  $r \times 1$  vector of parameter estimates from the ith imputation.
- $\bar{\theta} = \frac{1}{m} \sum_{i=1}^m \theta_i$  = Combined parameter estimates from the m imputations.
- $W_i$  be a  $r \times 1$  vector of within-imputation variances for the ith imputation.
- $\bar{W} = \frac{1}{m} \sum_{i=1}^m W_i$  = Within-Imputation Variance.
- $B = \frac{1}{(m-1)} \sum_{i=1}^m (\theta_i - \bar{\theta})^2$  = Between-Imputation Variance, dimension  $r \times 1$ .
- $T = \bar{W} + \left(1 + \frac{1}{m}\right) B$  = Total Variance of parameter estimates, dimension  $r \times 1$ .
- $\lambda$  = Proportion of observations with incomplete data.
- RE = Relative Efficiency = Ratio between variance of estimates based on m imputations to the asymptotic variance from infinite imputations.
- $RE = (1 + \lambda/m)^{-1}$ .

### Full Information Maximum Likelihood (FIML)<sup>7, 38</sup>

The FIML algorithm is the same as the ML (Maximum Likelihood) algorithm, except that all available information is used. As in multiple imputation, a key assumption is that the missing data mechanism is MAR or MCAR. In the Maximum Likelihood algorithm, an observation with data on 9 out of 10 variables would be excluded from the likelihood calculations. Whereas, FIML would use the 9 out of 10 variables on that record.

- Let  $S_n$  = Biased sample covariance =  $(n - 1)S/n$ , dimension  $r \times r$ .
- NOTE:  $X, Y, Z, \bar{z}, \mu, S, \Sigma$  same as previously defined in on page 2.
- $C$  = Number of complete observations.

$S$  and  $\bar{z}$  are estimated from the complete data with  $C$  observations. The initial estimates of  $\mu$  and  $\Sigma$  are estimated by maximum likelihood from the complete data. These estimates are updated with the data available in every observation to minimize the expression for  $F_{FIML}$  below.

Let  $z_{(i)}$  be the row vector for  $i$ th row of  $z$ . If the  $i$ th row of  $z$  has complete data, then  $z_{(i)}$  has dimension  $1 \times r$ . If  $k$  out of  $r$  variables on the  $i$ th observation contain data, then the dimension will be  $1 \times k$ .

Let  $\mu_i$  and  $\Sigma_i$  be the mean and covariance for the  $i$ th observation of  $z$ . If the  $i$ th observation of  $z$  has complete data, then  $\mu_i = \mu$  and  $\Sigma_i = \Sigma$ . Otherwise, if data for  $k$  of the  $r$  variables are present, then  $\mu_i$  will have dimension  $k \times 1$  and  $\Sigma_i$  will have dimension  $k \times k$ . Therefore,  $\mu_i$  is the sub-vector of  $\mu$  and  $\Sigma_i$  is the sub-matrix of for which data is present in the  $i$ th row. For example, if  $z$  contained 10 variables and 7 variables had data present on the 5<sup>th</sup> observation, the dimensions of  $\mu_5$  would be  $7 \times 1$ ;  $\Sigma_5$  would be a  $7 \times 7$  matrix, instead of a  $10 \times 10$  matrix.

The values in  $\mu_5$  would be the 7 means from  $\mu$  that corresponded to the positions in the 5<sup>th</sup> observation for which data was present. Similarly for  $\Sigma_5$ , no new data would be computed. The sub-matrix would be the values from  $\Sigma$  which corresponded to the non-missing position in the 5<sup>th</sup> observation.

Then, the discrepancy functions to minimize for ML and FIML are the following:

- $F_{ML} = \ln|\Sigma| - \ln|S| + tr(S\Sigma^{-1}) + (\bar{z} - \mu)^T \Sigma^{-1} (\bar{z} - \mu) - r$
- $F_{FIML} = \frac{1}{n} \sum_{i=1}^n \left( \ln(|\Sigma_i|) + (z_{(i)} - \mu_i^T) \Sigma_i^{-1} (z_{(i)} - \mu_i^T)^T + K_i \right)$ ;  $K_i$  is a constant.

If all  $n$  rows of  $z$  contain complete data, then the expression for  $F_{FIML}$  becomes the following, where  $K$  is a constant.

$$F_{FIML} = \ln|\Sigma| + tr(S_n \Sigma^{-1}) + (\bar{z} - \mu)^T \Sigma^{-1} (\bar{z} - \mu) + K.$$

### Results of Fitting a Path Analysis Model to the REACH Detroit Data.

The FIML estimation method was used to estimate the coefficients for all of the candidate models. Because FIML is maximum likelihood estimation, likelihood ratio tests can be used to compare nested models and AIC (Akaike Information Criterion) can be used to compare any two models. Four issues were addressed to select the final model:

- Transform to code FHA-accompanied doctor visits.
- Direct path from participation measures to Post-Intervention HbA1c.
- Effect of removing demographics, treatment group, participation measures.
- Self-management behavior as aggregate measure or include each variable used in computing the self-management score.

The first issue was how to transform FHA-accompanied doctor visits. In the original design of REACH-Detroit, clients were supposed to have one doctor visit accompanied by a FHA (Family Health Advocate). However, 93 of the 326 clients had two or more FHA-accompanied doctor visits. The AIC criterion was used to select the model with the transformation that best fit the data. Three transformations for doctor visits were considered and each produced a single variable to represent FHA-accompanied doctor visits. The coding systems that were considered were: {0, 1+}, {0, 1, 2, 3, 4+}, and square root(doctor visits). The reasoning behind the square root transformation was that zero

and one would retain their original values, while values above one would be weighted downwards. The AIC results for the three models were 22071.70 for 1+ or zero, 22638.11 for truncation at 4, and 22321.28 for the square root transformation. Because lower AIC indicates better fit, 1+ or zero is clearly the best coding for doctor visits and also fits best with the original design of REACH.

The second issue was whether the model should contain direct paths from the participation variables to post-intervention HbA1c. There were four participation variables: number of intervention classes, 1+ FHA-accompanied doctor visit, 1+ Journey to Health Group Class, and 1+ Self-Management Group Class. The LRT (Likelihood Ratio Test) was used to compare these two models by subtracting the  $\chi^2$  statistics between the full and reduced models. The Lrt statistic was  $(142.96 - 135.62) = 7.34$ , which produced a p-value of 0.1138. Based on the Lrt statistic, the direct path from the participation variables to post-intervention HbA1c was omitted from the model.

Third, what would be the effect of removing demographics and treatment group, along with each participation variable? Again, these questions were answered by using the likelihood ratio test and the results are shown in the Table 5.1.1 below. Based on the LRT tests, demographics, cohort, and treatment group do not add significant information to the model, although p-value was borderline significant a  $p < 0.10$ . This suggests that the demographics and treatment group have a small or negligible effect on Post-Intervention HbA1c.

Of the participation variables, both the number of intervention classes and the group versus one-on-one format were key variables. However, the FHA-accompanied doctor visit did not make much difference, possibly because this part of the intervention wasn't delivered to the clients as originally planned.

**Table 2: Likelihood Ratio Tests for Removing Selected Covariates**

Variables Removed	Degrees of Freedom	LRT Statistic	p-Value
Demographics, Cohort, Treatment Group	6	11.8	0.0667
Number of Intervention Classes	3	12.63	0.0055
1+ Journey to Health Group Class, 1+ Self-Management Group Class	6	14.47	0.0248
1+ FHA-Accompanied Doctor Visit	3	0.81	0.8475

### Comparison of the REACH Detroit Path Models Between FIML, ML, and MI .

According to the SEM literature, handling missing data with FIML is asymptotically equivalent to multiple imputation<sup>39, 40</sup>. The advantage to FIML is that the syntax is easier in software packages. For example, to use FIML estimation in SAS Proc CALIS, the only change is Method=FIML instead of Method=ML. FIML is the default estimation method in SPSS/AMOS. However, Multiple Imputation has the advantage of being able to include variables in the imputation model that are not in the Structural Equation Model.

To compare FIML, ML, and MI, the REACH structural equation model coefficients and standard errors were estimated with SAS Proc CALIS. For the FIML and ML estimations, the method option was set to ML and then to FIML. For the multiple imputation, 20 imputations were chosen because 42% of the observations in the dataset contained incomplete data and the Little & Rubin guidelines<sup>35</sup> indicated that 20 imputations would produce 98% relative efficiency. First, twenty imputations were conducted with SAS Proc MI. "Withdrawal" was included in the imputation, along with the other variables in the SEM, because "Withdrawal" was strongly correlated with missingness. Then, the SEM coefficients were estimated for each of the 20 imputations with Proc CALIS. Last, the estimated coefficients and standard errors from the 20 imputations were combined by using the equations on page 9. The SAS code for the missing data analysis is provided in the appendix.

The estimated coefficients and standard errors were compared between MI and FIML, and between ML and FIML by computing agreement ratios. The agreement ratio was defined as (estimate by other method) / (estimate by FIML). The agreement ratios were computed for both the point estimates and for the standard errors. Multiple imputation produced coefficient estimates with a mean agreement ratio to FIML of 0.98 and a median of 1.00. However, the standard errors from multiple imputation were a little larger than FIML with mean agreement ratio of 1.13 and a median of 1.03.

Using maximum likelihood instead of FIML produced estimates with larger variation. The average agreement ratio for the coefficient estimates was 1.14, while the median was 0.91. For the standard errors, the mean agreement ratio was 1.18 and the median was 1.14.

Of the three estimation methods, FIML produced the most stable estimates.

**Table 3: Agreement Ratios Between MI (Multiple Imputation) and FIML, ML (Maximum Likelihood) & FIML**

	Coefficient Point Estimate MI / FIML	Coefficient Standard Error MI / FIML	Coefficient Point Estimate ML / FIML	Coefficient Standard Error ML / FIML
<b>Average</b>	0.98	1.13	1.14	1.18
<b>Median</b>	1.00	1.03	0.91	1.14
<b>Min</b>	0.10	0.90	0.14	1.00
<b>Max</b>	2.09	2.25	3.40	1.83

### Interpretation of the Model.

The coefficients, standard errors, and significance levels from fitting the structural equation model to the REACH-Detroit data are shown in Figure 3. The model diagnostics are displayed in Table 4.

Because the covariance matrix was used for this analysis, the SEM model coefficients are interpreted in the same way as regression coefficients. A unit change in an exogenous variable, with all other variables held constant, corresponds to a mean change in the endogenous variable of the coefficient of the exogenous variable.

All of the endogenous variables: HbA1c, self-management behavior, knowledge of diabetes management, diabetes-related distress, and self-efficacy were strongly positively correlated with their pre-intervention values at  $p < .001$ .

**HbA1c.** As hypothesized, post-intervention HbA1c was predicted by post-intervention self-management behavior. A unit increase in self-management behavior was associated with -0.55 drop in post-intervention HbA1c ( $p < .001$ ). Although the majority of REACH participants were women, male gender was associated with a lower post-intervention HbA1c by -0.42 ( $p < .05$ ).

**SMB (Self-Management Behavior).** In the equation for post-intervention smb, the only significant predictor was a drop in diabetes distress. I.E., a drop in diabetes-related distress was associated with an increase in self-management behavior. While the coefficient for diabetes-related distress is significant at the  $p < .05$  level, the coefficient of -0.004 does not have a clear physical interpretation, although it indicates that lower distress is associated with better self-management. So, if diabetes distress dropped by 10 points, that would correspond to a .04 increase in self-management on a 1-to-4 point scale.

One possible reason for the low diabetes distress coefficient is that the effect is larger for African Americans than for Latino/as. From a multiple regression for the self-management equation in the sem model for African Americans, the diabetes distress coefficient was -0.01 ( $p < .001$ ) and self-efficacy was borderline significant with a point estimate of 0.006 ( $p \approx .05$ ).

**Knowledge of Diabetes Management.** Higher post-intervention knowledge was associated with better class attendance and attending classes in the group, rather than one-on-one, format ( $p < .05$ ). Knowledge was measured on a scale of 1-to-5, with higher values indicating better knowledge. For each intervention class attended the average increase in knowledge was 0.03 ( $p < .05$ ). Clients who received the intervention in group format had 0.25 greater increase in knowledge than clients who attended one-on-one with their FHAs. Therefore, if a client attended 10 classes and each class was in group format, the average increase in knowledge would be .55, which is approximately half a point on a four point scale.

**Diabetes-Related Distress.** Attending the "self-management" section of the intervention in group format was associated with a -5.86 ( $p < .01$ ) average drop in Diabetes distress. A six point drop in Diabetes distress on a 100 scale is considered a clinically significant, as well as a statistically significant improvement.

**Self-Efficacy.** Post-intervention self-efficacy increased on the average of 1.38 ( $p < .001$ ) for each intervention class attended. If a client attended 10 classes, average self-efficacy would increase by 13.8 on a 100 point scale.

**Goodness of Fit Indices.** In an ideal model, the overall  $\chi^2$  will be non-significant. As long as the other goodness of fit indices indicate reasonable fit, the model is considered acceptable, especially if the model coefficients have reasonable values.

Although the  $\chi^2$  for the REACH SEM is significant, GFI = 0.9928, meaning that the model explains 99.28% of the generalized variance. GFI (Goodness of Fit Index) is a measure of absolute fit and is analogous to  $R^2$  in linear

regression. GFI estimates the proportion of the sample covariance that is explained by the model. A GFI above .9 indicates a good absolute fit<sup>2,6</sup>.

The SRMR (Standardized Root Mean Square Residual) measures predictive fit and  $SRMR < .10$  is the goal<sup>4,16</sup>. For the REACH SEM,  $SRMR = 0.0464$ .

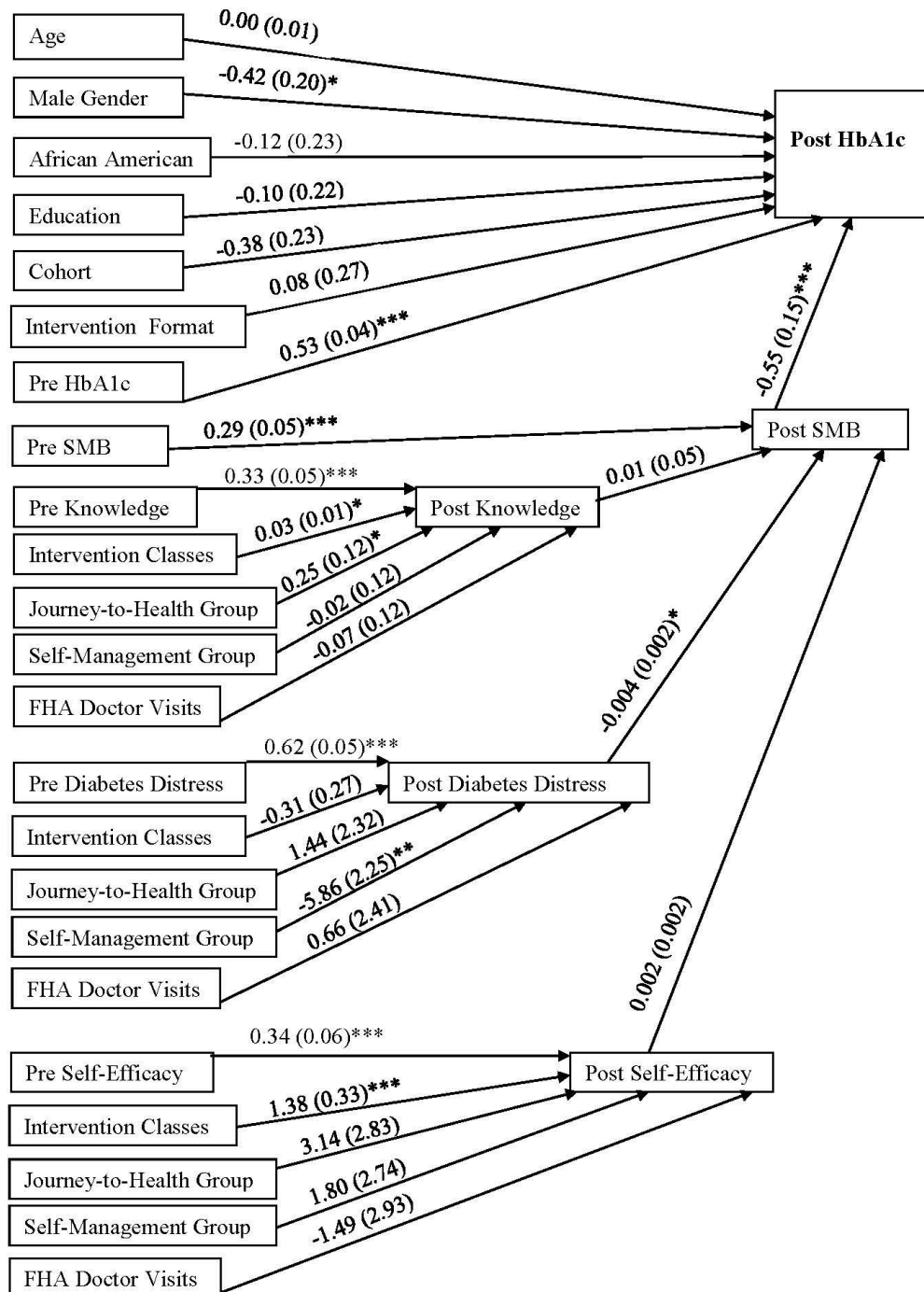
The RMSEA (Root Mean Square Error of Approximation) is a parsimony-adjusted index.  $RMSEA < .05$  is considered ideal and .05 to .08 indicates acceptable parsimony<sup>1,2,15</sup>. For the REACH SEM, RMSEA is 0.0670, with a 90% confidence interval of (0.0533, 0.0810).

CFI (Comparative Fit Index) is an incremental fit index that compares the hypothesized model to the null or intercept-only model. A value of CFI from .90 - .95 is considered acceptable, while above .95 indicates a better incremental fit<sup>2,14</sup>. For REACH, CFI is 0.9353.

**Table 4: Goodness of Fit Indices for the REACH-Detroit SEM**

<b>Fit Index</b>	<b>Value</b>
<b>Chi-Square</b>	142.9647
<b>Chi-Square DF</b>	58
<b>Pr &gt; Chi-Square</b>	<.0001
<b>Goodness of Fit Index (GFI)</b>	0.9928
<b>SRMR (Standardized Root Mean Square Residual)</b>	0.0464
<b>RMSEA Estimate</b>	0.0670
<b>RMSEA Lower 90% Confidence Limit</b>	0.0533
<b>RMSEA Upper 90% Confidence Limit</b>	0.0810
<b>Probability of Close Fit</b>	0.0222
<b>Bentler Comparative Fit Index</b>	0.9353

Figure 4: SEM Coefficients, Standard Errors, and Significance Levels



Note: \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$

## Comparison of SAS Proc CALIS, SPSS AMOS Module, and Mplus.

In SAS, Proc CALIS is the procedure for SEM. In SPSS, the AMOS module is used for SEM analysis. CALIS is an abbreviation for "Covariance Analysis and Linear Structural Equations" and AMOS stands for "Analysis of Moment Structures". SAS version 9.3 was compared to SPSS version 19 and Mplus version 6.1 in the same way that the FIML and ML estimates were compared earlier in this paper. Agreement ratios were calculated for each coefficient estimate and standard error by the formula (Other Software parameter) / (SAS parameter) and are displayed in Table 5.

The coefficient point estimates were nearly identical between SAS, SPSS, and Mplus with mean and median agreement ratios close to 1. However, the standard errors were slightly larger in AMOS and Mplus than in SAS.

**Table 5: Agreement Ratios Between SAS Proc CALIS, SPSS AMOS, MPLUS**

	<b>Coefficient Point Estimate SPSS / SAS</b>	<b>Coefficient Standard Error SPSS / SAS</b>	<b>Coefficient Point Estimate MPLUS / SAS</b>	<b>Coefficient Standard Error MPLUS / SAS</b>
<b>Average</b>	1.00	1.03	1.01	1.07
<b>Median</b>	1.00	1.00	1.00	1.01
<b>Min</b>	0.99	0.93	0.17	0.96
<b>Max</b>	1.00	1.38	1.54	1.55

## Conclusions.

- SEM is an effective method of modeling the process by which variables change. If a series of regression equations were used for outcomes  $Y_1$  to  $Y_q$ , SEM provides a method to model the sequence in the changes in  $Y_1$  to  $Y_q$ , rather than looking at each  $Y_i$  separately.
- HbA1c is the key outcome in Diabetes studies. An effective intervention will reduce HbA1c by at least .5. By looking at unadjusted HbA1c means from pre-intervention to post-intervention in REACH-Detroit cohorts 1 and 2, there are significant drops. SEM provides insight into the process by which HbA1c changes through the REACH behavioral intervention.
- One or more participation variables were associated with changes in knowledge, diabetes distress, and self-efficacy. When intervention format was significant, group format was always more beneficial than the one-on-one format, which is good for cost-effectiveness because group format is less expensive.
- When the endogenous variables are approximately normal, SAS Proc CALIS, SPSS/AMOS, and M-Plus produce the approximately the same results.

## REFERENCES

1. Center for Statistical Consultation And Research, University of Michigan. Applied structural equation modeling, may 10 - 14. 2010.
2. Kline RB. *Principles and Practice of Structural Equation Modeling*. 3rd ed. New York: The Guilford Press; 2011.
3. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*. 5th ed. New Jersey: Prentice Hall; 2002.
4. Kenny DA. Mediation. <http://davidakenny.net/cm/mediate.htm>. Accessed 1/6/2012.
5. Bollen KA. *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons, Inc.; 1989.
6. Joreskog KG, Sorbom D. Recent developments in structural equation modeling. *Journal of Marketing Research*. 1982;19(4):404-416.
7. SAS Institute. The CALIS procedure.  
[http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#calis\\_toc.htm](http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#calis_toc.htm). Accessed 7/26, 2011.
8. Gao S, Mokhtarian PL, Johnston RA. Non-normality of data in structural equation models. *Transportation Research Board's 87th Annual Meeting*.
9. Joreskog KG, Sorbom D. LISREL: Structural equation modeling. *Scientific Software International*. Chicago, IL:1996;8.
10. Hatcher L. *A Step-by-Step Guide to using SAS for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute; 2003.
11. Suhr DD. Exploratory or confirmatory factor analysis? *SAS Global Forum Proceedings*. 2006.
12. Estep D. Course notes for stat 560, Colorado State University. 2011;1:224-248.
13. West BT, Welch KB, Galecki AT. *Linear Mixed Models: A Practical Guide using Statistical Software*. Boca Raton, FL: Chapman Hall / CRC Press; 2007.
14. Bentler PM. Comparative fit indices in structural models. *Psychological Bulletin*. 1990;107(2):238-246.
15. Steiger JH. Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*. 1990;25(2):173-180.
16. Hu L, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 1999;6:1-55.
17. Loehlin JC. *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1992.
18. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*. 1996;1:130-149.
19. Fitzgerald JT, Davis WK, Connell CM, Hess GE, Funnell MM, Hiss RG. Development and validation of the diabetes care profile. *Eval Health Prof*. 1996;19:208-230.
20. Williams GC, Freedman ZR, Deci EL. Supporting autonomy to motivate patients with diabetes for glucose control. *Diabetes Care*. 1998;21:1644-1651.
21. Polonsky WH, Anderson BJ, Lohrer PA, et al. Assessment of diabetes-related distress. *Diabetes Care*. 1995;18(6):754-760.
22. Polonsky WH, Welch G. Listening to our patients' concerns: Understanding and addressing diabetes-specific emotional distress. *Diabetes Care*. 1996;9:8-11.
23. Toobert DJ, Hampson SE, Glasgow RE. The summary of diabetes self-care activities measure. *Diabetes Care*. 2000;23(7):943-950.
24. Agresti A, Mehta CR, Patel NR. Exact inference for contingency tables with ordered categories. *JASA*. 1990;85:453-458.
25. Anupama N, Watts D. Exact methods in the NPAR1WAY procedure. *SAS Users Group International*. 1996.
26. Balanda KP, MacGillivray HL. Kurtosis: A critical review. *The American Statistician*. 1988;42(2):111-119.



27. Mardia KV. Measures of multivariate skewness and kurtosis with applications. *Biometrika*. 1970;57(3):519-530.
28. Institute S. Multivariate analysis concepts. [support.sas.com/publishing/pubcat/chaps/56903.pdf](http://support.sas.com/publishing/pubcat/chaps/56903.pdf). Accessed 2/27/2012.
29. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. *Giorn. 1st. Ital. Attuari*. 1933(4):83-91.
30. Smirnov NV. Tables for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*. 1948;19:279-281.
31. SAS Institute. Proc univariate. [http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#univariate\\_toc.htm](http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#univariate_toc.htm). Accessed 7/27/2011.
32. Peng G, Lilly E. Testing normality of data using SAS, paper PO04. *SUGI*. 2004:1-6.
33. Bentler PM, Chou CP. Practical issues in structural modeling. *Sociological methods research*. 1987;16:78-117.
34. Little RJA. A test of missing completely at random for multivariate data with missing values. *JASA*. 1988;83:1198-1202.
35. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. New York: John Wiley; 2002.
36. Geldhof GJ, Selig JP. Using SAS for multiple imputation: The MI and MIANALYZE procedures. *Fall Colloquium*. 2007.
37. Welch K. Introduction to proc MI and proc MIAnalyze for multiple imputation of missing data in SAS. *Michigan SAS Users Group Conference*.
38. Yung YF, Zhang W. Making use of incomplete observations in the analysis of structural equation models: The CALIS procedure's full information maximum likelihood method in SAS/STAT 9.3. *SAS Global Forum Proceedings*. 2011.
39. Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*. 2001;8:430-457.
40. Schafer JL, Olsen MK. Multiple imputation for multivariate missing data problems: A data analyst perspective. *Multivariate Behavioral Research*. 1998;33:545-571.

## ACKNOWLEDGEMENTS

Thanks to Dr. Phil Chapman for his time and energy as my masters project advisor at Colorado State University, Fort Collins. Thanks to Dr. Myung-Hee Lee and Dr. Michael Lacy for serving on my masters project committee.

I would like to acknowledge the REACH Detroit Project Management Staff, Family Health Advocates, and clients for their innovative work. Thanks to my colleagues: Edith Kieffer, Michael Spencer, Gloria Palmisano, Michael Anderson, Michele Heisler, and Ann Marie Rosland. Thanks to the Center for Disease Control and Prevention (CDC) for funding the first two cohorts of REACH Detroit (Cooperative Agreement No. U50/CCU417409) and to the National Institutes of Health for funding the third cohort (Grant R18DK0785501A1).

On a personal level, I would like to thank my parents for raising their children to value education.

## CONTACT INFORMATION

Your comments and questions are welcome and encouraged. Contact the author at:

Ms. Brandy R. Sinco  
University of Michigan  
School of Social Work  
1080 S. University St.  
Box 183  
Ann Arbor, MI 48109-1106

Phone: 734-763-7784  
Fax: 734-763-3372  
E-Mail: [brsinco@umich.edu](mailto:brsinco@umich.edu)

## Appendix: SAS Code for Multivariate Normality, Multiple Imputation, Path Model for REACH Detroit

```
/* SAS Code to test for Multivariate Normality */
/* Compute Mahalanobis Distance, Diagnostic Statistics, Histogram & QQ Plots */
proc princomp data=SEMDData std out=out_prin;
  var Post_H1c Post_V110 Post_SMB Post_PAID Post_SelfEff_MH;
run;

data mahalanobis;
  set out_prin;
  d2 = uss(of prin:);  run;

Proc Univariate Data=Mahalanobis_endog;
  Var d2;
  Label mahalanobis_distance_to_mean="Squared Distance";
  Symbol1 V=Dot;

Histogram / Gamma(Alpha=2.5 Sigma=2 Theta=0 Fill)
  CFill=Blue Name="Data W/Post";  /* alpha = # variables / 2 */
  Inset Gamma(KSD KSDPval) / Header='Goodness of Fit Test'
    Position=(95,95) RefPoint=TR;

  QQPlot / Gamma(Alpha=2.5 Sigma=2 Theta=0 L=1)
    PctlMinor PCTLSCALE Name="5 Endogenous Variables";

Run;

/* SAS Code for Multiple Imputation */
*** Step 1: Impute with Proc MI ***;
*** Include WithFlag (Withdrawal as covariate) ***;
Proc MI Data=TestMissing NImpute=20 seed=11292011 out=ImputeSEM;
  Var WithFlag Age_BLIW GenderInd RaceInd Cohort2 Randomization2 HSGrad
  TotalClasses FHADrVisitsBinJrnyHlthGrp SelfMgtGrp
  Pre_V110 Pre_SMB Pre_PAID Pre_SelfEff_MH
  Post_V110 Post_SMB Post_PAID Post_SelfEff_MH
  Pre_H1cPost_H1c ;
Run;

*** Step 2: Combine estimates from Proc CALIS ***;
Proc Calis Covariance Method=ML VARDEF=N Data=ImputeSEM;
  LINEQS ... list line equations;
  VARIANCE List variances;
  VAR List variables;
  by _IMPUTATION_;
  ODS Output LINEQSEQ=LINEQSEQ;
Run;
/* To combine the estimates from the 20 imputations, I downloaded the estimates from
Proc CALIS and combined them in Excel by using the equations on page 9 */
```

```

*** SAS Proc CALIS Code for Path Analysis ***;
ods html newfile=proc path="c:\tempfiles";
ods graphics on;
Proc Calis Covariance Residual Modification Kurtosis Method=FIML plots=residuals
Data=SEMDData;
LINEQS
Post_V110 = P_PostV110_PreV110 Pre_V110 +
  P_PostV110_TotalClasses TotalClasses + P_PostV110_FHADrVisitsBin FHADrVisitsBin +
  P_PostV110_JrnyHlthGrp JrnyHlthGrp + P_PostV110_SelfMgtGrp SelfMgtGrp +
  E_PostV110,

Post_PAID = P_PostPAID_PrePAID Pre_PAID +
  P_PostPAID_TotalClasses TotalClasses + P_PostPAID_FHADrVisitsBin FHADrVisitsBin +
  P_PostPAID_JrnyHlthGrp JrnyHlthGrp + P_PostPAID_SelfMgtGrp SelfMgtGrp +
  E_PostPAID,

Post_SelfEff_MH = P_PostSelfEff_MH_PreSelfEff_MH Pre_SelfEff_MH +
  P_PostSelfEff_MH_TotalClasses TotalClasses + P_PostSelfEff_FHADrVisitsBin
FHADrVisitsBin +
  P_PostSelfEff_JrnyHlthGrp JrnyHlthGrp + P_PostSelfEff_SelfMgtGrp SelfMgtGrp +
  E_PostSelfEff_MH,

Post_SMB = P_PostSMB_PreSMB Pre_SMB + P_PostSMB_PostV110 Post_V110 +
  P_PostSMB_PostPAID Post_PAID + P_PostSMB_PostSelfEff Post_SelfEff_MH +
  E_PostSMB,

Post_H1c = P_PostH1c_PreH1c Pre_H1c + P_PostH1c_PostSMB Post_SMB +
  P_PostH1c_Age Age_BLIW + P_PostH1c_Gender Genderind +
  P_PostH1c_RaceInd RaceInd + P_PostH1c_Cohort2 Cohort2 +
  P_PostH1c_Rand Randomization2 + P_PostH1c_HSGrad HSGrad +
  E_Post_H1c;

VARIANCE
E_PostV110 = Var_PostV110,
E_PostSMB = Var_PostSMB,
E_PostPAID = Var_PostPAID,
E_PostSelfEff_MH = Var_PostSelfEff_MH,
E_Post_H1c = Var_PostH1c,
Pre_H1c = Var_PreH1c,
Pre_V110 = Var_PreV110,
Pre_SMB = Var_PreSMB,
Pre_PAID = Var_PrePAID,
Pre_SelfEff_MH = Var_PreSelfEff_MH,
Age_BLIW = Var_AgeBLIW,
GenderInd = Var_GenderInd,
RaceInd = Var_RaceInd,
HSGrad = Var_HSGrad,
Cohort2 = Var_Cohort2,
Randomization2 = Var_Randomization2,
JrnyHlthGrp = Var_JrnyHlthGrp,
SelfMgtGrp = Var_SelfMgtGrp,
FHADrVisitsBin = Var_FHADrVisitsBin,
TotalClasses = Var_TotalClasses;

VAR
Pre_H1c Pre_V110 Pre_SMB Pre_PAID Pre_SelfEff_MH
Age_BLIW Genderind RaceInd Cohort2 Randomization2 HSGrad
TotalClasses FHADrVisitsBin JrnyHlthGrp SelfMgtGrp
Post_H1c Post_V110 Post_SMB Post_PAID Post_SelfEff_MH;
ODS Output LINEQSEQ=LINEQSEQ SQMultCorr=SQMultCorr Fit=Fit;
Run;
ods graphics off;
ods html close;

```