PhysTEC Evaluation Phase II:
Predictive Modeling Using PTEPA Data

April 14, 2022

Dr. Remy Dou

This work was supported by the NSF under grant number PHY-1707990.
Executive Overview

The purpose of this report is to present statistical findings from data collected across 59 PhysTEC sites exploring possible contributing variables to teacher graduation outcomes. In particular, the analyses carried out were designed to confirm and build upon prior work\(^1\), which identified a parsimonious predictive model of teacher graduation rates based on general site characteristics (e.g., highest degree offered, prior teacher graduation rates, department size). The current analysis introduces data collected from the Physics Teacher Education Program Analysis (PTEPA) Rubric.

**Purpose:** The purpose of this analysis was to identify which variables from the PTEPA Rubric are predictive of teacher graduation rates while controlling for variables identified in prior work.

**Evaluation questions:**
1. What factors predict whether a site experiences a change in teacher graduates during funding? Post-funding? Total increase pre-to-post?
2. What factors predict the size of the change in teacher graduates during funding? Post-funding? Total increase pre-to-post?
3. What factors predict the number of teacher graduates during funding? Post-funding?

**Summary of Findings:**
1. The “Standards” and “Components” represented in the PTEPA Rubric were not predictive of teacher graduation rates during or post funding, nor predictive of changes in teacher graduation rates.
2. Statistical approaches to reducing or chunking PTEPA Rubric data were carried out, including principal component analysis, exploratory factor analysis, bipartite network centrality, and community clustering. None of these approaches led to statistically significant models.
3. A “shotgun” stepwise regression approach was run on PTEPA standards, components, and items. While further suggesting issues with implementation of the PTEPA Rubric, outcomes also suggest the following measures deserve a second look:
   a. **Component 6A** (positive association)
   b. **Item 2B-1** (partial positive association)
   c. **Item 2B-9** (partial positive association)
   d. **Item 2C-6** (partial positive association)
   e. **Item 3C-5** (partial negative association)
   f. **Item 5C-2** (partial negative association)

**Primary Conclusions:** Teacher graduation rates prior to funding continues to be the strongest positive predictor of teacher graduation rates during and after funding. Whether a site was a Comprehensive Site was at times a secondary predictor of teacher graduation rates. While additional data might improve some of the models, analyses pointed to potential issues in the application of the PTEPA Rubric across sites.

---

\(^1\) Regression Study: Graduation rates of PhysTEC Legacy Sites and Model Development; see Phase I Final Report.

Phase II: Predictive Modeling Using PTEPA Data
**Recommendations:** Program leaders might consider revisiting the statistically significant components and items identified through the shotgun analyses, as well as revisit the PTEPA Rubric protocol to identify potential issues with fidelity of implementation.

**Method Overview**

Prior data collection activities in relation to the 59 PhysTEC Legacy sites documented the number of total graduates and teacher graduates for specific years (Bachelor's degrees) in relation to PhysTEC funding (prior to, during, and post). These and related data were used to calculate the following outcome (i.e., dependent) variables of interest:

<table>
<thead>
<tr>
<th>Outcome Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgGradF1F3 (continuous)</td>
<td>Average number of teacher graduates during the first three years of funding.</td>
</tr>
<tr>
<td>AvgGradF1PF1 (continuous)</td>
<td>Average number of teacher graduates during all years of funding plus the first year post funding.</td>
</tr>
<tr>
<td>AvgGradPF2PF4 (continuous)</td>
<td>Average number of teacher graduates during the second to fourth year post funding.</td>
</tr>
<tr>
<td>AvgGradPF1PF12 (continuous)</td>
<td>Average number of teacher graduates during all years post funding.</td>
</tr>
<tr>
<td>AvgGradPF1PF12 (continuous)</td>
<td>Average number of teacher graduates during all years post funding.</td>
</tr>
<tr>
<td>AvgGradPF2PF12 (continuous)</td>
<td>Average number of teacher graduates during all years post funding with the exception of the first year post funding.</td>
</tr>
<tr>
<td>ChngPostPre (continuous)</td>
<td>Change in the average number of teacher graduates during all years post funding from the average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>ChangePost.1Pre (continuous)</td>
<td>Change in the number of teacher graduates during the first year post funding from the average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>ChangPF1.PF4Pre (continuous)</td>
<td>Change in the average number of teacher graduates during the first four years post funding from the average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>ChangPF2.PF4Pre (continuous)</td>
<td>Change in the average number of teacher graduates during the years two through four post funding from the average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>DidChngPostPre (binary)</td>
<td>Whether or not ChngPostPre increased by more than 0.50.</td>
</tr>
<tr>
<td>DidChangePost.1Pre (binary)</td>
<td>Whether or not ChangePost.1Pre increased by more than 0.50.</td>
</tr>
<tr>
<td>DidChangPF1.PF4Pre (binary)</td>
<td>Whether or not ChangPF1.PF4Pre increased by more than 0.50.</td>
</tr>
<tr>
<td>DidChangPF2.PF4Pre (binary)</td>
<td>Whether or not ChangPF2.PF4Pre increased by more than 0.50.</td>
</tr>
</tbody>
</table>

**Phase II: Predictive Modeling Using PTEPA Data**
Phase I

Predictive Modeling Using PTEPA Data

<table>
<thead>
<tr>
<th>NormChngPostPre</th>
<th>The normalized gains between the average number of teacher graduates during all years post funding and the average number of teacher graduates during the three years prior to funding.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>NormChangePost.1Pre</td>
<td>The normalized gains between the number of teacher graduates during the first year post funding and the average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>NormChangPF1.PF4Pre</td>
<td>The normalized gains between the average number of teacher graduates during the first four years post funding and the average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>NormChangPF2.PF4Pre</td>
<td>The normalized gains between the average number of teacher graduates during the years two through four post funding and the average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
</tbody>
</table>

Data modeling tested model iterations and variations against all outcome variables of interest.

The following variables were treated as control variables:

Table II. Control variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgGradF-1F-3</td>
<td>Average number of teacher graduates during the three years prior to funding.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>AvgDptSizePre</td>
<td>The average number of total graduates across the three years before funding.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>AvgDptSizePost</td>
<td>The average number of total graduates post-funding.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>ChngDptSizePreDuring</td>
<td>The average number of total graduates during funding minus AvgDptSizePre</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>DidDptSizePreDuring</td>
<td>Whether or not ChngDptSizePreDuring is greater than 0.50.</td>
</tr>
<tr>
<td>(binary)</td>
<td></td>
</tr>
<tr>
<td>NmbYrsSnceFndng</td>
<td>2020 minus the year funding began.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>AvgGradsPre</td>
<td>The average number of teacher graduates the three years prior to funding.</td>
</tr>
<tr>
<td>(continuous)</td>
<td></td>
</tr>
<tr>
<td>PhD Dum</td>
<td>Whether the highest degree granted by a site was PhD degrees [see “Exploratory” below].</td>
</tr>
<tr>
<td>(binary)</td>
<td></td>
</tr>
<tr>
<td>Mast Dum</td>
<td>Whether the highest degree granted by a site was Master’s degrees [see “Exploratory” below].</td>
</tr>
<tr>
<td>(binary)</td>
<td></td>
</tr>
<tr>
<td>Bach Dum</td>
<td>Whether the highest degree granted by a site was Bachelor’s degrees [see “Exploratory” below].</td>
</tr>
<tr>
<td>(binary)</td>
<td></td>
</tr>
<tr>
<td>Comp Dum</td>
<td>Whether a site is a Comprehensive Site.</td>
</tr>
<tr>
<td>(binary)</td>
<td></td>
</tr>
<tr>
<td>Targ Dum</td>
<td>Whether a site is a Targeted Site.</td>
</tr>
<tr>
<td>(binary)</td>
<td></td>
</tr>
<tr>
<td>Rec Dum</td>
<td>Whether a site is a Recruiting site.</td>
</tr>
<tr>
<td>(binary)</td>
<td></td>
</tr>
</tbody>
</table>

Phase II: Predictive Modeling Using PTEPA Data
Calculating PTEPA Variables

Institutions (N = 59) were evaluated on their implementation of the 6 PTEPA Rubric standards, which are organized into 19 components.

Standards:

1. Institutional Commitment (Components 1A – 1C)
2. Leadership and Collaboration (Components 2A – 2C)
3. Recruitment (Components 3A – 3D)
4. Knowledge and Skills for Teaching Physics (Components 4A – 4C)
5. Mentoring, Community, and Professional Support (Components 5A – 5C)
6. Program Assessment (Components 6A – 6C)

Each component was composed of a varying number of items developed to evaluate site traits during PhysTEC funding (i.e., the PTEPA Rubric). Several items made up a component and several components made up a standard. Evaluators would delineate an institution’s effectiveness on a particular item using the following hierarchical rating system:

- Not present
- Developing
- Benchmark
- Exemplary

For statistical analyses each of these ratings was assigned a numerical value ranging from zero for “Not present” to three for “Exemplary”. This data treatment choice has been discussed in a prior evaluation report. The values represented for each institution was then averaged out across the items that made up a particular component. This number was used as a numerical value representing the effectiveness of a particular institution at implementing the component in question. An institution’s score on a particular standard was measured as the average of the items making up that standard.

The following variables represent our predictor variables. They consist of measures recorded using the PTEPA rubric:

<table>
<thead>
<tr>
<th>Table III. PTEPA variables.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comp1A</strong></td>
</tr>
<tr>
<td><strong>Comp1B</strong></td>
</tr>
<tr>
<td><strong>Comp1C</strong></td>
</tr>
<tr>
<td><strong>Comp2A</strong></td>
</tr>
<tr>
<td><strong>Comp2B</strong></td>
</tr>
<tr>
<td><strong>Comp2C</strong></td>
</tr>
<tr>
<td><strong>Comp3A</strong></td>
</tr>
<tr>
<td><strong>Comp3B</strong></td>
</tr>
</tbody>
</table>

For information about the development of the rubric, specific rubric items, and its usage, see http://phystec.org/thriving.
<table>
<thead>
<tr>
<th>Comp3C</th>
<th>The average value of the items categorized under component 3C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp3D</td>
<td>The average value of the items categorized under component 3D.</td>
</tr>
<tr>
<td>Comp4A</td>
<td>The average value of the items categorized under component 4A.</td>
</tr>
<tr>
<td>Comp4B</td>
<td>The average value of the items categorized under component 4B.</td>
</tr>
<tr>
<td>Comp4C</td>
<td>The average value of the items categorized under component 4C.</td>
</tr>
<tr>
<td>Comp5A</td>
<td>The average value of the items categorized under component 5A.</td>
</tr>
<tr>
<td>Comp5B</td>
<td>The average value of the items categorized under component 5B.</td>
</tr>
<tr>
<td>Comp5C</td>
<td>The average value of the items categorized under component 5C.</td>
</tr>
<tr>
<td>Comp6A</td>
<td>The average value of the items categorized under component 6A. Note: Item 6A-1 was not included due to collinearity with outcomes variables.</td>
</tr>
<tr>
<td>Comp6B</td>
<td>The average value of the items categorized under component 6B.</td>
</tr>
<tr>
<td>Comp6C</td>
<td>The average value of the items categorized under component 6C.</td>
</tr>
<tr>
<td>Standard_1</td>
<td>The average value of components 1A through 1C.</td>
</tr>
<tr>
<td>Standard_2</td>
<td>The average value of components 1A through 1C.</td>
</tr>
<tr>
<td>Standard_3</td>
<td>The average value of components 1A through 1C.</td>
</tr>
<tr>
<td>Standard_4</td>
<td>The average value of components 1A through 1C.</td>
</tr>
<tr>
<td>Standard_5</td>
<td>The average value of components 1A through 1C.</td>
</tr>
<tr>
<td>Standard_6</td>
<td>The average value of components 1A through 1C.</td>
</tr>
</tbody>
</table>

### Analyses

**Dealing with missingness:** After data cleaning and variable development, data was tested for missingness. Missingness was a significant issue for this data set, both cross-sectional across sites during a given year or longitudinally across sites from the first year of funding through the current post funding period. Multiple imputation was attempted but the models failed to converge due to the large number of variables that needed to be predicted and the number of observations. Attempts were made to limit the number of variables needed for testing models, but none were successful.

**Permutations:** The models were run using a permutation algorithm from the *lmPerm* package in R to account for deviations from normality, inherent data dependency, and increased likelihood of Type II error from testing multiple models.

### Confirmatory

Prior to testing for the effect of the PTEPA variables, the model previously identified during Phase I was tested against the current data set. This was done to confirm the outcomes of the analyses carried out during Phase I given that additional data had been collected between this phase and the prior one.

The prior model is specified below:

**BASE MODEL**

\[
\text{OUTCOME} \sim \beta_1\text{AvgGradsPre} + \beta_2\text{CompDum} + \beta_3\text{BachDum} + e
\]
The Base Model was tested against all outcome variables listed in Table I. Outcome variables that take into account the average number of teacher graduates through the three years prior to funding (i.e., AvgGradF-1F-3) did not include this variable in the Base Model in order to avoid collinearity and dependency issues. Binary outcome variables were tested using logistic regression models. These models confirmed findings from Phase I. Statistically non-significant models were removed from further analyses. The following outcome variables were kept:

- AvgGradF1F3
- AvgGradF1F7
- AvgGradPF1PF12
- ChngPostPre
- ChangPF1.PF4Pre
- DidChangePost.1Pre
- NormChngPostPre
- NormChangePost.1Pre
- NormChangPF1.PF4Pre
- NormChangPF2.PF4Pre

Model Testing

Two primary sets of linear regression models were run against the outcome variables. The first set consisted of the base model plus measures of standards one through six calculated using data collected with the PTEPA Rubric as described above. The second set of regressions consisted of the base model plus measures of components 1A through 6C. All models were run with both centered and not-centered numeric variables.

**Results of Model Testing:** None of the models tested resulted in statistically significant models.

Alternative Analyses

Given the outcomes of model testing an attempt was made to reconsider the approach to model development. This reconsideration took into account potential error in the previously identified categorization structure of the items (i.e., their groupings into particular components and standards).

Principal Component Analysis

Principal component analysis is an approach to reducing data dimensionality, such as grouping a set of items into smaller subsets. A principal component analysis was used to test the likelihood of PTEPA Rubric items to load onto various components. A synthesis of the analysis outcomes led to testing five component model that resulted in capturing 54% of the variance captured in the dataset. Five variables were created by calculating the average of the items most likely to load onto each of the first five components across all institutions in the data (i.e., PC1, PC2, PC3, PC4, PC5). Regression models consisted of the base model plus PC1 through PC5 across all outcome variables. All models were run with both centered and not-centered numeric variables.

*Phase II: Predictive Modeling Using PTEPA Data*
Results of Model Testing: None of the models tested resulted in statistically significant models.

Bipartite Network Centrality

Techniques drawn from social network analysis were adopted to carry out an experimental approach to data reduction. This involved transforming the institutional PTEPA Rubric data into a bipartite network consisting of two types of nodes: PTEPA Rubric item and institution. The network was created by bidirectionally linking each institution to each item on which it had scored a “1” or higher. Items scored as “0” did not constitute a tie between the nodes.

Three centrality variables were calculated for each institution:

- Degree: the number of items tied to an institution.
- Weighted Degree: the number of items tied to an institution weighted by the value of the measure captured by each item.
- Betweenness: the likelihood of an institution to lie between sets of disconnected nodes.

Regression models were tested across all outcome variables and consisted of the base model plus the three centrality variables. All models were run with both centered and not-centered numeric variables.

Results of Model Testing: None of the models tested resulted in statistically significant models.

Community Clustering

Building on the approach described in the previous section, the bipartite network of items and institutions was examined for community clustering. Such an approach aims to identify a finite set of communities (or groupings) consisting of nodes in a network that are more centralized within one another. Membership in a particular community would then be treated as a categorical variable. In other words, each of the PhysTEC institutions would be categorized within one of the communities identified.

The “greedy” optimization algorithm in the R package igraph was used to identify communities. Three distinct communities were identified. Each institution’s particular community was captured used a nominal value of 1, 2, or 3 within a variable named “Community”. Regression models were tested across all outcome variables and consisted of the base model plus the “Community” variable.

Results of Model Testing: None of the models tested resulted in statistically significant models. Had membership in one or more communities predicted an outcome variable, the items clustered in that community would be examined for similarities.

Stepwise (“Shotgun”) Linear Regressions

Lastly, several outcome variables were tested against the Base Model plus variables introduced one at a time and kept in the model only if statistically significant. The outcome variables were AvgGradF1F3, AvgGradF1F7, and AvgGradF1PF1.
Three sets of variables were introduced against each of the outcome variables: the set of standards variables, the set of components variables, and the items variables. Component 6A (after removing item 6A-1 due to dependency issues) was predictive of all outcome variables. Five individual items were predictive of all outcome variables when treated as continuous or categorical variables. See Table III.

Table III. Statistically significant variables.

<table>
<thead>
<tr>
<th>Component 6A (positive association): “Program Outcomes”</th>
<th>In addition to the average number of teacher graduates prior to funding, values for Component 6A predicted each of the three outcome variables listed above (average $\beta = 0.36^{**}$).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 2B-1 (partial positive association): “Common vision among the PTE program team”</td>
<td>Scoring a “1” on item 2B-1 was positively associated with the outcome variables (average $\beta = 0.33^*$) but scoring a “2” was not.</td>
</tr>
<tr>
<td>Item 2B-9 (partial positive association): “Reputation of PTE program team for leading change”</td>
<td>Scoring a “1”, “2”, or “3” on item 2B-9 was positively associated with the outcome variables (average $\beta = 0.06^<em>$, $\beta = 0.41^{</em>**}$, $\beta = 0.09^*$, respectively).</td>
</tr>
<tr>
<td>Item 2C-6 (partial positive association): “Departmental representation”</td>
<td>Scoring a “1”, “2”, or “3” on item 2C-6 was positively associated with the outcome variables (average $\beta = 0.21^{**}$, $\beta = 0.13^<em>$, $\beta = 0.01^</em>$, respectively).</td>
</tr>
<tr>
<td>Item 3C-5 (partial negative association): “Exposure to K-12 teaching environments”</td>
<td>Scoring a “1” or “2” on item 3C-5 was negatively associated with the outcome variables (average $\beta = -0.26^{<strong>}$, $\beta = -0.28^{</strong>}$ respectively) but scoring a “3” was not.</td>
</tr>
<tr>
<td>Item 5C-2 (partial negative association): “Local physics teachers group”</td>
<td>Scoring a “1” on item 5C-2 was negatively associated with the outcome variables (average $\beta = -0.34^*$) but scoring a “2” or “3” was not.</td>
</tr>
</tbody>
</table>

Findings

The overall findings suggest there exists little to no association between institutional outcomes on the PTEPA Rubric and teacher graduation rates during or after funding. The only consistently predictive variable across nearly every model was the average number of teacher graduates during the three years prior to the beginning of funding (i.e., AvgGradF-1F-3). Even in the few models from the “shotgun” analyses that identified statistically significant PTEPA Rubric predictors, AvgGradF-1F-3 continued to have the highest effect size in relation to teacher graduation rates during or after funding.

The outcomes of the stepwise linear regressions identified one component and five items that predicted the average teacher graduation rate during the first three years of funding (i.e., AvgGradF1F3), the average teacher graduation rate during the first seven years of funding (i.e., AvgGradF1F7), and the average teacher graduation rate during all funding years plus the first year post funding (i.e., AvgGradF1PF1). The
components were introduced to the stepwise linear regression exclusive of item variables and vice versa. Due to the reduction in the number of item variables, they were treated as categorical once identified to preserve the natural structure of the data. The outcomes for each variable are discussed below:

**Component 6A:** This variable encompasses scores related to the annual recruitment at each PhysTEC site, the diversity of teacher candidates, and the career persistence of teacher graduates five years after the program. It was positively associated with teacher graduation rates, suggesting that focusing on improving these programmatic aspects may support increases in those rates.

**Item 2B-1:** Although scoring at the “Development” level for having a “common vision among the PTE program team” was positively associated with teacher graduation rates, scoring at higher levels was not. The pattern of association, given that “Benchmark” and “Exemplary” levels of meeting the PTEPA Rubric requirements were not statistically significant, suggests that the existence of a measurement issue, either in the item measurement specifications set by the rubric, error in the evaluation of rubric specifications either within or across sites, or a combination of both.

**Item 2B-9:** The “reputation of PTE program teach for leading change” was positively associated with teacher graduation rates, though scoring at the “Benchmark” level exhibited a nearly four times larger effect size than scoring at the “Exemplary” level.

**Item 2C-6:** The “departmental representation” was positively associated with teacher graduation rates, though scoring at the “Developing” level exhibited a larger effect size than scoring at either the “Benchmark” or “Exemplary” level.

**Item 3C-5:** “Exposure to K-12 teaching environments” was negatively associated with teacher graduation rates. Scoring at the “Developing” or “Benchmark” level exhibited similar effect sizes. Scoring at the “Exemplary” level was not associated with teacher graduation rates.

**Item 5C-2:** Scoring at the “Developing” level on traits related to the “local physics teacher group” was negatively associated with teacher graduation rates, scoring at higher levels was not. Specifications of the four evaluation categories seem to suggest a fundamental difference in having a score of “Developing” versus a higher score, that is, the presence of a local physics teacher group was negatively associated with teacher graduation rates.

**A Word on Comprehensive II Sites**

Several variables were included in each of the models to control for potential confounders and validate factors that likely contribute to graduation rates. Whether or not a site was designated as a comprehensive site (i.e., “CompDum”) was one those, but contextual aspects of the various PhysTEC sites suggest that those designated as “Comprehensive II” merited additional attention (see Table III). Shotgun analyses were replicated for just these sites. The outcomes closely mirrored those presented in the previous section but to a limited degree. The following variables were significantly associated with teacher graduation rates during funding in similar magnitude: **Item 2B-1, Item 2B-5, Item 2B-9, and Item 3C-5.**
Table III. Comprehensive II PhysTEC Sites.

<table>
<thead>
<tr>
<th>Arizona State University</th>
<th>Texas State University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston University</td>
<td>Towson University</td>
</tr>
<tr>
<td>California State University, Long Beach</td>
<td>University of Alabama</td>
</tr>
<tr>
<td>Colorado School of Mines</td>
<td>University of Central Florida</td>
</tr>
<tr>
<td>Florida International University</td>
<td>University of Cincinnati</td>
</tr>
<tr>
<td>Georgia State University</td>
<td>University of Missouri</td>
</tr>
<tr>
<td>James Madison University</td>
<td>University of West Virginia</td>
</tr>
<tr>
<td>Middle Tennessee State University</td>
<td>Virginia Tech</td>
</tr>
<tr>
<td>Rowan University</td>
<td></td>
</tr>
</tbody>
</table>

**Recommendations**

Program leaders might consider revisiting the statistically significant components and items identified through the “shotgun” analyses, as well as revisiting the PTEPA Rubric protocol to identify potential issues with fidelity of implementation and specification of “Developing”, “Benchmark”, and “Exemplary” item characteristics.
Appendix A:

R Script
# Libraries

library(lm.beta)
library(lmPerm)
library(psych)
library(Amelia)
library(ggplot2)
library(corrplot)
library(igraph)
library(tidyr)
library(data.table)

# Imported data; file name: "Master Grads Data Spreadsheet 122421-withPTEPA for Remy-122421-updated 012722"

# Renamed: PhaseII

# Creating Dummy Variables for Highest Degree Granted (HDG)

colnames(PhaseII)[3] = "HDG"
table(PhaseII$HDG)
PhaseII[31,3] = "NA"
PhaseII$HDG = factor(PhaseII$HDG)
str(PhaseII$HDG)

PhaseII$PhDum = PhaseII$HDG
PhaseII$PhDum = as.character(PhaseII$PhDum)
PhaseII$PhDum[PhaseII$PhDum == "PhD"] = 1
PhaseII$PhDum[PhaseII$PhDum == "Mast"] = 0
PhaseII$PhDum[PhaseII$PhDum == "Bach"] = 0
PhaseII$PhDum = as.factor(PhaseII$PhDum)
print(PhaseII$PhDum)

PhaseII$MastDum = PhaseII$HDG
PhaseII$MastDum = as.character(PhaseII$MastDum)
PhaseII$MastDum[PhaseII$MastDum == "PhD"] = 0
PhaseII$MastDum[PhaseII$MastDum == "Mast"] = 1
PhaseII$MastDum[PhaseII$MastDum == "Bach"] = 0
PhaseII$MastDum = as.factor(PhaseII$MastDum)
print(PhaseII$MastDum)

PhaseII$BachDum = PhaseII$HDG
PhaseII$BachDum = as.character(PhaseII$BachDum)
Phase II: Predictive Modeling Using PTEPA Data

Phase II: Predictive Modeling Using PTEPA Data

Phase II$BachDum[Phase II$BachDum == "PhD"] = 0
Phase II$BachDum[Phase II$BachDum == "Mast"] = 0
Phase II$BachDum[Phase II$BachDum == "Bach"] = 1
Phase II$BachDum = as.factor(Phase II$BachDum)
print(Phase II$BachDum)

Phase II$CompDum = factor(Phase II$CompDum, levels = c("1","0"))
levels(Phase II$CompDum)
str(Phase II$CompDum)

# Outcomes: Average Number of Teacher Graduates During

### Model 1: AvgGradF1F3
### Model 2: AvgGradF1PF1
### Model 3: AvgGradF1F7

# Predictors: AvgGradF.1F.3, CompDum, BachDum

Model1 = lm(AvgGradF1F3 ~ AvgGradF.1F.3 + CompDum + BachDum, data = Phase II)
summary(Model1) # AvgGradF.1F.3***
lm.beta(Model1) # 0.62

Model2 = lm(AvgGradF1PF1 ~ AvgGradF.1F.3 + CompDum + BachDum, data = Phase II)
summary(Model2) # AvgGradF.1F.3**
lm.beta(Model2) # 0.49

Model3 = lm(AvgGradF1F7 ~ AvgGradF.1F.3 + CompDum + BachDum, data = Phase II)
summary(Model3) # AvgGradF.1F.3***
lm.beta(Model3) # 0.54

# Outcomes: Average Number of Teacher Graduates Post Funding

### Model 4: AvgGradPF2PF12
### Model 5: AvgGradPF1PF12
### Model 6: AvgGradPF2PF4

# Predictors: AvgGradF.1F.3, CompDum, BachDum

Model4 = lm(AvgGradPF2PF12 ~ AvgGradF.1F.3 + CompDum + BachDum, data = Phase II)
summary(Model4) # CompDum*
lm.beta(Model4) # 0.46

Model5 = lm(AvgGradPF1PF12 ~ AvgGradF.1F.3 + CompDum + BachDum, data = Phase II)
summary(Model5) # CompDum*
lm.beta(Model5) # 0.49 ### BETTER ###

Model6 = lm(AvgGradPF2PF4 ~ AvgGradF.1F.3 + CompDum + BachDum, data = PhaseII)
summary(Model6) # Non-Sig
lm.beta(Model6) # Non-Sig

# Outcomes: Predicting Post Pre Change

### Model 7: ChangePost.1Pre
### Model 8: ChngPostPre
### Model 9: ChangPF1.PF4Pre
### Model 10: ChangPF2.PF4Pre

# Predictors: CompDum, BachDum

Model7 = lm(ChangePost.1Pre ~ CompDum + BachDum, data = PhaseII)
summary(Model7) # Non-Sig

Model8 = lm(ChngPostPre ~ CompDum + BachDum, data = PhaseII)
summary(Model8) # Non-Sig

Model9 = lm(ChangPF1.PF4Pre ~ CompDum + BachDum, data = PhaseII, center = FALSE)
summary(Model9) # Non-Sig

Model10 = lm(ChangPF2.PF4Pre ~ CompDum + BachDum, data = PhaseII, center = FALSE)
summary(Model10) # Non-Sig

# Outcomes: Predicting Post Pre Change of more than 0.50

### Model 11: DidChangePost.1Pre
### Model 12: DidChngPostPre
### Model 13: DidChangPF1.PF4Pre
### Model 14: DidChangPF2.PF4Pre

# Creating variables

PhaseII$DidChangePost.1Pre[PhaseII$ChangePost.1Pre > .49] = 1
PhaseII$DidChangePost.1Pre[PhaseII$ChangePost.1Pre < .50] = 0

PhaseII$DidChngPostPre[PhaseII$ChngPostPre > .49] = 1
PhaseII$DidChngPostPre[PhaseII$ChngPostPre < .50] = 0

PhaseII$DidChangPF1.PF4Pre[PhaseII$ChangPF1.PF4Pre > .49] = 1
PhaseII$DidChangPF1.PF4Pre[PhaseII$ChangPF1.PF4Pre < .50] = 0

PhaseII$DidChangPF2.PF4Pre[PhaseII$ChangPF2.PF4Pre > .49] = 1
# Running models

Model11 = glm(DidChgePost.1Pre ~ CompDum + BachDum, data = PhaseII)  
summary(Model11) # Non-Sig

Model12 = glm(DidChngPostPre ~ CompDum + BachDum, data = PhaseII)  
summary(Model12) # Non-Sig

Model13 = glm(DidChangPF1.PF4Pre ~ CompDum + BachDum, data = PhaseII)  
summary(Model13) # Non-Sig

Model14 = glm(DidChangPF2.PF4Pre ~ CompDum + BachDum, data = PhaseII)  
summary(Model14) # Non-Sig

# Outcomes: Normalized Gains - (Post-Pre)/(100-Pre)

### Model 15: Model7Norm  
### Model 16: Model8Norm  
### Model 17: Model9Norm  
### Model 18: Model10Norm

# Creating variables

PhaseII$Model7Norm = ((PhaseII$AvgGradPF2PF12 - PhaseII$AvgGradF.1F.3)/(100-PhaseII$AvgGradF.1F.3))  
PhaseII$Model8Norm = ((PhaseII$AvgGradPF1PF12 - PhaseII$AvgGradF.1F.3)/(100-PhaseII$AvgGradF.1F.3))  
PhaseII$Model9Norm = ((PhaseII$AvgGradPF1PF4 - PhaseII$AvgGradF.1F.3)/(100-PhaseII$AvgGradF.1F.3))  
PhaseII$Model10Norm = ((PhaseII$AvgGradPF2PF4 - PhaseII$AvgGradF.1F.3)/(100-PhaseII$AvgGradF.1F.3))

# Running models

Model15 = glm(Model7Norm ~ CompDum + BachDum, data = PhaseII)  
summary(Model15) # Non-Sig

Model16 = glm(Model8Norm ~ CompDum + BachDum, data = PhaseII)  
summary(Model16) # Non-Sig

Model17 = glm(Model9Norm ~ CompDum + BachDum, data = PhaseII)  
summary(Model17) # Non-Sig

Model18 = glm(Model10Norm ~ AvgGradF.1F.3 + CompDum + BachDum, data = PhaseII)  
summary(Model18) # Non-Sig
Preparing for Predictive Models

# Correlation table of outcomes:

CorTableOutcomes = cor(PhaseII_R[,c(59:75,102:106,110:117)], use = "pairwise.complete.obs")
CorTableOutcomes
corrplot(CorTableOutcomes, method="number")

# Correlation table of items, standards, and components:

CorTableItems = cor(PhaseII_R[,c(102:116,118:206)], use = "pairwise.complete.obs")
CorTableItems
corrplot(CorTableItems, method="number")

CorTableStandards = cor(PhaseII_R[,c(102:116,207:212)], use = "pairwise.complete.obs")
CorTableStandards
corrplot(CorTableStandards, method="number")

CorTableComps = cor(PhaseII_R[,c(102:116,213:231)], use = "pairwise.complete.obs")
CorTableComps
corrplot(CorTableComps, method="number")

# Description of items, standards, and components:

items = describe(PhaseII_R[,c(118:206)])
items

standards = describe(PhaseII_R[,c(207:212)])
standards

components = describe(PhaseII_R[,c(213:231)])
components

### Principal Component Analysis to reduce the items but removing 6A_1 since it would correlate with outcome variables

# Looking for zero variance columns

forpca = na.omit(PhaseII_R[,c(118:194,196:206)])
which(apply(forpca, 2, var)==0)

# Removing zero variance columns

forpca = forpca[ , which(apply(forpca, 2, var) != 0)]
items.pca <- prcomp(forpca, scale = T)
attributes(items.pca)
print(items.pca, rotate = "promax")
print(items.pca$x)
summary(items.pca)
biplot(items.pca, scale = 0)

# variance explained by each principal compnent

variance_explained = items.pca$sdev^2 / sum(items.pca$sdev^2)
qplot(c(1:16), variance_explained) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 1)

parallel <- fa.parallel(forpca) # parallel analysis

# number of components and number of factors - 2 factors, 5 components
# Looking at the first five principal components (explain 54% of the variance)

items.pca$rotation[1:86,1:5]
loadings <- items.pca$rotation[1:86,1:5]
round(loadings, 2)

# Creating variables out of the most positive ones

str(items.pca)
forpca.new <- cbind(forpca, items.pca$x[,1:5])
str(forpca.new)
forpca.new

names(PhaseII_R)
attach(PhaseII_R)
PhaseII_R$pca1 = ((X3D.1 + X6A.4 + X4A.1)/3)
PhaseII_R$pca2 = ((X1B.2 + X1C.1 + X1C.2 + X1C.3 + X1C.5 + X2A.3 + X2B.5 + X2C.2 + X2C.4 + X3A.1
  + X3A.5
  + X3B.3 + X3D.4 + X4B.3 + X4C.1 + X5C.3 + X6A.2 + X6C.1 + X6C.2)/18)
PhaseII_R$pca3 = ((X1A.2 + X1A.3 + X1A.4 + X1B.3 + X1C.4 + X2A.1 + X2C.3 + X2C.5
  + X2C.7 + X3B.5 + X3C.3 + X4B.5 + X4C.3 + X5A.2 + X5B.1 + X5B.4 + X6B.3 + X6B.4)/18)
PhaseII_R$pca4 = ((X1A.1 + X2A.2 + X2C.6 + X3B.4 + X3C.2 + X3D.3 + X4A.2 + X4B.2
  + X4C.4 + X5A.1 + X5B.5 + X5C.1 + X5C.2 + X5C.4 + X6B.1 + X6B.2 + X6C.4)/17)
PhaseII_R$pca5 = ((X1A.5 + X1A.6 + X1B.1 + X2A.4 + X2B.2 + X2B.3 + X2B.6
  + X2B.7 + X2B.8 + X2B.9 + X3A.2 + X3A.3 + X3A.4 + X3B.1 + X3B.2
  + X3C.5 + X3D.2 + X4A.3 + X4B.1 + X4B.4 + X5B.2 + X5B.3 + X6C.3)/23)
detach(PhaseII_R)
### Trying out an exploratory factor analysis

```r
forefa = forpca

# Checking the factorability of the data (KMO needs to be below 60)
KMO(r = cor(forefa)) # 0.50

# Testing whether factor analysis would be useful (Bertlett's Test of Sphericity)
cortest.bartlett(forefa) # Because p is not small, this suggest efa not valuable

det(cor(forefa)) # Positive determinant means factor analysis will probably run

# Determining number of factors (Looks like 4 - similar to pca)
fafitfree <- fa(PhaseII_R[,c(118:194,196:206)], nfactors = ncol(forefa), rotate = "none")
n_factors <- length(fafitfree$e.values)
scree <- data.frame(  
  Factor_n = as.factor(1:n_factors),  
  Eigenvalue = fafitfree$e.values)
ggplot(scree, aes(x = Factor_n, y = Eigenvalue, group = 1)) +  
  geom_point() + geom_line() +  
  xlab("Number of factors") +  
  ylab("Initial eigenvalue") +  
  labs(title = "Scree Plot",  
       subtitle = "(Based on the unreduced correlation matrix)"
)

parallel <- fa.parallel(forefa) # parallel analysis - Look at where red line crosses blue line for  
# number of components and number of factors - 2 factors -- 5 components

items.efa2 = factanal(forefa, factors = 2, scores = c("regression"), rotate="promax")
print(items.efa2)
```

---

**Phase II: Predictive Modeling Using PTEPA Data**

---

```r
PTEPA1 = lm(AvgGradF1F3 ~ AvgGradF.1F.3 + CompDum + BachDum + Standard_1  
             + Standard_2 + Standard_3 + Standard_4 + Standard_5 + Standard_6, data = Phasell_R, center  
             = FALSE)
summary(PTEPA1) # Non-Sig

PTEPA2 = lm(AvgGradF1F3 ~ AvgGradF.1F.3 + CompDum + BachDum + Comp_1A + Comp_1B  
             + Comp_1C + Comp_2A + Comp_2B + Comp_2C + Comp_3A + Comp_3B + Comp_3C +  
             Comp_3D  
             + Comp_4A + Comp_4B + Comp_4C + Comp_5A + Comp_5B + Comp_5C
```

---
Phase II: Predictive Modeling Using PTEPA Data

+ Comp_6A + Comp_6B + Comp_6C, data = PhaseII_R, center = FALSE)
summary(PTEPA2) # Non-Sig

PTEPA3 = lmp(AvgGradF1F3~ AvgGradF.1F.3 + CompDum + BachDum + pca1 + pca2
+ pca3 + pca4 + pca5, data = PhaseII_R, center = FALSE)
summary(PTEPA3) # Non-Sig

summary(PTEPA2) # Non-Sig

PTEPA4 = lmp(AvgGradF1F7~ AvgGradF.1F.3 + CompDum + BachDum + Standard_1
+ Standard_2 + Standard_3+ Standard_4 + Standard_5 + Standard_6, data = PhaseII_R, center = FALSE)
summary(PTEPA4) # Non-Sig

PTEPA5 = lmp(AvgGradF1F7~ AvgGradF.1F.3 + CompDum + BachDum + Comp_1A + Comp_1B
+ Comp_1C+ Comp_2A + Comp_2B + Comp_2C + Comp_3A + Comp_3B + Comp_3C + Comp_3D
+ Comp_4A + Comp_4B + Comp_4C + Comp_5A + Comp_5B + Comp_5C
+ Comp_6A + Comp_6B + Comp_6C, data = PhaseII_R, center = FALSE)
summary(PTEPA5) # Non-Sig

PTEPA6 = lmp(AvgGradF1F7~ AvgGradF.1F.3 + CompDum + BachDum + pca1 + pca2
+ pca3 + pca4 + pca5, data = PhaseII_R, center = FALSE)
summary(PTEPA6) # Non-Sig

summary(PTEPA3) # Non-Sig

PTEPA7 = lmp(AvgGradPF1PF12~ AvgGradF.1F.3 + CompDum + BachDum + Standard_1
+ Standard_2 + Standard_3+ Standard_4 + Standard_5 + Standard_6, data = PhaseII_R, center = FALSE)
summary(PTEPA7) # Non-Sig

PTEPA8 = lmp(AvgGradPF1PF12~ AvgGradF.1F.3 + CompDum + BachDum + Comp_1A + Comp_1B
+ Comp_1C+ Comp_2A + Comp_2B + Comp_2C + Comp_3A + Comp_3B + Comp_3C + Comp_3D
+ Comp_4A + Comp_4B + Comp_4C + Comp_5A + Comp_5B + Comp_5C
+ Comp_6A + Comp_6B + Comp_6C, data = PhaseII_R, center = FALSE)
summary(PTEPA8) # Non-Sig

PTEPA9 = lmp(AvgGradPF1PF12~ AvgGradF.1F.3 + CompDum + BachDum + pca1 + pca2
+ pca3 + pca4 + pca5, data = PhaseII_R, center = FALSE)
summary(PTEPA9) # Non-Sig

# Note - Also tried by removing AvgGradF.1F.3, CompDum, BachDum from models above
Phase II: Predictive Modeling Using PTEPA Data
summary(PTEPA17) # Non-Sig

PTEPA18 = glm(DidChangePost.1Pre ~ pca1 + pca2
    + pca3 + pca4 + pca5, data = Phasell_R, family = binomial)
summary(PTEPA18) # Non-Sig

# Normalized Gains - (Post-Pre)/(100-Pre)

# Creating variables

# Phasell$Model7Norm = ((Phasell$AvgGradPF2PF12 - Phasell$AvgGradF1F3)/(100-
# Phasell$AvgGradF1F3))
# Phasell$Model8Norm = ((Phasell$AvgGradPF1PF12 - Phasell$AvgGradF1F3)/(100-
# Phasell$AvgGradF1F3))
# Phasell$Model9Norm = ((Phasell$AvgGradPF1PF4 - Phasell$AvgGradF1F3)/(100-
# Phasell$AvgGradF1F3))
# Phasell$Model10Norm = ((Phasell$AvgGradPF2PF4 - Phasell$AvgGradF1F3)/(100-
# Phasell$AvgGradF1F3))

# Running models
PTEPA19 = lmp(Model7Norm ~ CompDum + BachDum + Standard_1
    + Standard_2 + Standard_3+ Standard_4 + Standard_5 + Standard_6, data = Phasell_R, center
    = FALSE)
summary(PTEPA19) # Non-Sig

PTEPA20 = lmp(Model7Norm ~ CompDum + BachDum + Comp_1A + Comp_1B
    + Comp_1C+ Comp_2A + Comp_2B + Comp_2C + Comp_3A + Comp_3B + Comp_3C +
    Comp_3D
    + Comp_4A + Comp_4B + Comp_4C + Comp_5A + Comp_5B + Comp_5C
    + Comp_6A + Comp_6B + Comp_6C, data = Phasell_R, center = FALSE)
summary(PTEPA20) # Non-Sig

PTEPA21 = lmp(Model7Norm ~ CompDum + BachDum + pca1 + pca2
    + pca3 + pca4 + pca5, data = Phasell_R, center = FALSE)
summary(PTEPA21) # Non-Sig

PTEPA22 = lmp(Model8Norm ~ CompDum + BachDum + Standard_1
    + Standard_2 + Standard_3+ Standard_4 + Standard_5 + Standard_6, data = Phasell_R, center
    = FALSE)
summary(PTEPA22) # Non-Sig

PTEPA23 = lmp(Model8Norm ~ CompDum + BachDum + Comp_1A + Comp_1B
    + Comp_1C+ Comp_2A + Comp_2B + Comp_2C + Comp_3A + Comp_3B + Comp_3C +
    Comp_3D

Phase II: Predictive Modeling Using PTEPA Data
Phase II: Predictive Modeling Using PTEPA Data
### Trying some bipartite analyses

```r
SNA = PhaseII_R[,c(2,118:194,196:206)]
SNA[SNA=="0"] = NA
SNA.DF=gather(SNA,person,weight,X1A.1:X6C.4,factor_key = TRUE)
SNA.DF=na.omit(SNA.DF)
gSNA.DF = graph_from_data_frame(SNA.DF, directed = FALSE)
SNA.DF.deg=degree(gSNA.DF,mode="all",loops=FALSE)
SNA.DF.str=strength(gSNA.DF,mode="all",loops=FALSE,weights = E(gSNA.DF)$weight)
SNA.F.bet=betweenness(gSNA.DF, directed=FALSE)
SNA.F.wbet=betweenness(gSNA.DF, directed=FALSE, weights=E(gSNA.DF)$weight)

SNA.DF.deg=as.data.frame(SNA.DF.deg)
SNA.DF.deg=setDT(SNA.DF.deg,keep.rownames=TRUE)
rownames(SNA.DF.deg)
SNA.DF.deg[SNA.DF.deg[c(-44:-131),]
colnames(SNA.DF.deg)[1]="IPEDS.ID"
names(SNA.DF.deg)

SNA.DF.str=as.data.frame(SNA.DF.str)
SNA.DF.str=setDT(SNA.DF.str,keep.rownames=TRUE)
rownames(SNA.DF.str)
SNA.DF.str[SNA.DF.str[c(-44:-131),]
colnames(SNA.DF.str)[1]="IPEDS.ID"
names(SNA.DF.str)

SNA.F.bet=as.data.frame(SNA.F.bet)
SNA.F.bet=setDT(SNA.F.bet,keep.rownames=TRUE)
rownames(SNA.F.bet)
SNA.F.bet[SNA.F.bet[c(-44:-131),]
colnames(SNA.F.bet)[1]="IPEDS.ID"
names(SNA.F.bet)

SNA.F.wbet=as.data.frame(SNA.F.wbet)
SNA.F.wbet=setDT(SNA.F.wbet,keep.rownames=TRUE)
rownames(SNA.F.wbet)
SNA.F.wbet[SNA.F.wbet[c(-44:-131),]
colnames(SNA.F.wbet)[1]="IPEDS.ID"
names(SNA.F.wbet)

SNA.DF.01 = merge(SNA.DF.deg,SNA.DF.str, all = TRUE, by = "IPEDS.ID")
SNA.DF.02 = merge(SNA.DF.01,SNA.F.bet, all = TRUE, by = "IPEDS.ID")
SNA.DF.03 = merge(SNA.DF.02,SNA.F.wbet, all = TRUE, by = "IPEDS.ID")

Phasell_R.SNA = merge(Phasell_R,SNA.DF.03, all = TRUE, by = "IPEDS.ID")
```

---

*Phase II: Predictive Modeling Using PTEPA Data*
Phase II: Predictive Modeling Using PTEPA Data

# MODELS

PTEPA39 = lm(AvgGradF1F3~ AvgGradF.1F.3 + CompDum + BachDum + SNA.DF.deg + SNA.DF.str + SNA.F.bet + SNA.F.wbet, data = Phasel_R.SNA, center = FALSE)
summary(PTEPA39) # Non-Sig

PTEPA40 = lm(AvgGradF1F7~ AvgGradF.1F.3 + CompDum + BachDum + SNA.DF.deg + SNA.DF.str + SNA.F.bet, data = Phasel_R.SNA, center = FALSE)
summary(PTEPA40) # Non-Sig

PTEPA41 = lm(AvgGradPF1PF12~ AvgGradF.1F.3 + CompDum + BachDum + SNA.DF.deg + SNA.DF.str + SNA.F.bet, data = Phasel_R.SNA, center = FALSE)
summary(PTEPA41) # Non-Sig

PTEPA42 = lm(ChngPostPre~ AvgGradF.1F.3 + CompDum + BachDum + SNA.DF.deg + SNA.DF.str + SNA.F.bet, data = Phasel_R.SNA, center = FALSE)
summary(PTEPA42) # AvgGradF.1F.3***

PTEPA43 = lm(ChngPF1.PF4Pre~ AvgGradF.1F.3 + CompDum + BachDum + SNA.DF.deg + SNA.DF.str + SNA.F.bet, data = Phasel_R.SNA, center = FALSE)
summary(PTEPA43) # AvgGradF.1F.3*

PTEPA44 = lm(DidChangePost.1Pre~ AvgGradF.1F.3 + CompDum + BachDum + SNA.DF.deg + SNA.DF.str + SNA.F.bet, data = Phasel_R.SNA, center = FALSE)
summary(PTEPA44) # AvgGradF.1F.3***

# Normalized Gains - (Post-Pre)/(100-Pre)

PTEPA45 = lm(Model7Norm ~ CompDum + BachDum + SNA.DF.deg + SNA.DF.str + SNA.F.bet, data = Phasel_R.SNA, center = FALSE)
summary(PTEPA45) # AvgGradF.1F.3**

# Community cluster detection
types <- V(gSNA.DF)$type
deg <- degree(gSNA.DF, mode="all", loops=FALSE)
str <- strength(gSNA.DF, mode="all", loops=FALSE, weights = E(gSNA.DF)$weight)
bet <- betweenness(gSNA.DF, directed=FALSE)
wbet <- betweenness(gSNA.DF, directed=FALSE, weights=E(gSNA.DF)$weight)

# Calculating communities / clusters
fc <- fastgreedy.community(gSNA.DF)
set.seed(123)

## layout (for graphing)
l <- layout.fruchterman.reingold(gSNA.DF, niter=1000)

## Values
membership(fc) # Who belongs in what community
sizes(fc) # Size of each community
str(fc) # Structure of my communities
print(fc)
print(fc$membership)
print(fc$post)

members = cbind(fc$names, fc$membership)
members = as.data.frame(members)

Phase II: Predictive Modeling Using PTEPA Data
Phase II: Predictive Modeling Using PTEPA Data

colnames(members)[1]="IPEDS.ID"
colnames(members)[2]="Community"
names(members)
members = members[-44:-131,]

Phasell_R.SNA = merge(Phasell_R.SNA,members, all = TRUE, by = "IPEDS.ID")

str(Phasell_R.SNA$Community)

# Checking which items are in each community
cl <- data.frame(name = fc$names, cluster = fc$membership, stringsAsFactors=F)
cl <- cl[order(cl$cluster),]
View(cl[cl$cluster==1,])
View(cl[cl$cluster==2,])
View(cl[cl$cluster==3,])

# Create a dataframe that includes the respondent and community they're in
clgraph <- data.frame(fc$names, fc$membership, stringsAsFactors=F)

################################################################################
########################## TESTING COMMUNITY MODELS ##########################
################################################################################

PTEPA29 = lmp(AvgGradF1F3 ~ AvgGradF.1F.3 + CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA29) # Non-Sig

PTEPA30 = lmp(AvgGradF1F7 ~ AvgGradF.1F.3 + CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA30) # Non-Sig

PTEPA31 = lmp(AvgGradPF1PF12 ~ AvgGradF.1F.3 + CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA31) # Non-Sig

PTEPA32 = lmp(ChngPostPre ~ CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA32) # Non-Sig
Phase II: Predictive Modeling Using PTEPA Data

PTEPA33 = lm(ChangPF1.PF4Pre ~ CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA33) # Non-Sig

PTEPA34 = lm(DidChangePost.1Pre ~ CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA34) # Non-Sig

# Normalized Gains - (Post-Pre)/(100-Pre)

PTEPA35 = lm(Model7Norm ~ CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA35) # Non-Sig

PTEPA36 = lm(Model8Norm ~ CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA36) # Non-Sig

PTEPA37 = lm(Model9Norm ~ CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA37) # Non-Sig

PTEPA38 = lm(Model10Norm ~ AvgGradF.1F.3 + CompDum + BachDum + as.factor(Community), data = Phasell_R.SNA, center = FALSE)
summary(PTEPA38) # Non-Sig

PTEPAexplore1 <- cbind(Phasell_R$AvgGradF.1F.3, Phasell_R$Comp_1C, Phasell_R$Comp_1A, Phasell_R$Comp_2C, Phasell_R$Comp_3C, Phasell_R$Comp_4B, Phasell_R$Comp_4C, Phasell_R$Comp_5B, Phasell_R$Comp_5C, Phasell_R$Comp_6A)
CorPTEPAexplore1 = cor(PTEPAexplore1, use = "pairwise.complete.obs")
CorPTEPAexplore1
corrplot(CorPTEPAexplore1, method="number") # Phasell_R$Comp_3C & Phasell_R$Comp_5B r = 0.65
### Stepwise Regressions

\[
P_{\text{TEPAX}1} = \text{lmp}(\text{AvgGradF1F3} \sim \text{AvgGradF.1F.3} + \text{Comp}_6A, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}1})
\]

\[
P_{\text{TEPAX}2} = \text{lmp}(\text{AvgGradF1F3} \sim \text{AvgGradF.1F.3} + \text{X2B.1} + \text{X2B.5} + \text{X2B.9} + \text{X2C.6} + \text{X3C.5} + \text{X5C.2} + \text{X6A.3}, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}2})
\]

\[
\text{lm.beta}(P_{\text{TEPAX}2})
\]

# Betas: AvgGradF.1F.3 (0.63), X2B.1 (-0.30), X2B.5 (0.22), X2B.9 (-0.46),
# X2C.6 (-0.20), X3C.5 (0.32), X5C.2 (0.23), X6A.3 (0.16)

### Other outcome variable

\[
P_{\text{TEPAX}4} = \text{lmp}(\text{AvgGradF1PF1} \sim \text{AvgGradF.1F.3} + \text{Comp}_6A, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}4})
\]

\[
\text{lm.beta}(P_{\text{TEPAX}4})
\]

\[
P_{\text{TEPAX}5} = \text{lmp}(\text{AvgGradF1PF1} \sim \text{AvgGradF.1F.3} + \text{X2B.5}, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}5})
\]

\[
P_{\text{TEPAX}6} = \text{lmp}(\text{AvgGradF1F7} \sim \text{AvgGradF.1F.3} + \text{Comp}_6A, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}6})
\]

\[
\text{lm.beta}(P_{\text{TEPAX}6})
\]

\[
P_{\text{TEPAX}7} = \text{lmp}(\text{AvgGradF1F7} \sim \text{AvgGradF.1F.3} + \text{X2B.1} + \text{X2B.5} + \text{X2B.9} + \text{X2C.6} + \text{X3C.5} + \text{X5C.2}, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}7})
\]

\[
P_{\text{TEPAX}8} = \text{lmp}(\text{AvgGradF1F7} \sim \text{AvgGradF.1F.3} + \text{X2B.1} + \text{X2B.5} + \text{X2B.9} + \text{X2C.6} + \text{X3C.5} + \text{X5C.2} + \text{as.factor(CompDum)}, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}8})
\]

# Taking only significant items and treating as factors

\[
P_{\text{TEPAX}7.1} = \text{lmp}(\text{AvgGradF1F7} \sim \text{AvgGradF.1F.3} + \text{as.factor(X2B.1)} + \text{as.factor(X2B.9)} + \text{as.factor(X2C.6)} + \text{as.factor(X3C.5)} + \text{as.factor(X5C.2)}, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}7.1})
\]

\[
\text{lm.beta}(P_{\text{TEPAX}7.1})
\]

\[
P_{\text{TEPAX}8.1} = \text{lmp}(\text{AvgGradF1F7} \sim \text{AvgGradF.1F.3} + \text{as.factor(X2B.1)} + \text{as.factor(X2B.9)} + \text{as.factor(X2C.6)} + \text{as.factor(X3C.5)} + \text{as.factor(X5C.2)} + \text{as.factor(CompDum)}, \text{data} = \text{PhaseII}_R, \text{center} = \text{FALSE})
\]

\[
\text{summary}(P_{\text{TEPAX}8.1})
\]

\[
\text{lm.beta}(P_{\text{TEPAX}8.1})
\]

---

**Phase II: Predictive Modeling Using PTEPA Data**

---
### Looking at other potential control variables

PTEPAX9 = lm(AvgGradF1F7 ~ AvgGradF.1F.3 + ChngDptSizePF2.12During.1, data = Phasel.R.SNA, center = FALSE) # I tried all from PTEPAX2
summary(PTEPAX9)
# Department size makes no difference

### NOTE: All of the above regressions were tried centered and uncentered

write.csv(Phasel.R.SNA, file = "Phasel.R.SNA.csv")