Croon’s Factor Score Regression Analysis

Department of Psychology, Louisiana Tech University

PSYC 720: AAMAIO

Dr. Steven Toaddy

November 17, 2020

Background

Over the past 100 years, structural equation modeling (SEM) has become one of the most popular and influential frameworks of quantitative methodologies in the social sciences (Kaplan, 2008). The recognition is likely due to the method's ability to assign relationships between latent variables and observable variables, as this is essential to psychometrics. Structural equation modeling assesses latent variables via measurement models, while concurrently computing the causal relationships of these variables via structural regression model (Bollen, 1989; Joreskog and Sorbom, 1993; Kline, 2016). Ironically, this process is simultaneously the framework's greatest strength and weakness. Misspecification, at any point in an SEM, can result in bias that proliferates throughout the model, leading to unreliable estimates of parameters as well as regression coefficients (Hayes and Usami, 2020). Misinformed structural relationships can be detrimental for empirical research, as latent construct relationships are often at the foundation of the social and behavioral sciences (Devlieger et al., 2016; Devlieger and Rosseel, 2017; Hancock and Mueller, 2011; Hoshino and Bentler, 2011; Lu et al., 2011). Hancock and Mueller argue that traditional methods are incapable of detecting these misspecifications (2011), which leaves researchers looking for a more reliable analysis.

In the quest to eliminate the bias from their models, an increasing amount of researchers are switching from the simultaneous estimation method to multistage procedures. Different types of factor score regression (FSR) and factor score path analysis are among the stepwise methodologies growing in popularity (Croon, 2002; Devlieger et al., 2016; Devlieger and Rosseel, 2017; Hoshino and Bentler, 2011; Lu et al., 2011). FSR is a two-step procedure in which each latent variable is measured separately in a structural regression model to create corresponding factor scores. The factor scores are typically computed using the regression or

Bartlett predictor. These corrected factors scores are subsequently analyzed in a linear regression, as if they were the true latent variable scores (Devlieger and Rosseel, 2017). Although this model has less issues with convergence than SEM, it is crucial to be careful in removing bias from this data, as factor scores are indeterminate, and thus, not fully reliable (Grice, 2001; Steiger and Schonemann, 1978). In order to account for this bias, several methods have been developed. The two most common approaches are the *bias-avoiding approach* (Skrondal and Laake, 2001) and Croon's *bias-correcting approach* (2002). Recent studies have found that Croon's approach outperforms Skrondal and Laake's, due to its standardized coefficients and its inability to be extended into a path analytic framework (Devlieger et al., 2016; Lu et al., 2011). For these reasons, this paper will focus on Croon's bias-correcting FSR approach.

Croon's approach to factor score regression is based on the premise that there is a difference between the variances and covariances of the factor scores and the variances and covariances of the true latent scores (2002). The approach uses an estimation of the variances and covariances of the true latent scores, rather than factor scores, in order to estimate the parameters (Croon, 2002). Between the studies of Croon (2002), Lu et al. (2011), and Devlieger et al. (2016), the approach has shown to be effective in producing unbiased parameter estimates in both populations and finite samples. When the sample size is large, the method was shown to be comparable in efficiency, mean square error, power, and type-1 error rate as SEM (Devlieger et al., 2016). In 2017, Devlieger et al. found that Croon's method performs just as well as SEM with regard to bias and convergence rate when path analysis is used. They also found evidence that Croon's method handles misspecifications better than SEM and requires a smaller sample size (Devlieger et al., 2017). Hayes and Usami (2020) found that Croon's method outperformed

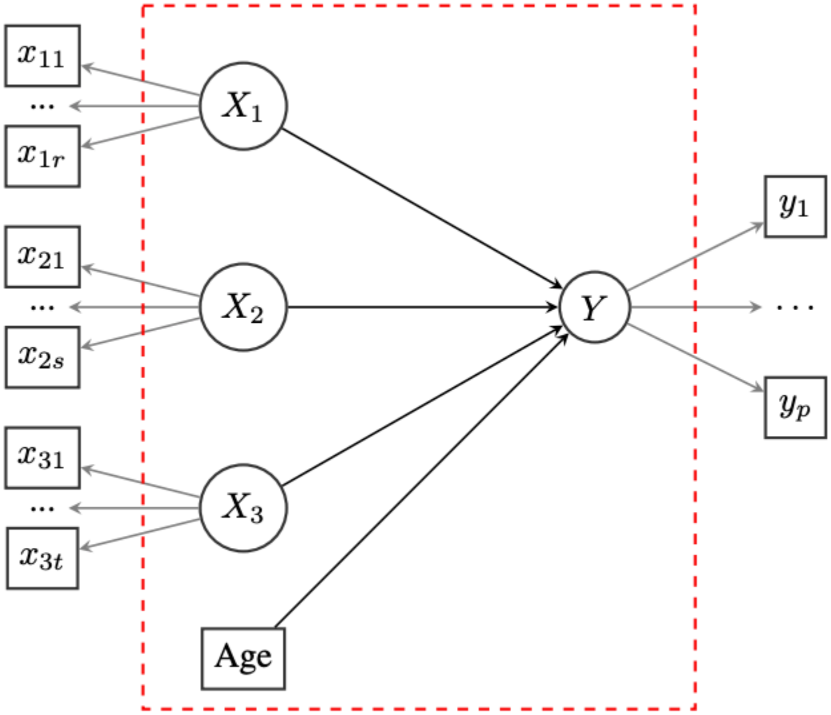
SEM using a standard specification of unique factor covariances. SEM performed comparably well when the unique factor covariances were specified but was outperformed again when the SEM specified the unique factor covariances but misspecified the structural model (Hayes and Usami, 2020). A recent study by Devlieger and Rosseel (2019) had shown that Croon's method outperforms SEM in the multilevel setting, especially when the number of between-level clusters is small. Evidence provided by the researchers, as mentioned above, gives reason to believe that Croon's FSR approach may be a suitable alternative for SEM in certain instances (e.g., misspecifications in parameter or structure, small sample size). In this paper, I will further describe Croon's FSR approach, utilize its methodology using simulated data, and assess its applicability in the field of Industrial-Organizational psychology.

Methodology

Croon's bias correcting factor score regression method was developed by Marcel Croon (2002) to combat the biases that are inherent in factor scores. It is based on the idea that there is a difference between the variances and covariances of the factor scores and the true latent variable scores. In order to eliminate this bias, Croon (2002) uses an estimation of the true latent variable scores, rather than the factor scores, when estimating the regression parameters (Devlieger and Rosseel, 2017). The stepwise approach is as follows: (1) Factor scores are first computed using the regression predictor (Thomson, 1934; Thurstone, 1935) or the Bartlett predictor (Bartlett, 1937; Thomson, 1938); (2) the variances and covariances of theses scores are calculated; which are then used (3) in Croon's proposed formulas (Appendix B) to compute the variances and covariances of the true latent variables; which are (4) used to calculate the regression coefficient between the variables. Because the variances and covariances are unbiased, so is the regression coefficient estimate (Devlieger et al., 2015). If you have a non-recursive model (i.e., models that contain feedback loops or reciprocal effects), you can use the unbiased variance and covariance estimates to perform a path analysis in the fourth step (Devlieger and Rosseel, 2017). Assuming that the measurement instruments that are being used are well established, the main focus of this method is the structural part of the model. In a regression model this emphasizes the independent and dependent variables (Rosseel and Devlieger, 2018; Figure 1). In a path analysis model, this focus is on mediating effects, feedback loops, etc. (Rosseel and Devlieger, 2018; Figure 2).

Figure 1

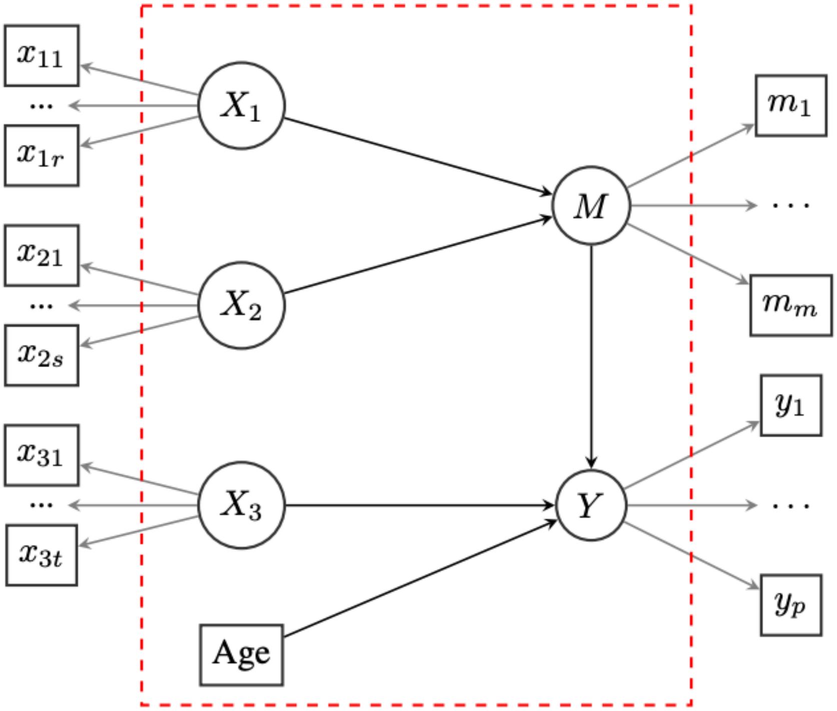
Structural Model: Regression Model



Note: From Rosseel and Devlieger, 2018

Figure 2

Structural Model: Path Analysis Model



Note: From Rosseel and Devlieger, 2018

The Four Steps Explained

The first step of any factor score regression method is to use measurement models to perform a factor analysis for each latent variable and calculate the corresponding factor scores (Devlieger and Rosseel, 2017). The factor score can be computed using either the regression predictor or the Bartlett predictor (Croon, 2002). In order to perform a factor analysis, the scales of the variables need to be fixed. According to Devlieger and Rosseel, this can be done by using standard parameterization, fixing the variance of the latent variable to 1, or unstandardized parameterization, fixing one factor loading per latent variable to 1 (2015).

The next step of any factor score regression is to calculate the variance and covariance of the factor scores. It is the third step of Croon’s method where the approach differs from other forms of factor score regression. The variances and covariance of the factor scores that were obtained in the second step are used to estimate the variances and covariances of the true latent variable scores. This is done by following Croon’s bias-correcting formulas (Appendix B), which takes factor loadings, factor score matrixes, and the covariance matrix of measurement errors into account. The fourth and final step of the factor score regression method is to use the newly found variance and covariance estimates to calculate the regression coefficient. If the model has a mediational relationship (full or partial), one can perform a series of linear regression analyses for each endogenous variable (Devlieger and Rosseel, 2017). However, multiple regression can only be used for recursive path models. If the model is non-recursive, a path analysis may be run to analyze factor scores without bias (Devlieger and Rosseel, 2017).

In order to use Croon’s method for hypothesis testing, it is necessary to have a corresponding significance test, which requires a standard error and a theoretical distribution. Devlieger and colleagues (2015) developed a method for calculating the standard error of Croon’s approach that corresponds with the corrected regression coefficient. This adjusted standard error is created by calculating the prediction error in the factor scores. The prediction error in the factor scores is the difference between the observed and corrected variance, thus allowing for standard error that coincides with the corrected regression coefficient. The proposed standard error formula is found in Appendix C.

The newest addition to factor score regression is a set of fit indices and a model comparison test. In order to inspect model fit, Devlieger and colleagues (2019) propose fit indices for factor score regression based on the Chi-Square, RMSEA, SRMR, and CFI fit indices used for structural equation modeling. Their newly proposed Chi-Square goodness of fit test can be used to conduct a model comparison test, which not only allows us to see how well the model fits Croon’s method, but how it holds up to other approaches (e.g., SEM, bias-avoiding method, etc.). The formulas for these fit indices can be found in Appendix D.

Limitations

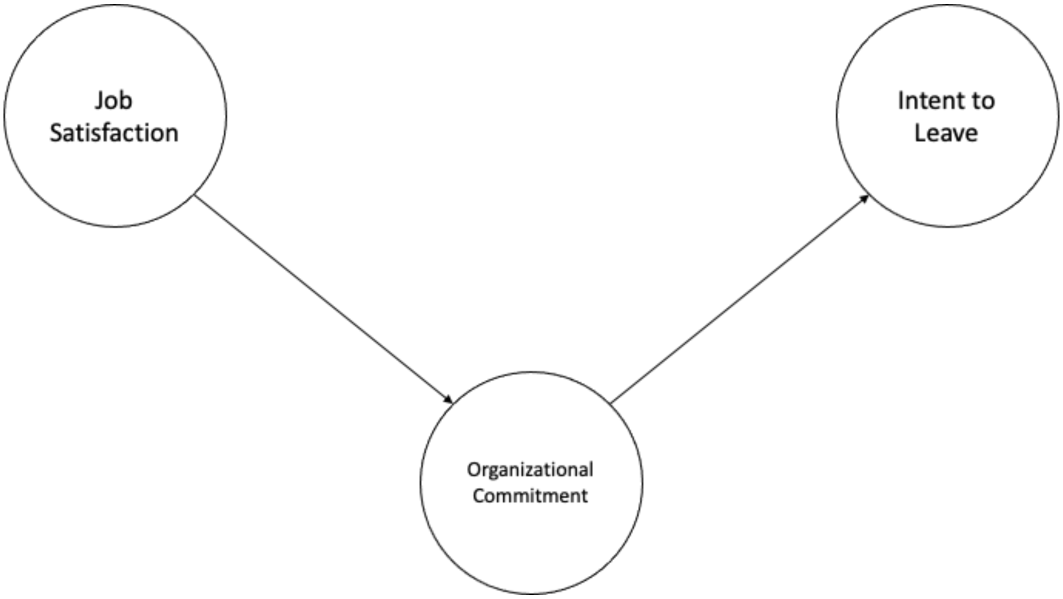
As this bias-correcting method continues to be explored, researchers have uncovered limitations to the approach that should be researched more in-depth. For example, factor score regression method only works when there are at least three items per latent variable. As of now, the Croon approach does not work for connected measurement models (e.g., models with cross-loadings or correlated residual errors). Devlieger and Rosseel (2017) are currently working to enhance the method’s capability of handling these types of models as well as extend the inferences made by the model. Because of the method’s maturity, there is very little software available to perform it – making it a tedious and challenging process. Many of the functions used for this method are still being developed in R by the aforementioned authors, and some functions are not yet available (i.e., matrix-oriented bias correction formulas for unique factor structures; Hayes and Usami, 2020).

Application

To test Croon’s method, I used a simulated dataset (N = 200) that I created to replicate Williams and Hazer’s turnover model (1986; Figure 1). This model looks at the following latent factors: job satisfaction, organizational commitment, and intent to leave the organization. In their paper, they use structural equation modeling to look at the relationship between these variables and subsequent turnover. Because I am looking at latent factor scores, I will only be utilizing the three previously mentioned latent factors of the model in my structured mediation model. Based on Williams and Hazer’s model, I hypothesize that organizational commitment mediates the relationship between job satisfaction and intent to leave.

Figure 1

Hypothesized Mediation Model



In this simulated model, I created five manifest variables for each latent construct (job satisfaction is measured by x1, x2, x3, x4, 5; organizational commitment is measured by, y1, y2, y3, y4, y5; and intent to leave is measured by z1, z2, z3, z4, z5; Table 1). These variables are placeholders for the scores of items used to measure the latent construct. In factor score regression analyses, the manifest variables are used to estimate the factor scores for each latent variable. I used standardized parameterization when creating these observed variables, this fixes the metric scale of the latent variable by setting the variance of the latent variable to 1 (Devlieger and Rosseel, 2015). I used the regression predictor as the method in my factor analysis because it provides factor scores that are maximally correlated to the estimated factor (i.e., most valid; DiStefano et al., 2009).

Table 1

Descriptives of Observed Variables

Observed variable	N	Mean	SD	Min	Max	Range	se
x1	200	0.01	1.07	-2.84	3.65	6.48	0.08
x2	200	0.08	0.98	-2.50	3.10	5.61	0.07
x3	200	0.05	1.00	-2.50	2.99	5.49	0.07
x4	200	0.04	0.99	-2.19	2.73	4.92	0.07
x5	200	0.00	1.07	-2.57	3.60	6.17	0.08
y1	200	0.00	1.49	-4.65	5.58	10.23	0.11
y2	200	0.10	1.43	-3.37	4.50	7.87	0.10
y3	200	0.03	1.36	-3.51	3.56	7.08	0.10
y4	200	0.01	1.27	-3.86	3.40	7.26	0.09
y5	200	-0.08	1.37	-3.75	4.87	8.63	0.10
z1	200	-0.10	1.15	-2.95	3.36	6.31	0.08
z2	200	-0.08	1.32	-3.63	3.84	7.47	0.09
z3	200	-0.04	1.29	-4.27	3.22	7.49	0.09
z4	200	-0.14	1.32	-3.44	3.16	6.60	0.09
z5	200	-0.14	1.20	-3.72	3.22	6.94	0.08

Analysis

I used Croon's bias-correcting factor score regression method to test the mediating relationship between three latent variables: job satisfaction, organizational commitment, and intent to leave. The data included simulated self-report scores on measures of each variable. Because the dataset was simulated, there were no missing data in the dataset. I fit the model using the Structural After Measurement approach in lavaan version 0.6-7 (Rosseel, 2012) in R version 3.60 (R Core Team, 2016). According to the creator of the package, the Structural After Measurement Approach is identical to Croon’s factor score regression approach when the “local” method is applied. As of now, this is the only software that has the capability of performing this method, and it is still in its beta stage. The approach used maximum likelihood estimation to fit the model. Regression factor scores were created for each latent variable. Subsequent steps of Croon’s factor analysis, including the final path analysis, were automated by the lavaan:::sam function. All of the R code can be found in Appendix A.

Results

The model converged normally after 17 iterations and was shown to significantly fit the simulated data. The results showed that job satisfaction directly impacts organizational commitment (R^2 = 0.71), and organizational commitment directly impacts the employees’ intent to leave the organization (R^2 = -0.52; See Figure 2). The standard errors of the regression coefficients were 0.105 and 0.065, respectively. The estimates of the regressions and factor scores can be found in Table 2. The standardized results are found in the last two columns, where standardized latent variable shows the results when the latent variables have a variance of 1 (this is the same as the “estimate” in the first column) and the standardized coefficient column shows the results when both the latent and observed variables have a variance of 1. The latter is reported in my model because it is most similar to regression coefficients, where x and y are both interpreted in terms of z-scores.

To determine the fit of the model, I followed the principles outlined by Rex Kline (2005): structural equation modeling should report the model chi-square, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Standardized Root Mean Square Residual (SRMR). The model chi-square test statistic was significant ($p < 0.05$; Hooper et al., 2008). The CFI (> 0.900; Hooper et al., 2008) and SRMR (< 0.08; Hooper et al., 2008) also met the cutoff for good fit. The model did not reach the RMSEA threshold for good fit. However, the RMSEA is prone to error in models with a small sample size (N) and low degrees of freedom (df). For this reason, Kenny, Kaniskan, and McCoach (2014) argue not to compute RMSEA for low df models. Therefore, this model fit index may not be the best indicator of my model fit, as the N = 200 and df = 1.

Figure 2

Path of Croon's Factor Score Regression Analysis

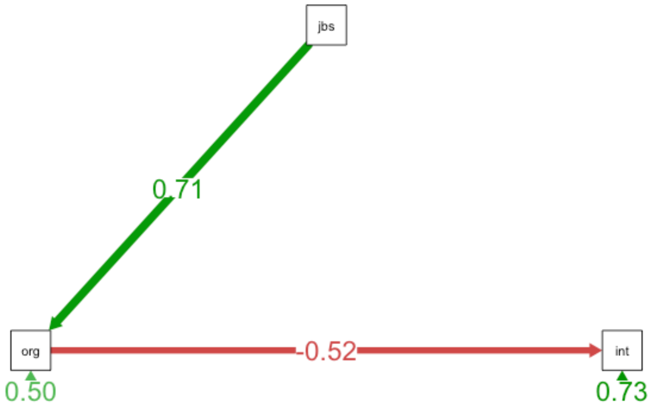


Table 2

Parameter estimates of model

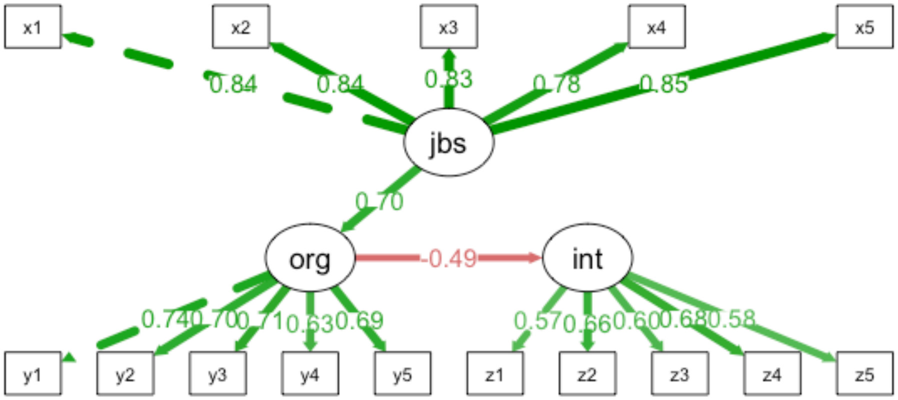
Regressions:	Estimate	Std. Error	Z-value	P value	Std. Latent Var.	Std. Coefficient
Org com ~ job sat	0.867	0.105	8.256	0.000	0.867	0.710
Intent to leave ~ org com	-0.322	0.065	-4.954	0.000	-0.322	-0.518
Variances:	Estimate	Std. Error	Z-value	P value	Std. Latent Var.	Std. Coefficient
Job satisfaction	0.793	0.11	7.113	0.000	0.793	1.00
Org. commitment	0.586	0.120	4.894	0.000	0.586	0.496
Intent to leave	0.334	0.088	3.815	0.000	0.334	0.732

FSR vs. SEM

After running the factor score regression for my model in R, I ran a structural equation model using lavaan's SEM function (Rosseel, 2012). The results of this model indicated similar regression coefficients as the FSR model. Job satisfaction and organizational commitment were positively correlated at R = 0.70 and organizational commitment and intent to leave were negatively correlated at R = -0.49 (See Figure 3). The standard error for these coefficients were 0.102 and 0.061, respectively, which is nearly identical to those of the factor score regression method.

Figure 3

Path of Structural Equation Model



Although the chi-squared test statistic was significant, the CFI, RMSEA, and SRMR did not reach the threshold of good fit. To compare this fit with the FSR method, I ran a nested model comparison test in R (compareFit; Jorgensen et al., 2020; Table 3). The Chi-Squared Difference Test analysis found the factor score regression method fit the model significantly better than the SEM. In conclusion, Croon's factor score regression method better fit my model, thus supporting prior evidence that factor score regression method is an efficient alternative to latent factor structure equation modeling.

Table 3

Model Fit Indices (Croon’s FSR vs. SEM)

	Chi-Square	Df	P value	CFI	TLI	AIC	BIC	RMSEA	SRMR
FSR	19.924*	1	0.000	.914*	.742*	1340.413*	1356.904*	.308	.076*
SEM	1059.910	88	0.000	.572	.489	8386.870	8492.416	.235*	.084

* indicates best fit index in comparison

Conclusion

The results of this analysis ultimately tell a story about how organizational commitment and job satisfaction affect employees’ intent to leave the organization. This could be crucial for an organization that is trying to understand patterns of turnover. The case that I have illustrated in my application of factor score regression analysis is just one example of how the method can be utilized in industrial-organizational psychology. Today it is turnover, tomorrow its perceptions of managerial support. Like most social and behavioral fields of science, psychology is a subject that is built on latent constructs. Methods like factor score regression and structural equation modeling give us the ability to see how these abstract constructs influence each other and various aspects of an organization. In order for our science to expand, we must have firm empirical evidence for the structured relationships that make up our paradigm. Psychometricians’ quest to create methodologies, like those highlighted in this paper, that allow us to potentially have more confidence in the nomological network of industrial-organizational psychology should encourage researchers and practitioners alike.

References

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

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis. *Methodology*, *13,* 31-38.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741-770.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological methods*, *6*(4), 430.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306-324.

Hayes, T., & Usami, S. (2020). Factor score regression in the presence of correlated unique factors. *Educational and Psychological Measurement*, *80*(1), 5-40.

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: guidelines for determining model fit.

Hoshino, T., & Bentler, P. M. (2011). Bias in factor score regression and a simple solution.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). semTools: Useful tools for structural equation modeling. R package version 0.5-3. Retrieved from https://CRAN.R-project.org/package=semTools

Kaplan, D. (2008). Structural Equation Modeling: Foundations and Extensions (2nd ed.). SAGE. ISBN 978-1412916240.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*, 486-507.

Kline, R. B. (2016). Principles and practice of structural equation modeling (4th ed. New York, NY: Guilford Press.

Lu, I. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing*, *28*(3), 258-268.

R Core Team. (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://r-project.org/

Roseel, Y. (2012) lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. Retrieved from

Roseel, Y. & Devlieger, I. (2018). *Why we may not need SEM after all*. Poster presented at the Meeting of the SEM Working Group, Amsterdam.

SchÖnemann, P. H., & Steiger, J. H. (1978). On the validity of indeterminate factor scores. *Bulletin of the Psychonomic Society*, *12*(4), 287-290.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563-575.

Williams, L. J., & Hazer, J. T. (1986). Antecedents and consequences of satisfaction and commitment in turnover models: A reanalysis using latent variable structural equation methods. *Journal of applied psychology*, *71*(2), 219.

Appendix A: R Code

```
library(lavaan)
library(semTools)

##generate model
demo.model <- '
org commitment ~ .8*job satisfaction  #strength of regression
with external criterion
intent to leave ~ -.5*org commitment

job satisfaction =~ .8*x1 + .8*x2 + .8*x3 + .8*x4 + .8*x5
#definition of factor f with loadings on 5 items
org commitment =~ .7*y1 + .7*y2 + .7*y3 + .7*y4 + .7*y5
intent to leave =~ .7*z1 + .7*z2 + .7*z3 + .7*z4 + .7*z5

x1 ~~ (1-.8^2)*x1 #residual variances. Note that by using 1-
squared loading, we achieve a total variability of 1.0 in each
indicator (standardized)
x2 ~~ (1-.8^2)*x2
x3 ~~ (1-.8^2)*x3
x4 ~~ (1-.8^2)*x4
x5 ~~ (1-.8^2)*x5

y1 ~~ (1-.8^2)*x1 #residual variances. Note that by using 1-
squared loading, we achieve a total variability of 1.0 in each
indicator (standardized)
y2 ~~ (1-.8^2)*x2
y3 ~~ (1-.8^2)*x3
y4 ~~ (1-.8^2)*x4
y5 ~~ (1-.8^2)*x5

z1 ~~ (1-.8^2)*x1 #residual variances. Note that by using 1-
squared loading, we achieve a total variability of 1.0 in each
indicator (standardized)
z2 ~~ (1-.8^2)*x2
z3 ~~ (1-.8^2)*x3
z4 ~~ (1-.8^2)*x4
z5 ~~ (1-.8^2)*x5
`

# generate data; note, standardized lv is default
set.seed(1234)
simData <- simulateData(demo.model, sample.nobs=200)
describe(simData, skew = FALSE)

#look at the data
```

```
View(simData)[,1:4]


model <-’
org commitment ~ job satisfaction # “~ is regressed on”
intent to leave ~ org commitment

job satisfaction =~ x1+ x2 + x3 + x4 + x5 # "=~ is measured by"
org commitment =~ y1+ y2 + y3 + y4 + y5
intent to leave =~ z1 + z2 + z3 + z4 + z5
`

##descriptive statistics for observed variables
describe(simData, skew = FALSE)

#factor scores
cfa <- cfa(model, simData)
lavPredict(cfa, method = "regression")


#FSR - SAM
fit.sam <- lavaan:::sam(model, data = simData,
                        sam.method = "local")
coef(fit.sam)
summary(fit.sam, standardized = TRUE)
semPaths(fit.sam, "std", style = "LISREL", edge.label.cex = 1.5)
parameterEstimates(fit.sam, add.attributes = TRUE, ci = FALSE)
fitmeasures(fit.sam, c("cfi", "rmsea", "srmr"))

#correlations
inspect(fit.sam, "cor.all")

#SEM
sem.fit <- sem(model, data=simData)
coef(sem.fit)
summary(sem.fit, standardized=TRUE)
semPaths(sem.fit, "std", style = "LISREL", edge.label.cex = 1,
residuals = FALSE)
fitmeasures(sem.fit, c("cfi", "rmsea", "srmr"))


compareFit(sem.fit, fit.sam)
```

APPENDIX B: Croon's Bias-Correcting Formulas

$$\widehat{cov(\xi,\eta)} = \frac{cov(F\xi, F\eta)}{\mathrm{A}_\xi \Lambda_x \Lambda_y \mathrm{A}_\eta}$$

$$\widehat{var(\xi)} = (var\left(F_\xi\right) - \mathrm{A}_\xi \Theta_\delta \mathrm{A}'_\xi)(\mathrm{A}_\xi \Lambda_x {\Lambda'}_x {\mathrm{A}'}_\xi)$$

APPENDIX C: Corrected Standard Error Formula

$$SE = \sqrt{\frac{S_{total}^{2}}{var\left(F_{\xi}\right)(n-1)}}$$

APPENDIX D: Modified Fit Indices

$$X_a^2 = n\mathrm{F}_{ML_a}$$

$$RMSEA_a = \sqrt{max\left(\frac{X_a^2 - df}{df(n-1)}, 0\right)}$$

$$SRMR_a = \sqrt{\frac{\Sigma_{i=1}\Sigma_{j=1}[(S_{ij} - \hat{\sigma}_{a_{ij}})/(S_i S_j)]^2}{k(k+1)/2}}$$

$$CFI_a = 1 - \frac{\max\,[(X_a^2 - df_a), 0]}{\max\,[(X_0^2 - df_0), 0]}$$