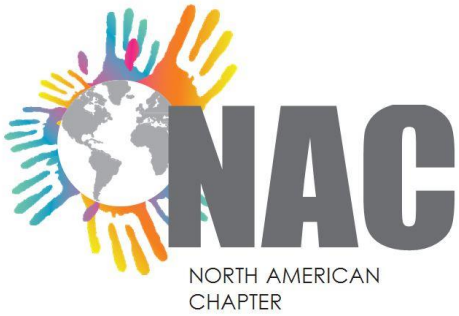
A Diagnostic Tool to Predict Performance on High-Stakes Multiple Choice Tests: An Attempt to Recognize Potential Discrepancies Due to Diversity Before the Test is Taken

This manuscript has been peer-reviewed, accepted, and endorsed by the North American.



North American Chapter
Journal of Interdisciplinary Education

Natalie N. Michaels, PT, Ed.D.
Belmont University

Ronald Barredo PT, Ed.D.
Edilberto Raynes, MD, Ph.D.
Deborah Edmondson, PT, Ed.D.
Elizabeth Kunnu, Ed.S., M.Ed., RHIA
Tennessee State University

Journal of Interdisciplinary Education, Vol. 15, No. 1 – January 2016
ISSN: 1092-7069 © 2016 North American Chapter

ABSTRACT

High-stakes testing can be detrimental for certain students who are confident in their knowledge of the test content, but simply cannot pass the examination. Detecting a reason for the difficulty could help these students develop strategies to overcome this hurdle. One potential reason for this difficulty is a misunderstanding of the test question that occurs when the wording of the question is different from the language variation typically used by the test-taker. (This research builds on prior research published by WCCI in 2011.)

A diagnostic tool was created to help determine if problems existed when test questions were reworded in accordance with Southern Caucasian and Southern African American language variations. When this language variation tool (LVT) was utilized in 2011 by these researchers with a group of doctoral physical therapy students in Nashville, TN, differences were found in student test scores when questions were reworded; but these differences were more evident in the African American student group when compared to the Caucasian group (with an intercept significance of p=.000).

The current study conducted in 2014, and the focus of this paper, demonstrated that there is predictive value to the LVT. Students who received a lower overall score on the diagnostic test also demonstrated lower scores on the practice board examination (Pearson correlation=.662; p=.000). This correlation was stronger than that of the grades from practice questions as originally worded, without the language variation component (correlation=.392; p=.029).

When there is a misunderstanding of the test question because wording of the question is different from the language variation typically used by the test-taker, there can be increased difficulty passing standardized tests. This study supports the utilization of this diagnostic tool (created to specifically match that of the high stakes testing instrument in question) to predict success on the high stakes examination. It is the opinion of the researchers that informing a test-taker of his or her limitations prior to an exam could potentially improve test performance, since this awareness gives the test-taker knowledge of how to better prepare for the exam. This contention would require further research.

Key Words: language variation, multiple choice testing, diversity

High-stakes testing can be a challenge for students who are confident in their knowledge of the test content, but simply cannot pass the examination. Detecting a reason for the difficulty could help these students develop strategies to overcome this hurdle. One potential reason for this difficulty is a misunderstanding of the test question; that happens when the wording of the question is different from the language variation typically used by the test-taker. This research builds on work that was previously published in the Journal of Interdisciplinary Education in 2011 (Housel et al., 2011).

The Problem

The faculty in the Doctor of Physical Therapy (DPT) program at Tennessee State University (TSU) observed that many students, who demonstrated advanced knowledge in course content, reported difficulty passing multiple choice examinations. It was also evident that the African American students were more likely to report this difficulty when compared to their Caucasian peers. Detecting a reason for this problem became paramount. It is well documented in the literature that the grades for multiple choice high-stakes tests are, on average, higher for Caucasian students than African American Students in the United States (Berlak, 2001; Davis et al., 2013; Johnson, Boyden & Pittz, 2001). In a study that was based on data from 28 states between 2005 and 2009, Nettles, Scatton, Steinberg, and Tyler (2011) found that Caucasian students scored higher than African American students on various standardized tests, with the Law School Admission Test (LSAT) and analytical components of the Graduate Record Examinations (GREs) showing the largest differential. The mean scores on the August, 2011 to June, 2014 Graduate Record Examinations (GREs) reported by the Educational Testing Service (ETS, 2014) were markedly lower for African American test-takers (Verbal Reasoning – 147.0, Quantitative Reasoning – 143.7, Analytical Writing – 3.3), than their Caucasian peers (Verbal Reasoning – 154.0, Quantitative Reasoning – 150.8, Analytical Writing – 3.9). The literature

supports the contention that there is an under-representation of minorities in the health care professions in the United States (Hamel et al., 2015; Nettles et al., 2011; Weintraub, 2015). Inability to pass the GREs, Medical College Admission Tests (MCATs), and standardized board examinations have been cited as potential reasons for this discrepancy (Davis et al., 2013). During this DPT program, students take many multiple choice examinations, but one of the major challenges is passing the test that comes at the end of the didactic curriculum. This is the board examination, also known as the National Physical Therapy Examination (NPTE). Utzman et al. (2007) found that African Americans were more than 200% likely to fail the NPTE when compared to their Caucasian counterparts. This information was cause for concern, especially in an educational environment that embraces diversity.

Language Variation

Although a multitude of interacting variables could lead to discrepancies in test scores, this research focused on English language variation of multiple choice questions. It is no surprise that language and communication skills have been cited as major contributors to student attrition (Bosher & Smalkoski, 2002; Bosher, 2003). A test comprised of fairly wordy items may actually begin to measure reading comprehension, as opposed to the subject of focus (Haladyna, 2004; Schellenberg, 2004). Although it has been documented in the literature that the level of cultural competency utilized when creating multiple choice testing instruments may affect student performance (Moule, 2005; Tellegen & Laros, 2004), a student's cultural background could also influence his/her ability to understand and comprehend test items, which may positively or negatively affect the student's performance. A student taking a test that is written in a way that is different from his or her own cultural proclivities might have difficulty simply understanding the questions.

The 2011 Study

In 2007, Housel and associates developed a diagnostic multiple choice testing instrument called the language variation tool (LVT), to help determine whether or not a student had this problem. The LVT was used as a pilot for the TSU DPT graduating classes of 2008, 2009, and 2010, and the results were published in a 2011 article (Housel, 2011). Although it was found that there were some Caucasian students who had difficulty answering questions when they were re-worded, the percentage of African American students with this problem was higher (with an intercept of p=.000). The purpose of this current study was to determine the actual predictive effectiveness of the LVT for future high stakes testing, specifically the NPTE. Butler and Roediger (2008) found that the provision of feedback after multiple choice testing led to a higher level of learning by increasing the proportion of correct reposes on later testing. Once success can be predicted, assistance can be provided to those who score lower by providing this valuable feedback, then determining strategies for potential success.

The Tool

The LVT was created utilizing the strategy outlined in *Rehabilitation Research: Principles and Applications, 3rd ed*. (Domholdt, 2004; with permission from the author). The steps included drafting, expert review, first revision, a pilot test, and a final revision. The creation of the LVT is quickly outlined here. More specific information on the creation of this tool was outlined in the 2011 article so that others could create a similar tool pertaining to their specific area of study. These tools could then be used as diagnostic instruments to detect students with problems with language variation on multiple choice testing.

Drafting

Test questions were obtained from the 2007 Scorebuilders preparatory guide for the NPTE (Giles, 2007; with permission from the author). Six students (3 Caucasian and 3 African American), volunteered to re-write questions as though they were talking to their peers. The reworded questions were then incorporated into a test of 60 questions where 30 were written as originally worded by Giles, 15 were re-worded in a southern Caucasian language variant, and the other 15 were re-worded as a southern African American language variant. Two versions of the test were created (Test A and Test B), with the second test switching the questions around so that the 30 originally worded questions were re-worded on the second test, and the other 30 questions changed back to the way they were originally worded by Giles. It was decided to mix the questions on two testing instruments to decrease the chance of the test-retest phenomenon.

Expert Review and First Revision

The re-written questions were then reviewed by the faculty research team to insure that the meaning of each question remained unchanged. The questions were then sent to two expert linguists for review. Walt Wolfram, PhD, is a William C. Friday Distinguished Professor of English Linguistics, North Carolina State University, Director of the North Carolina Language and Life Project, co-creator of the Interinstitutional Cooperative PhD English Linguistics Program at Duke University and author of *American English: Dialects and Variation* (Wolfram,1998). Lisa Green, PhD, is a Professor at the University of Massachusetts Amherst, and author of *African American English, A Linguistic Introduction* (Green, 2002). These experts reviewed the questions to insure that the items appropriately reflected the language variations.

After the questions were modified according to the linguist recommendations to more accurately reflect the language variation, the items were then sent back to the faculty to insure that the meaning of the questions remained unchanged. Each question was then checked by the

faculty for content validity by answering the following questions: (1) Were the major elements addressed? (2) Were the questions understandable within the limits of the dialect? (3) Were the terms defined satisfactorily? (4) Were the questions formatted appropriately? As a result of this review, some of the questions needed to be modified slightly. They were then returned to the expert reviewers to insure that the language variation remained unchanged.

Pilot Test and Final Revision

Approval for the pilot studies was granted by the TSU Institutional Review Board (IRB) in September, 2008. The pilot studies were conducted at TSU with the physical therapy graduating class of 2008 (*N=17*), the class of 2009 (*N=19*), and the class of 2010 (*N=22*). The LVT was administered to these students and group data were utilized. Both versions of the test were given to each student on two separate days. Half the students took Test A on day one through random assignment, and half took Test B, then alternated on day two. Differences were found in many student test scores for both Caucasian and African American students when questions were reworded, but the differences were more evident in the African American student group when compared to the Caucasian group (with an intercept of p=.000). Caucasians scored higher for everything except the African American worded questions (Table 1).

Table 1

Group data showing average results of two days of testing. The red circle depicts the area where the African American students scored higher than the Caucasian students.

	Caucasian	African American
Day One %	60.90	58.57
Day Two %	62.56	60.23
Total Score %	61.59	59.39
Caucasian % Score	59.45	56.19
African American % Score	57.23	64.27
Originally Written % Score	62.71	58.59

Students were debriefed after the pilot studies. Student recommendations were incorporated the tool; these consisted of formatting and spacing of questions for the final draft.

METHOD

IRB approval for the current study was granted by the TSU IRB on August 5th, 2013. The study was conducted in early 2014.

Sample

The sample consisted of 35 students from the TSU DPT graduating Class of 2014. Of these students, there were 3 African American students and 32 Caucasian students; 21 males and 14 females. Of the 35 students, only 31 had completed the Practice Examination and Assessment Tool (PEAT) during the time of this study. Of the remaining 31 students, there were 2 African American students, and 29 Caucasian students; 17 males and 14 females.

Instrumentation

The Practice Examination and Assessment Tool (PEAT) is a practice examination created by the Federation of State Boards of Physical Therapy (FSBPT, 2015). The Federation is the same entity that created the NPTE. The language of the PEAT is very similar to the language of the NPTE. The PEAT was found to correlate highly with the NPTE results by Barredo, Tan, and Raynes (2015, preliminary result). Because the scores from the PEAT were more convenient to access than the scores for the NPTE, and also because these researchers wanted to know how students would perform before the actual high stakes test, the PEAT was used in this study. These researchers wanted to know if student scores on the LVT would predict student scores on the PEAT.

Procedure

Student randomly drew a paper from a container to see if they were taking Test A or Test B. Students then took the LVT in a classroom setting with the timer set to 75 minutes, allowing for the same amount of time per question as the actual board examination. This test was given on two separate days but only the results of the first administration was used as a predictor for success on the PEAT, which was taken approximately three weeks later.

Results

The average score for the LVT was 66.5%. The average score for the PEAT was 68.7%. Students who received a lower overall score on the diagnostic test also demonstrated lower scores on the practice board examination (Pearson correlation=.662; p=.000). The correlation of the 60 questions taken together with the PEAT was stronger than that of the grades from the 30 original practice questions taken alone, without the language variation questions (Pearson correlation=.392; p=.029). The difference between these two correlations was found to be significant at the .05 level (p = .0216), utilizing methods suggested by Lea and Preacher (2013).

The correlation of the 60 questions taken together with the PEAT was also higher than the 30 re-written questions taken alone (Table 2).

Table 2

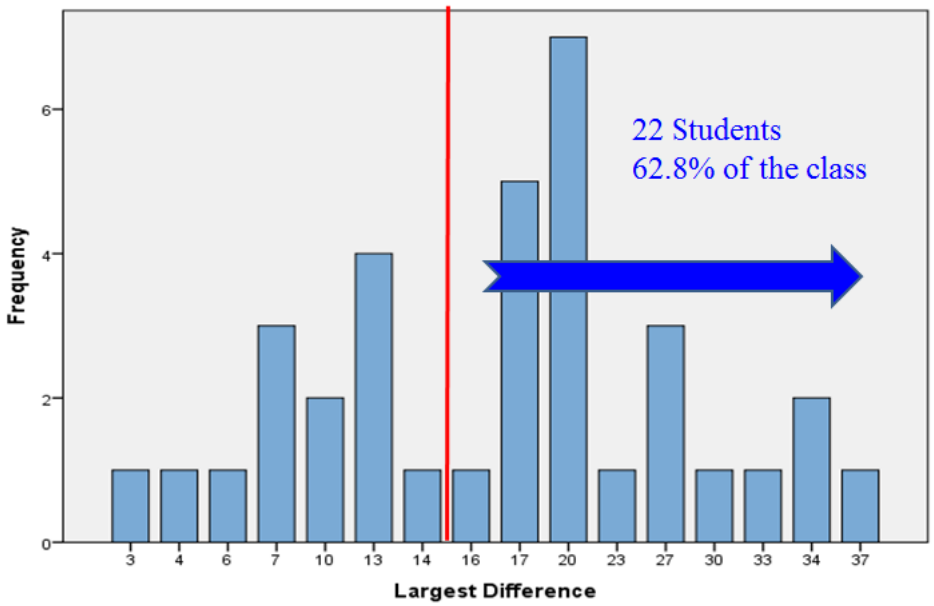
Results of Bivariate Correlations.

		First Original Grade	Southern AA worded Grade	Southern Caucasian worded Grade	Complete Language Variation Test	PEAT
	N	35	35	35	35	31
Complete LVT	Pearson Correlation	.646**	.646**	.789**	1	.662**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	35	35	35	35	31
PEAT#1	Pearson Correlation	.392*	.436*	.485**	.662**	1
	N	31	31	31	31	31
	Sig. (2-tailed) *p* value	.029	.014	.006	.000	

The cut-off for potential problems passing the board examination appeared to be when students demonstrated a 15% or higher score difference between questions as originally worded and questions that were re-worded. Of the 35 students who took the LVT, 22 scored a 15% difference or higher. This was 62.8% of the class (Figure 1).

Figure 1

The largest difference in scores when questions were re-worded.



Discussion and Conclusion

The LVT created specifically for DPT students illustrated in this paper appears to be moderately predictive of the PEAT test taken at a later date. Since the PEAT has been shown to predict performance in the NPTE by Barredo et al. (2015; preliminary result), it can be assumed that this tool might also be relatively predictive of the actual NPTE results. It could therefore be anticipated that if the students in this study did not do well on the LVT, they probably would not do well on the board examination, unless some form of intervention preceded that high stakes examination.

The average score for the LVT for the students in this study was 66.5%. The average score for the PEAT was 68.7%. Both averages are below the 70% passing score. Only 29% of the students in this study passed the PEAT. So, technically, many of the students from the Class of

2014 should have failed this examination, but this was not the case. Of the 35 students who later took the NPTE, all passed except one. This was a 97% pass rate. The reason that the students on average did not pass the LVT but did pass the NPTE was assumed to be due to the feedback provided, supporting the previously mentioned work of Butler and Roediger (2008). The researchers encouraged the students who had a problem with language variation to study both the content of the test and the language of the test. It is apparent that this technique helped a majority of the students to ultimately pass the NPTE. In some cases, it was almost as though the wording of the question was more important than the content of the question. If this contention is true, diagnosing this problem early could provide students with information needed to redirect their studying strategies when attempting to pass high stakes tests. This could provide them with valuable metacognitive information before the test is taken.

The researchers strongly support the usage of study guides like the one created by Scott Giles at Scorebuilders. The questions created by Giles (2007) and his team not only help students prepare for the content of the examination, but they also enable students to practice working through the rigors of multiple choice test taking. It was intriguing that the predictive value of the Scorebuilders questions increased when the reworded questions were added to the mix, and even more intriguing that the predictive value of the reworded questions was higher when combined with the Giles questions as originally worded. It appears that both the elements of content and language variation are needed to adequately predict success on the PEAT.

Limitations

The primary limitation of this study is that only 3 of the 35 students in the graduating class of 2014 study were African American, and only 2 of the remaining 31 who took the PEAT examination were African American. Although TSU is an Historically Black College/University (HBCU), the percentage of students in the program who are Caucasian has been gradually increasing. One reason might be because a component of admission has been GRE results, another high stakes test. This possibility is currently being explored. Another limitation of this study is the focus solely on language variation. Many other variables affect the outcome of high-stakes testing, from sociocultural backgrounds to the health of the student taking the test. Although language variation is clearly one of those variables, it is not the only determining variable.

Implications and Recommendations for Further Research

Over the past ten years, the TSU physical therapy department has seen an improvement in the pass rate on the NPTE from 50.0% to 97%. Although there have been many changes that could have assisted in this improvement (change in curriculum, hiring of more full-time faculty, changes in our admission standards, critical thinking workshops, etc.), many students have credited the use of the LVT and the feedback of the results as a strong reason for their ultimate success on the NPTE.

Utilizing the strategies suggested by Domholdt (2004) as outlined in this paper, it is the belief of these researchers that the LVT can be created for any subject area or professional degree program where high stakes testing is an issue. It is time consuming to create these tools, but they can provide valuable group data and enable individual students with the provision of more information about their own testing strategies. With this information, the students improve

their own metacognitive processes and create more effective studying strategies prior to taking a high stakes examination. Once the element of confusion related to language variation is removed, the test really will be testing knowledge of the content, and not reading comprehension, giving the results more validity. This could empower more students from various backgrounds to pass these examinations and truly create an educational environment that embraces diversity.

With the creation of more of these tools, this study could be replicated for other professions, enabling students who could not pass otherwise to be successful when taking their board examinations. This study could also be replicated in other countries as a way to predict student success on a high stakes tests, a way to improve the ultimate results on these tests, or as a way to increase the representation of diversity in the field in question. Only by focusing on the reasons for test taking difficulty and then determining strategies for success, can we create an educational culture of inclusion and equity.

Acknowledgements

This research team would like to thank the Tennessee State University (TSU) Doctoral Physical Therapy (DPT) Students from the Graduating Class of 2014, and Scott Giles for allowing us to utilize questions from Scorebuilders. We would also like to thank Walter Wolfram and Lisa Green for their expert review on the development of the LVT.

References

Barredo, R., Tan, J., and Raynes, E. (2015). GPA, PEAT scores, and the NPTE: A pilot Study. Abstract Submission 2017 CSM. American Physical Therapy Association. Available through Ron Barredo.

Berlak, H. (2001). Race and the achievement gap. Rethinking Schools On-line. Retrieved: http://www.rethinkingschools.org/archive/15_04/Race154.shtml

Betancourt, J.R. (2002). *Cultural Competence in Health Care: Emerging Frameworks and Practical Approaches*. Field Report for the Commonwealth Fund. Available: http://www.commonwealthfund.org

Bosher, S. (2003). Barriers to creating a more culturally diverse nursing profession: Linguistic bias in multiple-choice nursing exams. *Nursing Education Perspectives, 24*(1) 25-34.

Bosher, S., Smalkoski, K. (2002). From needs analysis to curriculum development: Designing a course in health-care for immigrant students in the USA. *English for Specific Purposes, 21*(1)59-79.Boukus, E.R., Cassil, A., O'Malley, A.S. (2008). A Snapshot of U.S. Physicians. Key findings from the 2008 Health Tracking Physician Survey. Center for Studying Health System Change. Retrieved. http://www.hschange.com/CONTENT/1078/

Butler, A.C., Roediger, H.L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616.

Davis, D., Dorsey, J.K., Franks, R.D., Sackett, P.R., Searcy, C.A., Zhao, X. (2013). Do Racial and Ethnic Group Differences in Performance on the MCAT Exam Reflect Test Bias? *Academic Medicine, 88*(5): 593-602.

Domholdt, E. (2004). *Rehabilitation Research: Principles and Applications, 3rd ed*. Saint Louis, Missouri: Elsevier Saunders.

ETS (2014). A snapshot of the individuals who took the GRE® Revised General Test: August, 2011-June, 2014. *Educational Testing Service.* Retrieved: http://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2014.pdf

FSBPT (2015). The Practice exam & assessment Tool (PEAT). Retrieved: https://www.fsbpt.org/OurServices/CandidateServices/PracticeExamAssessmentTool(PEAT).aspx

Giles, S.M. (2007). *PT EXAM: The Complete Study Guide*. Scarborough, Maine: Scorebuilders.

Green, L. J. (2002). *African American English: A Linguistic Introduction*. Cambridge, United Kingdom: Cambridge University Press.

Haladyna, T.M. (2004). *Developing and Validating Multiple-Choice Test Items (3rd. Ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Hamel, L.M., Chapman, R., Malloy, M., Eggly, S., Penner, L.A., Shields, A.F., Simon, M.S., Klamerus, J.F., Schiffer, C., Albrecht, T.L. (2015). Critical Shortage of African American Medical Oncologists in the United States. Journal of Clinical oncology, 33(32): 3697-3700.

Housel N, Barredo R, Edmondson E, Raynes R Kunnu E (2011). Social Justice for multiple choice examinations: Development of a diagnostic tool to detect student problems with language variation. *The Journal of Interdisciplinary Education*, *10*(1), 92-107.

Johnson, T., Boyden, J. E., & Pittz, W. J. (2001). *Racial Profiling and Punishment in U.S. Public Schools*: *How zero tolerance policies and high stakes testing subvert academic excellence and racial equity.* Report from the U.S. Department of Education: Office of Educational Research and Improvement. Oakland, CA: ERASE initiative: Applied Research Center.

Lea, I.A., Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common. Retrieved. http://quantpsy.org

Moule, J. (2005*). Cultural Competence: A Primer for Educators, 2nd Ed*. Belmont, CA: Wadsworth.

Nettles, M.T., Scatton, L.H., Steinberg, J.H., Tyler, L. L. (2011). Performance and passing rate differences of African American and white prospective teachers on Praxis™ examinations. A Joint Project of the National Education Association (NEA) and Educational Testing Service (ETS). Educational Testing Service. Retrieved. http://www.ets.org/research/contact.html

Schellenberg, S.J. (2004). *Test Bias or Cultural bias: Have We Really Learned Anything?* Paper Presentation for the Annual Meeting of the National Council for Measurement in Education. San Diego, CA. April 14, 2004.

Tellegen, P. J., & Laros, J. A. (2004). Cultural bias in the SON-R Test: Comparative study of Brazilian and Dutch Children. *Psicologia: Teoria e Pesquisa*, *20*(2) 103-111.

Utzman, R. R., Riddle, D.L., & Jewell, D.V. (2007). Use of demographic and quantitative admissions data to predict performance on the national physical therapy examination. *Physical Therapy, 87*(9) 1181-1192.

Weintraub, J., Walker, J., Heuer, L., Oishi, M., Upadhyay, K., Huang, V., Lindquist, C., Cushman, L.F., Ripp, J. (2015). Developing Capacity for the American Indian Health Professional Workforce: An Academic-Community Partnership in Spirit Lake, North Dakota. *Annals of Global Health, 81*(2): 283-289.

Wolfram, W. (1998). *American English: Dialects and Variation*. San Francisco, CA: Wiley Blackwell.