US-China Education Review B 6 (2011) 845-855
Earlier title: US-China Education Review, ISSN 1548-6613

Our Students Suffer From Both Lack of Knowledge and Consistency: A PPT (Potential Performance Theory) Analysis of Test-Taking

Stephen Rice, Kasha Geels, David Trafimow, Holly Hackett
New Mexico State University, Las Cruces, America

Test scores are used to assess one's general knowledge of a specific area. Although strategies to improve test performance have been previously identified, the consistency with which one uses these strategies has not been analyzed in such a way that allows assessment of how much consistency affects overall performance. Participants completed one of many exams over a variety of educational subjects. PPT (Potential Performance Theory) (Trafimow & Rice, 2008; 2009) was used to analyze their scores. Results indicated that consistency played a large role in affecting observed performance. Had participants been perfectly consistent, their observed performances would have been significantly higher. Analysis of sample individual data revealed that different individuals need to work specifically on either consistency, content knowledge, or both.

Keywords: potential performance theory, education, test-taking

Introduction

In educational settings, tests are often used as a tool to evaluate and assess individuals' learning, knowledge and aptitude. These evaluations and assessments can carry a lot of weight. For instance, a student's final course grade is typically a combination of their test performances throughout the course's duration. Also, standardized test scores (e.g., scholastic aptitude test and graduate record examination) are often used in deciding whether or not an individual will be accepted into a college or university. Given the extreme value placed on test-taking performance, the ability to perform well on tests becomes very important. One valuable question that arises is what influences an individual's test-taking performance.

Our objective is to expand upon the test-taking performance literature by implementing Trafimow and Rice's (2008; 2009) recent contribution of PPT (Potential Performance Theory) (Trafimow & Rice, 2008, 2009; Hunt, Rice, Geels, & Trafimow, 2010; Hunt et al., 2011; Rice, Trafimow, & Hunt, 2010; Rice, Trafimow, Keller, Hunt, & Geels, 2011; Trafimow, Hunt, Rice, & Geels, in press; Trafimow, & Rice, in press). In agreement with previous suggestions regarding test-taking performance (Cohen, 2006), PPT also suggests that consistency influences task performance. The great advantage to applying PPT to the domain of test-taking performance is that it not only allows us to effectively estimate individuals' consistency across tests, but it also allows us to demonstrate how an

Stephen Rice, Ph.D., associate professor, Department of Psychology, New Mexico State University.
Kasha Geels, Department of Psychology, New Mexico State University.
David Trafimow, Ph.D., professor, Department of Psychology, New Mexico State University.
Holly Hackett, Department of Psychology, New Mexico State University.

individual's consistency or inconsistency, influences their actual observed performance.

Test-Taking Performance

Fortunately, some factors and strategies have been identified that can influence test performance. These factors and strategies can either be applied to the test-preparation process or test-taking process. For instance, some influential factors that can be implemented when an individual is preparing for a test include improving retention and retrieval of information and using study strategies appropriately. However, other influential factors that can be implemented when an individual is actually taking a test include engaging in specific test-taking behaviors and applying test-taking strategies consistently.

One obvious way to improve test performance is to increase the knowledge of the test material. From this aspect, memory literature can easily be applied. Deep processing involves close attention to information and focusing on the meaning of the information, whereas, shallow processing involves little attention to the meaning of information. Deep versus shallow processing leads to improved memory (Craik & Tulving, 1975). Similarly, practices that implement deeper processing, such as active and organizational note taking on material (Peverly, Brobst, Graham, & Shaw, 2003; Pressley, Yokoi, Van Etten, & Freebern, 1997), repeated testing over material (Wheeler, Ewers, & Buonanno, 2003; Butler & Roediger, 2007; Roediger & Karpicke, 2006) and practicing retrieval of material (Karpicke, 2009; Pilotti, Chodorow, & Petrov, 2009) lead to better retention and later retrieval of the material compared to simply reviewing or rereading material. Greater retention and retrieval in turn leads to improve test performance.

Another factor that can influence test performance is using appropriate study strategies (Pressley, Yokoi, Van Etten, & Freebern, 1997). In fact, research has found that simply providing appropriate study strategies to students that lack established study strategies can improve their test performance (Fleming, 2002). Simply implementing a strategy is more beneficial that not using any type of strategy at all (McDougall & Gruneberg, 2002). However, it is also important for individuals to know when to apply specific study strategies. For example, different test forms require different preparations (Pressley, Yokoi, Van Etten, & Freebern, 1997). The use of inappropriate strategies for specific tests can lead to lower test performance (Balch, 2007). Beyond the skill of implementing correct study strategies, research has also discovered that simply allocating time for studying can influence test performance. Distributing study time versus cramming has shown to enhance delayed test performance (Smith & Rothkopf, 1984; Reder & Anderson, 1982; Kornell, 2009). In other words, knowing appropriate study strategies and implementing the study strategies in an appropriate manner can increase test performance.

Interestingly, an individual can improve their test performances during the test-taking process regardless of the study strategies that were employed previously or the knowledge that was brought to the test. It has been discovered that some individuals are simply better test takers than others. This ability to take a test well, regardless of the knowledge of test material, and improve test performances has been termed test-wiseness (Millman, Bishop, & Ebel, 1965; Rogers & Yang, 1996).

Several test-taking behaviors or strategies used by test-wise individuals have been identified. For instance, one behavior identified is reading the test questions carefully and completely (Cohen, 2006). Also, test-wise individuals tend to anticipate the answer to the question prior to proceeding to the available options or attempting to complete the question (McClain, 1983). Furthermore, these individuals will review all of the available options completely prior to answering the question. Test-wise individuals also demonstrate good use

of their test time. They tend to skip questions when the question appears to be difficult or when the answer is not known, rather than using excess time focusing on a single question (Hong, M. Sas, & J. C. Sas, 2006). When returning to the skipped questions, these individuals use reasoning and the process of elimination to make educated guesses (McClain, 1983). The final behavior of test-wise individuals includes reviewing their work or their selected answers.

Researchers have successfully demonstrated that training individuals to use one of the foregoing strategies, or a combination of them, can enhance test performance. For example, individuals diagnosed with learning disabilities, emotional or behavioral disorders often demonstrate a lack of test-taking strategies (Scruggs & Tolfa, 1985; Scruggs & Mastropieri, 1986). Numerous experiments have found that when these populations of individuals, whether in elementary school, middle school or college, are provided with useful test-taking strategies and training to utilize the strategies, their test performances improve (Scruggs & Mastropieri, 1986; Therrien, Hughes, Kapelski, & Mokhtari, 2009; Ritter & Idol-Maestas, 1986; Holzer, Madaus, Bray, & Kehle, 2009; Hughes, Deshler, Ruhl, & Schumaker, 1993). These experiments would suggest that the strategies and behaviors demonstrated by test-wise individuals can, in fact, assist with the test-taking process and contribute to better test performance.

Even though being provided with test-taking strategies is valuable and can improve test performance, there is another influential factor worth mentioning. An individual's ability to successfully apply test-taking strategies and apply the strategies with high frequency may also influence test performance (Cohen, 2006). For instance, researchers have found that individuals who applied most of the trained test-taking strategies improved their test performance more than individuals who did not apply the majority of the test-taking strategies (Therrien, Hughes, Kapelski, & Mokhtari, 2009; Holzer, Madaus, Bray, & Kehle, 2009). These findings suggest that even if an individual is aware of test-taking strategies, applying the strategies inconsistently and inappropriately may prevent the individual from improving their test-taking performance to their full potential. Fortunately, we now have an appropriate procedure to investigate and address this observation by utilizing PPT (Trafimow & Rice, 2008; 2009).

PPT (Potential Performance Theory)

As previously stated, PPT suggests that consistency plays an influential role in one's observed task performance. It is worth mentioning that consistency in terms of PPT is across blocks of trials which should not be confused with any other sense of the term. Different fields of research have used the term "consistency" to refer to several different phenomena. For instance, attribution research uses consistency to refer to a person responding similarly across different situations (Kelley & Michela, 1980; Orvis, Cunningham, & Kelley, 1975). According to work on the Lens model (Brunswik, 1952), consistency is used to refer to the correlation between predicted judgments and actual judgments, or the predicted criterion value and the actual criterion value. Furthermore, personality research uses consistency to refer to stability of personality traits. PPT uses yet another sense of the term, where consistency refers to the correlation across two blocks of matched trials.

The following paradigm demonstrates the function of consistency on task performance. Suppose students take a 50 true-false question history test in Block 1. Then, after completing a short irrelevant task, the students take the same 50 true-false question history test again in Block 2. This procedure produces 50 pairs of answers for each student. Given this design, correlation coefficients across the two blocks for each student can easily be computed to obtain each student's consistency coefficient. Naturally, if there is more randomness (i.e., student

provides different answers on corresponding questions across the blocks), then the consistency coefficient decreases; whereas, if there is less randomness (i.e., students provide the same answers on corresponding questions across blocks), the consistency coefficient will increase. In other words, the consistency coefficient can be viewed as an inverse measure of randomness where zero randomness would result in a perfect consistency coefficient.

When an individual's base level performance is above chance level, introducing randomness will undoubtedly decrease their actual observed performances. Similarly, decreasing randomness should increase actual observed performances. PPT takes advantage of this statistical fact. For example, using the PPT paradigm of dichotomous items, observed task performance will shift towards 50% when inconsistency (i.e., randomness) is present. Similarly, observed task performance will shift towards 100% when randomness is decreased. By taking an individual's actual observed performance and consistency score, PPT provides a procedure to effectively estimate how that individual would have performed in the absence of randomness (i.e., individual responded with perfect consistency), which is termed the individual's potential performance.

Once an individual's observed performance and consistency is obtained using the aforementioned paradigm, an individual's potential performance can be computed using a step-by-step procedure proposed by PPT. The first step involves converting an individual's observed performance into a correlation coefficient. To reiterate the PPT proposed paradigm, individuals are forced to choose between two answers (e.g., true or false) with the correct answer being one of the two choices. Using this paradigm, each individual's observed performance across the trials can easily be summarized by a 2 × 2 table (see Table 1). The cell frequencies are noted as $a, b, c,$ and d, the row frequencies are noted as r_1 and r_2, and the column frequencies are noted as c_1 and c_2. Once this actual performance table is completed, implementing Equation (1) will convert this table into a correlation coefficient.

$$r = \frac{|ad-bc|}{(a+b)(c+d)(a+c)(b+d)} = \frac{|ad-bc|}{r_1 r_2 c_1 c_2} \tag{1}$$

Table 1

A 2 (Individual Choice) × 2 (Correct Choice) Frequency Table Where a, b, c, and d Indicate the Actually Observed Cell Frequencies, r_1 and r_2 Are the Row Frequencies, and c_1, and c_2 Are the Column Frequencies

Correct choice	Individual's choice		Row margin
	True	False	
True	*a*	*b*	r_1
False	*c*	*d*	r_2
Column margin	c_1	c_2	

The second step of obtaining an individual's potential performance involves adjusting the aforementioned correlation coefficient for attenuation due to inconsistency. This can be accomplished by using the famous formula that was originally derived from classical true score theory, though it can also be derived from more modern theories (Allen & Yen, 1979; Cohen & Swerdlik, 1999; Crocker & Algina, 1986; Gulliksen, 1987; Lord & Novick, 1968). In Equation (2), the "True" correlation coefficient (*R*) is expressed as a function of the correlation obtained in Equation (1) (*r*) and the consistency coefficient of the dependent variable across blocks of trials ($r_{xx'}$).

$$R = \frac{r}{\sqrt{r_{xx'}}} \tag{2}$$

The next step in obtaining an individual's potential performance is to convert the "True" correlation coefficient (*R*) value obtained in Equation (2), into "True" cell frequencies. In other words, we want to build another 2 × 2 table with the "True" cell frequencies (see Table 2). These "True" cell frequencies represent the best estimates of cell frequencies if an individual was perfectly consistent. By utilizing Equations (3)-(6), one can obtain an individual's "True" cell frequencies (Trafimow & Rice, 2008). Similar to a Chi-square test or a Fisher's exact test, PPT also fixes the margin frequencies at the obtained levels (Trafimow & Rice, 2009). Given this assumption and the fact that we are now using "True" cell frequencies versus actual cell frequencies, row frequencies are noted as R_1 and R_2, column frequencies are noted as C_1, and C_2 and the cell frequencies are noted as $A, B, C,$ and D.

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

$$B = \frac{R_1(R_1 + R_2) - \left(R\sqrt{R_1 R_2 C_1 C_2} + C_1 R_1\right)}{(R_1 + R_2)} \tag{4}$$

$$C = \frac{C_1 R_2 - R\sqrt{R_1 R_2 C_1 C_2}}{(R_1 + R_2)} \tag{5}$$

$$D = \frac{C_2(R_1 + R_2) - \left[R_1(R_1 + R_2) - \left(R\sqrt{R_1 R_2 C_1 C_2} + C_1 R_1\right)\right]}{(R_1 + R_2)} \tag{6}$$

Table 2

A 2 (Individual Choice) X 2 (Correct Choice) Frequency Table Where $A, B, C,$ *and* D *Indicate the True Cell Frequencies,* R_1 *and* R_2 *Are the Row Frequency, and* C_1*, and* C_2 *Are the Column Frequencies*

Correct choice	Individual's choice		Row margin
	True	False	
True	*A*	*B*	R_1
False	*C*	*D*	R_2
Column margin	C_1	C_2	

To summarize, the PPT steps to obtain an individual's potential performance are as follows. First an individual's answers are obtained across two blocks and placed into a 2 × 2 actual performance table. Then, using Equation (1), the table can be converted into a correlation coefficient. Second, using Equation (2), the obtained correlation coefficient is corrected for attenuation due to inconsistency. Next, using Equations (3)-(6), estimates of "True" cell frequencies (i.e., cell frequencies that would be obtained in the absence of randomness) are obtained and placed into a new 2 × 2 table. Finally, using Equation (7), an individual's potential performance, or the performance an individual would have had if they had been perfectly consistent across the two blocks, is obtained.

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

Current Study

It is evident from the previous literature that test-taking performance can be influenced by both the knowledge one brings to a test and the test-taking strategies one employs while taking a test. It is no surprise that the more knowledge one has of the test material, the better their performances will be on the test. More interesting is the suggestion and previous observations that greater consistency can improve one's performance on a test. Recognizing that both knowledge and consistency can influence peoples' observed test performance

is beneficial. However, effectively identifying the extent to which one's knowledge, consistency, or both are actually influencing their observed test performance is a valuable tool.

PPT now provides us with a procedure to effectively identify how knowledge and consistency influence observed performance. If one can identify whether or not and to what extent knowledge and consistency are influencing a person's observed performance, then effective training may be implemented to improve that person's performance. The current study provides a demonstration of how PPT may be utilized as a useful tool in the area of test-taking performance.

Experiment

Participants

A total of 346 (244 females) undergraduate students from a large Southwestern university participated in the experiment for partial course credit. The mean age was 21.09 (standard deviation = 4.09). All participants were tested for normal or corrected-to-normal vision.

Materials and Stimuli

The experimental display was presented via E-prime 1.1 on a Dell PC (personal computer) with a 22-monitor using 1,024 ×768 resolution and a refresh rate of 65 Hz. A series of 50 questions were generated for each of 13 subjects/topics (algebra, grammar, biology, astronomy, US history, world history, spelling, analogies, geography, range problems, state capitals, digit memory and spatial memory). These questions were presented in random order. In order to facilitate a PPT analysis, the same questions were presented again in a second block and again in random order. All questions were presented as "True/False". For example, in the algebra subject test, a sample false statement was (if $-9x + 4 = -20$; $x = 3$). A sample true statement in the world history test was "In the Second World War, Italy fought with Germany".

Procedure

Participants first signed a consent form and were then seated comfortably in a chair facing the experimental display. Instructions were presented on-screen and participants were given an opportunity to ask questions before beginning the test. Each trial began with a fixation point, which was a small black "+" in the middle of the screen lasting 1,000 msec. Following this was the question display, where the test statement was presented in black Arial font size 14 across the middle of the display for 5,000 msec. Lastly was the choice display, whereby participants were asked to press "J", if they thought the statement was "True" or "False", if they thought the statement was false.

Design

A between-participants design was used by which different participants answered each of the different exam topics.

Results

Based on observed scores and consistency coefficients, PPT provides potential scores that estimate how participants would perform if they were perfectly consistent. Because PPT is sufficiently flexible or both group level and individual level analyses, both will be presented below. The subsection on group analyses provides results that are averaged across the participants, whereas the subsequent subsection provides a few cases of particularly interesting results for individuals.

Group Analyses

There was considerable variation both in the observed scores and in the consistency coefficients across the 13 test areas. Observed scores were largest for the digit memory test (M = 0.79) and smallest for the geography test (M = 0.61), and the mean across test areas was 0.72. Consistency coefficients were largest for the astronomy test (M = 0.78) and smallest for the spatial memory test (M = 0.38), and the mean across test areas was 0.61. These data, in combination with PPT, allow for the computation of potential scores and also the difference between observed and potential scores. Potential scores were largest for the digit memory test (M = 0.99) and smallest for the world history test (M = 0.66), and the mean across test areas was 0.80. But the most important statistics, from the practical perspective of assessing the likely impact of interventions designed to increase consistency, concern the differences between observed and potential scores. This difference was largest for the digit memory test (difference = 0.20) and smallest for the astronomy test (difference = 0.03), for an average difference across the eleven test areas of 0.08. The straightforward interpretation is that training people to respond more consistently could potentially improve performance by an average of 20% for the digit memory test, whereas this figure would only be 3% for the astronomy test. Thus, education efforts with respect to digit memory should be heavily weighted towards improving consistency, whereas education efforts with respect to astronomy should be weighted towards improving actual knowledge.

Individual Analyses

An advantage of PPT is that it is not necessary to depend on group analyses, such as those performed in the foregoing subsection. A few individual analyses will illustrate the potential for improving observed scores merely by improving consistency.

We first consider the algebra test, where the average difference between observed and potential scores was 0.13. In contrast, considering one participant with an observed score equal to only 0.68 but after accounting for consistency, which made a whopping difference of 0.32, this person’s potential score was precisely 1.00. The implication is that if this participant could be trained to respond to algebra questions with perfect consistency, the result would be perfect observed performance. The individual analysis illustrates the importance of not depending only on group data, as the importance of consistency was much greater for this participant than that for the group mean.

Let us now consider the world history test, where the average difference between observed and potential scores was a relatively small, but nevertheless important, 0.04. But one participant’s observed score was 0.66 and the potential score was 0.83, for a difference of 0.17—more than fourfold the average difference for the group. Again, we see that averaged data poorly represent some of the individuals. Even when group data suggest that consistency is of relatively less importance, on average, for a particular test, consistency nevertheless can be extremely important for particular individuals. There are many more examples of particular individuals who were poorly represented by the group data, but the two example presented are sufficient for illustration.

Discussion

The foregoing analyses indicate that tests in different subjects can vary widely in observed scores, consistency coefficients and potential scores. There also is much variance across subjects in the differences between potential and observed scores, thereby similarly indicating that merely training students to be more consistent is likely to provide much stronger benefits in some subjects than in others. Possibly more important,

there were substantial deviations from group data when individuals were analyzed. These deviations indicated that even in cases where consistency is not very important at the group level (e.g., the astronomy test), it nevertheless is important for some students in the group.

As with most studies, the most obvious limitations pertain to generalizing from the findings. Although there were 346 participants in total, when the fact that there were 13 subjects was considered, it becomes clear that there was an average of less than 27 participants per subject area. Also, our participants were undergraduate students in the Southwest and we doubtless would have obtained different findings with undergraduates from other areas or people who are not undergraduates. In addition, we used items with only two choices (“True” or “False”) and most academic tests involve more than two choices. Because of these considerations, we advise caution in attempting to generalize from the obtained findings.

An interesting difference between the present study and many other studies in the education area is that we used PPT, whereas it is more common to use IRT (item response theory). Although this is not the place to engage in a detailed comparison between the two theories, it seems worthwhile to present a brief contrast. Lord and Novick (1968) noted that some measurement theories are based on strong assumptions, whereas others are based on weak ones. The advantage of strong assumptions is that it is easier to draw strong conclusions that can be usefully applied. The advantage of weak assumptions is that they are more likely to be true. PPT and IRT provide an interesting case in point. IRT makes at least two strong assumptions that are unnecessary for PPT (Hulin, Drasgow, & Parsons, 1983; Lord, 1970). First, there is a single latent trait underlying test scores. Second, the items are locally independent, which means that the only reason for their inter-correlations is by dint of the latent trait. If the effects of the latent trait were taken away, the items would not correlate with each other at all. We believe neither of these assumptions is likely to be true with respect to the present paradigm. It is difficult for us to believe that only a single latent trait (e.g., knowledge of the subject area) underlies test scores. Rather, we believe that additional systematic factors, such as general intelligence, test taking skills, effort and so on, which are likely to influence test scores. In addition, if several systematic factors influence test scores, then it seems obvious that the assumption of local independence is also incorrect. The removal of only one of these influences, if it could be done, would seem unlikely to eliminate inter-correlations among the items. It is important to be clear that we are not criticizing IRT, which is an extremely important theory, but rather to point out that there might be times when the researcher might be unwilling to accept the strong assumptions that are necessary to make the IRT machinery run. In such cases, PPT might be preferred. The foregoing analyses, along with Figures 1 and 2, demonstrate that PPT can generate useful conclusions, even with weaker assumptions.

One advantage of PPT, from an applied perspective, is that it dovetails nicely with some of the current thinking in education pertaining to the importance of individualizing training to fit the idiosyncratic characteristics of each student. Because PPT provides observed scores, consistency coefficients and potential scores for each student, it facilitates the development of individualized training programs. The difference between potential and observed scores can be computed easily for each student and training can be brought to bear according to these differences. When the difference is large, then training for consistency is likely to cause a large increase in observed performance. But when the difference is small, training for consistency is unlikely to have much of an effect on observed performance. Therefore, large difference scores indicate that training should be devoted towards increasing consistency, whereas small difference scores indicate that training should be devoted towards increasing potential scores (e.g., the focus could be on increasing content knowledge).

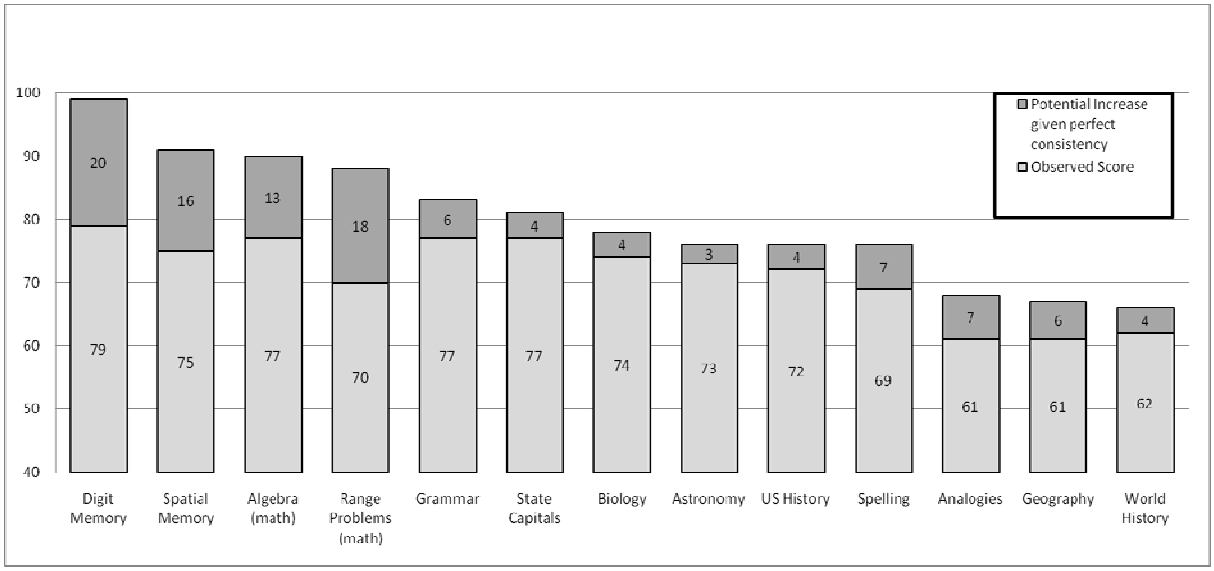

Figure 1. Group data.

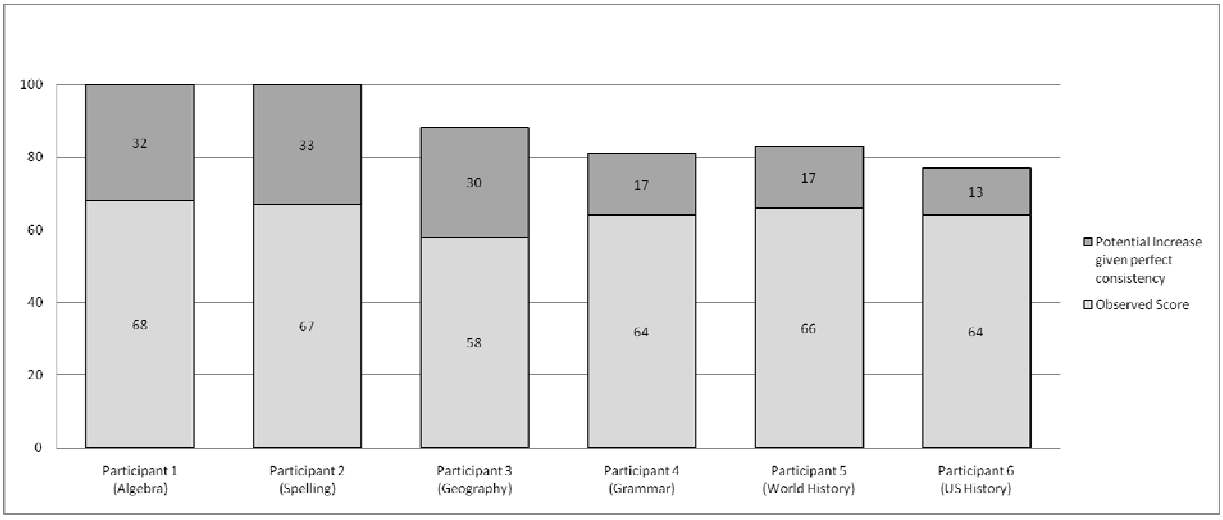

Figure 2. Sample individual data.

There is an additional way of considering the issue of training, which is to explore what is unlikely to work. Recall that one of the participants received a score of 0.68 but that person's potential score was 1.00. The normal conclusion would be that a student with such a low observed score simply does not know the material. The potential score of 1.00 contradicts this conclusion. If all inconsistency could be eliminated, the student would perform perfectly. Thus, training for knowledge would be unlikely to substantially help this student. Rather, the focus should be on training for consistency.

As we stated earlier, given the aforementioned limitations, we make no claims that the present findings would generalize. However, we do believe that the method of using PPT to analyze test performance does generalize. Whatever the idiosyncratic characteristics happen to be with respect to the population of interest, the test being used and the PPT paradigm will provide numbers that are useful for designing training for students. Such training programs can be designed at the group or individual levels. The present study is the first demonstration and illustration of this intriguing possibility.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, C. A.: Brooks/Cole Publishing Company.

Balch, W. R. (2007). Effects of test expectation on multiple-choice performance and subjective ratings. *Teaching of Psychology, 34*(4), 219-225.

Brunswik, E. (1952). *The conceptual framework of psychology: International encyclopedia of unified science*. Chicago: University of Chicago Press.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514-527.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Learning Assessment Quarterly, 3*(4), 307-331.

Cohen, R. J., & Swerdlik, M. E. (1999). *Psychological testing and assessment: An introduction to tests and measurements* (4th ed.). Mountain View, C. A.: Mayfield.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology, 104*(3), 268-294.

Crocker, L., & Algina, J. (1986). *Introductions to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Fleming, V. M. (2002). Improving students' exam performance by introducing study strategies and goal setting. *Teaching of Psychology, 29*(2), 115-119.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale: N. J.: Lawrence Erlbaum Associates, Publishers.

Holzer, M. L., Madaus, J. W., Bray, M. A., & Kehle, T. J. (2009). The test-taking strategy intervention for college students with learning disabilities. *Learning Disabilities Research and Practice, 24*(1), 44-56.

Hong, E., Sas, M., & Sas, J. C. (2006). Test-taking strategies of high and low mathematics achievers. *Journal of Educational Research, 99*(3), 144-155.

Hughes, C. A., Deshler, D. D., Ruhl, K. L., & Schumaker, J. B. (1993). Test-taking strategy instruction for adolescents with emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders, 1*(3), 189-198.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, Illinois: Dow Jones-Irwin.

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, C. A..

Hunt, G., Rice, S., Trafimow, D., Schwark, J., Sandry, J., Busche, L., & Geels, K. (2011). Visual vs. auditory memory in an aviation task: A potential performance theory analysis. Proceedings of *the 16th Annual International Symposium of Aviation Psychology*. Dayton, O. H..

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology, 138*(4), 469-486.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology, 31,* 457-501.

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*(9), 1297-1317.

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—A confrontation of Birnbaum's logistic model. *Psychometrika, 35,* 43-50.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley Publishing Company.

McClain, L. (1983). Behavior during examinations: A comparison of "A", "C", and "F" students. *Teaching of Psychology, 10*(2), 69-71.

McDougall, S., & Gruneberg, M. (2002). What memory strategy is best for examinations in psychology? *Applied Cognitive Psychology, 16,* 451-458.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*(3), 707-726.

Orvis, B. R., Cunningham, J. D., & Kelley, H. H. (1975). A closer examination of causal inference: The roles of consensus, distinctiveness, and consistency information. *Journal of Personality and Social Psychology, 32*(4), 605-616.

Peverly, S. T., Brobst, K. E., Graham, M., & Shaw, R. (2003). College adults are not good at self-regulation: A study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology, 95*(2), 235-346.

Pilotti, M., Chodorow, M., & Petrov, R. (2009). The usefulness of retrieval practice and review-only practice for answering conceptually related test questions. *Journal of General Psychology, 136*(2), 179-203.

Pressley, M., Yokoi, L., Van Etten, S., & Freebern, G. (1997). Some of the reasons why preparing for exams is so hard: What can be done to make it easier? *Educational Psychology, 9,* 1-38.

Reder, L. M., & Anderson, J. R. (1982). Effects of spacing and embellishment on memory for the main points of a test. *Memory and Cognition, 10*(2), 97-102.

Rice, S., Trafimow, D., & Hunt, G. (2010). Using PPT to analyze sub-optimal human-automation performance. *Journal of General Psychology, 137*(3), 310-329.

Rice, S., Trafimow, D., Keller, D., Hunt, G., & Geels, K. (2011). Using PPT to correct for inconsistency in a speeded task. *The Journal of General Psychology, 138*(1), 12-34.

Ritter, S., & Idol-Maestas, L. (1986). Teaching middle school students to use a test-taking strategy. *Journal of Educational Research, 79*(6), 350-357.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255.

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment, 12*(3), 247-259.

Scruggs, T. E., & Mastropieri, M. A. (1986). Improving the test-taking skills of behaviorally disordered and learning disabled children. *Exceptional Children, 53,* 63-68.

Scruggs, T. E., & Tolfa, D. (1985). Improving the test-taking skills of learning-disabled students. *Perceptual and Motor Skills, 60,* 847-850.

Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction, 1*(3), 341-358.

Trafimow, D., & Rice, S. (2008). Potential performance theory: A general theory of task performance applied to morality. *Psychological Review, 115*(2), 447-462.

Trafimow, D., & Rice, S. (2009). Potential performance theory (PPT): Describing a methodology for analyzing task performance. *Behavior Research Methods, 41*(2), 359-371.

Trafimow, D., & Rice, S. (in press). Using a sharp instrument to parse apart strategy and consistency: An evaluation of PPT and its assumptions. *Journal of General Psychology.*

Trafimow, D., Hunt, G., Rice, S., & Geels, K. (in press). Using potential performance theory to test five hypotheses about meta-attribution. *Journal of General Psychology*.

Therrien, W. J., Hughes, C., Kapelski, C., & Mokhtari, K. (2009). Effectiveness of test-taking strategy on achievement in essay tests for students with learning disabilities. *Journal of Learning Disabilities, 42,* 14-23.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571-580.