THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

OFFICE OF THE EXECUTIVE
VICE CHANCELLOR AND PROVOST

104 SOUTH BUILDING
CAMPUS BOX 3000
CHAPEL HILL, NC 27599-3000

T 919.962.2198
F 919.962.1593
www.unc.edu/provost

Friday, April 11, 2014

Executive Summary of External Independent Review

In January 2014, public claims were made regarding the reading ability of some of UNC-Chapel Hill's football and men's basketball players. These students were screened for learning disabilities between 2004 and 2012 using parts of the Scholastic Abilities Test for Adults (SATA), a test that is unrelated to the widely-known Scholastic Aptitude Test (SAT). The University conducted an internal review of the data set, and found that the analysis and claims were not supported by the data. (https://uncnews.unc.edu/?p=38025)

To further review the data set cited as the basis for the public claims, the University sought outside experts. Based on their independent analyses, all three experts found that the data set did not support the public claims. They found no evidence in the data to support the claims of widespread low literacy levels among tested UNC student-athletes. They also determined that the majority of the students referenced in the public claims scored at or above college entry level on the SATA Reading Vocabulary subtest (a 25-question multiple choice vocabulary test). The data set was based on these scores.

The roughly 182 athletes in revenue sports (generally considered football, men's basketball and women's basketball) who were mentioned in the public claims represent a small fraction of approximately 1,800 student-athletes who attended UNC in the eight-year period between 2004 and 2012, and they represent only 0.5% of the more than 35,000 undergraduate students who attended the University during this time.

This document provides a summary of the findings of the three external experts. These experts were provided with the data set given to UNC's Executive Vice Chancellor and Provost, who is the Chief Academic Officer and the administrator responsible for overseeing, among many areas, research. The three experts independently examined the data and reported separately on the validity of the analysis used to make the public claims. **This executive summary was produced by the Office of the Provost; the three experts reviewed and approved this summary.**

Contents:

Background
Internal Review of the Data Set
External Review of the Data Set
Charge to the External Experts
Key Findings and Results

Background

Between 2004 and 2012, UNC's Academic Support Program for Student-Athletes (ASPSA) contracted with an outside neuropsychologist to administer screening tests to some student-athletes to determine possible learning differences and/or learning disabilities. This is common practice at many NCAA Division I universities.

The results of these subtests were used by the outside neuropsychologist and ASPSA to identify whether additional testing would be necessary to determine a student's need for accommodations under federal laws prohibiting discrimination on the basis of disability. The neuropsychologist was not contracted to determine literacy levels or reading grade levels.

The screening tests were administered when the student-athletes first arrived at UNC, either during summer session II or fall semester of their freshman year. The student-athletes who took these tests, like other students tested for learning differences and/or disabilities, did so voluntarily with the expectation that the results would be treated confidentially and would be used solely to strengthen their educational experience.

The tests used to screen these students were subtests of the Scholastic Abilities Test for Adults, or SATA (the SATA and the SAT, or Scholastic Aptitude Test, are different tests, owned and administered by different entities). Not all of the SATA subtests were used for the screenings. For example, the SATA Reading Comprehension and Writing Composition subtests were not used for screening.

The following screening tests were used:

1. SATA Reading Vocabulary (RV) subtest (a 25-question multiple choice vocabulary test)
2. SATA Writing Mechanics (WM) subtest (a test of spelling, punctuation and capital letters)

3. SATA Math Calculation (MC) subtest

Internal Review of the Data Set

In January, public claims were made regarding UNC student-athletes' reading ability based on a review of scores on the SATA (taken by some student-athletes after they were admitted to the University). These claims stated that of 183 athletes in revenue-generating sports admitted to UNC between 2004 and 2012:

- Between 8% and 10% were reading below a third grade level.
- About 60% were reading between the fourth and eighth grade levels.

In order to assess the validity of these and other similar public claims, UNC's Provost requested that the data set used to make these claims be shared with him.

The Provost received the data set on January 13, 2014. It included scores from the SATA Reading Vocabulary, SATA Writing Mechanics, and SATA Math subtests for 176 student-athletes entering the University, beginning with the academic year 2004-05 and ending 2012-13. The public claims have stated that there were 183 students in the data set; however, there were 182 student names listed beneath a header row. There were no SATA subtest scores for six of the students listed. The data set contained identifiable information about students, including names, the year they entered the University and their sport, as well as SAT scores and GPA information that was added later, after the original screening.

An internal team from the Provost's Office and ASPSA compared the January 13 data set to the outside neuropsychologist's records that had been retained in ASPSA since 2004. They found that the January 13 data set did not include a complete list of student-athletes from revenue sports who were screened between 2004 and 2012. Rather, the January 13 data set included some (but not all) members of the football, men's and women's basketball, baseball and volleyball teams. The data set did not include information about student-athletes from non-revenue sports who were tested during the same period. Also, the data set listed some members of the baseball team as members of the men's basketball team.

Upon receipt of the data set on January 13, the Provost requested that the internal team review and attempt to replicate the findings that led to the public statements. On the basis of their analysis and a review of the SATA Examiner's Manual, the internal team had significant concerns, including:

- According to the SATA Examiner's Manual, **the SATA Reading Vocabulary subtest is not a valid test of reading.** "Any standardized test purporting to provide a comprehensive measure of reading that does not assess sentence or passage comprehension should be considered inadequate." (Wiederholt & Bryant (1987) p. 96, quoted in SATA Examiner's Manual p. 28)

- **The SATA Examiner's Manual strongly advises against using grade equivalents to report on students' reading ability.** "Before using grade equivalents as evidence of scholastic accomplishment, examiners are urged to read about their limitations. We strongly urge SATA users to report grade equivalents only when they have to." (SATA Examiner's Manual, p. 20)

- **Even if used against the Manual's advice, the grade equivalents calculated during the internal review did not match the public statements made about UNC student-athletes' reading levels.** The University's internal review found that more than 60% of the 176 students tested above a 12th grade equivalent. Less than 6% tested below an 8th grade equivalent, and even these results are questionable because the SATA does not have test score information for anyone below 10th grade: "Because our test was normed only on Grades 10-16 in the school-aged population (including postsecondary school) we had to extrapolate downward...This process results in sometimes spurious grade equivalents." (SATA Examiner's Manual p. 19)

- The column headings related to SATA data in the January 13 data set were RV, WM and Math. There was no column heading in the January 13 data set for grade equivalents. The data under the column heading for RV matched exactly the RV *standard scores* in ASPSA's records. **This was a significant concern because the public statements referred to these data as grade equivalents.** Standard scores are not the same as grade equivalents[1], they are not interchangeable. Standard

[1] Explanation of SATA subtest scores (from the SATA Examiner's Manual, pp. 18-20)

Raw scores are the total number of items scored correct for a subtest.

Standard scores allow examiners to make comparisons across subtests. Based on the distribution with a mean of 10 and standard deviation of 3, standard scores are converted from raw scores and assigned ranges for interpretation (very superior, superior, above average, average, below average, poor, very poor).

Grade equivalents are derived scores that indicate performance based on the average raw scores of persons at a particular age. Because the test was normed only on Grades 10-16 in the school-aged

scores reported as grade equivalents would represent a miscalculation and would be significantly low.

External Review of the Data Set

The Provost's Office engaged three external experts to provide an independent analysis of the January 13 data set. The external experts were identified and retained based on their expertise in adult literacy, assessment and measurement in education, and multivariate analysis. These individuals were identified based on recommendations obtained by Dr. Debbi Clarke, a consultant in the Provost's Office who serves on the Provost's Student-Athlete Academic Initiative Working Group. Dr. Clarke holds advanced degrees in higher education administration from Vanderbilt University and the University of Pennsylvania. She worked previously as Director of the MBA Program at UNC's Kenan-Flagler Business School.

These three external experts were hired:

Nathan Kuncel, Distinguished Professor of Psychology, University of Minnesota (https://www.psych.umn.edu/people/facultyprofile.php?UID=kunce001);

Lee Alan Branum-Martin, Associate Professor, Department of Psychology and Co-Investigator at the Center for the Study of Adult Literacy at Georgia State University (http://www2.gsu.edu/~wwwpsy/branummartin.html); and

Dennis Kramer, Assistant Professor of Higher Education at the University of Virginia (http://curry.virginia.edu/academics/directory/dennis-a.-kramer-ii).

The external experts worked independently of one another and were each given a copy of the data set provided to the Provost on January 13, as well as the analysis completed by the internal team. Because the information in the data set was personal and sensitive, the University took appropriate steps to preserve and protect the privacy of the students and the confidentiality of their records.

The University asked the external experts to focus specifically on the use of tests to measure reading ability and the use of grade levels to explain the results of these tests.

---

population (including postsecondary), [SATA] had to extrapolate downward...resulting in sometimes spurious grade equivalents.

Charge to the External Experts

1. Is the SATA Reading Vocabulary (RV) subtest a true reading test?

2. Were SATA RV standard scores mistakenly interpreted as grade equivalents to report on UNC student-athletes' reading grade levels?

3. Can you verify that the SATA Reading Vocabulary (RV) data in the January 13 data set is in the form of standard scores?

4. Can you verify that SATA RV standard scores are not the same as grade equivalents?

5. The SATA test is normed against a sample of 1005 people with the following demographics:

   Demographics (SATA sample N=1005)
   46% male
   54% female
   86% white
   8% black
   6% other

   The UNC student-athletes who took the SATA RV subtest have the following demographics:
   86% male
   14% female
   24% white
   59% black
   17% other

   Based on your professional expertise and your analysis of the data set, would this difference in demographics have an impact on reporting results?

6. To what extent should the SATA RV subtest be considered a measure of literacy? Should these results be reported in terms of grade level?

Key Findings and Results

Although the external experts were hired and worked independently of one another, they reached similar conclusions in three separate reports. Excerpts of those reports appear below; the full reports are available at http://carolinacommitment.unc.edu/:

1. **The SATA RV subtest, a 25-question multiple choice vocabulary test, is not a true reading test and should not be used to draw conclusions about student reading ability.**

"There does not appear to be conclusive evidence that the SATA RV subtest is an accurate measure of literacy." (Kramer)

"The SATA Reading Vocabulary subtest is a test of reading vocabulary, not of reading comprehension. The SATA includes a subtest for reading comprehension, but that was not reported." (Branum-Martin)

"The Reading Vocabulary (RV) scales assess a person's vocabulary knowledge. It cannot be viewed as a reasonable or comprehensive assessments of adult literacy." (Kuncel)

2. **The data do not support the public claims about the students' reading ability.**

"This report could not find any analytical approach that produced the 60% reported from the data provided." (Kramer)

"Figure 2.1 also makes clear that, at least based on estimated total scores, the majority (n=109) of the 176 students with valid scores in the sample had Reading Vocabulary grade equivalents above 12th grade. It should be remembered, however, that these are grade equivalents based on vocabulary, not on comprehension of reading connected text." (Branum-Martin)

"The grade equivalent for a raw score of 14 is 13.3th grade not 8th grade (Table D-1). Therefore, the 60% below average group is not anchored at the high end by 8th grade but by 13.3th grade, freshman in college." (Kuncel)

3. Reading ability should not be reported as grade equivalents.[2]

[2] See http://oms.umn.edu/mstp/tests_and_services/glossary.php for information regarding raw scores, standard scores, and grade equivalents.

"Standard scores for the SATA, which range from 1 to 20, do not represent grade level any more than SAT scores, which range from 200-800, represent grade level." (Kuncel)

"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. 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." (Kramer)

4. The difference in demographics between the SATA test norm and the demographics of the UNC student-athletes is important to understanding conclusions that can be drawn from the data.

*"*It is certainly true that the national norms have a different gender and racial composition than the student-athlete sample. The national norms were sampled to approximate the demographic profile of the United States. In contrast, the student-athlete data has more male and African-American students than the general population." (Kuncel)

"The UNC sample has relatively more males and African-American students than did the SATA norm sample. Contemporary test development usually includes sensitivity analyses of their test items to "bias" (e.g., differential item functioning) across demographic groups. Such subgroup bias analyses were not reported in the 1991 SATA Examiner's Manual (Raju, 1995). Examining test items for bias should be done both in terms of content as well as statistical models. Based on the age of the test, the lack of rigorous subgroup analyses, and the demographic differences between the SATA norm group and the UNC sample, results should be interpreted with caution. Moreover, because the grade levels tested in the SATA norm group only covered tenth grade and above, statements regarding grade level would seem to require other sources of evidence besides the SATA Reading Vocabulary subtest." (Branum-Martin)

"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 [public claims] is football and men's basketball players were admitted with significantly lower reading levels as compared to 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." (Kramer)

5. **The SATA subtests were administered in low-stakes settings, meaning that the result of the test had relatively unimportant consequences to the test taker. Low stakes settings are thought to influence test results.[3]**

"As is typical for counseling interventions, students were examined under low-stakes settings at UNC-Chapel Hill. An extensive literature has documented that assessments in low-stakes settings tend to result in lower scores than would be obtained if the test taker is highly motivated (e.g., seeking admission to college, competing for money) (Duckworth et al., 2011). In fact, even small financial incentives significantly and substantially increase test taker motivation and performance. Therefore it is likely that the SATA scores underestimate the maximal performance of the student-athletes." (Kuncel)

6. **While SATA RV (the 25-question multiple choice vocabulary subtest) results can be informative as part of screening for learning differences and/or disabilities, they are not accepted by the psychological community as an appropriate measure of reading grade level and literacy.**

"Typically, when people talk about literacy, they are most concerned with reading comprehension, or more broadly, student ability to handle academic reading and writing tasks. As noted previously, vocabulary is essential, but not sufficient for reading comprehension. In this respect, the SATA Reading Vocabulary subtest could be informative regarding reading related skills (esp. word knowledge), but may not be sufficient for broad judgments of student literacy. Ideally, other sources of information should be included, especially tests of reading comprehension." (Branum-Martin)

[3] See http://www.jmu.edu/assessment/wm_library/Examinee_Motivation.pdf for more information regarding low-stakes testing.

"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." (Kramer)

"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." (Kramer)

To read the external experts' full analyses, see http://carolinacommitment.unc.edu/.