Analysis of Materials Regarding UNC Student Athletes

by

Lee Branum-Martin, Ph.D.

Report prepared for the office of the Provost of

The University of North Carolina at Chapel Hill

March 23, 2014

Overview of Findings

This report contains responses to the specified questions from the Office of the Provost. There were two Key questions and six Specific questions. Some overall findings of this report across the questions include the following:

1. The SATA is a rather dated test, especially when compared to other achievement tests for college-aged students. The SATA was not designed to provide accurate grade equivalent scores below grade ten. Data from the reading comprehension subtest of the SATA were not provided. Data from other tests of reading-related skills were also not provided, but would help to form a more complete profile of student literacy capabilities.
2. Standard scores alone are incomplete to fully evaluate the valid administration of a test. Standard scores alone on vocabulary and writing are also insufficient to evaluate claims of broad reading capabilities of students. Examinee ages and total scores should be provided, including item responses. The full records of the students' exam performance including item responses should be obtained, not only on the SATA, but other available tests of reading-related abilities (e.g., the SAT verbal). Based on the limited information provided, the spreadsheet appears to contain standard scores.
3. From student ages, reading vocabulary total scores were estimated for the UNC student sample, and grade equivalents derived from the SATA Examiner's Manual. Thirteen percent of the 176 vocabulary scores were estimated to be below the SATA reading vocabulary grade equivalent of ten. While SATA vocabulary grade equivalents are not dependable below grade ten, low levels of performance merit further investigation in order to ensure accuracy and appropriate services to those students.

Key Question 1:
Is the SATA Reading Vocabulary (RV) subtest a true reading test?

The framing of this question is dangerously simplistic. Investigating test quality and the nature of reading ability is a difficult endeavor, currently ongoing with leading scholars and much national investment (National Research Council, 2012). A more reasonable question would be:

Is the SATA Reading Vocabulary subtest reasonably valid for use among college athletes to determine their reading abilities?

The SATA was developed for use among adults in 1991. At the time of its development, it was well-designed and showed promise (Raju, 1995; Smith, 1995). However, the test does not appear to have been updated and it is possible that it does not adequately reflect contemporary educational and social influences. There are a wide range of contemporary tests designed for use among young adults which have larger bodies of validity evidence (e.g., Woodcock tests, Wide Range Achievement Test, Peabody Individual Achievement Test).

The SATA Reading Vocabulary subtest only measures word knowledge. While word knowledge is a crucial skill for reading (National Research Council, 2012; Phythian-Sence & Wagner, 2007), the SATA Reading Vocabulary subtest does not directly measure understanding main ideas or literal comprehension, as does the Reading Comprehension subtest of the SATA. It is therefore strange that a test of reading comprehension was not included in the provided spreadsheet.

Vocabulary is an important skill to measure for understanding students' capabilities with respect to reading, especially to distinguish problems in word knowledge from problems in comprehension. As reported in the SATA Examiner's Manual, the Reading Vocabulary subtest is highly related to the Reading Comprehension subtest, and both are related to an overall factor of Verbal proficiency. Ideally, vocabulary scores would be used in conjunction with comprehension scores to make a more complete portrayal of student skills related to literacy. Information from other reading-related tests would provide more evidence regarding the stability and level of student literacy.

The SATA is perhaps dated and lacks the ongoing validity evidence that other achievement tests have accumulated for college-aged students. The SATA Reading Vocabulary subtest was reasonably well developed as a test of vocabulary, but if we wish to understand student *reading* levels, additional information regarding reading and reading-related skills should be provided (National Research Council, 2012).

Key Question 2:
Did Mary Willingham mistakenly use RV standard scores instead of grade equivalents to report on UNC student-athletes' reading grade levels?

Answer:

The given scores are not sufficient to substantiate grade levels with respect to overall reading ability. Other information would be required.

The given two columns of scores cannot be validated without student total (raw) scores or item responses. SATA total scores are needed to determine grade equivalents and standard scores (as noted in the SATA Examiner's Manual), but total scores were not given in the data file. It seems plausible that the scores in the spreadsheet are standard scores. However, other information (ideally, the full battery of educational tests administered) would be needed to give a better portrait of student reading.

Rationale:

Based on the information provided, the spreadsheet appears to contain standard scores. In order to calculate grade equivalents, total (raw) scores are needed (Table D-1 in the SATA Examiner's Manual), but those did not appear in the spreadsheet. This is strange, because the normal procedure using the SATA would have been:

1. Record examinee item responses.
2. Total the number of correct items.
3. Using the total score, look up the standard score (age-based), grade equivalent, and other possible scores (e.g., percentile) for each examinee.

While grade equivalents are given in another sheet in the Excel workbook, total scores do not appear there, either. Therefore, the data cannot be independently verified.

The reason validation is difficult is that on the SATA, the three types of scores have very similar ranges (total, standard, and grade equivalent scores). Moreover, item responses were not provided. The RV and WM scores reported in the spreadsheet appear to be standard scores, because WM extends to 20 and SATA grade equivalents are bounded at 16.0. We have further evidence from the other spreadsheet (“data compared – Table 1”) which was labeled as standard scores. The two sets of RV values match numerically, but differ in patterns of missingness among the 211 observations. Where RV scores are reported in these two sources, they are identical, but some scores are missing from one column, and others are missing from the other column. It is unclear why there would be such a different pattern. A summary of the two sources of RV scores is shown in Table K.1. The means are not statistically different (*t* = 0.71; *p* = 0.48).

Table K.1: Descriptive statistics for the RV variables from the two columns, MW & LJ

RV Source	n	mean	SD	min-max
MW	176	9.6	2.5	5-16
LJ	179	9.4	2.5	4-16

As for the media statements, those could have been made from student data not reported here, since total scores are required to determine both standard scores and grade equivalents. The two columns of data in the spreadsheet do not provide enough information to be independently evaluated or to shed much light on the media statements, largely because the SATA has scores which are all on very similar metrics: item totals, standard scores, and grade equivalents. It is recommended that UNC build a full database of all athlete test scores, including item responses, in order to ensure appropriate allocation of services to student skills as well as to evaluate the quality of the tests being used.

Specific Question 1:
Can you verify that the Reading Vocabulary (RV) data in Mary Willingham's spreadsheet is in the form of standard scores?

<u>Answer:</u>

The RV scores do appear to be standard scores, but this cannot be verified definitively without item responses or total (raw) scores, along with student ages at time of testing. The standard scores from elsewhere in the spreadsheet numerically match the given RV scores, but differ in patterns of missingness. It would therefore seem plausible, but not certain, that the given RV scores are standard scores.

<u>Rationale:</u>

As noted in the SATA Examiner’s Manual and as is typical of educational tests, item responses are recorded (25 items on the RV subtest), a total score is calculated, and then age-based standard scores can be looked up. Grade equivalents can also be looked up, using the total scores. The spreadsheet was missing item responses, total scores, and examinee age.

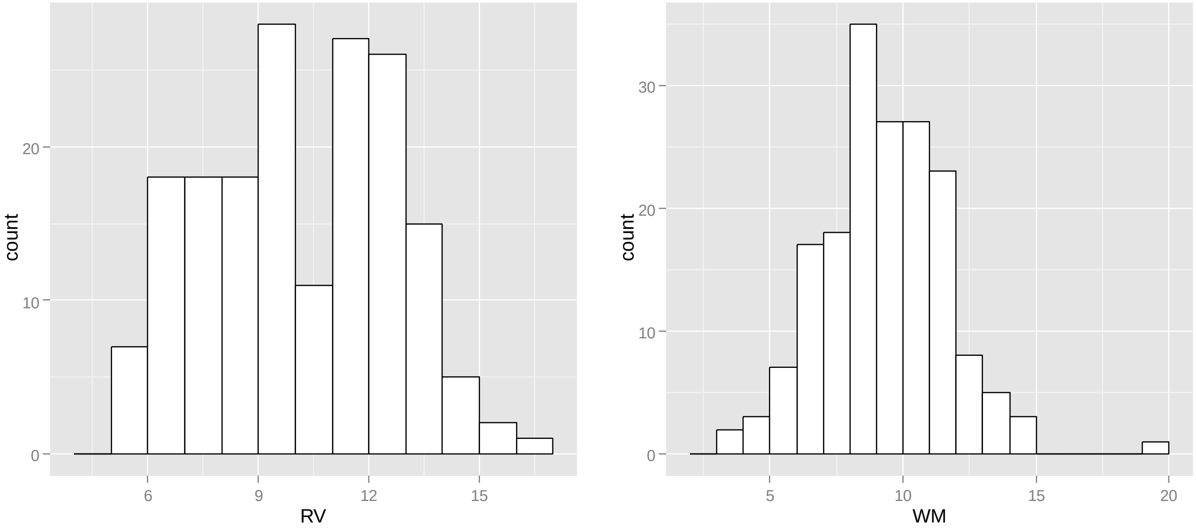
Table 1.1: Descriptive statistics for the RV and WM variables

Variable	n	mean	SD	min-max
RV	176	9.63	2.54	5-16
WM	176	8.80	2.34	3-19

The spreadsheet column labeled “RV” falls within the ranges given by the SATA Examiner’s Manual for standard scores (mean = 10; SD = 3; range = 1 to 20) as well as for grade equivalents (range = 3-16; Appendix D) and total scores (range = 0 to 25). However, because the Writing Mechanics scores ranged from 3 to 19, it could be assumed that the RV scores also have this range and are therefore most likely standard scores. The RV scores elsewhere in the Excel workbook, labeled as standard scores, seem to support that the given RV scores are standard scores.

The summary statistics in Table 1.1 above do not adequately capture the shape of the distributions. Figure 1.1 below shows histograms for the RV (left) and WM (right) variables. The vertical axis in each graph shows the number of students who received that score. The horizontal axis shows the level of the score. Both variables fall within the acceptable range of SATA standard scores. One observation on the WM measure falls outside the upper grade equivalent of 16.0. Assuming this one observation is correct, the WM scores should be standard scores, not grade equivalents.

Figure 1.1: Histograms for the provided RV and WM variables



The scores in the spreadsheet appear to standard scores, but this cannot be validated without total scores (and preferably with item responses). It is standard educational testing practice, reinforced by IRB regulations, that such record forms are retained in a secure manner. These records should be obtained.

Specific Question 2:
Can you verify that treating standard scores as grade equivalents is a serious error in data analysis?

Answer:

Generally, yes, treating standard scores as grade equivalents would be a serious error. On the SATA, however, total scores, standard scores, and grade equivalents all have similar numerical values. More importantly, however, the SATA has explicit warnings about the limitations of its grade equivalent scores below the tenth grade.

Rationale:

For many tests, the metric of standard scores (e.g., mean = 100, SD = 15) is obviously different from grade equivalents. However, the metric of standard scores for the SATA (from 1 to 20) is quite similar to grade equivalents (from 3 to 16). For the SATA, confusing the standard scores for grade equivalents could create problems, but they might be small and difficult to determine. If total scores were provided, the meaning of the given scores would be easy to validate. If item responses were provided, the psychometric functioning of the SATA as a valid test for this sample could also be investigated (Hambleton, Swaminathan, & Rogers, 1991; McDonald, 1999).

There are two important issues with respect to this Question. First, the grade equivalents provided by the SATA are limited. Second, with examination dates provided, we can calculate examinee ages, estimate total scores (Table A-4 in the Examiner's Manual) and then determine grade equivalents (Table D-1 in the Examiner's Manual), even though they were not provided in the original data set. Each of these issues will be detailed next.

Grade levels in the SATA norm sample. It is important to note an explicit limitation of SATA grade equivalents:

> "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. Thus, these scores must be interpreted with caution. We strongly urge SATA users to base their test interpretations on the standard scores that are reported."
>
> (SATA Examiner's Manual, p.19)

Because no students below grade ten were tested, grade equivalents from the SATA below ten are extrapolations of unknown validity. Moreover, judging reading ability and learning disabilities especially is a difficult task, requiring multiple forms of evidence (Fletcher, Lyon, Fuchs, & Barnes, 2007). Two tests administered at a single time point would provide only limited information regarding reading capabilities (especially without a measure of reading comprehension), and would not provide rigorous evidence of learning disability without other sources of information.

Estimating total scores to get grade equivalents. The spreadsheet was missing item responses, total scores, and examinee age. Year of testing was provided in a separate file (Debbi Clarke, personal communication), which enabled approximate ages to be calculated. With examinee ages, observed scores were reverse-looked up in Table A-4 in the SATA Examiner's Manual. Where total scores in Table A-4 included ranges of 2-3 possible total scores, the mean total score in that age range was given to that student. Using this method, total scores would be expected to deviate from actual scores by up to one point. Grade equivalents based on this method would be expected to be accurate within about a single grade, based on Table D-1 in the SATA Examiner's Manual.

Figure 2.1: Grade equivalents based on total scores calculated from examinee age

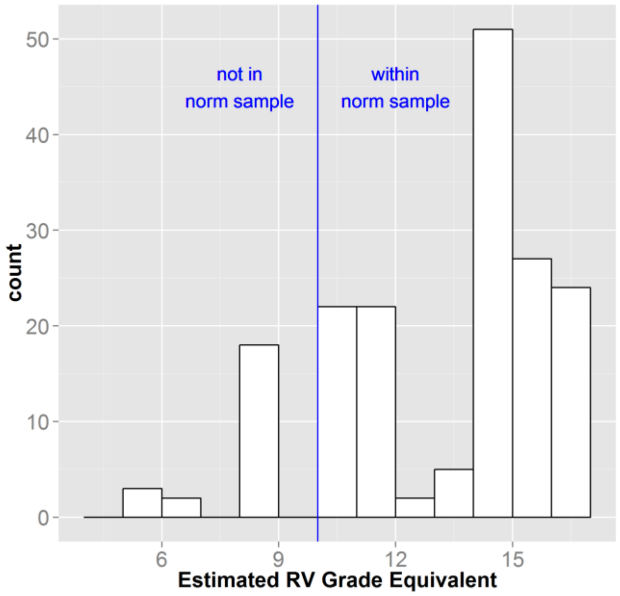


Figure 2.1 shows a histogram of the number of students who scored at each estimated grade equivalent for the reading vocabulary subtest. The blue line in Figure 2.1 divides the grade equivalents into grade levels observed in the SATA norm sample above grade ten, and grade levels not tested to the left, below grade ten. Thus the left-hand portion of grade equivalents in Figure 2.1 represents extrapolations, not observed in the norm sample examinees. Therefore, it would seem questionable to rely on the

SATA for grade equivalence judgments. Ideally, such judgments should be corroborated by other sources, such as other tests and performance assessments relevant to the grade levels for which the students should be evaluated (such as relevance to college-level reading and coursework).

Figure 2.1 shows that of the 176 student observations with valid RV scores, 23 students (13%) would have received a RV grade equivalent below grade 10. As stated in the SATA Examiner's Manual on p19, these scores are extrapolations. 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.

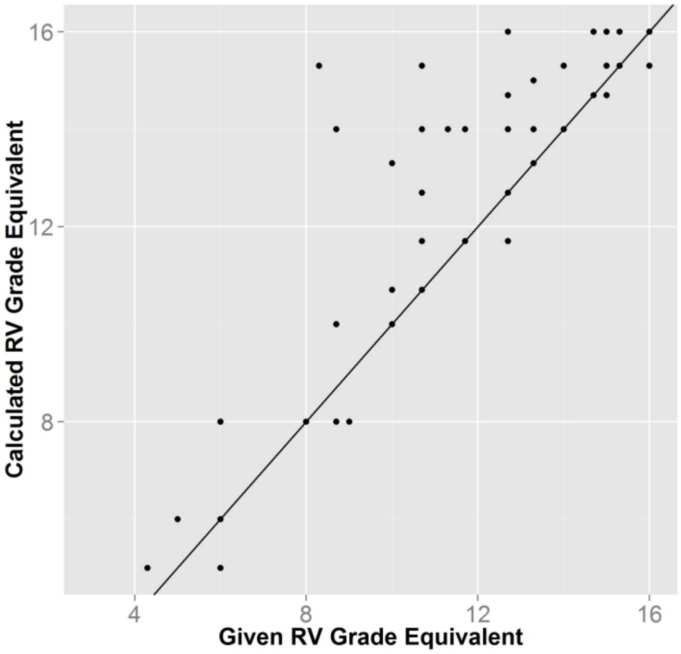
Sixty-seven of the students (38%) scored below the 12th grade level. This level of performance would suggest that that these students may need additional assessment to better understand their complete reading skills and their capacities for college-level academic reading.

Note on the Grade Equivalent Calculations. One of the spreadsheets given had grade equivalents for the RV scores. Because the current procedure involved a reverse-lookup in Tables A-4 and D-1, and the tables sometimes give a range of scores (e.g., 13-15), there is a small amount of imprecision compared to grade equivalents that would have resulted from actual total scores. However, these scores can be compared to evaluate their precision, as shown in Figure 2.3.

Figure 2.3 shows a scatterplot of the given RV grade equivalents (horizontal) against the grade equivalents calculated for the current report (vertical). The diagonal line represents equality, where grade equivalents would be identical. Ideally, all points should lie on or very close to the diagonal line, indicating that grade equivalents were equal to those calculated in the present report.

Because the Examiner's Manual Tables A-4 and D-1 frequently reference a range of scores (e.g., a total score of 14-16), some points could be slightly above the line of equality (i.e. over-estimated by the current procedure), or slightly below (i.e., under-estimated). A small amount of scatter (e.g., 1-2 grade levels) would be expected, given the small amount of uncertainty in the ranges of the table.

Figure 2.3: Calculated versus given grade equivalents for Reading Vocabulary scores



However, Figure 2.3 shows more points above the diagonal line than below. These values were verified to include errors, suggesting that some RV grade equivalents assigned to students in the given spreadsheet were lower than what they should have been, if the lookup rules had been followed. While such mistakes cannot be definitively proven without actual total scores, the calculated points in this report lie substantially above the line of equality in Figure 2.3. This implies that if the given grade equivalents in the spreadsheet were previously used, several grade equivalents were too low. On average, the discrepancy was .69 grade equivalent units lower for the given grade equivalents than for the grade equivalents calculated in the current report.

Specific Question 3:
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?

Yes, such an assignment would be possible, but should include appropriate caveats regarding unreliability of tests (Fletcher et al., 2007), standard errors of measurement (Smith, 1995), and a discussion of the quality and amount of different pieces of information available. Having other data such as SAT verbal scores could provide helpful information in estimating students' practical levels of performance. However, such scores were not provided in the given spreadsheet. One-on-one work with students involves professional opinion, but could hopefully be corroborated with a variety of objectively scored tests, performance assessments, and supportive documentation from professionals familiar with the students (e.g., teachers, tutors).

Putting reading performance on a practical scale for non-researchers is a challenge. Grade equivalents are rather coarse (e.g., not all tenth graders read equivalently). Consequently, a range of tests are usually needed. While all such evaluations involve a certain degree of professional judgment, the reporting of results should stand on the amount and quality of evidence, rather than on testimony and reputation. Reports and underlying documentation should be open to review by the scientific and educational community.

Specific Question 4:
If yes, what level of professional expertise would be needed to do so?

It is possible that grade levels could be assigned reasonably well by a person with Master's level training in assessment. The SATA Examiner's Manual recommends that examiners have "some formal training in assessment" as well as in statistics and test administration. Such topics are typically covered in Master's and doctoral level courses in educational psychology or educational measurement.

However, a scientific record of examinee performance should stand on the evidence provided, and less on personal qualifications or expert opinion. The given spreadsheet only had scores for vocabulary and writing. Reading comprehension scores were not provided. The full records of the SATA items and scores and other reading measures should be obtained for the sake of serving the student athletes and for the evaluation of the programs which serve those athletes.

Specific Question 5:
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?

The differences between the SATA norm sample and the UNC sample are potentially problematic and suggest caution in interpretation. There are two major issues of concern: population differences over time and demographic differences. Each of these two issues will be discussed in turn.

Differences over time. The SATA test was normed in 1991 on a sample of 1,005 examinees with the demographics listed in Table 5.1. Table 5.1 shows the percent of students in each category, both for the SATA norm sample as well as the UNC sample.

Table 5.1: Demographics of the SATA norm sample vs. the UNC sample (percent)

Group	Male	Female	White	Black	Other
SATA norm (n = 1,005)	46	54	86	8	6
UNC athletes (n = 211)	86	14	24	59	17

The UNC data were collected in 2004 to 2012, 13 to 21 years after the SATA was normed. Because populations change due to shifts in educational and societal influences, test norms typically require updating. The SATA does not appear to have been updated by the publisher since 1991. It is not clear how valid its normative age-based standard scores would be for today’s students.

Demographic differences. 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. While the RV test was carefully designed (Raju, 1995; Smith, 1995), the item content could be evaluated for the extent to which it might provoke test-irrelevant responses, especially for speakers of African American dialect. Also, item responses could be checked for differential functioning across groups, such as men versus women, and white versus black students. The SATA does not appear to have been subjected to tests of such differences (Raju, 1995). Consequently, there is no evidence to suggest that scores on the RV test are equally valid for a group which is so different from the norm sample (there is also no direct evidence to suggest that the items on the RV test are biased in any particular way).

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.

Specific Question 6:
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?

Literacy. 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 in the spreadsheet.

Table 4.12 in the SATA Examiner's Manual suggests that the reading vocabulary and reading comprehension subtests are both highly related to a general verbal ability (i.e., a latent statistical factor which was distinguished from other subtests which were related to a general quantitative ability in a two-factor solution). Conceptually and empirically, vocabulary is highly related to reading comprehension.

Typically, when people talk about literacy, they are most concerned with reading comprehension, or more broadly, students' 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, especially 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.

Grade levels. Grade equivalents represent an attempt to put a practical label on potentially complicated scoring of tests. Grade equivalents are coarse, and care must be taken that the levels assigned correspond to clear standards of text difficulty and tasks in which students are expected to be proficient. There are efforts to link reading test performance to clearly defined difficulty features of texts (e.g., the Lexile framework of MetaMetrics). Frequently, grade equivalents in tests do not have such rigorous standards associated with them.

Great care should be taken when reporting results in terms of grade level. The SATA tests are clearly limited in how accurate their grade equivalents should be taken, especially below the tenth grade. If accurate measurement of vocabulary and reading skills are desired, especially for those who may struggle at the college level and need assistance with skills which may be more common below the tenth grade, a different set of vocabulary and reading tests should be chosen which have rigorous evidence to cover this range of performance.

References

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: From identification to intervention*. New York: Guilford.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

National Research Council. (2012). Improving adult literacy instruction: Options for practice and research. In A. M. Lesgold & M. Welch-Ross (Eds.), *Committee on Learning Sciences: Foundations and Applications to Adolescent and Adult Literacy*. Washington, DC: The National Academies Press.

Phythian-Sence, C., & Wagner, R. K. (2007). Vocabulary acquisition: A Primer. In R. K. Wagner, A. E. Muse & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 1-14). New York: Guilford Press.

Raju, N. S. (1995). [Review of the Scholastic Abilities Test for Adults]. *The twelfth mental measurements yearbook* (Vol. 12:344). Lincoln, NE: Buros Institute of Mental Measurements.

Smith, D. K. (1995). [Review of the Scholastic Abilities Test for Adults]. *The twelfth mental measurements yearbook* (Vol. 12:344). Lincoln, NE: Buros Institute of Mental Measurements.

About the author

Lee Branum-Martin is an Associate Professor in the Department of Psychology at Georgia State University. He received his Ph.D. in Educational Psychology at the University of Houston in 2004. His research focuses on measurement issues in language and literacy. His work and qualifications can be found on Google Scholar, ResearchGate.net, and Academia.edu.