

Using Potential Performance Theory to Assess How to Increase Student Consistency in Taking Exams

Stephen Rice¹, David Trafimow¹ & Keegan Kraemer¹

¹ Potential Performance Technologies, LLC, Las Cruces, New Mexico, U.S.A.

Correspondence: Stephen Rice, Potential Performance Technologies, LLC, Las Cruces, New Mexico, U.S.A. Tel: 1-575-571-7750. E-mail: srice67@gmail.com

Received: July 20, 2012 Accepted: August 13, 2012 Online Published: November 28, 2012

doi:10.5539/hes.v2n4p68

URL: <http://dx.doi.org/10.5539/hes.v2n4p68>

Abstract

It has been a concern among educators and academics that U.S. students suffer from a lack of knowledge about the world around them. This is reflected in low history scores, particularly in world history. The common explanation for this is that there is some systematic deficiency in American students, in that they either do not know the material or have poor testing strategies. We offer a different way of looking at this problem using Potential Performance Theory (PPT). With PPT, we assessed the consistency with which students answered test questions and show how much performance would improve if a student were perfectly consistent. Furthermore, we show how much improvement there is in consistency over multiple sessions. Participants were given a short world history test six times in a row. The results were interesting. Consistency did improve with practice, but the systematic factors that students employed (e.g. strategies) were poor enough to counter-act the improvement due to rising consistency levels.

Keywords: PPT, potential, performance, theory, world history, exams

1. Introduction

In 1983, urgency was brought about in the national crisis for educational reform via the “A Nation at Risk” report (Gardner, 1983). Only 18 months earlier, the Secretary of Education, T. H. Bell, instructed the National Commission on Excellence in Education to examine the quality of education in the United States. The resulting report compared schools and colleges in the United States with other nations and identified educational programs that resulted in “notable student success in college” (Gardner, 1983, p. 7). The report made clear that the gains that the United States had made in education have been squandered and that we are no longer the leader in international commerce.

A 2002 report from the U.S. Department of Education concludes that the United States is *not* among the top ranking nations for a variety of educational factors. From the years 1990 to 1997, developing countries saw a substantial increase in postsecondary education with enrollment for Asia increasing from an estimated 23,314 in 1990 to 34,844 in 1997. During this same time period, postsecondary enrollment in North America, including Canada, increased from an estimated 15,628 to 16,038, a sizable difference from that of Asia. Other nations also spend more on their students than the United States. As of 1999, Luxembourg spent an estimated \$19,436 per 12th grade student, while the U.S. spent approximately \$8,157 for the same demographic (Snyder & Hoffman, 2003). As such, American students are trailing behind students from developing nations, such as Cyprus and South Africa, specifically in math (Bush, 2002).

In an attempt to rectify this “genuine national crisis,” President George W. Bush passed the “No Child Left Behind (NCLB) Act” in 2002 (Bush, 2002). He deemed it a bipartisan solution to educational reform that would ensure all children, no matter their circumstances, would receive equal educational opportunities within the public school system. The educational blueprint consisted of “increasing accountability for student performance, focusing on what works, reducing bureaucracy and increasing flexibility, and empowering parents” (p. 2). Three years past the enactment of NCLB, the National Assessment of Educational Progress (NAEP) reported a persistent pattern of elevated state test results for fourth-grade math scores (Fuller, Gesicki, Kang, & Wright, 2006). However, not all scores showed positive results. Fuller et al. (2006) found that even though individual states showed incremental increases in reported yearly point averages for reading, the results from the NAEP showed no discernible change post NCLB. The changes in the educational system see “strong adoption and

implementation but not strong institutionalization” (Fullan, 2000, p. 1). This leads to a need for continuity within the educational system.

Students at all levels of education experience anxiety and suboptimal test-taking abilities due to factors such as poor time management skills, fatigue, or lack of topic familiarity (Swearingen, 1998). Brown (1999) suggests that parents and school counselors alike take action to intervene to improve overall student achievement and school climate through avenues such as time management training, study skills groups, and achievement motivation. Such interventions may mitigate limitations in academic performance. Studies indicate several cognitive and psychological factors jeopardize students’ overall test-taking performance, including anxiety, attitudes towards the test subject matter and general test-taking, as well as test-taking strategies (Dodeen, 2008). A lack of institutionalization of changes in the educational system leaves educators little time for implementing performance improvement for test-taking abilities for all individuals, but rather focuses attention on those with definitive issues (Kubistant, 2001).

Test-taking strategies have demonstrated strong relationships among academic scores and “can improve overall validity of test scores” (Dodeen, 2008, p. 411). Kubistant (2001) outlined four general areas that researchers may look to in order to improve test-taking performance: a) knowledge of tests, b) experience, c) mental and emotional preparation, and d) allowing, flowing and doing. The latter refers to doing rather than passive planning of the action. Furthermore, students should be aware that preparation takes place before, during, and after the test, as well as ensuring time management of learned strategies. Time management and self-testing are also strong indicators of academic performance (West & Sadoski, 2011). Students that utilize test-taking strategies have shown increased positive attitudes regarding testing, lower levels of anxiety, and achieve better test scores (Vattanapath&Jaiprayoon, 1999).

Increasing test performance through test-taking strategies is not a dilemma with a one-size-fits-all solution. It has become apparent that the United States needs to step up when it comes to educating its students, both traditional and non-traditional. Strategies need to match appropriately the student’s ability level and preparation style in order to improve testing accuracy and validity, reduce testing anxiety, and improve student’s overall attitudes towards test-taking (Dodeen, 2008). Educators may also encourage high academic achievement through promotion of skills related to organizing the synthesizing testing materials (Weinstein & Gipple, 1974). Larsen, Butler, and Roedigger (2008) further suggested that reviewing test material in a formative manner enhances the acquisition and retention of knowledge. In testing situations where students are asked to produce answers, such as short answers, monitored comprehension and self-testing techniques also improve academic performance (West & Sadoski, 2011).

1.1 Consistency in Performance

When assessing student performance, educators fail to appreciate that performance is influenced not only by systematic factors (e.g., strategy, knowledge, motivation, and so on) but also by consistency. Although the importance of consistency has been known at least since Spearman’s (1904) seminal work, more than a century later, educators nevertheless seldom consider it. Put simply, consistency refers to a person’s tendency to make similar responses on similar items. It is a simple fact of mathematical regression that a lack of consistency pushes scores towards the chance level; less consistency, keeping all other factors constant, implies a decrease in performance so long as the base level of performance exceeds the chance level. An easy way to see this is to assume that answers to some of the items are decided by a coin toss, regardless of the person’s level of knowledge, motivation, and so on. As the number of answers decided by coin tosses increases, overall performance will be increasingly closer to the chance level.

Because the class of systematic factors and consistency both matter, it implies a possible effect that may, at first, seem counterintuitive. But let us commence with what is intuitive. Suppose a person is exposed to items measuring mathematics knowledge multiple times. In addition, suppose that such exposure increases the person’s ability to respond to items in a particular domain. In that case, the person’s performance should increase. In addition, the person may learn to recognize that similar items should be performed in similar ways, thereby increasing performance consistency. An increase in the favorability of systematic factors or an increase in consistency both should push performance higher.

But the foregoing scenario is not the only one possible. Suppose that repeated exposure increases consistency but actually decreases the favorability of systematic factors. For example, suppose that repeated exposure causes people to be increasingly biased. Participants’ performances might decrease due to the increase in bias, but might be positively influenced due to using the biases in a more consistent way. In more general terms, the decrease in the favorability of systematic factors (e.g., bias) might be counterbalanced by an increase in consistency, thereby

resulting in little change in the level of performance that an educator actually would observe. The natural conclusion would be that repeated exposure causes no changes whereas the truth of the matter would be that it cause two changes, but in opposite directions. Normally, there would be no way to test this possibility of counterbalancing effects, but a recent advance by Trafimow and Rice (2008; 2009), termed *potential performance theory* (PPT), provides a theory-based way to do so. This theory has been supported by multiple empirical studies in recent years (Hunt, Rice, Trafimow & Sandry (in press); Rice, Geels, Hackett, Trafimow, McCarley, Schwark, & Hunt, 2012; Rice, Geels, Trafimow & Hackett, 2011; Rice & Trafimow, in press; Rice, Trafimow & Hunt, 2010; Rice, Trafimow, Keller, Hunt & Geels, 2011; Trafimow, Hunt, Rice & Geels, 2011; Trafimow, MacDonald & Rice, in press; Trafimow & Rice, 2008; 2009; 2011).

1.2 Potential Performance Theory

As we already have seen, observed performance is influenced by the favorability of systematic factors, which PPT terms *potential performance* or *potential scores*, and by consistency. In one kind of PPT paradigm (e.g., Hunt, Rice, Geels & Trafimow, 2010; Trafimow & Rice, 2009), participants complete two or more sessions, with two blocks of similar trials within each session. The reason for having two blocks of trials within each session is to enable the researchers to compute a correlation coefficient, for each participant, across the two blocks of trials; i.e., a *consistency coefficient* that measures the person's consistency across the two blocks of trials. Thus, across sessions, it is possible to determine whether each person's consistency increases, decreases, or does not change.

Based on the combination of observed performance and consistency, it is possible to compute each person's potential score. A person's potential score represents the totality of systematic factors that influence that person's performance, in the absence of any inconsistency whatsoever. Put another way, a person's potential score indicates how that person would perform if he or she were perfectly consistent. Assuming that a person's base level of performance is better than chance, increasing consistency increases performance, and so potential scores tend to exceed observed scores.

PPT computations are not difficult to make. Assuming multiple two-block sessions of dichotomous items, each person can make one choice or the other on each item, and the correct answer can be one choice or the other. Thus, there are frequencies, for each person, of four possibilities that arbitrarily can be labeled a , b , c , and d . These frequencies can generate row and column frequencies, r_1 , r_2 , c_1 , and c_2 . Given that all of these have been obtained, it is easy to convert the two-by-two matrix into a correlation coefficient, as Equation 1 below demonstrates.

$$r = \frac{|ad-bc|}{\sqrt{r_1 r_2 c_1 c_2}} \quad (1)$$

In addition, by using a version of Spearman's famous formula, it is possible to correct the correlation coefficient obtained in Equation 1 for the effects of inconsistency, using Equation 2 below. In Equation 2, R denotes the corrected or potential correlation coefficient and $r_{XX'}$ denotes the consistency coefficient across the two blocks of trials.

$$R = \frac{r}{\sqrt{r_{XX'}}} \quad (2)$$

Using the result from Equation 2, Equations 3-6 provide the cell frequencies that would be obtained in the absence of randomness. Because Equations 3-6 are concerned with potential scores, we use upper case letters throughout. Thus, A , B , C , and D refer to the potential cell frequencies corresponding to a , b , c , and d , respectively. Also, similar to a Fisher's Exact Test or a Chi-Square test, we assume fixed margin frequencies, designated by R_1 , R_2 , C_1 , and C_2 .

$$A = \frac{R\sqrt{R_1 R_2 C_1 C_2} + C_1 R_1}{(R_1 + R_2)} \quad (3)$$

$$B = R_1 - A \quad (4)$$

$$C = C_1 - A \quad (5)$$

$$D = R_2 - C \quad (6)$$

Based on the potential cell frequencies, Equation 7 renders the *potential performance* or *potential score*.

$$\text{potential performance} = \text{potential score} = \frac{A+D}{A+B+C+D} \quad (7)$$

1.3 Current Study

Let us now return to the issue at hand. Suppose a person is exposed to 3 two-block sessions of dichotomous world history items. We used world history items because we expected people to become more biased in their responses. We hypothesized that the greater bias with more exposure would have counterbalancing effects on potential performance and consistency, thereby leading to very little in the way of change in observed performance. That is, although we expected observed performance to not change much across sessions, we also expected potential performance to decrease but consistency to increase.

2. Experiment

2.1 Participants

Twenty-six participants were recruited from a large southwestern university. The mean age was 21.65 ($SD=2.54$). All participants had successfully completed high school.

2.2 Materials

Participants were asked to answer 30 true-false world history statements that spanned several thousand years and various countries. These statements can be found in Appendix A. Half of the statements were true and half were false.

2.3 Procedure

Participants first gave written consent and then proceeded with the experiment. The statements were presented online and participants were given as much time as they needed to answer each one. Importantly, in order to conduct PPT analyses on the results, participants were given 6 blocks of the identical statements. The first two blocks represented Session 1, the second two blocks represented Session 2, and the final two blocks represented Session 3. In each block, all of the statements were randomized in order of presentation. Participants were given short breaks in between each block. The experiment took an average of approximately 20 minutes to complete. Upon completion, participants were debriefed and dismissed.

2.4 Design

A within-participants design was employed by which all participants answered all 6 blocks of questions.

3. Results

PPT analyses were conducted on each of the 3 sessions to determine the observed scores, the potential scores, and the consistency coefficients for each participant. Figure 1 presents these data. The differences in observed scores across sessions was not significant, $F(2, 50)=1.40, p=0.17$. This appeared to be due to the fact that while the difference between the consistency coefficients was significant, $F(2, 50)=8.25, p=0.001$, the potential scores fell enough to counter the improvement in consistency scores.

The consistency scores improved significantly from Session 1 to Session 2, $t(25)=3.10, p=0.004$, two-tailed, and from Session 1 to Session 3, $t(25)=3.25, p=0.003$, two-tailed, but the difference between Session 2 and Session 3 was not significant, $t(25)=1.28, p=0.21$, two-tailed, although it was in the same direction as the Session 1 to Session 2 change.

4. Discussion

In the introduction, we suggested that educators appreciate, insufficiently, the importance of consistency in affecting observed performance. Consequently, educators also fail to realize that potential performance and consistency interact to determine observed performance. The fact of this interaction implies the interesting possibility that potential performance and consistency can go in opposite directions, with repeated item exposure, to render no effect on observed performance. Normally, it would be impossible to distinguish two types of lack of change in observed performance from each other—lack of change due no changes whatsoever versus lack of change due to two underlying changes in opposite directions. However, PPT renders it possible to distinguish these two possibilities. Our goal was to demonstrate the latter, and more interesting, possibility. In fact, that is what we have done. Potential performance decreased, consistency increased, and these two changes balanced each other so that there was insignificant change in observed performance across sessions of exposure.

To our knowledge, this is the first demonstration of counterbalancing changes in potential performance and consistency rendering insignificant change in observed performance in the education domain. To be sure, Trafimow and Rice (2009) demonstrated a similar effect, but their effect was limited in two ways, at least from the present point of view. The most important limitation is that although they obtained the effect for particular individuals, they did not obtain it across a set of participants. An additional limitation is that they used a visual

search task with limited educational relevance.

4.1 Practical Applications

The practical applications of this data are important to note. With PPT, teachers and parents are able to assess why their students and children, respectively, are not doing as well as they would like. It could be that the students are scoring poorly on exams for systematic reasons (e.g. lack of knowledge, poor test-taking skills, etc.), or it could be a case of performing inconsistently, or both. PPT allows educators to parse the two factors (non-random and random) and determine for each individual student, where the practice/training should be focused. If the student shows poor consistency in test-taking, then the current data show that it could be as simple as increasing the hours of practice in that subject for the student to improve in consistency, and thus improve overall performance as well.

4.2 Limitations

As with all studies, this experiment has limitations that should be discussed. First, the sample size was not very large. With a larger sample size, it would not only give researchers more power in determining differences between groups, but also more generalizability. Second, the students were all from the southwest United States, which also limits generalizability. More research should follow in order to replicate these findings with other sample groups in order to increase generalizability. Third, we only used world history questions. It may be the case that other types of tests would reveal different results.

4.3 Conclusion

The purpose of this study was to examine how to improve students' test-taking consistency via multiple sessions (i.e. practice). Students took a history test six times over three sessions so we could obtain consistency and potential scores. While their consistency improved significantly over multiple sessions, their observed scores did not change significantly because of the counter-acting effect of falling potential scores. The present research constitutes the first demonstration that consistency and potential scores can counter-act each other at the group level.

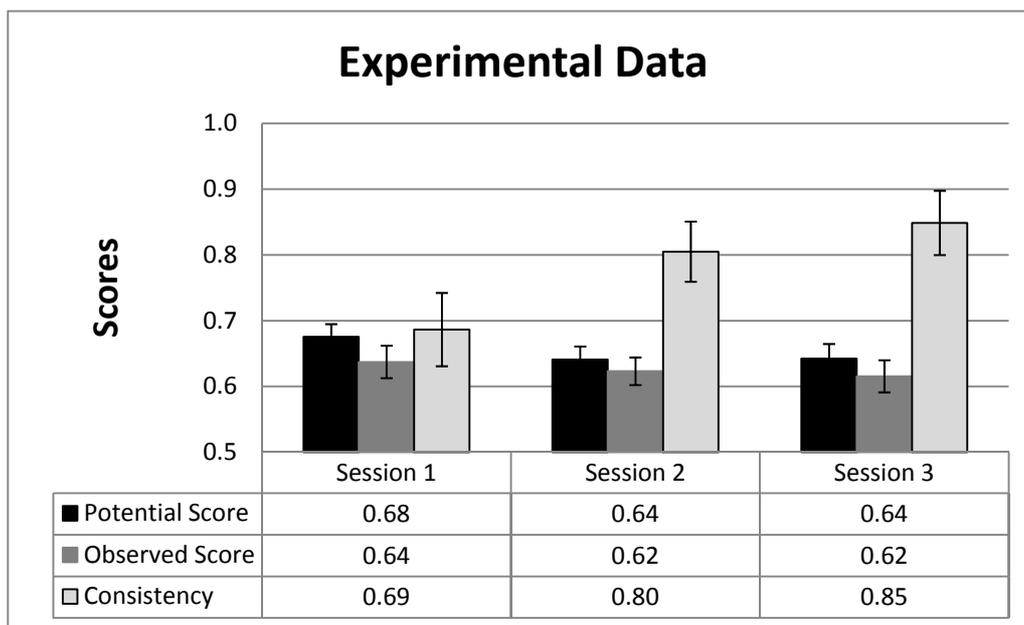


Figure 1. Data from the experiment (SE bars are included)

References

Brown, D. (1999). *Improving Academic Achievement: What School Counselors Can Do*. ERIC Digest. Retrieved from ERIC database. (ED435895)

U. S. Department of Education. (2002). *No child left behind*. Retrieved from

- <http://12.20.128.93/docs/nche/000005/00000528.pdf>
- Dodeen, H. (2008). Assessing test-taking strategies of university students: Developing a scale and estimating its psychometric indices. *Assessment & Evaluation In Higher Education*, 33(4), 409-419. <http://dx.doi.org/10.1080/02602930701562874>
- Fullan, M. (2000). The three stories of education reform. *Phi Delta Kappan*, 81, 581-584. Bloomington, Indiana: Phi Delta Kappa International. Retrieved from <http://mathacts.mspsnet.org/index.cfm/9540/>
- Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the no child left behind act working? The reality of how states track achievement.* (Working paper, University of California). Retrieved from <http://eric.ed.gov/PDFS/ED492024.pdf>
- Hunt, G., Rice, S., Geels, K., & Trafimow, D. (2010). Analyzing sub-optimal human-automation performance across multiple sessions. *Proceedings of the 54th Annual Meeting of the Human Factors and Ergonomics Society*, Santa Monica, CA.
- Hunt, G., Rice, S., Trafimow, D., & Sandry, J. (in press). Using Potential Performance Theory to analyze systematic and random factors in enumeration tasks. *American Journal of Psychology*.
- Kubistant, T. (1981). Test performance: The neglected skill. *Education*, 102(1), 53.
- Larsen, D. P., Butler, A. C., & Roedigger, H. L., III. (2008). Test-enhanced learning in medical education. *Medical Education*, 42, 959-66. <http://dx.doi.org/10.1111/j.1365-2923.2008.03124.x>
- National Commission on Excellence in Education. (1983). *A Nation at Risk: The Imperative For Educational Reform.* Government Printing Office, Washington, DC.
- Rice, S., & Trafimow, D. (in press). Time pressure heuristics can improve performance due to increased consistency: A PPT methodology. *Journal of General Psychology*.
- Rice, S., Geels, K., Hackett, H., Trafimow, D., McCarley, J. S., Schwark, J., & Hunt, G. (2012). The harder the task, the more inconsistent the performance: A PPT analysis on task difficulty. *Journal of General Psychology*, 139(1), 1-18. <http://dx.doi.org/10.1080/00221309.2011.619223>
- Rice, S., Geels, K., Trafimow, D., & Hackett, H. (2011). Our students suffer from both lack of knowledge and consistency: A PPT analysis of test-taking. *US-China Education Review*, 1(6), 845-855.
- Rice, S., Trafimow, D., & Hunt, G. (2010). Using PPT to analyze sub-optimal human-automation performance. *Journal of General Psychology*, 137(3), 310 - 329. <http://dx.doi.org/10.1080/00221301003645236>
- Rice, S., Trafimow, D., Keller, D., Hunt, G., & Geels, K. (2011). Using PPT to correct for inconsistency in a speeded task. *Journal of General Psychology*, 138(1), 12-34. <http://dx.doi.org/10.1080/00221309.2010.531791>
- Swearingen, D. L. (1998). Person-fit and its relationship with other measures of response set. Paper presented at the *Annual Meeting of the American Educational Research Association*. San Diego, CA.
- Trafimow, D., & Rice, S. (2008). Potential Performance Theory: A general theory of task performance applied to morality. *Psychological Review*, 115(2), 447-462. <http://dx.doi.org/10.1037/0033-295X.115.2.447>
- Trafimow, D., & Rice, S. (2009). Potential Performance Theory (PPT): Describing a methodology for analyzing task performance. *Behavior Research Methods*, 41(2), 359-371. <http://dx.doi.org/10.3758/BRM.41.2.359>
- Trafimow, D., & Rice, S. (2011). Using a sharp instrument to parse apart strategy and consistency: An evaluation of PPT and its assumptions. *Journal of General Psychology*, 138(3), 169-184. <http://dx.doi.org/10.1080/00221309.2011.574173>
- Trafimow, D., Hunt, G., Rice, S., & Geels, K. (2011). Using potential performance theory to test five hypotheses about meta-attribution. *Journal of General Psychology*, 138(2), 1-13. <http://dx.doi.org/10.1080/00221309.2010.540591>
- Trafimow, D., MacDonald, J., & Rice, S. (in press). Using PPT to account for randomness in perception. *Attention, Perception and Psychophysics*.
- U. S. Department of Education, National Center for Education Statistics. (2003). *Digest of education statistics 2002* (NCES 2003-060). Retrieved from <http://nces.ed.gov/pubs2003/2003060.pdf>
- Vattanapath, R., & Jaiprayoon, K. (1999). An assessment of the effectiveness of teaching test-taking strategies for multiple-choice English reading comprehension test. *Occasional Papers*, 8, 57-71.

Weinstein, P., & Gipple, C. (1974). The relationship of study skills to achievement in the first 2 years of medical school. *J Med Educ*, 49, 902–5.

West, C., & Sadoski, M. (2011). Do study strategies predict academic performance in medical school?. *Medical Education*, 45(7), 696-703. <http://dx.doi.org/10.1111/j.1365-2923.2011.03929.x>

Appendix

Plato was a philosopher from Greece

The last man to successfully conquer England was from Normandy

In the Second World War, Italy fought with Germany

The Ancient Egyptians were ruled by Pharaohs

The book 1984 was written by George Orwell

Charles Darwin conducted research that supported the concept of Evolution

Karl Marx is the father of Communist thought

The code name for Germany's invasion of the Soviet Union during World War II was Operation Barbarossa

Napoleon was the leader of France

Julius Caesar conquered Gaul in the First Century BC

The French helped the Americans during the Revolutionary War in the 18th Century

Genghis Khan ruled the Mongol Empire

Saddam Hussein was executed by his own people

The Code of Hammurabi was a system of laws

Islam was founded in the 7th century AD

Augustus Caesar ruled the Roman Empire before Julius Caesar

Written history goes back about 160,000 years

Einstein won the Nobel Prize in history

France never had a colony in Australia

At the end of World War 2, the Allied conference in China decided the partition of Germany

Christopher Columbus was from India

The Hanging Gardens of Egypt were one of the 7 Ancient Wonders of the World

Stamford Raffles is the founder of Tibet

The Church of England was founded by King Edward Lackshanks

Ernesto 'Che' Guevara died in Brazil

Virginia was once an independent country

Pearl Harbor was attacked by the Vietnamese on Dec 7, 1941

Bacteria was discovered by Madame Curie in 1898

The ancient city of Alexandria lay within the boundaries of Turkey

George Washington was America's first king