Paper 4837-2020

SAS® BI Platform to Improve Use of Analytics in Higher Education: Do the Stats Match the Intent?

Sean W. Mulvenon, University of Nevada, Las Vegas

ABSTRACT

A national trend for all university and college administrators is increased pressure from legislators and other education advocates to increase persistence and graduation rates of all students. Learning analytics (LA) models have been implemented nationwide with the use of Learning Management Systems (LMS) software to identify access patterns of students to support materials, identify those who might be performing below expectations, and to use this information to be more effective in providing educational support. A quick review of data structures and the common metrics employed revealed numerous statistical anomalies that might be problematic in use of LA and LMS in higher education. For example, one dubious metric identified was "duration", which was an aggregate of the time a student used to complete practice exams. If a student required 50 minutes to complete three practice exams, the information provided was 50 minutes and the final score on the last exam, with no information provided on the number of attempts. A student requiring 20 minutes to score 5/10, then 15 minutes to score of 7/10, and finally 5 minutes to obtain 9/10 is a very different pattern of progression, and is important in order to understand student persistence. This session presents an integrated data model using SAS® Business Intelligence Platform to improve the accuracy and interpretation of analytics in order to improve student persistence and graduation rates in higher education.

INTRODUCTION

Postsecondary institutions must endure consistent pressure to increase *persistence* and *completion* of students by parents, business, educational advocacy groups and legislators. The days of postsecondary institutions relying on the personal responsibility of college students to attend class, complete homework or seek assistance on their own are long gone. The new expectation is the postsecondary institution has a fiduciary responsibility to actively support and guide students to academic success. As such, postsecondary institutions are constantly seeking academic resources to proactively support their efforts with students to increase persistence and completion. Learning Management Systems (LMS) have become popular on postsecondary campuses as a method to identify and provide the necessary resources for students.

A challenge with use of LMS at UNLV has been in identifying and understanding the definitions of many of the variables selected. Additionally, numerous labels employed in postsecondary education, such as *persistence* and *completion*, have nebulous interpretations and meanings. The use of LMS data and use of various labels to describe performance raises the question “What is really being measured?” and “What are you attempting to measure?” The answer to both of those questions, and to demonstrate the challenges of improving higher education analytics, is demonstrated through application/development of a model to improve persistence and completion at UNLV. The purpose of this paper is to present a proactive model referred to as the “Elevator Model” that demonstrates a “Data Lake to Dashboard” approach using SAS as a single source solution; and that actively resolves challenges associated with undefined or nebulous labels in higher education.

THE "ELEVATOR MODEL" TO IMPROVE PERSISTENCE AND COMPLETION OF STUDENTS AT UNLV

What is an "Elevator Model?" The term "Elevator Model" was developed to describe a process of engineering a SAS single source solution for transitioning information from the UNLV "Data Lake" to educational "Dashboards" for students and faculty. Figure 1 represents a conceptual model of how information "flows" from one level (i.e., floor of a building) to the next level; and how at each level there are actions completed to improve/clean/analyze or interact with the data to improve use in educational support systems.

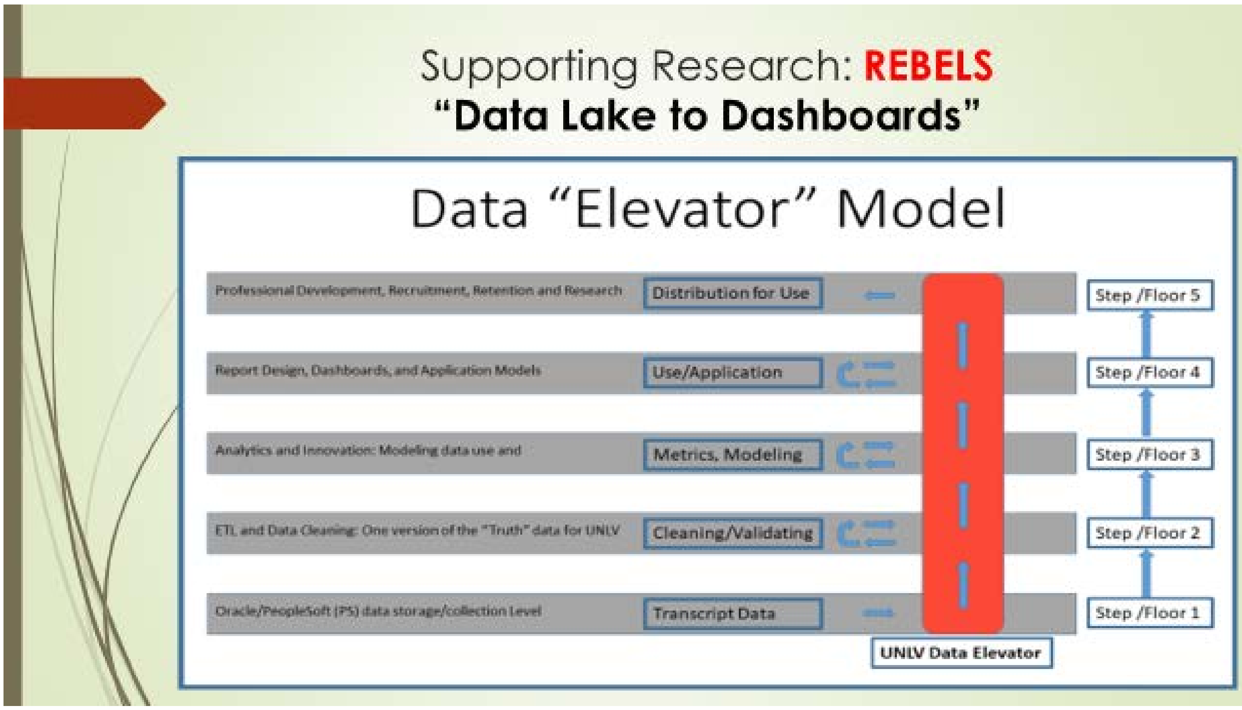

Figure 1. ***UNLV "Elevator" Model***

Data in Figure 1 moves from "floor to floor" up the system with the distribution for use by students via educational dashboards. I will concede this model is simplistic, but it is effective in describing how data moves through required steps before it can be used by consumers. A seminal focus of this paper is the Step/Floor 3 or Metrics and Modeling.

RESEARCH, EVALUATION AND BUILDING EFFECTIVE LEARNING SYSTEMS (REBELS)

It is important to develop prototypes of any theoretical work, but more importantly in academe, it is essential to design pilot studies to provide a proof of concept. The REBELS program was developed as a National Science Foundation (NSF) "proof of concept" proposal that examines use of LMS data to increase STEM persistence and completion; and applies the approach of an "Elevator Model."

Learning Analytics (LA) models have been implemented nationwide with use of LMS software to record access by students of support materials, interactions with instructors and course curriculum materials (Adzharuddin & Ling, 2013). LMS data are linked to progress within specific courses and this information is used to identify and provide students performing below expectations with additional academic support. A review of UNLV LMS data structures and the common metrics employed revealed numerous statistical anomalies that are problematic in use of LMS and the exploratory data models that have been employed to improve persistence in STEM courses. The seminal issues revolve around limited definition of the metrics or statistical challenges related to their use in analytics. For example, duration is a metric for "time to complete" a pre-test quiz and is recorded with a "Total Score." However, "Duration" is the total time required for a student complete ALL attempts of the pre-test quiz. "Total Score" is only the best score obtained by the student on the pre-test quiz. For example, Student A completes the pre-test in 10 minutes with a score of 90%; and has "Duration" of 10 minutes and "Total Score" of 90% recorded. Student B attempts the pre-test on three occasions with a "Duration" of 60 minutes (time to complete all three attempts) and a "Total Score" of 90%. "Duration" and "Total Score" represent different metrics in this example. The LMS data of the progress of Student B to obtain a score of 90% is not captured; and the overall command of Student A is not evaluated. The use of these metrics increases statistical/measurement error and create issues with interpretability of these variables in statistical models.

DEFINITIONS ALIGNED WITH RESEARCH/POLICY GOALS

Persistence implies a student continues to enroll and pursue a degree; and graduation represents success in completing a sequence of courses. Use of cross-sectional data models with single courses to predict success of a student is incongruent with the definitions of both *persistence* and *graduation*. Additionally, what constitutes "success" in a STEM course? Current models have used a dichotomous outcome of "B or Higher" and "C or Lower" which is incongruent with institutional policy of a "C" as a passing grade representing success. **REBELS** examines the longitudinal sequencing of courses required in the computer engineering curricula as well as the overall progress of students from initial enrollment, persistence and their "pathways" to completion.

To better understand the challenges of consistency of definitions in reporting, research, and educational support systems the term "Data Science" was evaluated relative to education. What is "Data Science" in education? I really believe there are two definitions:

Data Science in Education

Definition #1: Use of statistical methods to extract information from BIG DATA in education to evaluate student performance, curriculum and instruction.

Definition #2: Manipulation of data using ***ALL possible*** ways to find something that is positive to report.

Clearly, the first definition represents an effort to formally describe data science in education; and the second definition represents a reality that is far too common in educational data science. When defining persistence, is completion of a single course representative of persistence? Is persistence a subset of the larger expectation of a pathway to completion of a degree? The expectation of progression toward a degree clearly means successfully completing a course, but it also implies completing a sequence of courses which also leads to completion of a degree in higher education. Too often our definitions are created "fit" the policy or research goals, but are incongruent with computing meaningful analytical models.

The example of "duration" is one that contributes many challenges when these data are used in research models. What is the impact of this metric in a regression or growth model? How large is the measurement error. Many of the LMS variables have either dubious or no definitions to identify the actual goal/use of the data. Incongruent metrics with policy or research goals is a systemic problem in education (Mulvenon, 2015). More importantly, the point to remember it isn't the statistics that is the problem, but the policy that dictates how metrics are computed. Figure 2 provides a few examples of metrics used in educational accountability systems that are incongruent with the policy or research goals.

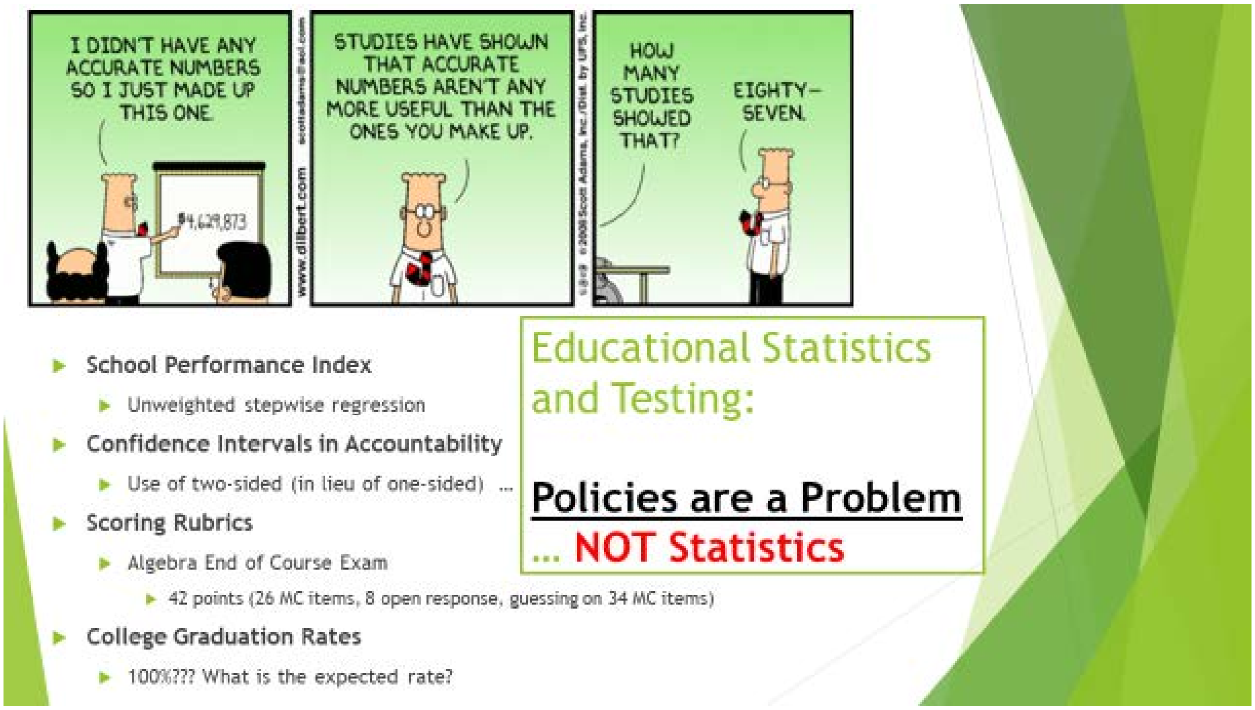

Figure 2. ***Examples of incongruent metrics***.

Each of the metrics in Figure 2 have severe methodological challenges which raise the question "What is really being measured?" The last example in Figure 2, college graduation rates, is an example of an interesting metric in higher education that seems simple. But is it? A few challenges that exist with this metric are associated with data rules. For example, do you compute graduation rates based on all students enrolled in a college? Do you track institutional graduation rates? In most cases you do have metrics computed for both the college and institution. However, a significant question is whether a goal of 100% graduation rate is reasonable. A metric proposed within REBELS is an innovative method to compute and Expected Graduation Rate (EGR). Table 1 provides simulated data for graduation rates based on academic performance prior to enrolling in a college degree program. The graduation rates of students within the various groups are used to develop an EGR as provided in formula 1.

Table 1. ***Computing an Expected Graduation Rate.***

Group	Graduation Rate (N)	Graduates (N)	Goal for REBELS*
ACT 30–36, GPA > 3.5	90% (100) = n_1*pr(X_1)	90	100% (100)
ACT 30–36, GPA 3.0 - 3.5	80% (80) = n_2*pr(X_2)	64	90% (72)
…	…	…	…
ACT < 20, GPA 2.0 – 2.5	25% (1,200) = n_{12}*pr(X_{12})	300	35% (120)
Totals	38% (6,000) =Exp Grad Rate	2,280	48% (600)

** Represents a 10% increase in each category.

$$Expected\ Graduation\ Rate\ (EGR) = n_1 * pr(X_1) + n_2 * pr(X_2) + \cdots + n_{12}pr(X_{12}) \qquad (1)$$

The EGR provides a more representative metric of the historical performance of students within a degree program toward completion. If the EGR is 60% and a degree program improves to 66% this would represent a significant gain as opposed to an interpretation that they are underperforming the 100% benchmark. This type of metric will provide much more accurate information on the success of various degree programs and benchmark them to more reasonable expectations.

Use of BIG DATA in education is commonplace and driven by various educational policy initiatives. Figure 3 outlines several Federal policies designed to improve educational outcomes.

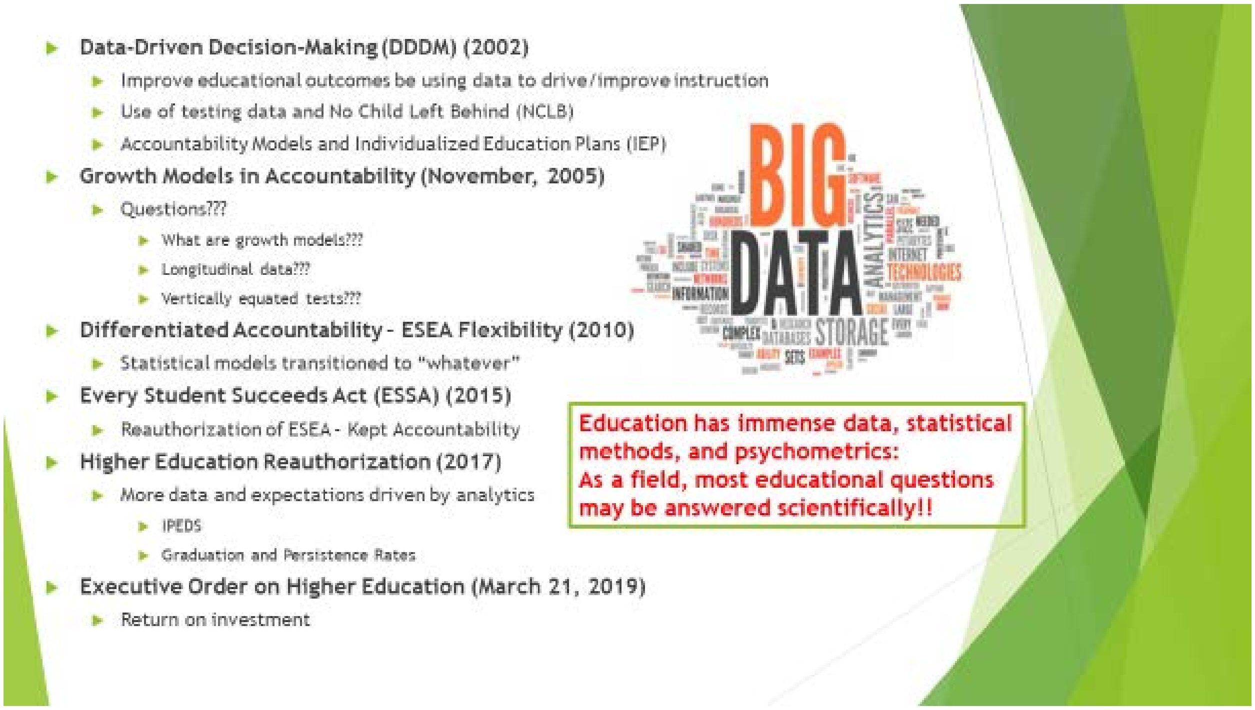

Figure 3. ***Education Policy Initiatives Focused on use of BIG DATA***

The last example in Figure 3 is the Executive Order on Higher Education signed by President Trump (March 21, 2019). All postsecondary institutions are to provide a report to students on what I've defined as an Economic Return on Investment (EROI). I developed a metric to compute an EROI in 2017 (Mulvenon, 2017) that is a method to help students understand four key metrics: (1) time to completion, (2) student loans, (3) median income for degree, and (4) cost to attend. The simplistic metric of graduation rates does not provide this key information in determining a degree to pursue or the relative cost/benefit of the degree.

CONCLUSION

The goal of REBELS is to use BIG DATA from the UNLV "Data Lake" to create the elevator model for improving persistence and completion in STEM programs; and to provide information using LMS data and innovative metrics such as the EROI and EGR. Figure 4 outlines key data elements and the SAS BI components to be employed. Each of these elements is employed within the "Elevator Model" and is representative of the power of SAS. However, as important as the ability to use a single source solution, is the need to be able to develop and identify your key metrics. The EGR and EROI are new metrics developed as part of these projects, but many of the LMS metrics and other variables collected as part of the Data Lake at UNLV require additional attention to improve the effectiveness of their use in education.

A final element of REBELS will be to automate the process of data extraction for populating individualized student dashboard reports to help support students. This is a key element to improve educational outcomes while reducing the workload and efforts of professionals in

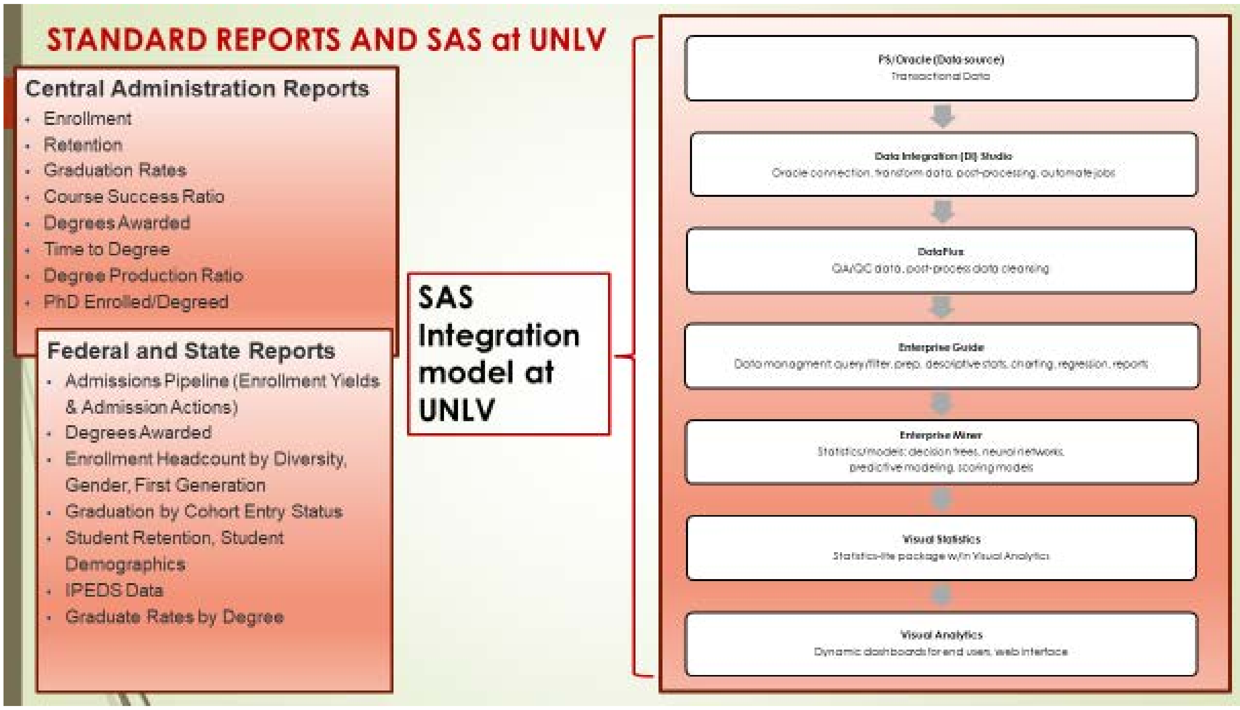

Figure 4. ***Sample data and SAS BI tools in REBELS***

the Office of Information Technology at UNLV. The use of the "Elevator Model" also represents the proof of concept efforts to transition data from storage to usage as a first effort. The success of this process provides a roadmap for future projects to develop dashboards for administrators (e.g., admissions, deans, etc.) to support use of educational data to improve student persistence and completion campus wide. Figure 5 represents a final thought on use of educational data and a seminal question of "do the analytics match the intent?"

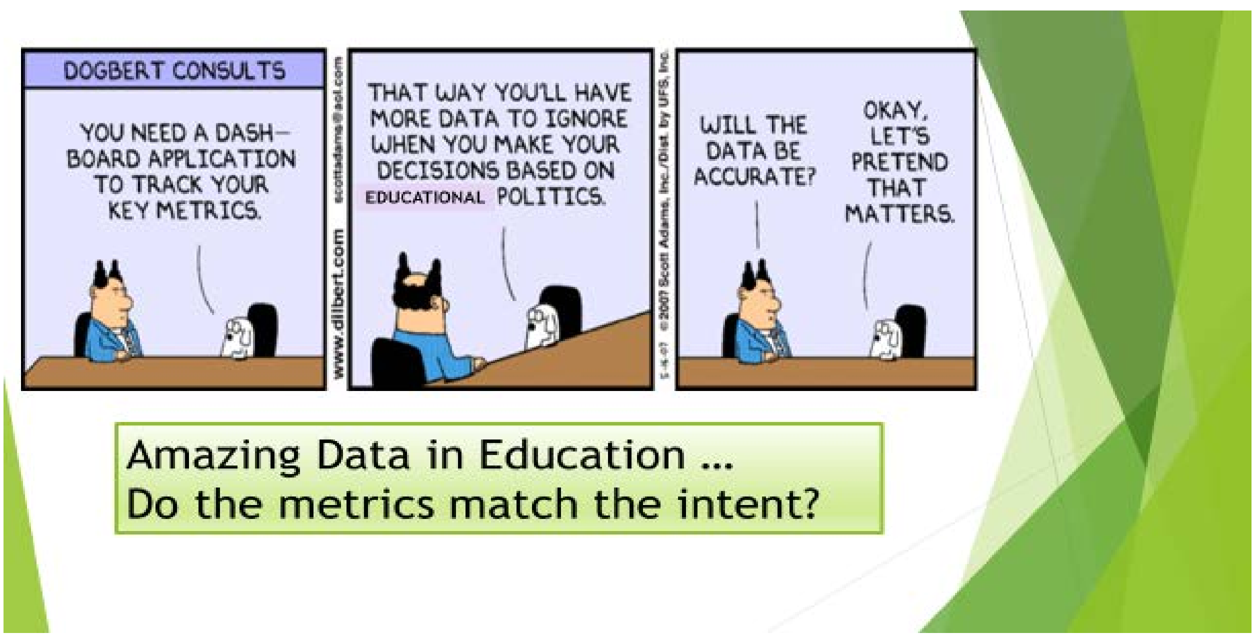

Figure 5. ***Do the metrics match the intent?***

REFERENCES

Adzharuddin, N. & Ling, Lee (2013). Learning Management System (LMS) among University Students: Does it Work? *International Journal of e-Education and e-Learning,* vol 3, n3, pg 248 -252.

Mulvenon, S. (2017). Improving the Evaluation of Higher Education: Understanding the Myths, Methods, and Metrics. Manuscript published in the proceedings of the SAS Global Forum 2017, Orlando, FL. Available online at: https://support.sas.com/resources/papers/proceedings17/0826-2017.pdf

Mulvenon, S. & Bowman, S. (2015). An Evaluation of how the "Policies of K-12 Testing" Impact the Effectiveness of Global Testing Programs. Smith, W. (Editor) *The Global Testing Culture: Shaping Education Policy, Perceptions and Practice.* Oxford Press: Oxford, UK.

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *Fundamentals of Programming in SAS®*
- *BIG DATA, DATA MINING, and MACHINE LEARNING*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sean W. Mulvenon, Ph.D.
University of Nevada, Las Vegas
Phone: (702) 895 – 4647
Email: sean.mulvenon@unlv.edu