Paper 1442-2015

Confirmatory Factor Analysis Using PROC CALIS: A Practical Guide for Survey Researchers

Lindsey M. Philpot, PhD; Crystal Carel, MPH; Sunni A. Barnes, PhD

Baylor Scott & White Health

ABSTRACT

Survey research can provide a straightforward and effective means of collecting input on a range of topics. Survey researchers often like to group similar survey items into construct domains in order to make generalizations about a particular area of interest. Confirmatory Factor Analysis is used to test whether this pre-existing theoretical model underlies a particular set of responses to survey questions. Based on Structural Equation Modeling (SEM), Confirmatory Factor Analysis provides the survey researcher with a means to evaluate how well the actual survey response data fits within the a priori model specified by subject matter experts. PROC CALIS now provides survey researchers the ability to perform Confirmatory Factor Analysis using SAS®. This paper provides a survey researcher with the steps needed to complete Confirmatory Factor Analysis using SAS. We discuss and demonstrate the options available to survey researchers in the handling of missing and "not applicable" survey responses using an ARRAY statement within a DATA step and imputation of item non-response. A simple demonstration of PROC CALIS is then provided with interpretation of key portions of the SAS output. Using recommendations provided by SAS from the PROC CALIS output, the analysis is then modified to provide a better fit of survey items into survey domains.

INTRODUCTION

Survey assessments and questionnaires can be used in a variety of settings to ascertain customer, employee, and patient feedback. Survey questionnaires can be easy to develop and administer, but the analytics of survey responses can be more challenging. Fortunately social scientists have developed analytic and modeling techniques to allow for greater understanding of survey-based data. In this example we focus on Confirmatory Factor Analysis (CFA), a type of Structural Equation Modeling (SEM) technique, to further understand the results of data generated from a survey instrument.

During the development of a survey tool, subject matter experts may be engaged to assign individual survey items to construct domains. Construct domains are used to group survey questions into themes by grouping survey questions that seek to gather information related to a central concept or idea (Schumacker & Lomax, 2010). Statistically speaking, we call these domains "factors" or "latent variables". Confirmatory Factor Analysis seeks to confirm that the survey item assignment to construct domains is supported by the variance – covariance observed in the survey data.

Throughout the context of this paper we will use a single real-life example (Xiao et al., 2014). In 2011 a group of healthcare leadership and administrators sought to understand the impact of a newly implemented electronic health record (EHR) system on hospital-based, direct care nursing staff. A novel survey tool was developed based on literature review. Subject matter experts were then engaged to assign individual survey items into one of five pre-determined concept domains (Table 1). Further information on the development of the survey tool can be found elsewhere (Xiao et al., 2014). During the data collection period, a total of 1,301 nurses responded to the 25-item survey tool. The example procedures were performed using SAS Enterprise Guide, Version 6.1 (BASE 9.3).

Concept Domains (Factors)
Training/Competency
Usability
Usefulness
Infrastructure
End User Support

Table 1. Concept Domains (Factors) for the Baylor Health Care System EHR End User Experience Survey

USING PROC CALIS TO PERFORM CONFIRMATORY FACTOR ANALYSIS ON SURVEY RESPONSE DATA

PREPARING YOUR SURVEY RESPONSE DATA

Once construct domain assignments have been made by subject matter experts, and survey response-data has been collected, you are ready to explore the fit of your *a priori* construct domain assignments. However, before we jump into defining our model using PROC CALIS, we may need to perform a couple of DATA steps to prepare our survey response data. This paper will discuss two common survey respondent data issues: missing/incomplete data and selection of a "not applicable" answer choice. To help facilitate this conversation, Table 2 indicates the survey response labels with associated response coded values for our EHR User Experience Survey example.

Survey Response Label	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Not Applicable
Coded Response Value	1	2	3	4	5	6

Table 2. Survey Response Labels and Survey Coded Values for Baylor Health Care System EHR End User Experience Survey

Handling Item Non-Response

PROC CALIS uses complete-case only data during the modeling. In other words, if a particular survey respondent chose to skip survey items or did not complete their entire survey, and therefore has missing data elements, the PROC CALIS procedure will not include the respondent in the analysis. A survey researcher may decide to perform their Confirmatory Factor Analysis based on complete case-only data. The researcher may also decide to impute the mean item responses to those cases that are missing, or to impute the "neutral" survey item. Deciding between imputation techniques is beyond the scope of this paper, so the following DATA step with an ARRAY statement illustrates how the researcher would impute the neutral response for missing item values. The value of "3" is imputed as the neutral value as the coded responses from our survey response data is as follows in Table 2. The DATA step including ARRAY statement for this process is:

```
data work.ehrnurse2;
set work.ehrnurse1;
array replace_missing [25]
Q5_a Q5_b Q5_c Q5_d Q5_e Q5_f Q5_g Q5_h Q5_i Q5_j Q5_k Q5_l Q5_m Q5_n
Q9_a Q9_b Q9_c Q9_d Q9_e Q9_f Q11_a Q11_b Q11_c Q11_d Q11_e;
do a=1 to 25;
if replace_missing [a]=. then replace_missing [a]=3;
end;
run;
```

Handling Response Selection of "Not Applicable"

Another common occurrence when analyzing survey respondent data is the response option of "Not Applicable". This response option may be provided on all or a subset of survey items. This response option is often provided when there is a chance that survey respondents are not involved or have no exposure to the topic of the survey question. For example, in the EHR User Experience Survey (Xiao et al., 2014), survey respondents may answer "Not Applicable" to the question "The training I received related to the EHR was effective". This answer choice was provided in case survey participants did not receive training related to the EHR.

As noted in Table 2, when a survey respondent selects "Not Applicable" to a survey item, the coded value generated is a "6". The selection of "Not Applicable" is intentional on behalf of the survey respondent. The survey respondent is indicating that the survey question is not applicable to his/her situation, and that he/she therefore is unable to provide an opinion. This should be considered a different situation than if a respondent leaves a survey question blank, as this may be intentional or unintentional.

There are three options that are commonly deployed when a "Not Applicable" response is selected: set the response to missing, set the response to neutral, or set the response to the question answer mean. For this example, the survey researcher has chosen to set the "Not Applicable" responses to missing, even though the PROC CALIS procedure will no longer include the survey response in the model. An ARRAY statement within a DATA step is used to perform this action:

```
data work.ehrnurse2;
set work.ehrnurse1;
array replace_na [25]
Q5_a Q5_b Q5_c Q5_d Q5_e Q5_f Q5_g Q5_h Q5_i Q5_j Q5_k Q5_l Q5_m Q5_n
Q9_a Q9_b Q9_c Q9_d Q9_e Q9_f Q11_a Q11_b Q11_c Q11_d Q11_e;
do a=1 to 25;
if replace_na [a]=6 then replace_na [a]=.;
end;
run;
```

Of our original 1,301 survey responses, 1,187 remained after those with any "Not Applicable" responses were not included due to coding the value of "6" to missing data.

SETTING UP YOUR PROC CALIS STATEMENTS

Once your survey response data is ready for Confirmatory Factor Analysis, you will need to program your PROC CALIS statements. Here is a simplistic example of a CFA model using PROC CALIS:

```
proc calis data=work.ehrnurse2 [1]nobs=1187 [2]modification;
[3]factor
      [4]FTraining       ---> [5]Q5_a Q9_d Q5_d Q5_b,
      FUsability         ---> Q11_e Q9_a Q5_c Q9_b Q5_f Q5_i Q5_j,
      FUsefulness        ---> Q5_e Q5_g Q5_k Q5_l Q5_m Q9_f,
      FInfrastructure ---> Q11_a Q11_b Q11_c Q11_d,
      FUserSupport       ---> Q9_c Q9_e Q5_h;
[6]pvar
FTraining FUsability FUsefulness FInfrastructure FUserSupport = 5 * 1;
[7]cov
FTraining FUsability FUsefulness FInfrastructure FUserSupport = 5 * 0;

run;
```

[1]Specify the number of survey respondents in your dataset
[2]The 'Modification' option tells SAS you want recommendations to create a better fitted model
[3]Factor is where you outline the factor – variable relationships you seek to evaluate
[4]F[Concept domain name] is how we labeled our factors
[5]List of survey items associated with a factor
[6]Specifies the parameters for factor variances and error variances of manifest variables
[7]Specifies any suspected correlations between factors

In the above example statement, we have included the total number of observations within our dataset, and requested the 'modification' option. By using the 'modification' option, SAS will generate a series of suggested modifications you can make to your model in order to make a better fit. Next, we have outlined the five construct domains listed in Table 1, then associated the individual survey items with the construct domains using "arrows". In this sense, we have outlined the factor – variable relationships that we seek to evaluate in our model. Within the 'pvar' option, we indicate the factor and error variables of the model. In this example we use 5*1 to indicate 5 factors with a variance of 1. This allows us to fix the variances in order to perform model identification.The researcher can use this option to specify which, if any factors may be correlated. For this initial example, we have not included the statement to indicate that we expect no correlation.

The final option we will discuss in this paper is the 'cov' option. For this initial example, we set the correlation between factors to zero (5 * 0) in order to allow the PROC CALIS function to generate recommended correlations between our factors. In your example, you will want to think about whether you factors may be correlated, and assign them properly.

INTERPRETING YOUR PROC CALIS RESULTS

The PROC CALIS procedure provides the survey researcher with a wealth of information and output. Since the purpose of this paper is to provide a basic understanding of the use of PROC CALIS, we will highlight some of the output but will not provide a comprehensive overview. Those with further interest in the results of PROC CALIS are encouraged to reference SAS documentation at: http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_calis_sect077.htm .

The PROC CALIS procedure provides a couple of initial tables depicting the characteristics of the data set used (Figure 1), variable and factor information (Figure 2), factor loading matrix (Figure 3), and additional variable and factor statistics. Through review of these initial output tables, the survey researcher can ensure that the correct number of observations was used, that the correct variables were selected, and that the correct survey items were assigned to the proper factor.

Modeling Information	
Data Set	WORK.EHRNURSE2
N Records Read	1301
N Records Used	1187
N Obs in Dataset	1187
N Obs Used	1187
Model Type	FACTOR
Analysis	Covariances

Figure 1. Basic Modeling Information from PROC CALIS Output

Variables in the Model	
Variables	Q5_a Q5_b Q5_c Q5_d Q5_e Q5_f Q5_g Q5_h Q5_i Q5_j Q5_k Q5_l Q5_m Q9_a Q9_b Q9_c Q9_d Q9_e Q9_f Q11_a Q11_b Q11_c Q11_d Q11_e
Factors	FTraining FUsability FUsefulness FInfrastructure FUserSupport
Number of Variables = 24 Number of Factors = 5	

Figure 2. Variable and Factor Information from PROC CALIS Output

Initial Factor Loading Matrix					
	FTraining	FUsability	FUsefulness	FInfrastructure	FUserSupport
Q5_a	.	0	0	0	0
	[_Parm01]				
Q5_b	.	0	0	0	0
	[_Parm04]				
Q5_c	0	.	0	0	0
		[_Parm07]			

Figure 3. Factor Loading Matrix from PROC CALIS Output (abbreviated)

Once the survey researcher has reviewed the initial PROC CALIS output to ensure that the data was loaded properly and that the survey items were assigned to the correct construct domain, a review of the model statistics will indicate how well the *a priori* model was fit. Figure 4 provides the fit summary statistics for the model.

Fit Summary		
Modeling Info	N Observations	1187
	N Variables	24
	N Moments	300
	N Parameters	58
	N Active Constraints	0
	Baseline Model Function Value	11.6300
	Baseline Model Chi-Square	15212.0712
	Baseline Model Chi-Square DF	276
	Pr > Baseline Model Chi-Square	**<.0001**
Absolute Index	Fit Function	1.5475
	Chi-Square	2024.1157
	Chi-Square DF	242
	Pr > Chi-Square	**<.0001**
	Z-Test of Wilson & Hilferty	34.0168
	Hoelter Critical N	**214**
	Root Mean Square Residual (RMSR)	0.0830
	Standardized RMSR (SRMSR)	0.0624
	Goodness of Fit Index (GFI)	**0.8766**
Parsimony Index	**Adjusted GFI (AGFI)**	**0.8470**
	Parsimonious GFI	0.7686
	RMSEA Estimate	**0.0750**
	RMSEA Lower 90% Confidence Limit	**0.0720**
	RMSEA Upper 90% Confidence Limit	**0.0781**
	Probability of Close Fit	<.0001
	ECVI Estimate	1.6379
	ECVI Lower 90% Confidence Limit	1.5301
	ECVI Upper 90% Confidence Limit	1.7515
	Akaike Information Criterion	2140.1157
	Bozdogan CAIC	2498.3828
	Schwarz Bayesian Criterion	2440.3828
	McDonald Centrality	0.5063
Incremental Index	Bentler Comparative Fit Index	0.8807
	Bentler-Bonett NFI	**0.8669**
	Bentler-Bonett Non-normed Index	0.8639
	Bollen Normed Index Rho1	0.8482
	Bollen Non-normed Index Delta2	0.8810
	James et al. Parsimonious NFI	0.7601

Figure 4. Model Fit Summary from PROC CALIS Output

When looking at the Modeling Info, the researcher can see that the Baseline Model Chi-Square estimate is less than 0.05. Additionally, the Chi-Square estimates for the Absolute Index are also less than 0.05. This indicates that the *a priori* model may not be the best fit for the survey response data.

Under the Absolute Index, the researcher can see that the Hoelter Critical N is greater than 200, indicating that the sample size for the model is adequate to assess the model fit.

The Goodness of Fit Index (GFI) and the Adjusted GFI (AGFI) are interpreted similar to an R^2 estimate, in that an estimate closer to one indicates a better fitting model. In this example, we see that the GFI is estimated at 0.8766 and the AGFI is estimated at 0.8470. The Goodness of Fit estimates here indicate a relatively well-fitted model.

Under the Parsimony Index, PROC CALIS estimates Root Mean Square Error of Approximation (RMSEA). Interpretation of this statistic is that the closer to zero, the better the model fit. Less than 0.05 is preferable, which the example model does not meet.

The final model statistic that will be highlighted in this paper is the Bentler-Bonett Normed Fit Index (NFI). The Incremental Index statistics are testing whether the fitted model is a better fit than an independence model. The preferable estimate is one that is greater than 0.80, which this model achieves.

Based on the model fit statistics, the survey researcher may conclude that the *a priori* model may not be the best fitted model based on our survey respondent data. Fortunately, through the PROC CALIS option 'modification', SAS generates alternative factor assignments, covariance structures, and error variances and covariances that the survey researcher may consider to create a better fitted model.

ADJUSTING YOUR MODEL BASED ON PROC CALIS RECOMMENDATIONS

When the survey researcher specifies the 'modification' option within the PROC CALIS procedure, SAS generates a series of tables at the end of the SAS output that provides recommended changes to the factor – variable relationships (Figure 5), and the correlations among factors (Figure 6). It is important to note that the survey researcher should evaluate the logic behind the reassignment of survey items and factors characteristics based on the respondent data alone.

Rank Order of the 10 Largest LM Stat for Factor Loadings

Variable	Factor	LM Stat	Pr > ChiSq	Parm Change
Q9_d	**FUserSupport**	**9.55763**	**0.0023**	**2.18756**
Q11_e	FInfrastructure	2.14354	0.0672	0.55587
Q5_c	FTraining	1.56507	0.1387	0.82440
Q5_c	FUsefulness	1.29458	0.1838	-0.60088
Q5_b	FUserSupport	1.11787	0.2548	-0.29672
Q5_h	FUsefulness	1.02898	0.2999	0.56647
Q11_a	FUserSupport	0.72243	0.4338	0.35738
Q11_a	FUsefulness	0.70244	0.4400	0.32909
Q11_a	FUsability	0.54121	0.6988	0.32689
Q11_a	FTraining	0.67732	0.8114	0.28568

Figure 5. PROC CALIS Recommended Item Assignment Model Adjustments

In the above example, PROC CALIS lists data-driven recommendations for the assignment of individual survey items into different construct domains/factors (Figure 5). The first of the recommended item assignments in Figure 5 may have significant impact on the performance of the overall model, given the predicted Chi Square impact p-values (PR>ChiSq) less than 0.05. This adjustment can be made by altering the original PROC CALIS code to reassign individual survey items to different concept domains/factors. See example code below, where survey item Q9_d is reassigned from the original code to factor FUserSupport.

```
proc calis data=work.ehrnurse2 nobs=1187 modification;
      factor
            [4]FTraining       ---> Q5_a Q5_d Q5_b,
            FUsability       ---> Q11_e Q9_a Q5_c Q9_b Q5_f Q5_i Q5_j,
            FUsefulness      ---> Q5_e Q5_g Q5_k Q5_l Q5_m Q9_f,
            FInfrastructure ---> Q11_a Q11_b Q11_c Q11_d,
            FUserSupport     ---> Q9_c Q9_e Q5_h Q9_d;
      pvar
FTraining FUsability FUsefulness FInfrastructure FUserSupport = 5 * 1;
run;
```

Rank Order of the 3 Largest LM Stat for Covariances of Factors

var1	var2	LM Stat	Pr > ChiSq	Parm Change
FTraining	**FUserSupport**	**8.95268**	**0.0028**	**0.44165**
FUsability	FInfrastructure	7.70790	0.0708	0.40132
FUsability	FUsefulness	4.61896	0.1216	0.30411

Figure 6. PROC CALIS Recommended Factor Correlation Model Adjustments

Figure 6 shows the second output table generated by SAS when the 'modification' option is used in PROC CALIS. This table lists the three largest correlated construct domains/factors from our survey example. As you can see, PROC CALIS suggests that the model assume FTraining and the FUserSupport factors are correlated in order to improve the performance of the model, as the predicted Chi Square p-value is less than 0.05 (Pr>ChiSq=0.0028). In order to make this particular adjustment, the survey researcher utilizes the cov option within PROC CALIS, as seen in the example code below:

```
proc calis data=work.ehrnurse2 nobs=1187 modification;
      factor
            FTraining       ---> Q5_a Q5_d Q5_b,
            FUsability      ---> Q11_e Q9_a Q5_c Q9_b Q5_f Q5_i Q5_j,
            FUsefulness     ---> Q5_e Q5_g Q5_k Q5_l Q5_m Q9_f,
            FInfrastructure ---> Q11_a Q11_b Q11_c Q11_d,
            FUserSupport    ---> Q9_c Q9_e Q5_h Q9_d;
pvar
FTraining FUsability FUsefulness FInfrastructure FUserSupport = 5 * 1;
cov
FTraining FUserSupport = 2 * 1.,
FUsability, FUsefulness, FInfrastructure;
run;
```

CONCLUSION

PROC CALIS is a powerful tool to allow survey researchers to evaluate the accuracy of construct domain assignments made by subject matter experts. PROC CALIS is a data-driven method that analyzes the assignment of survey items to construct domains based on common variance among survey items. The example within this text walks the survey researcher through a basic application of PROC CALIS based on real survey data. The 'modification' feature of PROC CALIS has SAS make recommended alterations to the *a priori* SEM model; however the survey researcher should use logic to determine whether data-driven recommendations are appropriate to the context of the survey tool and factor model.

REFERENCES

Schumacker, R. E., & Lomax, R. G. (2010). Confirmatory Factor Models *A Beginner's Guide to Structural Equation Modeling* (Third Edition ed., pp. 163-177). New York, NY: Routledge.

Xiao, Y., Montgomery, D. C., Philpot, L. M., Barnes, S. A., Compton, J., & Kennerly, D. (2014). Development of a Tool to Measure User Experience Following Electronic Health Record Implementation. *Journal of Nursing Administration, 44*(7/8), 423-428. doi: 10.1097/nna.0000000000000093

CONTACT INFORMATION

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

Lindsey M. Philpot, PhD MPH
Baylor Scott & White Health
Lindsey.Philpot@baylorhealth.edu