ADDRESSING STANDARDIZED TESTING THROUGH A NOVEL ASSESMENT MODEL

Catherine C. Schifter and Martha Carey
Temple University, Philadelphia, PA, USA

ABSTRACT

The No Child Left Behind (NCLB) legislation spawned a plethora of standardized testing services for all the high stakes testing required by the law. We argue that one-size-fits all assessments disadvantage students who are English Language Learners, in the USA, as well as students with limited economic resources, special needs, and not reading on grade level. The SAVE Science project was developed to explore whether and how contextually driven assessments support these students demonstrate their understanding of science content in middle grades in the USA. Preliminary findings from this 6-year study suggest that situating assessment in virtual contexts does in fact help students in answering multiple choice questions correctly and also helps students better understand their own science knowledge and learning process.

KEYWORDS

Standardized testing, virtual environments, critical theory.

1. INTRODUCTION

The ubiquitous standardized tests developed by a small number of educational services companies and used in American public schools contain questions culled from the cultural experiences of, and based on the language abilities of, the test content developers. All students are expected to be familiar with this content, but in truth it is often far removed from the experiences and skills of actual students. And this expectation automatically disadvantages groups of students, particularly English Language Learners in the USA, students with limited economic resources (which can constrain exposure to varied cultural experiences) and students with special needs. Students taking such tests experience an existential dislocation: they must answer questions in a formal and rigid way, questions that may call for cultural acuity or information they may not have, and questions written by unseen experts for whom this information is often intuitive. This can turn test taking into an Escher-esque endless loop of disconnectedness.

A Principal of a K-5 public school in New York penned an op-ed in the New York *Times* recently which touched on that disconnection, noting that English Language Arts standardized test content (developed by Pearson Publishing for the State of New York) presented students with questions that were "confusing" and "developmentally inappropriate…There was a strong emphasis on questions addressing the structure rather than the meaning of texts. There was also a striking lack of passages with an urban setting." (Phillips 2014) These tests carry high stakes for both the schools and the teachers. This shift to high stakes testing is yet another national educational policy change that, as educational researcher Michael Apple notes, is an outcome of ongoing cultural and political conflicts at the macro level. (Apple 2007, p. 165) At the classroom level, in order to make these tests "count," teachers must work within a new and permanent professional contradiction -- they are trained to differentiate instruction to meet the needs of students where they are but they then must standardize testing.

Standardization has been embraced by school reformers and educational policy makers in the USA as a means of tracking the performance and accountability of schools, teachers, and students alike, which leaves the ethical educator with few options. To counter the uniform application of these standards of knowledge to students who have varying skills, experiences, and language abilities, some refuse to give such tests, as a cluster of teachers in the city of Seattle, Washington did in 2013. Some have actively protested the test content and the standards aligned with them, most recently a group of teachers, parents, and administrators in New York City. And still others have worked within the standardization framework to create test

environments and test content that minimizes that endless loop. One goal of the SAVE Science study is to address this problem in testing and lessen that disconnect between test content and student experience through the creation of a new kind of assessment tool for middle school science students, where tests are taken by navigating virtual game environments and students use visual cues and inquiry skills to solve contextual problems. These assessments are proximal tests, directly linked to curricular concepts but delivered in a new context, but also incorporate test content derived from distal measures – in this case, the statewide Pennsylvania System of School Assessment (PSSA) tests. (Geier, et al 2008, p. 923) This paper presents a critical theory view of the importance of the SAVE Science project and what it brings to the dialogue around high stakes testing and differentiated instruction/learning.

2. SAVE SCIENCE

The SAVE Science study is an NSF-funded project to create, implement, and evaluate computer-based assessment modules for science content in the middle grades. The modules are designed to enable students at varying skill levels and language abilities to perform a series of problem-solving tasks in a virtual world, tasks that provide data about how those students apply content knowledge related to classroom curricula. These alternative assessments address several of the conditions needed for better science assessments, chief among them contextualization. Because students have a difficult time applying their understandings of science content and their own experiences to the decontextualized questions found on multiple choice written tests, assessments should contextualize questions. One example of this, from a recent PSSA test, is a question about freshwater fish and how they adapt to live in weedy areas of lakes, a question that urban students may have to answer with *no* lived experience to draw from. A SAVE Science test module would provide the student with context by offering an immersive virtual environment (IVE) where this question was accompanied by an active rendering of that type of fish swimming in its environment, and students could observe the fish in its habitat before answering the question. Another example: the SAVE Science module aligned with content about gas laws provides students with an IVE where basketballs are played with an indoor basketball gym, and then are used on an outdoor basketball court in cold weather. (see Ketelhut, et al 2013 for more detail) Students must determine why the basketballs bounce differently between inside and outside by gathering empirical data, analyzing that data, and reporting back to the appropriate character their hypothesis and evidence. These activities are then followed by 3 standardized test items that correlate with the high stakes test questions about the same content.

Evidence centered design principles (Mislevy, 2011) were used to develop each of the SAVE Science modules. As Mislevy noted, "One challenge is that the development of a valid simulation-[virtual-] based assessment requires the expertise from disparate domains come together to serve the assessment's purpose (typically including subject matter knowledge, software design, psychometrics, assessment design, and pedagogical knowledge)." The SAVE Science team was comprised of 1) an expert in science content, science teaching, and assessment design for science content; 2) an expert in designing virtual environments for assessment; and 3) a psychometrician. This team, along with 12 science teachers, four science education doctoral students, one science education post-doctoral fellow, and one qualitative research specialist, designed five assessment modules (2 for 7^{th} grade, 3 for 8^{th} grade), plus two introductory modules (1 for 7^{th} grade, 1 for 8^{th} grade).

Working with two senior science teacher leaders from a large urban school district, the team identified specific areas of middle grade science curriculum that was determined to be difficult to assess through the high stakes objective assessments. Given these assessments are objective in nature requiring reading of English (in the U.S.) on grade level, evidence centered design allowed the designers to create a virtual world where the students were put into a context (e.g., a basketball course inside and outside) to gather data around a problem presented as urgent, but within the context recently taught in the curriculum. Using Vygotsky's (1978) zone of proximal development theory, each assessment module was designed to be just beyond the capabilities of the students, but close enough to not be too complex. The evidence gathered by the system included a trace of every non-player character encountered, data gathered using the science tools built into each module, and a 3-dimensional time-stamped map of each student's movements within the virtual environment module. In the end, students answered both objective questions and open-ended explanations of evidence, and giving a solution to the original problem/question posed.

The results are a rich set of data that can be used to determine whether the student understood the question posed and data collected sufficiently to successfully respond to the assessment questions at the end of the module. But analyzing these disparate bits is not simple. Most teachers are not taught data-driven decision-making using multiple data points/sets. Evidence from SAVE Science initially suggests it is possible to identify those students who clearly understand the science content in each module from those who clearly do not (Sil, Shelton, Ketelhut, & Yates, 2012). We continue to refine the analysis toward a prediction model.

3. DISCUSSION

These efforts are one direct attempt to reduce the disconnection traditional standardized tests can cause among students, which is an issue that permeates education in the accountability era. Today approximately 21% of the public school population in the USA is made up of English Language Learners (ELL) and only 3% of these students reached proficiency or above on the 2009 National Assessment of Educational Progress (NAEP) reading assessment, as compared to 35% of native English speaking students. (Lara-Alecio, et al 2012, p. 987) One aspect of ELL students' lack of proficiency on these tests is the "unnecessary linguistic complexity" of test items they encounter on such tests. (Abedi & Gandara 2006, p. 39) These students are directly disadvantaged by the test content they are required to master, test content developed by native English speaking test developers presented out of context and accompanied by minimal visual cues. Along those same lines, NCLB legislation ushered in an era of testing "focusing solely on student outcomes" as a means of improving schooling, but ELL students "disproportionately attend high-poverty schools with limited resources, and fewer schools offer bilingual education programs than did before the passage of NCLB." (Menken 2010, p. 127)

Recent research has shown that urban districts in the USA overall are suffering the consequences of accountability systems based on test scores, and that "academically disadvantaged students in large cities are currently being left behind because the use of proficiency counts in NCLB does not provide strong incentives for schools to direct more attention to them." (Neal & Schazenbach 2010, p. 280) Compounding this issue is the fact that the AYP measure is intended to reflect a rising minimum threshold for improvement, so that "schools that begin with low test scores, typically urban schools with a high percentage of children living in poverty in the USA, can have improving test score results, but because they do not rise above the minimum threshold, remain classified as failing...[and] because NCLB requires that by 2014 essentially all students need to pass every test, almost all the schools in the USA will be found to be failing." (Hursh 2013, p. 577) This conundrum reveals the limitations of, if not the fallacy of, an uncritical reliance on high stakes testing at the broadest levels. At the student level, research has shown that learning through a type of digital gaming similar to that developed by SAVE Science can be directly linked to learning outcomes, and that contextualized information in game environments allows us to "measure [students'] growth across time, and track different trajectories to mastery." (Herold 2013)

As noted above, teachers in the USA are inculcated with the concept of differentiating instruction based on student ability, including, but not limited to, English language proficiency, special needs, reading level, and prior learning. The age of accountability, symbolized by high stakes objective type tests which include extensive reading passages to convey science content, poses a conundrum for teachers: Do they follow ethical teaching and differentiate instruction, knowing that not all students will achieve the same level of proficiency or even basic knowledge, or do they teach to the test? SAVE Science was design to challenge the notion of what is a test in science, putting the questions in to context where the student can demonstrate through multiple ways their understanding of content and scientific inquiry.

4. CONCLUSION

Two practical goals of the SAVE Science study were to develop new types of computer-based assessments that integrated and contextualized science content for students and to enhance understanding of students' use of inquiry processes in science through the use of such alternative tests. A key motivating question for the Principal Investigators of the project was "Can we create something that's reliable and valid as an alternative to traditional testing?" And recent research supports the contention that assessments situated in virtual

environments can also offer insights into student understanding not easily captured with other assessment methods and provides information about students' strategies in problem solving. (Ketelhut, 2007) So at its base level the SAVE Science study has been exploring how to improve students success in understanding and answering required test questions correctly. But the broader aim of the study has been about developing alternative forms of assessment which provide contextualized information for students and opportunities for them to demonstrate the ability to identify problems, collect data, and find solutions in a manner that does not alienate them or punish them for not intuiting knowledge that is not part of their lived experience. Preliminary findings from this 6-year study suggest that situating assessment in visual, virtual contexts does in fact help students in answering multiple choice questions correctly and also helps students better understand their own science knowledge and learning process. These promising results provide initial evidence that situating assessments in IVEs and situating test questions in context can play a role in improving standardized high stakes tests, and contribute to the ongoing conversation about what such tests measure, and how they are used. SAVE Science provides us with data about the use of virtual environment assessments with contextualized questions that may mitigate that endless loop of disconnection for students -- and provides clear examples for test content developers to use in designing assessments that do a better job of meeting students where they are.

ACKNOWLEDGEMENT

We acknowledge the support of this paper by Dr. Diane Jass Ketelhut, who is the Primary Investigator on this NSF funded project. This material is based upon work supported by the National Science Foundation under Grant No. 0822308.

REFERENCES

Abedi, J and Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: interaction of research and policy. *Educational Measurement: Issues and Practice*, 25(4), pp. 36-46.

Apple, M. (2007) Social movements and political practices in education. Theory and research in education, 5(161), pp. 161-170.

Geier, R. et al (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, 45(8), pp. 922-929.

Herold, B. (2013). Researchers see video games as testing, learning tools. *Education Week*, 32(37), pp. 14-15.

Hursh, D. (2013) Raising the stakes: high-stakes testing and the attack on public education in New York. *Journal of Education Policy*, 28(5), pp. 574-588.

Ketelhut, D. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City, a multi-user virtual environment. *Journal of Science Education & Technology*, 16 (1), pp 99-111.

Ketelhut, D. et al (2013). Improving science assessments by situating them in virtual environment. *Education Sciences*, 3, pp. 172-192.

Lara-Alecio, R. et al (2012). The effect of an instructional intervention on middle school English learners' science and reading achievement. *Journal of Research in Science Teaching*, 49(8), pp. 987-1011.

Menken, K. (2010). NCLB and English language learners: Challenges and consequences. *Theory into Practice*, 49(2), pp. 121-128.

Mislevy, R.J. (2011). Evidence-centered design for simulation-based assessment. CRESST Report 800. Los Angles, CA: CRESST, UCLA. Accessed from http://www.cse.ucla.edu/products/reports/R800

Neal, D. and Schanzenbach, D. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92(2), pp. 263-283.

Phillips, E. (2014). We need to talk about the test. New York *Times online*, April 9 2014. Retrieved from: http://www.nytimes.com/2014/04/10/opinion/the-problem-with-the-common-core.html?_r=0

Sil, A et al. (2012). Automatic grading for scientific inquiry. In: Proceedings of the NAACL-HLT 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7), Montreal, QC.

Vygotsky, L. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.