Review of Incoming Student-Athlete Readiness

Prepared for the University of the North Carolina, Chapel Hill

March 2014

TABLE OF CONTENTS

Executive Summary and Key Findings

Introduction

The following report provides an investigation into a recently published article about the reading proficiency levels of incoming student-athletes at the University of North Carolina, Chapel Hill. It is important to note that this report is in no way biased towards or against the validation of the claims of either party, but is an independent third-party review of not only the technical calculations, but also the instrumentation.

The calculations performed within the original study are basic. This report will replicate such calculations and provide further insight into potential strengths and weaknesses of the data provided. A full review of the Scholastic Abilities Test for Adults (SATA) assessment will provide additional insight into the appropriateness of the instrument to measure student-athlete readiness.

Summary Responses to Key Questions:

- Can you verify that the Reading Vocabulary (RV) data in the provided spreadsheet is in the form of standard scores?

  Answer: Yes, I can validate that the data provided is standard scores rather than raw scores. This is based on a distribution analysis and an understanding of the SATA instrument.

- Can you verify that treating standard scores as grade equivalents is a serious error in data analysis?

  Answer: In my opinion, grade equivalents should never be presented as a measure of performance. There are a number of drawbacks associated with grade equivalents, they are difficult to interpret. For the SATA test specifically, I find that the grade equivalents lack statistical rigor in their construction. Finally, the use of Standard Scores (as discussed within the SATA testing manual) is not the same as grade equivalents; to present it as such would be a serious error.

- Based on your professional expertise, is it possible to assign a reading grade level to a student based on a combination of SATA RV and writing subtests, SAT verbal scores, and one on one work with students?

  Answer: As discussed later in this report, grade-level equivalents are rarely an appropriate mechanism to report performance. It is my opinion that there are many misconceptions about the interpretation of grade-level equivalents and the technical merits of their generation are suspect. As someone who has developed assessments within both the K-12 and postsecondary enrollment, grade-level equivalents do not provide an accurate measure of student readiness or ability.

- Based on your professional expertise and your analysis of the data set, would this difference in data set population vs the sample norm present a problem in reporting results?

  Answer: In theory, random sampling of individuals should create an unbiased measure. However, I am not convinced that the SATA instrument is based on such. Additionally, the SATA instrument uses a relatively low number of participants to validate its reliability and item difficulty as compared to other leading ability test.

  The greater concern is the overrepresentation of revenue-generating student-athletes within the sample. As discussed below, revenue generating sport athletics represent approximately 18% of the total UNC student-athlete population (according to the EADA dataset – http://oped.ed.gov/athletics) but represent 81% of the individuals within the data provided.

- Finally, to what extent should the SATA RV subtest be considered a measure of literacy? And if so, should these results be reported in terms of grade level?

  Answer: In my opinion, I do not find conclusive evidence that the SATA RV subtest is an accurate measure of literacy. Reading and literacy are not synonymous. Literacy is traditionally defined as knowing how to read, write, and speak formally. Reading is the oral interpretation of written language.

  Again, grade-level equivalents are rarely an appropriate mechanism to report student ability.

ABOUT THE AUTHOR

Prior to reading this review, it is important to gain perspective on the author. This review is authored by Dennis Kramer. Kramer is an Assistant Professor of Higher Education at the University of Virginia and the program advisor for the Intercollegiate Athletic master's concentration within the higher education program. Kramer specializes in advanced quantitative and econometric research methods for studying such issues as the impact of financial aid policies on secondary and postsecondary success, and the economic influences on the college choice process. He also has a strong research interest in institutional decision-making and financial trends in college athletics. Kramer's work on intercollegiate athletics has been featured in reports published by the Knight Commission on Intercollegiate Athletics, the Chronicle of Higher Education, Inside HigherEd, and numerous academic publications and news media outlets. Prior to his University of Virginia post, Kramer worked as the senior research and policy analyst for the Georgia Department of Education, where he managed Georgia's policy development and evaluation research.

In the Fall 2014, Kramer will be joining the faculty at the University Florida as a tenure-track Assistant Professor and Associate Director of the Institute of Higher Education within the University of Florida's higher education administration program.

SECTION I: OVERVIEW OF THE INSTRUMENT

The following section provides an overview of the Scholastic Abilities Test for Adults (SATA) assessment as its current use in practice and the academic literature supporting its use. The section concludes by presenting concerns around the construction of the instrument. Additional questions are also presented to provide further clarity on assumptions.

OVERVIEW OF THE SATA ASSESSMENT

SUMMARY OF PURPOSE

According to Mental Measurement Yearbook (MMY) – the annual report of the technical aspects of the assessments used in individual diagnostic settings -- the SATA assessment instrument is suitable as a "general measure of scholastic accomplishment" (p.919). The MMY specifies that two of the major purposes of the instrument are "determination of an individual's relative strengths and weaknesses in both ability and achievement areas" and "documentation of an individual's progress over time in an intervention program" (Smith, p. 921). No mention is given about the ability for the SATA instrument to accurately generate grade-level equivalencies.

RELIABILITY

Internal consistency reliability estimates for all subtest and composite scores are presented in SATA's test manual. A majority of SATA's reliability coefficients are over 0.80. This generally signifies reliable assessment.

However, the concern is the lack of information provided regarding test-retest reliability. A concern regarding SATA's reliability is the relatively small sample size (n=23) in which the reliability coefficients were estimated. The development of the more robust and nationally normed assessment, test-retest reliability coefficients are generated from a larger subsample of initial testers. It is then validated with a randomly selected subgroup of participants outside of the initial group. This does not appear to be the case with the SATA assessment. While the internal reliability coefficient of 0.80 is adequate the underlying data generating such appears to be limited. This means the measurement of student-athletes' performance at UNC may not be reliable across time.

VALIDITY

Content-, criterion- and construct-validity information is provided within the SATA test manual. Correlations between the SATA and other tests (such as the ACT) are given as evidence of criterion validity. These figures appeared to suggest soundness in this area. Average performance for three unique groups of examinees (college and high school students with LD and students of adult basic education) was in expected ranges, further demonstrating construct validity. The use of the correlations as evidence of construct validity is a weak approach and additional test validity measures should be implemented to ensure the SATA measures reading meaningfully.

Other Test Information

The SATA test can be administered to groups or individuals. The amount of time allowed is 2 hours for the former situation and 1to 1-½ hours for the latter. The MMY recommends that the test users have "some formal training in assessment" (Smith, p. 922) but do not specify the level of training needed. This instrument has only been in use for approximately 10 years. Further development and refinement is warranted prior to publication of a revised test manual. The testing manual also provides the ability for individual subtests to be assessed. It was not clear from the data the length of time provided to each student-athlete to complete the assessment. Also, no information was provided about the level of SATA-based training completed by the individuals who administered the instrument to the student-athletes.

Additional Test Construction Concerns

In reviewing the SATA testing manual, a number of concerns regarding the test construction are identified. The following provides an overview of three concerns.

Sample Size for Item Development and Normalization

On page 31 of the SATA testing manual, a sample size of 1,005 individuals from 19 different states served as the sample in which item-level analyzes were performed. The subsequent tables provide "evidence" that the SATA test is not only a valid but reliable measure of reading ability. However, the sample of 1,005 is smaller than other nationally used tests measuring student abilities. In fact the Wide Range Achievement Test (WRAT) and WRAT-R – the most widely used test and one in which SATA includes in their correlations to other test section on page 37 – uses over 15,000 participants. The SATA assessment uses less than 10% of that sample size, which produced questions on the claims made in the testing manual.

Subtest Reliability

In general, the reliability of the SATA assessment is sufficient; however, the subtest for the RV – the subtest used in the analysis – appears to have mixed results. It is important to note that for each age group the reliability coefficient was greater than 0.80 – indicating a strong reliability coefficient. However, for test-takers under the age of 18, the RV provided the least reliable measure amongst reading / verbal subtests. For 18 and 19 year old test-takers, the RV subtest was comparatively moderate. For test-takers in their 20's, RV was the most reliable reading / verbal subtest. The differences in the relative reliability make understanding the performance of student-athletes who differ in age difficult. In reviewing that provided data, this was not taken into consideration while interpreting the results.

Subtest Discriminating Power & Item Difficulty

Subtest discriminating power and item difficulty are two important aspects of psychological assessments. In particular, these two components describe the ability of the instrument to detect real differences between individuals and the appropriate level of difficulty. Similar to the reliability coefficients, item difficulty and discriminating power appear to be different based on the test-taker's age. These differences further complicate the interpretation of student performance.

Current Assessment Uses

After reviewing the current uses of SATA within higher education, it appears to be some utilization of the SATA assessment within institutions of higher education. In particular, the SATA assessment is designated as an approved assessment for the identification of students with disabilities – or relative performance amongst peers. This report could not identify a single institution that exclusively utilizes SATA; however, a number of them provided it as one option. The lack of institutional commitment to the SATA assessment is a cause for concern to be used as a single test of students' ability.

Support and Limitations within the Academic Literature

Concerns Discussed within the Literature:

While there appears to be support of SATA within learning support settings, there are limited academic studies that produced defensible data on the ability of the instrument to predict abilities. In a survey of 728 of the nation's leading neuropsychologists, Stevens and Rice (1999) found that the most widely used assessment was the Wide Range Achievement Test (WRAT) for assessing the reading abilities of students. Less than 10 respondents (less than 1%) reported using the SATA instrument in any capacity and fewer reported using it for the assessment of reading abilities. The lack of acceptance of the SATA assessment by the psychological community produces yet another concern with the use of the SATA assessment as the only measure of student ability.

SECTION II: TECHNICAL REPLICATIONS & ADDITIONS

The following provide both a descriptive and analytical overviews of the data provided. The replication of grade equivalency claims made from this data is the primary aim of this section. Additional considerations around Standard Error of Measurement will assist in future interpretation of the results.

RECONSTRUCTION OF GRADE EQUIVALENTS

The following is an attempt to replicate the grade equivalent data reported from the provided data.

PROCEDURAL OVERVIEW

To ensure accuracy and future replication, this section provides an overview of the steps and assumptions used in the creation of grade equivalency. Information on the conversion from standard scores to the grade equivalency is informed by the SATA testing manual.

Step 1 – Convert Standard Scores into Raw Scores

Within the SATA testing manual, grade equivalency tables only provide a concordance between raw scores and grade equivalency. Page 49 of the SATA testing manual provides the information needed to convert standard scores into raw scores. An important consideration is the age of the student when tested, as age impacts the appropriate conversion.

In the event that a standard score was mapped to a range of raw scores, the lowest possible raw score was selected. This was done purposefully to ensure a "worst-case scenario" in terms of reading ability. In the end, this decision could potentially underestimate the actual reading level of student-athletes.

Step 2 – Converting Raw Scores into Grade Equivalencies

Using the conversion table on page 72 of the SATA testing manual, the raw score for each student was converted into a unique grade equivalency.

Step 3 – Accuracy Check

After converting raw scores into grade equivalencies, the calculation was re-run to ensure the accuracy in not only the assignment of the standard scores to raw scores, but also the incorporation a student's age into the assignment of raw scores. Finally, a second examination was conducted on the raw score to grade equivalency conversation.

DESCRIPTIVE STATISTICS

Using the approach described above, Table 3.1 provides an overview of the descriptive statistics related to grade-level reading performance. On average, the tested student-athletes within the sample provided performed at reading level equivalent to grade 12. On average, student-athletes participating in football and men's basketball test at an equivalent grade-

level reading level at 12.17, while student-athletes participating in all other sports demonstrated a 12.87 average reading grade-level equivalency.

Table 3.1: Descriptive Statistics on Grade-Level Reading Ability (by gender and sport type)

	Average Grade Equivalency	Minimum Grade Equivalency	Maximum Grade Equivalency	Standard Deviation
All Revenue Sports	***12.17***	***3.00***	***16.00***	***2.84***
Male	*12.17*	*3.00*	*16.00*	*2.84*
Female	*N/A*	*N/A*	*N/A*	*N/A*
Non-Revenue Sports	***12.87***	***8.30***	***16.00***	***2.02***
Male	*12.74*	*8.30*	*16.00*	*2.09*
Female	*13.25*	*10.00*	*16.00*	*1.82*
All UNC Sports	***12.31***	***3.00***	***16.00***	***2.71***
Male	*12.23*	*8.30*	*16.00*	*2.09*
Female	*13.25*	*10.00*	*16.00*	*1.82*

Table 3.2 provides the distribution of the grade-level reading performance by sport type. The initial distribution appears to refute the findings of initially reported distribution of performance. Additionally, it appears that the original data incorrectly assigned the grade-level reading levels.

Table 3.2: Grade-Level Reading Distribution (by sport type)

	Football & Men's Basketball	All Other Sports	All Sports
No Score Reported	5	1	6
Grade 3	1	0	1
Grade 4	2	0	2
Grade 5	1	0	1
Grade 6	3	0	3
Grade 7	2	0	2
Grade 8	4	1	5
Grade 9	18	1	19
Grade 10	16	3	19
Grade 11	12	5	17
Grade 12	8	3	11
Grade 13	35	8	43
Grade 14	23	7	30
Grade 15	27	9	36
Grade 16 or Higher	14	2	16

% Below Postsecondary Level	**40.4%**	**33.3%**	**39.0%**
% 4th to 8th Grade Level	**7.8%**	**2.6%**	**6.8%**

Conclusion on Grade Equivalency Accuracy

A recent CNN articled stated, "60% [of UNC athletes] read between fourth- and eighth-grade levels. Between 8% and 10% read below a third-grade level." The initial comments gained national traction due to the severity of the claims. However, this analysis finds, rather than 60% of UNC student-athletes possessing a 4th to 8th grade reading level, only 6.8% of student-athletes read at such level.

Given the severity in the differences between the findings of the initial claims and this report, additional attempts were made to replicate the initial findings. It does not appear the original data equated either standard scores or raw scores as grade equivalents. Additionally, this report could not find any analytical approach that produced the 60% reported initially from the data provided. It is unclear if a different dataset was used to generate these claims; however, the data provided for this report could not substantiate the claims.

Predictors of SATA Performance

In an effort to verify the original claim, this review conducted a linear fixed-effect ordinary least squares regression model. This model attempted to quantify the potential impact of student-level participation to verify the claims. Table 3.3 provides an overview of the regression output. The results presented within Table 3.3 should be cautioned as the data provided was limited. However, results indicate that performance on the SATA assessment was significantly predicted by race and gender, but not sport participation or age of entry. In particular, it appears that males performed two points lower than their female counterparts. Additionally, African Americans performed 2.3 points lower than their White counterparts regardless of their age or sport participation.

The common discourse around the original report and its data is that football and men's basketball players were admitted with significantly lower reading levels as compared to student-athletes in non-revenue generating sports. However, it appears that the SATA assessment is biased downward for males and African Americans rather than football and men's basketball participants. Given that African American males are highly represented within these two sports, it stands to reason that the potential gender and racial biases of the SATA assessment are leading to lower scores for that particular population. While further data is needed to validate these claims, Table 3.3 provides a basis for future inquiry into the potential bias.

Table 3.3: OLS Regression Output

Number of obs	176
F (7 , 160)	5.820
Prob > F	0.000
R-squared	0.207
Adj R-squared	0.132
Root MSE	2.545

	Coefficient	Standard Error	*t*	*p*	95% Confidence Interval	
Male	-2.067**	0.850	-2.430	0.016	-3.745	-0.389
African American	-2.267***	0.425	-5.340	0.000	-3.106	-1.428
Bi-Racial	-1.062	0.801	-1.330	0.187	-2.643	0.519
Other	-0.692	0.862	-0.800	0.423	-2.393	1.010
Tri-Racial	-1.525	2.500	-0.610	0.543	-6.462	3.412
Sport: Non-Revenue	-1.306	0.768	-1.700	0.091	-2.823	0.210
Age at Entry	-0.240	0.216	-1.110	0.269	-0.666	0.187
Constant	20.298	3.817	5.320	0.000	12.760	27.835

Notes:
1: Model includes a year of entry fixed-effect; race coefficients are referenced to the White subgroup population
2: A full technical write-up on the analytical model can be provided upon request of the author

STANDARD ERROR OF MEASUREMENT (SEM) CONSIDERATION

The standard error of measurement is defined in the Standards for Educational and Psychological Testing (1985) as "the standard deviation of errors of measurement that is associated with the test scores for a specified group of test takers" (p. 94). Page 34 of the SATA test manual provides an overview of the instruments SEM. In addition, Table 4.6 provides an overview of SEM estimates by age and subtest. The primary concern related to initial findings is that the estimates of grade-level equivalency did not include consideration of the SEM. This means that, in reality, any given student-athlete might have an actual performance level higher than what was achieved on the RV subtest and their lower performance on that particular administration may be directly related to the construction of the test rather than their ability. While the SEM estimates for the RV subtest are relatively small (1 point for each), the resulting impact could be as significant as a 1.3 increase in grade-level equivalency.

SECTION III: ADDITIONAL PERSPECTIVES

The next section provided two additional perspectives to consider when reviewing the original claims based on the provided data.

CAUTION ON THE USE OF GRADE-LEVEL EQUIVALENTS

The misconceptions of grade equivalents have been well documented by research organizations, most noting that the metrics create more confusion than clarity (AERA/APA/NCME, 1985; Airasian, 1994; Miller, Linn and Gronlund, 2009; Stiggins, 2009). In 1981, for example, the International Reading Association (IRA) crafted a resolution on the misuse of grade equivalents. In it, the organization "strongly advocates that those who administer standardized reading tests abandon the practice of using grade equivalents to report performance of either individuals or groups of test takers" (1981).

Grade equivalents (grade.month) do not represent a grade-level curriculum standard. For example, a grade equivalent of 5.9 does not indicate the desired level of performance for all fifth graders. Rather, it only represents the group's median score, or projected score, for fifth graders in their ninth month of schooling. Achieving the same score as the average student in the referenced group may not be an appropriate goal for all students. As discussed previously, the small sample used to derive such grade-level equivalents leads to potential bias and misrepresentation of actual reading levels.

ENTRANCE READING LEVEL VS. VALUE-ADDED

Admission within higher education is traditionally viewed as a holistic exercise where students' applications are viewed in totality. Often-times, universities (even those with the academic reputation of the University of North Carolina, Chapel Hill) admit students who possess exemplary talents not captured within traditional measures of academic abilities. These admits typically include students who excel in the area of arts and music, have unique skillsets in the area of technology, or have a commitment to community service and activism. Student-athletes are viewed in much of the same way. They possess a talent that is not captured by any of the current pre-college academic assessments.

For student-athletes in particular – especially those in revenue-generating sports – athletics serves as a primary access point to a postsecondary education. Often-times these athletes are the first person in their family to attend college, many of them the first to graduate high school. The academic literature on the biases of standardized academic measures for minority and first-generation college students is well documented. There is no evidence to the contrary for the SATA assessment. Finally, the academic literature provides a robust discussion on the many benefits student-athletes provide to their general student body peers.

While this report finds some instances where student-athletes were admitted to the UNC-Chapel Hill with a "reading" grade-level equivalency less than grade 12, the interest should not be level of entrance but the value-added provided during their time in college. The incentives around student-athlete eligibility and academic performance create an environment

where athletic departments provide academic support to student-athletes who need such to succeed. If a student-athlete is admitted with a reading proficiency level below college-readiness, but exits their four-year of college with a reading level approaching or exceeding what would be expected for a college senior, then the university provided the necessary value-added to that student-athlete.

On the surface, it appears that UNC Chapel Hill provides the necessary value-added to student-athletes – as measured by the NCAA's Graduation Success Rate (GSR). Appendix A provides an overview of recent performance on the GSR. In summary UNC Chapel Hill routinely outperforms its public peer institutions in: 1) all sports; 2) football; and 3) men's basketball. In many cases, UNC demonstrates high performance on the GSR measure.
.

Section IV: Concluding Thoughts and Recommendations for Practice

The following section provides some concluding thoughts and recommendations for future inquiry on this topic. Many of these points represent the opinion of the author and should be taken as such. Few of these recommendations can be completed with the current set of data; however, future consideration of these would provide further evidence on the validity of the claims.

Triangulation of UNC Data

The most logical way to test the validity of the claims would be administer two (2) different assessments to future student-athletes at a single time. Given the discussion in the prior section about the drawbacks of the SATA assessment, it would benefits the researchers to validate their claims through multiple assessments. Additionally, course-based performance in first-semester English courses may provide additional information on ability to complete college-level work.

General Student Population Estimates

The claims made from the initial data appear to articulate a contrast in the incoming abilities of student-athletes as compared to their general student-body peers. However, no data is provided on the performance on the general students on the SATA assessment. This claim appears to be conjecture rather than actual data. Additionally, the claims made on total student-athletes performance appears to be misleading as the data's sample are biased toward student-athletes participating in revenue-generating sports and not representative of the entire student-athlete population.

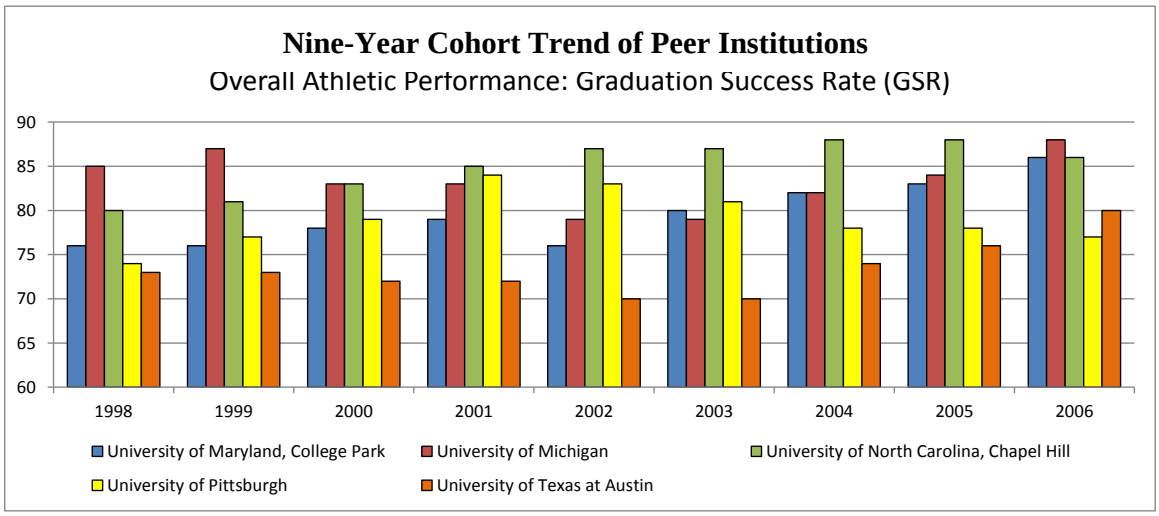
Concluding Thoughts

In conclusion, it appears the technical calculations reported from the initial data were inaccurate. This analysis finds that less than 7% of UNC student-athletes possessed a reading level between 4th and 8th grade. Additionally, this report provides a number of concerns around the sample used within the initial report and technical concerns around the SATA instrument. Further information is needed on the testing environment and time provided to complete the assessment. Finally, it is important to interpret the results with caution given the concerns outlined within this report.

APPENDIX A: GSR PERFORMANCE TRENDS

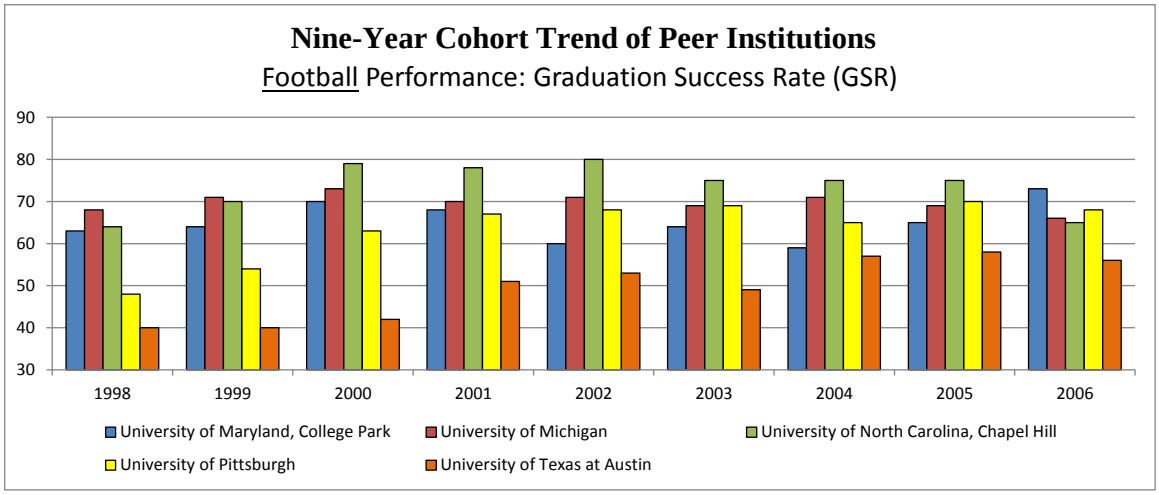
OVERALL GSR:

Overall, UNC Chapel Hill student-athletes had a GSR that was greater than the majority of the four identified public institution peers. In fact, cohorts from 2000 – 2005 outperformed each of the other peer institutions.



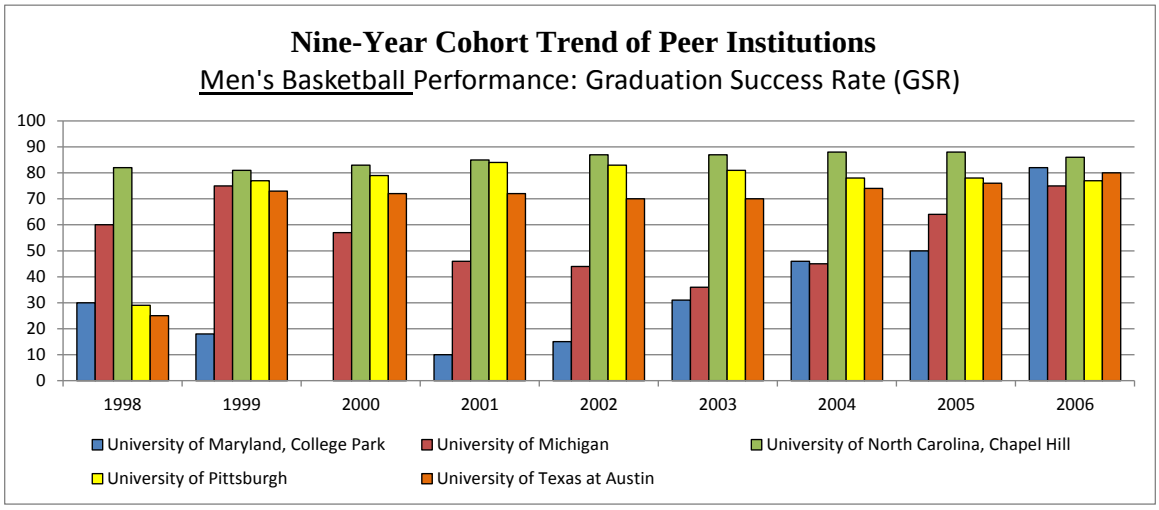
FOOTBALL GSR:

For eight (8) of the last nine (9) reported cohorts, the University of North Carolina, Chapel Hill football program finished first and second of the list of public peer institution.



MEN'S BASKETBALL GSR:

For each of the last nine (9) reported cohorts, the University of North Carolina, Chapel Hill outperformed each of the four identified peer public institutions.



CAVEAT

The author has used his best efforts in preparing this brief. The author makes no representations or warranties with respect to the accuracy or completeness of the contents of this brief and specifically disclaim any implied warranties of fitness for a particular purpose. There are no warranties which extend beyond the descriptions contained in this paragraph. No warranty may be created or extended by representatives of the author or its marketing materials. The accuracy and completeness of the information provided herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every member. Neither the publisher nor the author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Moreover, the author is not engaged in rendering legal, accounting, or other - professional services.