Evaluation and Discussion of Student Athlete Testing Data at the University of North Carolina at Chapel Hill

Nathan R. Kuncel, Ph.D.
Department of Psychology
University of Minnesota

March 3, 2014

Overview

In this report, I examine data provided to me by the University of North Carolina at Chapel Hill (UNC-Chapel Hill) to answer some specific questions about assessments of student athletes, the reasonable interpretation of these scores, and the meaning of these scores placed in the broader context of student assessment and success in higher education.

Scholastic Abilities Test for Adults (SATA)

I read the examiner's manual for the SATA, the *Mental Measurement Yearbook* entries for the SATA, and conducted a literature search of research that appeared relevant to the current questions. The SATA contains a battery of subtests used to assess basic reading, writing, and mathematical skills. Only some of the subtests were used in student assessment and most of the public discussion of student athletes has focused on the Reading Vocabulary scale.

Interpretation of Reading Vocabulary as "Literacy" or a "Reading Test"

The Reading Vocabulary (RV) scales assess a person's vocabulary knowledge. It cannot be viewed as a reasonable or comprehensive assessments of adult literacy. Literacy assessments frequently demand that a person use multiple kinds of text documents to answer different kinds of questions or accomplish different goals. This approach is consistent with the two most recent national studies of adult literacy that defined literacy as, "Using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential." (p.2, Kirsch et al., 1993; NCES, 2006). The most recent national study, designed for national survey work, consisted of multiple types of documents (newspaper, order form, magazine article) and asked adults to perform a variety of tasks with the material. These measures provide a much clearer picture of literacy than a short vocabulary test.

However, it should be noted that all assessments of human cognitive abilities and skills are positively correlated to some degree. This means that if a person does well on one assessment (e.g., vocabulary tests) they are likely to do well on another assessment (e.g., college admissions tests). Similarly, if a person struggles with one they are likely to struggle with another. Given the evidence that weak admission test scores are associated with being at risk of poor performance in college (e.g., Kobrin, 2007; Sackett, Kuncel, Beatty, Shen, Rigdon, & Kiger, 2012), low scores on a test like reading vocabulary would make it more likely, but not certain, that a student may struggle in college. This relationship is consistent with the use of the Reading Vocabulary measure as a screening tool to identify students who may need additional academic support or have a learning disability. On its own, however, low RV scores are not sufficient to label a student as unable to benefit from a college education. As discussed later, a direct and comprehensive assessment of the student's academic preparedness, motivation, support, and study habits, attitudes and skills would be necessary to make the best evaluation of a student.

Testing in Low Stakes Settings

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.

This effect would be especially pronounced if students viewed it as boring, unnecessary, or a distraction. It is also worth noting that the Reading Vocabulary scale has some of longest instructions of any of the scales and requires the test taker to identify pairs of words with the same meaning or opposite meaning or the situation where there is no association among the words. Test takers need to both identify the words and label the type of association or label the absence of any association among the words. Given these complexities, if students are distracted or unmotivated, errors could occur further lowering scores. I do not have additional information on student motivation or interest in this specific case, however, test motivation may have affected scores.

Interpretation of Grade Equivalents

It is my professional judgment that the grade equivalent scores are not particularly meaningful and should not be interpreted. Grade equivalents compare a person's test score with the typical (median) performance of people at different grade levels. For example, if on a test a score of 62 was typical for test takers in 7th grade, the grade equivalent would be 7 for a person to obtain a 62. ***However, the SATA norms do not have testing score information for anyone below the 10th grade***. Grade equivalents for earlier grades were established by linearly extrapolating from scores for people in grades 10 to 16. In other words, grade equivalents for 2nd to 9th graders were based on the performance of later high school and college aged test takers and not 2nd to 9th graders.

> "Because our test was normed only on Grades 10-16 in the school aged population (including postsecondary school) we had to extrapolate downward." (Bryant et al., 1991,p. 19).

This process assumes that the relationship between scores and grade level for 10-16th graders holds for 3rd graders up to 9th graders. This assumption has not been evaluated for this extrapolation process. The SATA test developers note:

> "This process results in sometimes spurious grade equivalents." (Bryant et al., 1991, p. 19).

Finally, grade equivalents in general present challenges for interpretation as long noted in many psychometrics and test theory texts. For example, Crocker and Algina (1986) note the "severe limitations in these scores" (p. 450). Given the problems with extrapolated grade equivalents, they should not be interpreted. My conclusion is actually echoed by the SATA test developers who note in two places:

> "We strongly urge SATA users to base their test interpretations on the standard scores that are reported." (Bryant et al., 1991, p. 19).
>
> "SATA users are reminded that grade equivalents are easily misinterpreted and are useful only when federal or state departments mandate them." (Bryant et al., 1991, p. 32).

Despite their superficial intuitive appeal, grade equivalents should generally be avoided. If the goal is to provide accurate grade equivalents for literacy, a longer and more comprehensive assessment would be needed. This assessment should have score information available from students at all grade levels to permit accurate assignment of grade equivalent scores. Using professional judgment is not a desirable alternative to create grade equivalent scores. The student population a person works with will affect that person's impression of typical language skills at a grade level. For example, if a person generally works with fairly capable young readers, their perception of average performance will tend to be skewed upward. If expertise must be used to produce grade equivalent levels of literacy, the expert should have considerable experience evaluating and working with students in a wide range of grade levels.

Evaluating The Data File Scores

It was necessary before conducting additional analysis of the data to verify and check the scores listed in the file I was provided. Scores from two sources as well as a set of grade equivalent scores were present in the file. The grade equivalent scores were from an internal analysis conducted by UNC-Chapel Hill. The following steps were taken to check these scores for use in other analyses. First, the estimated age equivalents were clearly labeled in the file. Although these are problematic for interpretation, they correspond perfectly with raw scores allowing for a comparison with the other scores in the data file. That is, an age equivalent score can be converted into a raw score (number correct) without needing student age information. In four cases, the recorded age equivalent was not a possible value for Reading Vocabulary. The relatively small number of items means that the scale is fairly course. Multiple causes might produce these impossible values including errors on the part of the test scorer in looking up the grade equivalent, placing values from another test in the wrong spreadsheet column or a decision to modify the scores to a value that is not possible from the Reading Vocabulary Raw Score.

To test the correspondence between the grade equivalent scores and the scores from two sources, I used an estimate of student age and the raw score to determine the standard score using the table on p. 49 of the testing manual. This is important because the age of the test taker affects the standard score assigned. The data file contained scores reported from two sources for most of the student athletes in the file. These were consolidated into a single value to maximize the information available. The file also contained birth month and year as well as year entering college for most, but not all, of the student athletes. Information from UNC-Chapel Hill indicated that students were assessed during the summer session prior to Fall semester. Therefore, tests were assumed to be given during July. Students with birth months of July or earlier were treated as having had their birthday by the time of testing. Those with birth months from August to December were treated as not having had their birthday for that year yet.

There is considerable correspondence between the scores listed in the file and the standard scores I generated from the raw scores and the student age at testing information (as estimated by year/month of birth and year of entry into college). The match between the two indicates that the grade equivalent scores that were calculated internally were done so fairly accurately. In a minority of cases, there was not a perfect correspondence between the two. Most of the discrepancies can be attributed to age at testing because a single year of age can change the standard score. Assumptions about date of testing had to be inferred from year of entry to college as information about specific testing dates was available. A few of the cases (6) suggest recording errors on the part of the test score recorder. In these cases, the age equivalent score indicated a raw score that cannot produce the standard score reported in the data file (even at different ages). In these few cases, it appears that the raw score associated with the age equivalent score was recorded instead of the correct converted standard score.

Evaluation of Specific Data Claims for Reading Grade Level

I was asked to evaluate specific public claims made about grade reading level scores for 183 UNC student athletes including:

"..60% read between fourth- and eighth-grade levels." (CNN, 2014)

As already noted, the Reading Vocabulary score should not be taken as a reading level measure and the grade equivalent scores should not be interpreted. Similarly, 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.

Setting this aside, I attempted three methods of reproducing the above claim. First, the data file has 183 lines of data, but this includes a variable name line, dropping the total number of cases to 182. Of these, 176 have test scores. With these I examined:

1. The percent of sample with standard scores of 4-8 to see if standard scores had been incorrectly treated as grade equivalent scores.

2. The percent of the sample with actual grade equivalent scores between 3 to 8th grade.

3. The percent of the total sample with below average scores and the upper grade level associated with this cutoff.

Method 1. Approximately 35% (61 students) have standard scores between 4-8 and 65% (115 students) have scores between 9 and 16. Again, standard scores are not grade equivalents but if misinterpreted as such this would mislead us to saying 35% are 4th to 8th grade. If the scores are assumed to be raw scores (and not standard scores) and then converted to grade equivalents, the data still don't support an argument of 60% between 4-8th grade. Using the more complete data from two sources with 205 students, the proportions remain nearly the same with 36.5% with standard scores between 4-8 and 63.5% with scores between 9-16. This does not appear to be the source of the claim.

Method 2. To examine the claim based on the grade equivalent scores in the data, I calculated frequencies of grade equivalent scores for the full 205 students. There are 13 students with scores from "3rd grade" to "8th grade" which is 6% of the total sample. ***If we accept the grade equivalents as meaningful (which I do not believe) then 6% of the total group would be the most accurate figure for students with grade equivalent vocabulary between 3rd to 8th grade***. This use of the data, however, does not appear to be the source of the claim.

Method 3. Finally, by taking everyone who is at or below average, a standard score of 10 or less, I am able to produce a group that constitutes 59.5% of the sample, a match with the statement of 60% below average. It is true that this group is at or below the national average in vocabulary. However, a standard score of 10 for an 18 year old is associated with raw score of 14 (Table A-4). 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. This use of the data provides a match to the above claim of 60% below average but does not indicate that this group has vocabulary scores between a 4th to 8th grade level.

National Norms and Average SATA Scores for Student Athletes

I was asked to evaluate whether the composition of the national norm sample for the SATA presented a problem for reporting student athlete results. 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. Making comparisons with norm groups can be problematic in at least two ways.

First, if the norm group is poorly constructed, any inferences we might try to make could be misleading. This does not appear to hold in this case as the SATA norms appear to be adequate although based on somewhat small samples by age group (a comment echoed in Raju's critique of the SATA). The sampling plan was somewhat driven by convenience but a good effort was made to get a representative sample. It reasonably reflects the general population with respect to gender, geographic region, race and ethnicity.

The second way comparisons with a norm group can be problematic is entirely dependent on the comparison or question of interest. If the goal is to examine how the tested student athletes compare with the general population, then the SATA norm group is perfectly reasonable. The 50% standard score is set at 10 according to the test manual. Across all students in the data, as presented to me, the average standard score is 9.46. For just the football and men's basketball (n =166) the average is 9.78. So all of the student athletes who were tested are about a sixth standard deviation lower than the general population. This is not large but is still meaningful. If, however, the comparison of interest is how do these students compare to "student athletes at other universities" or "college students at major public universities", then the test norms would be inappropriate. The testing manual does present norms for a small sample of college students with learning disabilities with an average Reading Vocabulary score of 9.4, a slightly lower average than the student athlete sample from UNC-Chapel Hill. Overall, the normative data suggest that the student athletes in the sample are below average, but not dramatically, compared to the general population in reading vocabulary.

Placing Student Scores in Context

There is no doubt that academic preparedness as measured by verbal and math skill tests are important predictors of academic success in college and beyond (Kuncel, Hezlett, & Ones, 2004; Kuncel & Hezlett, 2007; Sackett, et al., 2012). At the same time, verbal and mathematical reasoning skills are not the only characteristics associated with student success in higher education. In addition, situational features and interventions can also improve academic performance and not only compensate for weaker academic preparedness but can even help develop those basic skills. Clearly additional education is something of long term value to students. Therefore, low scores, although a risk factor and a cause for concern, do not guarantee failure. Other information can and should be used to determine if a student is likely to be successful in college. Considerable research has demonstrated that several additional characteristics are important and there are multiple methods for assessing them. In other words, when looking for the best indication of a student's likelihood of success, multiple predictors of success in higher education should be considered. A brief review follows to illustrate the range of information that can be collected.

Effective assessment of a student's likelihood of success is complex but multiple tools can be used. Prior academic record as measured by high school GPA is consistently the single strongest predictor of future academic success (Sackett, et al., 2012). A student’s study habits, attitudes and skills (SHAS) are also especially strongly associated with academic performance (Crede & Kuncel, 2008) and can be taught through SHAS training. That is, students who are less academically prepared can benefit from being taught better study habits which will improve their likelihood of succeeding in college. Conscientiousness and drive are important predictors of student grades and are independent of verbal and mathematical reasoning scores (Poropat, 2009). In other words, a high level of drive and conscientiousness in students can help offset weaker basic math and reading skills. In fact, such a personality is actually associated with gains in learning or training environments (Campbell & Kuncel, 2001). Assessing these characteristics can come from letters of recommendation, interviews, personal statements, peer reports and self-report surveys. Letters of recommendation (Kuncel, Kochevar, & Ones, 2013) are not exceptional tools for predicting student success but do contain some useful information particularly about a student's drive and determination. Interviews can yield very valuable information when structured and ideally conducted by multiple interviewers (Huffcutt & Arthur, 1994). Personal statements are predictive of student grades but are typically weak predictors as normally implemented (Murphy, Klieger, Borneman, & Kuncel, 2009). Peer and other source reports (classmates, teammates, coaches) of a person's personality are solid predictors of academic performance (Connelly & Ones, 2010). Finally, as noted above, self-report of study habits and personality are predictive of academic performance (Crede & Kuncel, 2008; Poropat, 2009). In conclusion, student performance is complex and the ideal assessment includes many pieces of information combined with support of the student by the university as well as friends and family.

References

Campbell, J. P., & Kuncel, N. R. (2001). Individual and team training. In N. Anderson, D. S. Ones, H. Sinangil, & C. Viswesvaran, (Eds.), *Handbook of Industrial, Work, and Organizational Psychology* (pp. 287-. London, UK: Sage Publications.

CNN, 2014 retreived online at: http://www.cnn.com/2014/01/07/us/ncaa-athletes-reading-scores/

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092-1122.

Crede, M., & Kuncel, N. R. (2008). Study habits, study skills, and study attitudes: A meta-analysis of their relationship to academic performance among college students. *Perspectives on Psychological Science, 3*, 425-453.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classic and Modern Test Theory*. New York: Holt, Rinehart, & Winston.

Duckworth, A., L., Quinn, P. D., Lynam, D. R., Loeber, R., &, Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *PNAS, 108*, 7716-7720.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry level jobs. *Journal of Applied Psychology, 79*, 184-190.

Kirsch, I. S., et al. (1993). *Adult literacy in America: A first look at the results of the national adult literacy survey*. Washington, DC: National Center for Educational Statistics.

Kobrin, J. (2007). *Determining SAT benchmarks for college readiness*. College Board Research Note RN-30. New York: The College Board.

Kuncel, N. R. & Hezlett, S. A. (2007). Standardized tests predict graduate student's success. *Science*, *315*, 1080-1081.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology* [*Special Section, Cognitive Abilities: 100 Years after Spearman (1904)], 86*, 148-161.

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions. Reasons for hope. *International Journal of Selection and Assessment, 22*, 101-107.

Murphy, S. R. , Klieger, D. M., Borneman, M., & Kuncel, N. R. (2009). The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. *College and University, 84,* 83-88.

NCES (2006) National Assessment of Adult Literacy: A first look at the literacy of America's adults in the 21st Century. http://nces.ed.gov/NAAL/PDF/2006470.PDF

Poropat, A. E. (2009). A meta-analysis of the five factor model of personality and academic performance. *Psychological Bulletin, 135*, 322-338.

Raju, N. Review of the Scholastic Abilities Test for Adults. *Mental Measurement Yearbook*.

Sackett, P. R., Kuncel, N. R., Beatty, A. S., Shen, W., Rigdon, J., & Kiger, T. B. (2012). The Role of Socio-Economic Status in SAT-Grade Relationships and in College Admissions Decisions. *Psychological Science, 23*, 1000-1007.

Nathan R. Kuncel, Ph.D. is the Marvin D. Dunnette Distinguished Professor of Industrial-Organizational Psychology at the University of Minnesota where he also earned his doctorate in Industrial-Organizational Psychology. Prior to returning to the University of Minnesota he was faculty at the University of Illinois. Nathan's research focuses broadly on how individual characteristics (intelligence, personality, interests) influence subsequent academic, work, and life success as well as efforts to model and measure success. His research has appeared in *Science, Psychological Bulletin, Review of Educational Research, Perspectives on Psychological Science, Psychological Science*, among others. He edited the Industrial and Organizational section of the 3 volume *APA Handbook of Testing and Assessment in Psychology*. Nathan received both the Cattell Early Career Research Award from the Society of Multivariate Experimental Psychology and the Anne Anastasi Early Career Award from the American Psychological Association – Division 5, Evaluation, Measurement, and Statistics.