CCHE Index Score Decomposition and Predictive Utility Assessment

Daniel Howard, Research Analyst, 2/29/16

Purpose

This report explores the Colorado Commission on Higher Education (CCHE) Index score, a measure of academic preparation that is assigned to each high school student who applies to CSU. The CCHE Index is a composite score derived from a non-linear combination of ACT score plus either High School Rank Percentile or High School GPA. This report investigates the degree to which each of the Index's component variables contribute to the Index score, across different student groups. The report also investigates the predictive validity of the Index composite for several common measures of undergraduate success (First Fall GPA, Retention, Six Year Graduation), compared with predictive models that omit Index in favor of using the Index's constituent component variables as independent predictors.

Data

The data set used for this report includes all First Time Full Time (FTFT) freshman cohorts starting at CSU in Fall 2005, Fall 2006, Fall 2007, Fall 2008, and Fall 2009. Only domestic students are included in this study, since very few international applicants provide all data points required to calculate Index and High School GPA standards vary greatly internationally. Students who deceased before attaining a degree are also excluded from the study. ACT score is the greater of the student's ACT composite score or the student's SAT score converted to an ACT composite score (see ACT website for methodology).

Methodology

This study uses linear regression to assess the impact of predictor variables on a dependent outcome variable. In the first portion of the study, the Index decomposition, the three index components (High School GPA, High School Rank Percentile expressed as a value between 1 and 100, and ACT score) are used to predict the CCHE Index Score. This portion of the study allows us to assess which component variables most strongly influence the final index score, and allows us to investigate whether there are differences in the components' predictive behavior across demographic groups of interest.

The remainder of the study compares the predictive utility of three common undergraduate success regression models using the CCHE Index Score as an independent variable, versus otherwise-identical models that use the Index components as predictor variables. This design allows us to directly compare the Index model to the component models to determine whether modeling with the composite Index score adds or decreases explanatory power compared to using Index components components. Similar to the first part of the study, these comparisons are conducted for all studied freshman students and for at-risk student (URM, First Generation) to assess whether effects vary across key demographic groups.

Regression Results: CCHE Index Component Decomposition

The CCHE index score is designed to represent a single measure of a high school graduate's academic preparation for college coursework. The index is calculated as a composite of the student's ACT score and either the student's High School Rank Percentile or the student's cumulative High School GPA, whichever ranks relatively higher; see CCHE's official documentation for specifics on calculating CCHE index score for an individual student.

A linear regression model regressing the index components on the Index score was created to investigate the statistical contribution that each of the three component variables makes toward the overall index. A model was created investigating the composition of the index for the entire Freshman student population, in addition to models studying First Generation students versus non-First Generation students and Underrepresented Minority (URM) students versus non-URM students. Table 1 below displays the model R^2 values for each studied population, plus the standardized beta weights for the three component variables (all five models are significant at the $p < .01$ level).

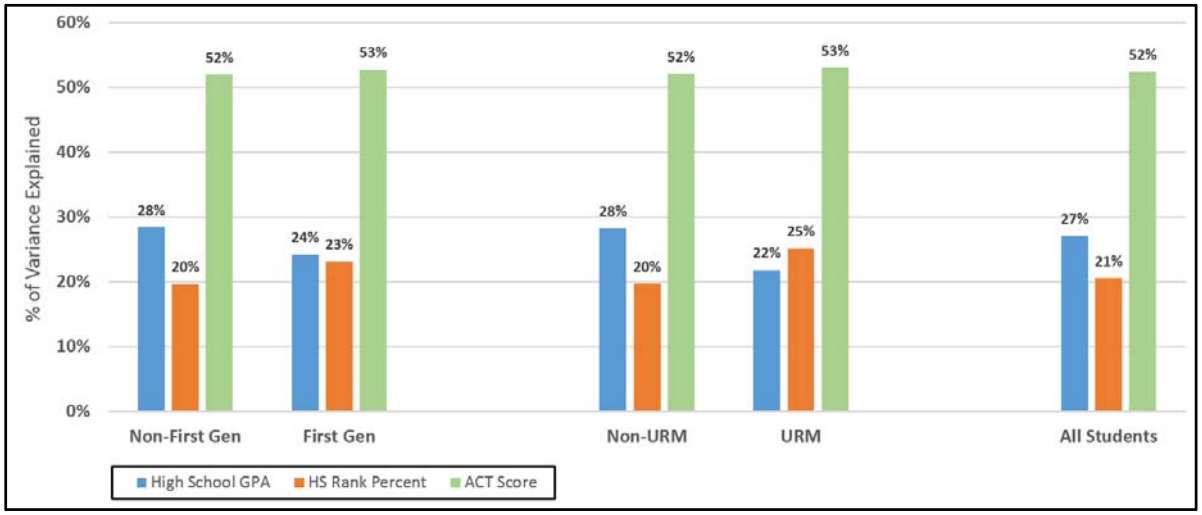
Table 1: CCHE Index Component Regression, FA05-FA09 FTFT Freshman Cohorts, by URM Status and FGEN Status

Regression Model: DV = Index, IVs = Components					
	Non-FGEN	FGEN	Non-URM	URM	All
Model R Square	.979	.979	.980	.976	**.979**

Index Component: Standardized Beta Weights					
	Non-FGEN	FGEN	Non-URM	URM	All
High School GPA	.347	.303	.345	.284	**.333**
HS Rank Percent	.240	.289	.241	.328	**.253**
ACT Score	.635	.660	.637	.692	**.646**

The three index components account for 97.9% of the overall variance in Index score for all students, with Index variance for each of the four studied subgroups explained at a rate of 97.6% or better. These large proportions of explained variance makes sense given that the three predictor variables are the only inputs used to calculate the Index score. The small deviation from 100% of variance explained is due to the linear nature of regression modeling and the non-linear nature of CCHE's indexing process. Figure 1 provides a graphical representation of the proportion of explained variance that is accounted for by each of the individual component variables.

Figure 1: Proportion of CCHE Index Variance Explained by Predictor Variables, by URM/FGEN Status



For each subgroup and for all students, ACT Score accounts for just over 50% of variance in Index score, a logical result given that ACT contributes to the final Index score for every student. For most groups, a majority of the remaining variance is accounted for by High School GPA, with the least amount of variance account for by High School Rank Percentile. The exception to this pattern is the URM group, in which a greater proportion of Index variance is accounted for by High School Rank than GPA. This interaction suggests that URM students' index scores are more likely to be influenced by High School Rank, while the overall pattern of findings suggests that GPA contributes more greatly to Index for the other four studied student groups.

Regression Results: First Fall GPA

A primary goal of this investigation is to assess the utility of the CCHE Index for Colorado State University's assessment and planning purposes. In particular, is the Index useful for predicting future student outcomes? Is it differentially predictive for at-risk student groups (in particular, First Generation and URM students)? And does the composite CCHE Index add predictive utility compared to models using the individual Index component variables instead of Index? To investigate these questions, the predictive utility of the CCHE Index Score is contrasted with the predictive utility of its constituent components (all three components, and GPA/ACT only) for three commonly-assessed undergraduate student outcomes: First Fall GPA, Freshman Retention to the second Fall semester, and Six Year Graduation from CSU in any College.

Linear regression models are constructed to assess the relationship of the Index and its components to First Fall GPA. Each model includes several demographic variables known to correlate with undergraduate student success at CSU including STEM major status, Gender, Pell Eligibility, Colorado Residency, First Generation Status, and URM status. The CCHE Index model includes the above demographic variables plus the computed Index score; the All Components model includes the demographic variables plus High School GPA, High School Rank Percentile, and ACT score; and the GPA/ACT only model includes the demographic variables plus High School GPA and ACT score. Each model is applied to five Freshman student populations: First Generation students, Non-First Generation students, URM students, non-URM students, and All Students.

Table 2 displays the regression model R^2 results for each model and each studied population. Each of the 15 displayed models is significant at the $p < .01$ level. Comparisons should be made primarily within student group, represented by values in the same column of the table.

Table 2: First Fall GPA Regression Model R^2, by URM Status and FGEN Status

First Fall GPA: Regression R Square					
	Non-FGEN	FGEN	Non-URM	URM	All
CCHE Index	.258	.216	.254	.206	**.252**
All Components	.284	.244	.281	.232	**.278**
GPA/ACT only	.278	.243	.276	.232	**.274**

For the All Students group the regression model using Index score, but no components, accounts for 25.2% of the variability in First Fall GPA. By contrast, the two component models each account for over 27% of the variability in First Fall GPA, a difference of over 2 percentage points. These differences in explained variance are consistent across the four studied sub-groups, with the All Components model and GPA/ACT only models explaining about 2 percentage points more variance than the CCHE Index model. Table 3 shows these differences in explained variance for the CCHE Index and GPA/ACT only models as a proportion of the explanatory power of the All Components model.

Table 3: Proportion of Variance Explained versus All Component Model, by URM Status and FGEN Status

First Fall GPA: Model R Squares, relative to All Components					
	Non-FGEN	FGEN	Non-URM	URM	All
All Components	100.0%	100.0%	100.0%	100.0%	100.0%
CCHE Index	90.6%	88.5%	90.6%	88.9%	90.7%
GPA/ACT only	97.8%	99.7%	98.4%	100.2%	98.6%

Across the five studied populations, the First Fall GPA regression model that only uses CCHE Index explains between 88.5% and 90.7% of the variance that is explained by a model using all three of the Index's components. This is far below the proportion of variance explained by the GPA/ACT only model – a model that uses only two of the Index's components – which explains 97.8% or more of the variance accounted for by the All Components Model.

Table 4 displays the percentage of model explained variance (R^2) that is accounted for by each independent variable in the All Components model, across the four studied subgroups (each column sums to 100%). Variables that are non-significant are represented by ---, while variables not included in a subgroup's model are noted with N/A. For all groups of students, High School GPA and ACT Score are the two strongest predictors of First Fall GPA, with STEM Major status and Gender explaining the most variance among the demographic predictors. It may be of interest to note that High School Rank Percentile is not a significant predictor of First Fall GPA for URM or First Generation students, yet also serves as the second-strongest component of the CCHE index for URM students and accounts for over 20% of Index composition for every studied student group.

Table 4: IV % of Explained Variance in All Component Model, by URM Status and FGEN Status

First Fall GPA: % of Explained Variance by IV					
	Non-FGEN	FGEN		Non-URM	URM
High School GPA	39%	48%		39%	47%
HS Rank Percentile	8%	---		7%	---
ACT	22%	23%		21%	24%
Stem Major	13%	12%		12%	14%
Gender	9%	8%		9%	9%
URM Status	3%	---		N/A	N/A
First Gen Status	N/A	N/A		3%	---
Pell Status	3%	5%		6%	5%
CO Residency	2%	4%		2%	---

Logistic Regression Results: Freshman Retention

Freshman persistence to the second fall term (or "Retention") is another key measure of student success studied in this report. Historically, logistic regression models predicting retention at CSU yield small model fit values, as the school experiences high retention rates that limit our ability to predict retention strongly. Despite these low model fit values, provided that all tested logistic regression models are significant then we may compare between Pseudo R^2 model fit values to determine which set of independent variables yields a better fit for predicting Retention outcomes.

Table 5 displays Nagelkerke pseudo R^2 estimates of model fit for each of the 15 logistic regression models created using the 3 independent variable sets on the left side of the table across the five student groups at the top of the table. All 15 models are significant at $p < .05$ level. All 15 models include the studied demographic variables.

Table 5: Retention Logistic Regression Pseudo R^2, by URM Status and FGEN Status

Retention: Logistic Regression Nagelkerke Pseudo R2					
	Non-FGEN	FGEN	Non-URM	URM	All
CCHE Index	.034	.041	.038	.045	**.040**
All Components	.040	.049	.047	.050	**.047**
GPA/ACT only	.037	.048	.043	.049	**.045**

Similar to the results observed when comparing First Fall GPA regression models, we observe a trend within each of the five studied groups wherein the All Components model outperforms the GPA/ACT model, which in turn outperforms the CCHE Index model. The absolute differences between the proportions of variance predicted by the CCHE model versus the All Components model appear to be small, ranging from 0.5 percentage points to 0.9 percentage points. However, these discrepancies represent larger percent differences than those observed in the First Fall GPA study, with the CCHE Index accounting for between 82.4% and 89.2% of variability compared to the All Components model (see Table 6 for proportion of variance explained by each model relative to All Components).

Table 6: Proportion of Retention Variance Explained versus All Component Model, by URM Status and FGEN Status

Retention: Model Pseudo R Squares, relative to All Components Model					
	Non-FGEN	FGEN	Non-URM	URM	All
All Components	100.0%	100.0%	100.0%	100.0%	100.0%
CCHE Index	84.2%	84.1%	82.4%	89.2%	85.3%
GPA/ACT only	93.0%	98.0%	91.9%	98.1%	94.4%

Table 7 shows the log odds for each independent variable in the Retention All Components model, for each of the five studied student groups. For each variable, --- represents a non-significant finding at the $p < .05$ level, N/A indicates that the variable was not used in that model, and the numerical value indicates the change in log odds associated with a one-unit change in the independent variable (log odds under 1 correspond to an inverse correlation). For all five studied groups, High School GPA and ACT Score serve as significant predictors of retention. Colorado Residency and First Generation Status were also significant across all models in which those variables were included. It

may be interesting to note that Rank Percentile is only a significant predictor of retention for First Generation and URM students, whereas for First Fall GPA that predictor was significant for the opposite subgroups only (non-First Generation and non-URM students). Overall, we observe similar patterns of significance among the component variables across Index-only versus All Components models with the exception of the gender variable.

Table 7: IV Log Odds by URM Status and FGEN Status, Index-Only and All Components Models

Retention: Log Odds, Index-Only Model				
	FGEN	URM	Non-FG/URM	All Students
Index (standardized)	1.38	1.42	1.36	**1.37**
Stem Major	1.28	---	1.14	**1.14**
Gender (Male=1)	---	---	---	**1.05**
URM Status	---	N/A	N/A	---
First Gen Status	N/A	0.77	N/A	**0.75**
Pell Status	---	---	0.73	**0.86**
CO Residency	1.89	1.84	1.65	**1.71**

Retention: Log Odds, Components Model				
	FGEN	URM	Non-FG/URM	All Students
HS GPA (stand.)	1.57	1.55	1.36	**1.43**
HS Rank % (stand.)	0.87	0.82	---	---
ACT (standardized)	1.11	1.18	1.12	**1.13**
Stem Major	1.20	---	---	**1.11**
Gender (Male=1)	---	1.27	1.16	**1.13**
URM Status	---	N/A	N/A	---
First Gen Status	N/A	0.79	N/A	**0.75**
Pell Status	---	---	0.69	**0.86**
CO Residency	1.95	1.85	1.72	**1.79**

Logistic Regression Results: Six Year Graduation

Six year graduation rate is defined as the proportion of Freshman students who graduate with a bachelor's degree from CSU within 6 full years of initial matriculation. As with studies that assess retention, our predictive models are limited in their capacity to explain six year graduation rates, although our ability to accurately predict graduation is substantially greater. Table 8 shows the three logistic regression models run across the five studied student groups, with Six Year Graduation as the dependent variable. Again, each of the 15 studied models is significant at the $p < .05$ level.

Table 8: Graduation Logistic Regression Pseudo R^2, by URM Status and FGEN Status

Six Year Grad: Logistic Regression Nagelkerke Pseudo R2					
	Non-FGEN	FGEN	Non-URM	URM	All
CCHE Index	.064	.077	.073	.067	**.075**
All Components	.084	.106	.098	.088	**.098**
GPA/ACT only	.084	.103	.095	.087	**.097**

As with the prior regression model comparisons, we observe that the All Components model tends to explain the highest estimated proportion of variance, while the GPA/ACT only model explains the next greatest proportion (except in the case of non-First Gen students, for whom GPA/ACT only slightly outperforms the All Components Model). Once again, the CCHE index model explains a far lesser portion of the variance than the component models, at a rate that is at least 2 percentage points lower than the All Components across all student groups. These differences are more noticeable when expressed in percentage form (Table 9), with the CCHE Index models explaining between 72.4% and 76.4% of the variance in six year graduation explained by the All Components model.

Table 9: Proportion of Graduation Variance Explained versus All Component Model, by URM Status and FGEN Status

Graduation: Model Pseudo R Squares, relative to All Components Model					
	Non-FGEN	FGEN	Non-URM	URM	All
All Components	100.0%	100.0%	100.0%	100.0%	100.0%
CCHE Index	75.9%	72.4%	74.5%	75.8%	76.4%
GPA/ACT only	100.5%	96.7%	97.6%	98.6%	98.2%

Table 10 shows the log odds for each independent variable included in the All Components models predicting Six Year Graduation, across the five studied student groups. As a reminder, --- represents a non-significant finding at the $p < .05$ level, N/A indicates that the variable was not used in that model, and the numerical value indicates the change in log odds associated with a one-unit change in the independent variable (log odds under 1 correspond to an inverse correlation).

Table 10: IV Log Odds by URM Status and FGEN Status, Index-Only and All Components Models

Six-Year Grad: Log Odds, Index-Only Model					
	FGEN	URM		Non-FG/URM	All Students
Index (standardized)	1.65	1.54		1.51	**1.55**
Stem Major	---	0.80		---	**0.92**
Gender (Male=1)	0.89	---		0.80	**0.84**
URM Status	---	N/A		N/A	**0.84**
First Gen Status	N/A	0.82		N/A	**0.72**
Pell Status	0.87	---		0.69	**0.78**
CO Residency	1.60	1.68		1.45	**1.50**

Six-Year Grad: Log Odds, Components Model					
	FGEN	URM		Non-FG/URM	All Students
HS GPA (stand.)	2.09	1.85		1.67	**1.78**
HS Rank % (stand.)	0.84	0.81		---	---
ACT (standardized)	1.10	1.17		1.05	**1.08**
Stem Major	---	0.73		0.91	**0.90**
Gender (Male=1)	---	---		---	---
URM Status	---	N/A		N/A	**0.83**
First Gen Status	N/A	---		N/A	**0.70**
Pell Status	0.85	---		0.64	**0.74**
CO Residency	1.55	1.69		1.42	**1.47**

Similar to the retention logistic regression model, we observe that High School GPA, ACT Score, and Colorado Residency are all significant predictors of graduation across all student groups. Also similar to the retention study, we observe that High School Rank Percentile is only a significant predictor of graduation for two subgroups, First Generation and URM students. Gender is the only studied variable that is not a significant predictor of six year graduation for any studied populations. The high log odds ratio for the standardized High School GPA variable underscores the unique predictive importance of high school GPA for six-year graduation success at CSU, particularly for at-risk First Generation and Underrepresented minority students.

The observed interaction between URM status and First Generation status is one relationship that may be of interest. For First Generation students, URM status does not have a significant effect on graduation, but for non-First Generation students we observe a substantial effect of URM status on graduation. Likewise, for URM students there is no effect of First Generation status but for non-URM students there is a significant decrease in likelihood of graduation associated with being a First Generation student. The fact that we do not observe additive effects suggests that a student's membership in multiple higher risk graduation groups – i.e. URM and First Generation – is of minimal concern relative to a student's membership in at least one higher risk group.

Conclusions

The CCHE Index is commonly used in CSU descriptive reports and predictive models as a singular measure of academic preparedness. While the index may hold some value for decision makers and modelers who require a single-number proxy, this study provides strong evidence that the Index is less useful for predictive modeling of student success than models that substitute the Index's constituent components as IVs. The losses in predictive validity associated with using Index rather than its components are consistently present across all studied outcome variables and student subgroups, with relative proportions of explained variance ranging from 90.7% at the high end to 72.4% at the low end in these three studies.

Perhaps of greater interest is the fact that a limited components model that uses only two of the Index Components as IVs – GPA/ACT only – also performs substantially better than a model using the Index composite. These GPA/ACT only models explain slightly less variance than the All Components models in 13 of 15 cases, and actually explain a descriptively higher (though statistically similar) proportion of variance in 2 of the 15 observed group-model combinations. This latter finding may seem counterintuitive at first, but makes sense in light of the fact that Rank Percentile is minimally predictive for the studied outcome variables and may actually contribute noisy data that conflicts with the more accurate GPA and ACT predictors for some students.

Practically speaking, the fact that the GPA/ACT only models strongly outperform the CCHE Index models holds substantial implications for predictive modeling at CSU. Institutional data on incoming freshmen often omits High School Rank Percentile, but High School GPA and ACT Score are almost always available for accepted domestic students. Per the findings in this report, substituting High School GPA and ACT Score into predictive models where we currently use Index score as a proxy for academic preparedness would likely yield stronger model fits and improve our ability to predict and interpret student outcomes. Moving toward use of these two index components in predictive models, rather than the CCHE index score proxy, may also help us better understand how those two components interact with the experience of First Generation students, URM students, and other student groups of interest in a more nuanced fashion than the unitary CCHE Index score currently allows.