**An Alternative to the Carnegie Classifications: Using Structural Equation Models to Identify Similar Doctoral Institutions**

*Paul Harmon, Sarah McKnight, Laura Hildreth, Ian Godwin, and Mark Greenwood*

# Introduction: Institutional Classifiers

## The Rising Importance of Institutional Classifications

Systems for institutional classification are increasingly important to administrators, faculty, and students at institutions of higher education**(CITE)**. The Carnegie Classifications for Institutions of Higher Education**(CITE)**, the US News Ranking **(Cite)**, and other metrics of institutional quality are utilized not only by students in deciding where attend school, but by a range of users from prospective faculty to administrators interested a given campus’ academic reputation. The Carnegie Classifications are intended to be used for identification of peer institutions, not to create a ranking system; however, they are often misconstrued as metrics of institutional quality (CITE).

For doctoral-granting institutions, the Carnegie Classifications sort institutions into one of three groups: Highest, Higher and Moderate research activity (cite). For decision makers at schools in the top group, maintenance of status is the priority – schools that are in the “Highest Research Category” (also called R1 status) have the potential to market their status as a quality that makes them desirable to perspective students, faculty, and funding. The effect of institutional classifications may be more evident in the “High Research Activity” (R3) and “Higher-Research Activity” (R2) categories – several such schools have implemented policy goals and timelines explicitly oriented towards achieving R1 status. As such, these metrics are sometimes used to direct institutional policy. At least two institutions in the ‘Higher research activity’ group have explicitly set out policy goals directed towards improving their standing in the Carnegie Classifications, including Montana State University (Montana State University Strategic Plan, 2015) as well as the University of Idaho (Idaho, 2015).

While the Carnegie Classifications, US News World Rankings, and other metrics of institutional characteristics provide interesting and useful data for decision makers at institutions, they lack reproducibility and transparency as they either are not well documented or are proprietary. The purpose of this paper is to illustrate the concerns associated with the current available arsenal of institutional classifications, and present two multivariate alternatives using mixture modeling on scores obtained from a Structural Equation model, or SEM (Bollen, 1989). Much of the research pertaining to processes for institutional classification deals with identifying the data and content that most accurately describe each university(CITE), however we are more concerned with the statistical methodology used than the data themselves. Rather than proposing that new variables be added to the dataset to improve classifications, we propose an alternate classification system that utilizes the same dataset as was used by the Carnegie Classifications and provides an easier model to interpret.

# Methodology of the Carnegie Classifications

## The Data

The data used in the Carnegie Classifications are published with each update and can be obtained from the Carnegie Classifications website at http://carnegieclassifications.iu.edu/. The data from the 2015 update, with which this analysis is primarily concerned, contain information pertaining to institutions at every degree-granting level. However, this analysis focuses specifically on doctoral-granting schools since the methodologies assessed here only apply to those such institutions. Similar methods would be applicable to other tiers of the classification system as well, so this could be replicated for non-doctoral institutions as well. Information contained in the Carnegie Dataset come from a variety of sources, including the Integrated Postsecondary Education Data System (IPEDS) , and the National Science Foundation (NSF), and are collected in a single snapshot rather than over a period of multiple years.

The data are separated into two distinct sets of variables, an aggregate and a per-capita dataset. The aggregate dataset contains the following 7 variables:

* **STEM PhDs**: Counts of doctorates awarded in STEM fields, based on IPEDS classification
* **Humanities PhDs**: Counts of doctorates awarded in non-STEM fields, based on IPEDS classification
* **Social Science PhDs**: Counts of doctorates awarded in Social Science fields, based on IPEDS classification
* **Other PhDs**: Counts of doctorates awarded in education (EdD or PhD), based on IPEDS classification
* **STEM research expenditures**: Expenditures spent on STEM-related research fields, in thousands of dollars
* **Non-STEM research expenditures**: Expenditures spent on non-STEM research fields, in thousands of dollars
* **Research Staff size**: Non-tenurable research faculty and Post-doctoral researchers, based on IPEDS classification

The per-capita dataset contains the last three variables mentioned in the above list, but divides them by the size of the tenured/tenure-able faculty. They are listed below.

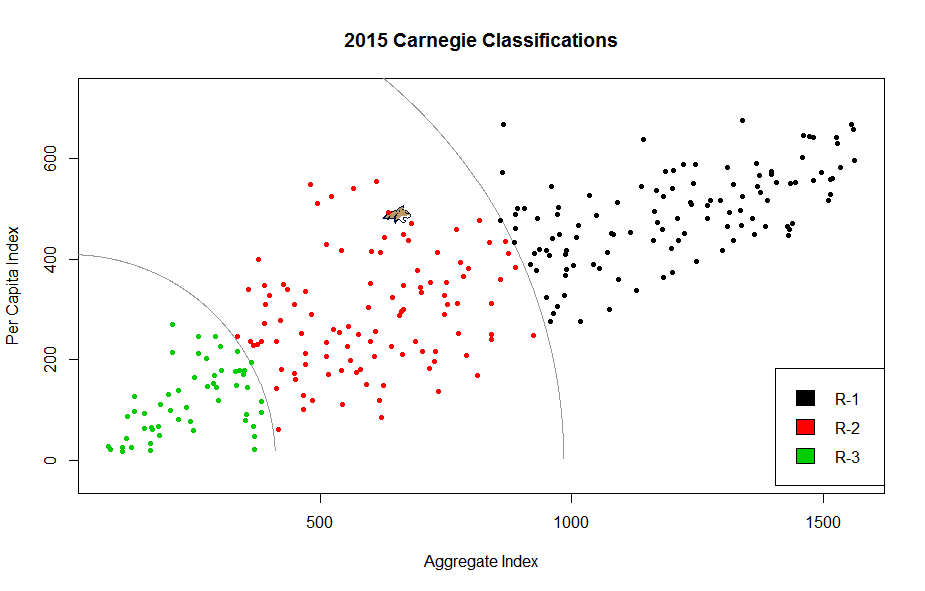
* **STEM research expenditures per capita**: Expenditures (in dollars) spent on STEM-related research fields, divided by the headcount of tenured/tenure-track faculty.
* **Non-STEM research expenditures per tenured/tenure-track faculty headcount**: Expenditures (in dollars) spent on non-STEM research fields divided by the headcount of tenured/tenure-track faculty
* **Research Staff size per capita**: Non-tenurable research faculty and Post-doctoral researchers, based on IPEDS classification, divided by the headcount of tenured/tenure-track faculty

## The Methodology

Although their exact methodology is not documented, the Carnegie Classifications can be generally replicated via the following general process. Data are first ranked from smallest to largest, and s This is done because the size of some of the largest universities dwarfs the size of many of the smallest institutions in the dataset, many of the variable distributions are highly right-skewed.

The ranked datasets are put into two principal component analyses (PCA), which use eigenvector decompositions of each set of variables to create orthogonal axes that explain the most variation in the underlying variables (CITE). The first components explain the most variance in the underlying variables; these are from each of the PCAs are taken to form a single aggregate index that explains most of the variation in the original 7 aggregate variables. Similarly, a per-capita index using the first principal component to explain the variation in the per-capita variables. In previous iterations of the Carnegie Classifications, these indices only explained between 68% and 72% of the variation in the underlying data. In the 2015 update, these were 70% and 72% for the aggregate and per-capita scales, respectively.

The per-capita scores (y-axis) are then plotted against the aggregate scores (x-axis) for each school. The plot of scores is then partitioned into thirds by hand– this is the most subjective part of the Carnegie Classifications – by drawing concentric circles that separate the three clusters of institutions **(CITE)**. Schools in the bottom left corner of the plot are in the “Moderate Research” category, and those in the top-right corner are in the “Highest Research” category, with the “Higher Research” category in the middle. Because these scores and boundaries can change in different years, it is possible for schools to move between categories from one update to the next.



## Problems with the Carnegie Classifications

The Carnegie Classifications are intended to be tools for institutional comparisons, not for ordinal rankings of schools. Based on the metrics used in their calculation, they have little to say about undergraduate education or outcomes. However, they have the potential to be used by policymakers on campuses to drive institutional goals and academic development.

Universities that prioritize moving from one category to another, however, must shoot at a moving target. Because the data used to calculate the Carnegie Classifications in any given year are based on ranked snapshots at a single time point, the weighting of a single factor can change from year to year. This is not always drastic; however, if large changes were to occur in the characteristics of many of the universities in the calculation, it is possible that the loading for a given variable could be different in the next release of the classifications.

Thus, the is thatthe efficacy of their interpretation because variables that carry a great deal of weight in one year may not carry the same amount of weight in another year. For instance, a school may determine based on the weights used in the 2015 calculation of the classifications that it needs to gain a certain number of PhDs in STEM and Social Science and increase STEM expenditures by a substantial amount. That school might implement those changes, only to find out that because of the changes in the underlying PCA that generates each school’s score, those policy goals were not actually aligned with the most important variables in the updated 2018 classifications.

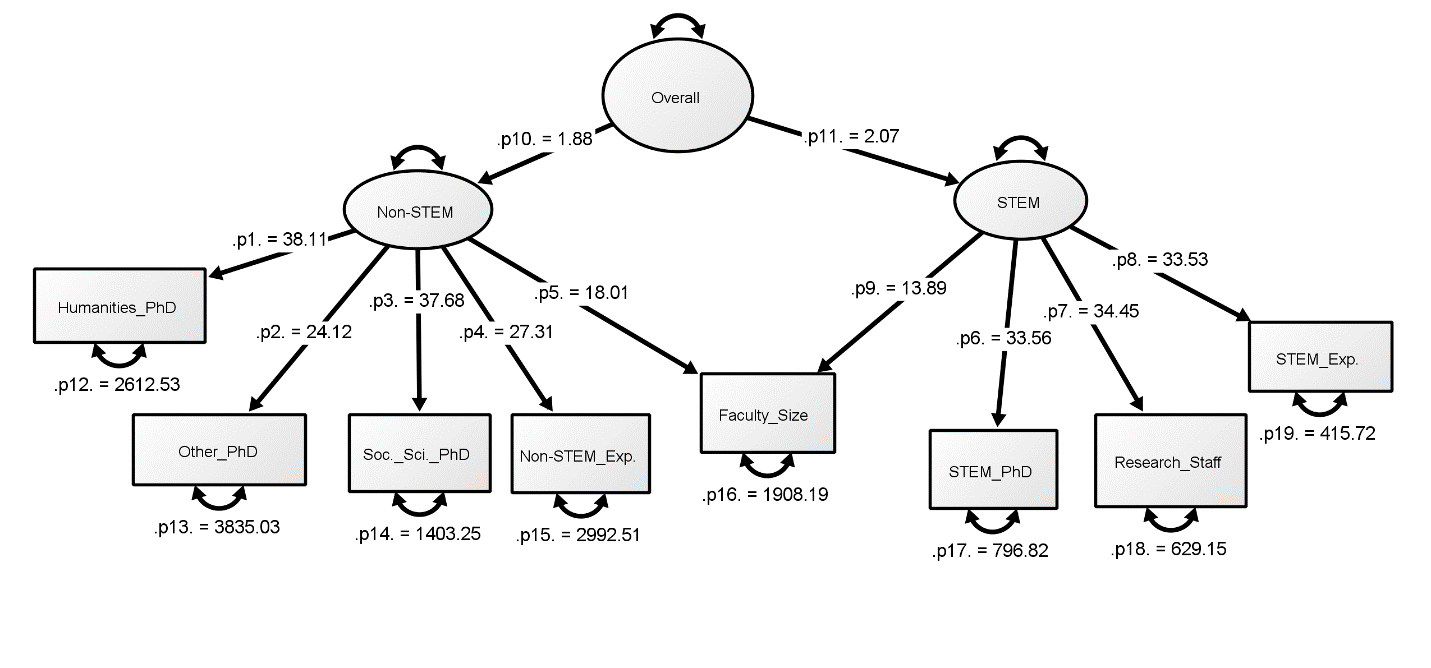
Second, another problem with the methodology of the Carnegie Classifications is that the PCA-based methodology removes information in a way that cannot be controlled. The creation of a single aggregate index from seven variables necessarily leads to information loss; in 2010 this was roughly 30 percent of the information in the underlying variables and in 2015 it was roughly 29 percent(). However, using the first principal component is not guaranteed to explain 70 percent of the variation in the underlying variables. It is possible that if the amount of correlation in the underlying variables was somehow much smaller in a subsequent year, the first PC might explain substantially less variation. Similarly, if the data were more correlated, the new index would explain far more than 70 percent of the information contained in the underlying variables. Thus, there is not currently a way to directly compare the results from one edition of the Carnegie Classifications to another.

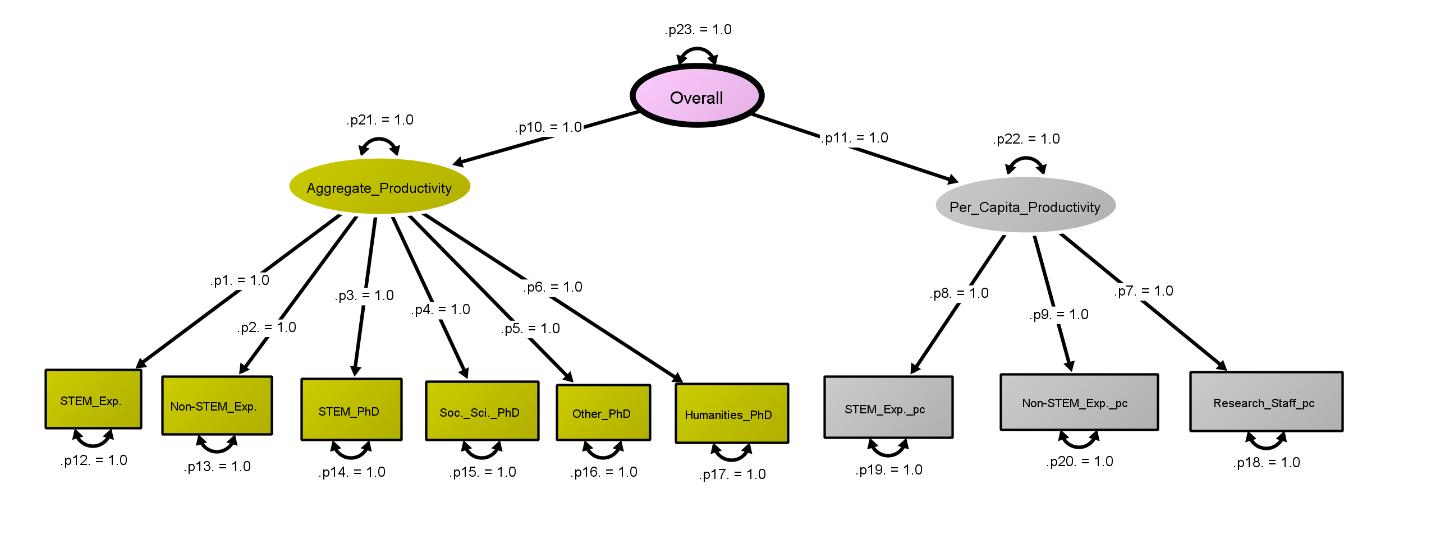
Finally, the lack of transparency of the Carnegie Classifications makes their exact classifications difficult to replicate. The SEM-based method, on the other hand, is easily reproduced using the methodology outlined in this paper.

## Introducing the Next Section: SEM-Based Classifications

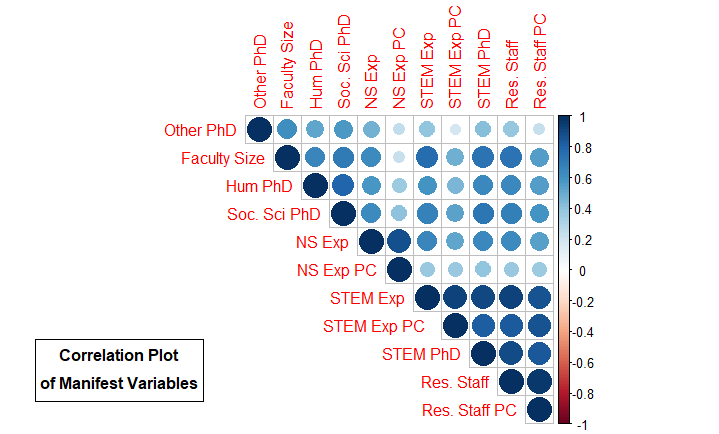
We propose using structural equation modeling (Bollen, 1989) to obtain factor scores which can then be used to classify institutions. Structural equation modeling is a statistical methodology that allows for the modeling of simultaneous equations, the use of latent or unobserved variables, and variables to be measured with error. A structural equation model (SEM) consists of two parts: the latent variable model which describes the relationships among latent variables and the measurement model which relates the latent variables to their indicators or items.

For the latent variable model, as depicted in Figure 1, we use a second order latent factor model. The first order latent factors are STEM productivity and non-STEM productivity. These latent factors are assumed to be measures of the second order factor overall productivity. We chose to use STEM and non-STEM productivity as latent factors as opposed to two factors for aggregate productivity and per capita productivity, the two factors used by Carnegie Classifications in their PCA, as this model formulation is more intuitive making the model easier to interpret. A second order factor for overall productivity was included as STEM and non-STEM productivity are distinct but related concepts that can be accounted for by one underlying factor (Chen, Sousa, and West, 2005). This also allows for easier interpretation of this model and allows us to obtain a single score for productivity as opposed to two scores used by the Carnegie Classifications.





The measurement model relates the STEM and non-STEM latent factors to their items. As shown in Figure ??, the items for STEM productivity are STEM PhDs produced, STEM expenditures, and research staff size while the items for non-STEM productivity are humanities PhDs produced, social science PhDs produced, other PhDs produced, and non-STEM research expenditures. The number of tenured and tenure-track faculty is an item for both STEM and non-STEM productivity as the number of tenure/tenure-able faculty is a measure of productivity in STEM and non-STEM fields. We opted to use research staff size as a measure of STEM productivity but not non-STEM productivity as research staff are predominantly employed in STEM fields. We chose to use a specific variable for an item of a given latent factor as the choice intuitively makes sense. Further, when examining the correlation matrix of the items in Table/Figure ??, the items of the latent factor for STEM productivity are highly correlated and the items for the latent factor for non-STEM productivity are also highly correlated while items of different factors are at most moderately correlated. The number of tenure and tenure-track faculty is moderately to highly correlated with items for both latent factors.



We fit the hypothesized model using R (R Development Core Team, 2008) using the lavaan package (citation). Standardized parameter estimates are displayed in Figure/Table ??. These results indicate that the hypothesized model fits the data moderately well (chi-square statistic = 110.024 with 17 df, RMSEA = 0.141, CFI = 0.958). The standardized factor loadings are all above 0.7, with the exception of number of tenure and tenure-track faculty as it cross-loads on both latent factors (factor loadings are 0.482 and 0.400 for STEM and non-STEM productivity, respectively) and number of other PhDs produced (0.639(, which indicates that at least half of the variability in each of the items is explained by its associated latent factor. The path coefficients relating overall productivity to STEM and non-STEM productivity are 0.900 and 0.883 respectively. This indicates that the variability of STEM and non-STEM productivity is largely explained by overall productivity. Overall these results are consistent with what is expected.

To compare institutions, ideally we would compare the values of the latent factor for overall productivity. Because latent factors by definition are unobserved these values must be estimated. This is done by creating factor scores which can then be used in subsequent analyses (DiStefano, Zhu, and Mindrila, 2009). Factor scores are computed using a weighted average of the items with a number of options available for weighting. The most common method used to calculate factor scores is Bartlett’s method (Bartlett, 1937) as it leads to unbiased estimates of the true factor scores. In subsequent analyses we use the factor scores created using Bartlett’s method.

## Determining Scores: Univariate Clustering

In contrast with the Carnegie Classifications, which used two indices to determine cluster membership for each university, the SEM-based rankings could be based on each university’s single factor score. This gives a single number for each school, so a single dimension and ordering of institutions can be obtained. Determination of the cluster partitions may be the most subjective aspect of the Carnegie Classifications because the lines dividing groups are hand drawn each year (CITE). In years when the data are poorly separated, determining an optimal place for the lines to be drawn is ambiguous at best. Rather, we seek an objective method for defining both the number of optimal groups and cluster membership.

## Model Based Clustering

[fill in section with some discussion of MBC method and talk about the results that we had]

[include plots from Mark’s output]

# Sensitivity Analysis

## R-Shiny Applications

It is of interest to determine how sensitive both metrics are to changes in the underlying data, especially since institutions are driving policy intended to move up in Carnegie rank. We developed a Shiny (CITE) application in R designed to allow the user to select a school and assess the sensitivity of that school’s classification to changes in the underlying variables for both the Carnegie and SEM-based methods. The user could select a school and then use a slide bar to either increase or decrease the number of PhDs awarded in each category, research staff size, or research expenditures. Changes can be made to either a single variable, all of them, or just a select group. The application takes the user input and re-calculates the PCA-based indices and SEM-based results on the new dataset, and shows where the university would be relative to other schools in that update.

The applications can be found at: \_\_\_INSERT URL HERE\_\_. For administrators and decision makers at institutions, they can be used to assess sensitive spots in both classification systems and inform growth-oriented policy. Institutions can test hypothetical policy actions to determine their efficacy on both classification systems. Moreover, the applications can be used to compare the sensitivity of changes in the underlying variables across classification systems.

# Discussion

## Problems Addressed by the SEM Model

The Carnegie Classifications are undoubtedly a useful tool for identifying and quantifying differences between academic institutions. However, they are not perfect. Our proposed model addresses some of the problems associated with the Carnegie methodology. However, the SEM-based model is also subject to some of the same issues.

The first problem addressed is that of the unstable loadings. In the Carnegie Classifications, the loadings of each variable on the two indices can change from year to year based on the variability in the ranked manifest variables. The SEM-based model addresses this issue in a different way – the manifest variables are \_\_\_.

Secondly, the SEM-based model better uses all of the information in the underlying manifest variables than the PCA-based method does. The Carnegie Classifications’ indices are not guaranteed to explain a large or consistent proportion of the variation in the original variables; the amount of variation explained could differ wildly based on changes in the underlying data. The SEM functions in a different statistical paradigm, using manifest variables to explain latent STEM and non-STEM factors at each institution and modeling them as a single factor of factors without substantial loss of information.

Thirdly, the SEM-based classifications system relies on an automated method for determining classifications rather than hand-drawn delineations. While any automated method for determining both the number of clusters and cluster membership has the possibility of selecting either an overly complicated solution or choose too few groups, the mixture model resulted in a reasonable three-group solution.

Finally, this method of institutional classification is well documented and reproducible. It can be applied to new datasets and consistently compared.

## Further Research

This research addresses a statistical question, not a qualitative one. The Carnegie Classifications have changed over time as more data have become available, and it is likely that they will change in the future. The efficacy of including different variables into the SEM-based classification system is not something that we assessed; however, the method discussed here would still work if new variables were included in a reasonable way.

Moreover, other methods could be applied to the classification piece. Mixture modeling is not the only method for determining clusters with univariate data, even though it does provide a reasonable number of clusters. Further work could focus on comparing different univariate clustering on the SEM scores or focus on bivariate clustering algorithms to the Carnegie indices.

References added by Laura

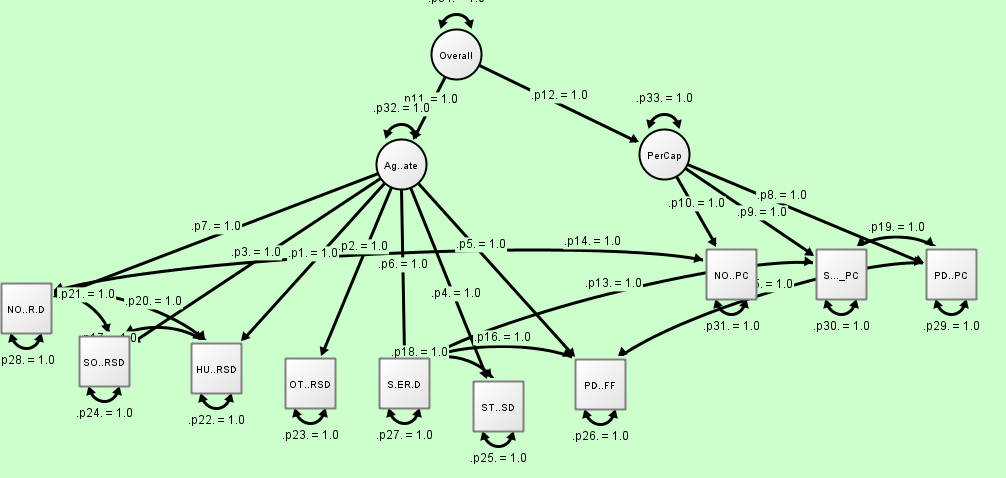
Bartlett, M. S. (1937). The statistical conception of mental factors. *British journal of Psychology*, *28*(1), 97-104.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural equation modeling*, *12*(3), 471-492.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 2.

Some additional stuff

Using a PCA-based approach is not the only way to develop an index for institutional characteristics. Alternatively, we propose a classification system built on Structural Equation Models, or SEMs (Bollen 1989). All models here were fit using the statistical software package R (R Core Team, 2018) and the package Lavaan (CITE). Rather than creating two individual indices of institutional characteristics, an SEM allows for modeling of latent traits that can be built from a set of manifest variables that are measured. Using the same dataset, we proposed fitting an SEM-based analogue to the Carnegie Classifications.

The manifest variables would be mapped to two latent traits in the same way as before, with the Aggregate latent trait comprising the counts of PhDs awarded, Research Staff, STEM-based expenditures, and non-STEM research expenditures. In the same vein, the per-capita latent trait would similarly be based on the per-capita versions of the research expenditures and Research Staff. Structural Equation Modeling allows for modeling of covariances between the manifest variables loaded onto different latent traits, so we proposed a model that allowed for correlation between the per-capita and raw versions of each variable. Finally, the two latent traits would be combined to form a single index, called a factor of factors (LAURA???) . Scores would be used to cluster schools based on a more objective method than splitting into three groups; rather, we proposed to use univariate mixture model to determine the optimal groupings of schools along this single new dimension. 

Initially, the goal of the research was to use and SEM that directly replicated the construction of the Aggregate and Per-Capita indices from the Carnegie method. When the model failed to converge, this suggested a model misspecification, namely that the two latent factors were too correlated to be separated. This makes sense because the two latent factors, the Aggregate and Per-Capita scales, literally comprise the same (slightly transformed) manifest variables. Although the ranked per capita manifest variables are not exactly correlated with their ranked aggregate counterparts, the correlations are very close to 1 (“STEM Research Doctorates”: 0.92, “Non-Stem Research Doctorates”: 0.88, Research Staff: 0.96). *This is a serious issue in SEM because we are unable to identify how much variability in the data is from each manifest variable (****citation****).*

## An Alternative Model

A new model was constructed which loaded these variables to two latent factors, cross-loading total number of tenured/tenure-able faculty onto both latent factors to emulate the per-capita variables without constructing per-capita variables from them directly. The two latent factors were then loaded onto a factor of factors to form a single index on which institutional comparisons could be made. Because this model does not have the correlation in the two latent factors, it is able to converge and we are able to obtain Huber-White robust standard errors.

## Model Fit Assessment

Model fit for the STEM/non-STEM model was fair (RMSEA: 0.141, CFI: 0.958, Chi-Square: 6.5), with fit characteristics that suggest a good relative fit compared to a null model. However, the absolute fit that may need improvement. That is, the model is comparatively better than an alternative with only an intercept, but does a middling job recreating the variability of the data. The proportion of the variability in the manifest variables explained by the latent traits was above 0.7 for all variables with the exceptions of “Other Doctorates,” “Non-Stem Research Expenditures,” and number of faculty.

(**Table(s) of variability proportions??**).

It is important to note that this model gives a different interpretation than the model utilized in the Carnegie Classifications. Scores are computed as averages of the ranks and weights associated with each manifest variable. Each university receives three scores, a non-STEM score, a STEM score, and an overall weighted average index of the two. The final overall score, created as a factor of factors from the SEM, is utilized in the final clustering algorithm.