Methods
Past analyses of FYSS data relied on preserving the factor structure of the scales used. This analysis instead seeks to first analyze individual question items for evidence of correlation with success outcomes, and then to regress the best predictors onward based on common themes. (A formal factor analysis is not used, we will refrain from calling them “factors.”) All analysis was performed in R.

Student data were gathered from UK’s data warehouse. Three outcome variables were used:
- RETENTION_FALL_2017: Indicating that the student has enrolled in their 2nd fall term.
- SPRING_2017_CUMULATIVE_GPA: The student’s cumulative grade point average.
- GRADE: The grade in each class during the 2016-17 academic year, on a four-point scale.

Six variables were selected as indicators of college retention and GPA:
- HS_GPA: Weighted grade point average from the student’s last attended high school.
- ACT/SAT: The students complete maximum ACT score, converted with converted SAT.
- UNMET_NEED: Total unmet financial need, family income present, and receipt of federal aid, in their cohort term. UNMET_NEED is imputed as $0 when FAFSA data is missing.
- ON_Campus: Indicates that the student lived in on-campus housing in their cohort term. Most students who live off-campus during their first year are Lexington residents.
- KY_RESIDENT: Indicates that the student’s permanent address is in Kentucky.
- FIRST_GEN: Indicates that neither parent attended a post-secondary institution. This data is collected on the spring applications of all students.

Survey questions were largely 4-point Likert scales, e.g., “Disagree” to “Agree” for attitudes, and “Not at all Confident” to “Completely Confident” for self-efficacy. Other questions used custom ordinal scales: “How many of your close friends will be attending college?”, “None”, “Some”, “Most”, “Almost All.” A few were dichotomous: “Are you or any of your family members experiencing any of the following?”… “Serious financial difficulties”. “You” (check box), “Family member or close friend” (check box). The survey was administered via email through Qualtrics, using a personalized sampling, to the entire Fall 2016 cohort, with a gift card incentive. Several bi-weekly e-mail reminders were sent. Among the 5,505 invites, 36% provided complete responses, 4% declined to consent, and 7% started, but quit the survey before it was completed.

First, all 73 question items were grouped independently for pairwise correlation with each of the two outcomes, retention (Kendall’s $\tau$) and grades (Pearson’s $r$). In the interest of parsimony, items with correlations less than 0.05 are excluded. The qualifying items were then grouped into six simple thematic groups: belonging, crises, financial, grit, home, generation, study, and all. 30 items were used for retention and 23 for grades. 45 items were dropped.

After questions items were grouped, item scores were created by first standardizing, then averaging values weighted by the correlation coefficient with the respective outcome. Due to different variables being selected for retention and GPA, and different correlations computed, group scores differ. Also, because some questions were negatively associated with outcomes (and were correspondingly negatively coded in their scales of origin) their values were reversed before averaging. Thus all group scores can be interpreted as positively correlated. In the example table, there was only one negatively associated question, which was then negated, resulting in a score that is negative for a student who indicated they were experiencing a personal crisis. Likewise, items correlating first-generation status were negatively coded in generation, and related variables of education were positively coded. Correlation between group scores ranged from very low (0.04) to moderate (0.58) with the highest correlations between belonging, ties, and generation.

Within this framework, the primary goal was to assess whether, and to what degree, models of retention and GPA including survey variables (Survey Model) out-perform those without access to survey variables (“null model”). Statistical models were constructed using a hybrid method combining classical linear models with algorithmic optimization. To model retention, logistic regression was used, and was modeled by linear regression. In both cases, models optimization was achieved using genetic algorithm (9 package MQL/XTAND from the R Project for Statistical Computing). Candidate models were limited to linear terms and 2nd order interactions. For regression analysis, cases, i.e. single students, were deleted listwise on all candidate predictors, and no additional imputation was performed. This resulted in losses due to missingness of 2% in the retention model and 10% in the GPA model from the total completed survey sample of 1,849.

**Exploratory Results**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HS_Belonging</td>
<td>Satisfied in my college.</td>
<td>0.16</td>
<td>0.02</td>
<td>** 0.05</td>
<td>** 0.0001</td>
<td>0.05</td>
<td>0.05</td>
<td>** 0.0001</td>
<td>** 0.0001</td>
</tr>
<tr>
<td>First_Gen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ky_Resident</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT/SAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNMET_NEED</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT/SAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS_GPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS_GPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ky_Resident</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KS_RESIDENT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT/SAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS_GPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Regression Results**

- **Comparison of Logistic Regression Models of Retention Excluding and Including Survey Factors**
  - **HS_GPA:** 0.341, Adj. $R^2$: 59.0%
  - **UNMET_NEED:** 0.109,
  - **HS_GPA-FIRST_GEN:** 0.497
  - **FIRST_GEN:** -0.044,
  - **KFY_RESIDENT:** 0.055,
  - **ACT/SAT:** 0.231,
  - **UNMET_NEED:** 0.359

**Comparison of Regression Models of GPA Excluding and Including Survey Factors**

- **HS_GPA:** 0.363, Adj. $R^2$: 59.0%
  - **UNMET_NEED:** 0.109
  - **HS_GPA:** 0.543
  - **HS_GPA-FIRST_GEN:** 0.303
  - **FIRST_GEN:** -0.044

**Correlation**

The genetic algorithm converged for all four models. Both models of retention and GPA were improved by the inclusion of survey variables. Numerical predictors and GPA were re-scaled (divided by SD) for ease of comparison between standardized coefficients. In the null model for retention, high HS_GPA and low UNMET_NEED are associated with retention. In the survey model, both belonging and no-crisis are associated with retention. The effect of UNMET_NEED is of higher magnitude.

**Discussion**

Many of the relationships described in these models are consistent with the motivations that included of psychometric scales in the FYSS. It is noted that self-efficacy, grit, and intelligence mindset were largely absent from the optimized models. While the theories behind these instruments may be sound, the psychological traits they attempt to measure may be more directly measured by long-term behavior such as high school performance.

The predictive contributions of generation, belonging, and ties speak to the importance of social processes in student success. Often school’s only measures of this type are similar to FIRST_GEN: a single question on the application asking students to indicate their parents are college graduates. If students are educated and motivated to volunteer accurate responses, survey instruments such as the FYSS may be a way to tap into these otherwise invisible social processes, to provide the needed data for novel interventions.

Overall, adaptation of psychometric scales shows promise for prediction of undergraduate student success. However, the volunteer response to the instrument likely introduces bias into analyses such as these. Furthermore, incomplete student level data means incomplete coverage of interventions based on it. Changing the survey to a compulsory assessment would solve these problems. In order to make a more rigorous and effective survey instrument, further analyses (such as factor analysis, mediation analysis, machine learning algorithms like random forests, and more fully leveraging institutional data) and literature review must iteratively refine the framework of the survey instrument.

**Acknowledgments**

Many thanks to Dr. Ellen Usher, Dr. Christia Brown, and Dr. Jane Jensen for assisting with construction of the 2016 survey, Kathryn Wong for project management, Virginia Lacefield for just about everything, the AA team for data flow, and Dr. Craig Rudick for letting us spend so much effort on a survey!

For references, see handout.