SUMMATIVE ASSESSMENT STRATEGIES FOR STATISTICAL LEARNING: DEVELOPMENT, ADMINISTRATION, AND SCORING OF AUTHENTIC AND PERFORMANCE ASSESSMENTS

DICKINSON, Wendy B.
Ringling College of Art and Design
USA

ONWUEGBUZIE, Anthony J., HINES, Constance, HALL, Bruce W.
University of South Florida
USA

This paper describes three original assessments developed for use in undergraduate- and graduate-level mathematics and statistics courses: (a) the context-dependent item set (undergraduate); (b) the visual data display project (undergraduate); and (c) the statistics notebook (graduate). The goal of each assessment was to measure student learning of statistical concepts and methodology, gauge the student's ability to apply those concepts in context, and provide an opportunity for students to appreciate statistics as a way to investigate, summarize, and explain phenomena of interest. Examples of all three assessments and rubrics are available for presentation.

PERFORMANCE ASSESSMENT AND AUTHENTIC ASSESSMENT: DEFINITIONS AND CONCEPTUALIZATIONS.

Some years ago, Newmann and Archbald (1992) noted that "what counts for success in school is often considered trivial, meaningless, and contrived—by students and adults alike" (p. 71). They noted that when we ask students to recognize the difference between verbs and nouns, to match authors with their works, or to label correctly rocks or body parts, we are asking for the mere reproduction of knowledge and, thereby, failing to challenge students with assessments measuring outcomes that represent "appropriate, meaningful, significant, and worthwhile forms of human accomplishment" (Newmann & Archbald, 1992, p. 72). According to these authors, students should be required to engage in disciplined inquiry that produces knowledge that has value in their lives. Mastery of this sort would be demonstrated through the completion of long-term projects that involve discourse, performances, and products of interest to students, their peers, and the public at large (Newmann & Archbald, 1992).

At all levels of the educational enterprise, from elementary school through advanced programs of higher education, many educators have begun to argue in favor of increased authenticity in the assessment of students, primarily through the use of *performance assessments or authentic assessments.* As described by McMillan (2004), "simply put, a *performance assessment* is one in which the teacher observes and makes a judgment about the student's demonstration of a skill or competency in creating a product, constructing a response, or making a presentation" (p. 198). He adds, "the emphasis is on the students' ability to perform tasks by producing their own work with their knowledge and skills" (p. 198). According to Stiggins (1994), in the case of the demonstration of a skill " . . . evidence of achievement is seen in the respondent's ability to carry out the proper sequence of activities or to do something in the appropriate manner. It is the doing that counts" (p. 86). In the case of the evaluation of a product, Stiggins (1994) notes that " . . . the respondent creates a complex achievement-related product that is intended to have certain attributes. The assessor examines the tangible product to see if those attributes are indeed present" (p. 86). He also notes that in product evaluation the focus typically is on the attributes of the product itself, although the process of creating the product can also be evaluated (Stiggins, 1994, p. 86).

Authentic assessment, according to McMillan (2004), " . . . involves the direct examination of a student's ability to use knowledge to perform a task that is like what is encountered in real life or in the real world" (p. 198). Authenticity is judged in terms of the nature of the task completed as well as the context of the task. Like any performance assessment, the students “plan, construct, and deliver an original response, and explain or justify their answers” (McMillan, 2004, p. 198). According to Newmann and Archbald (1992), the most critical attribute for authentic achievement is that it has “*aesthetic*, *utilitarian*, or *personal value* apart from documenting the competence of the learner” (p. 73). They contend that authentic achievements are those that have a special value that is missing in tasks that are contrived only for the purpose of assessing knowledge (Newmann & Archbald, 1992, p. 73). Haladyna (2002) adds that “cognitive learning theory and the social/constructivist learning theory seem to favor teaching and testing that happen in a natural context where students see the merits of what they are learning” (p. 194). Nitko and Brookhart (2007) provide the following perspective on authenticity:

> . . . the ‘authentic’ in *authentic assessment* usually means presenting students with tasks that are directly meaningful to their education instead of indirectly meaningful. For example, reading several long works and using them to compare and contrast different social viewpoints is directly meaningful because it is the kind of thoughtful reading educated citizens do. Reading short paragraphs and answering questions about the ‘main idea’ or about what the characters in the passage did, on the other hand, is indirectly meaningful because it is only one fragment or component of the ultimate learning target of realistic reading. (p.253)

Those who advocate the increased use of performance assessments and authentic assessments tend to agree that these types of assessments require that scoring criteria or rubrics be carefully developed prior to the application of the assessment and, moreover, that these criteria or rubrics be shared with the students who are to be assessed (e.g., Ryan, 1994; Stiggins, 1994; Elliott, 1995; McMillan, 2004). Nitko and Brookhart (2007) describe a scoring rubric as “a coherent set of rules you use to assess the quality of a student’s performance” (p. 245). These rules guide the judgments of the assessor and ensure that the judgments are applied consistently. The authors assert that performance activities that lack such scoring rubrics cannot qualify as assessments (Nitko & Brookhart, 2007, p. 245).

From the above discussion, it should be evident that performance assessment and authentic assessment have much in common, but it should also be evident that some performance assessments may not be authentic, and those that are may vary in the degree of authenticity they attain (McMillan, 2004; Nitko & Brookhart, 2007). As noted by Hanna and Dettmer (2004), “unfortunately, authenticity has formidable costs; when the context for performance assessment is highly authentic, it often is *not* uniform and/or is *not* economical to assess” (p. 212). They add that “authenticity (i.e., realism), economy (i.e., practicality), and reliability (i.e., consistency) are difficult to pursue concurrently; they often vie for our favor” (p. 212). This leads them to suggest the use of simulation as a desirable compromise for classroom assessment--in effect the tolerance of some degree of artificiality in the content or setting of the assessment (Hanna & Dettmer, 2004, p. 216).

The assessment dilemmas faced by teachers throughout the educational enterprise are no less evident within higher education, and specifically within college-level statistics courses. Onwuegbuzie and Leech (2003) provide a thorough discussion of the issues related to attaining *statistical authenticity* in instruction and assessment within the statistics classroom, and describe the role that both performance assessment and authentic assessment can play in monitoring and assessing students’ statistical outcomes. They argue that both types of assessment provide statistics instructors with useful tools to evaluate both the processes and the products that result from student performance.

EXAMPLES OF PERFORMANCE AND AUTHENTIC ASSESSMENTS

Context-dependent Item Sets

“Conventional tests typically only ask the student to select or write the correct response – irrespective of reason” (Wiggins, 1990, p. 1). According to Chance (2002), meaning is provided when numbers are interpreted in context. By developing context-dependent item sets that require the students to graph, analyze, and interpret data, we can assess higher-order thinking skills and instill a sense of “connectedness” between the data itself and the world we inhabit. An example of a performance assessment is the context-dependent item set shown in Appendix A.

The context-dependent item set developed for the undergraduate statistics course provides the student with data describing sumo wrestlers, their weights, and the respective number of tournament wins for each wrestler. By graphing the data, and interpreting the resultant display, students can observe the relationship between weight and number of wins for these wrestlers. Providing the context of sumo wrestling engaged students as they crafted a written and visual response, and advanced the idea of statistics and mathematical reasoning as a way to summarize and explain phenomena of interest.

Visual Data Display Project

The visual display project (Appendix B) is an example of an authentic assessment.
Images can be visual renditions or representations of ideas, dimensions, and events (Dickinson, 2001). Creating an image of quantitative data provides an opportunity for students to summarize, examine, and interpret the variables of interest. Wainer (2005a) notes that by collecting their own data, students regard a project as more “real” than when using a dataset from a textbook or other source. Students are able to document their data visually, providing a narrative design to communicate quantitative information.

The updated Bloom’s Taxonomy, modified by Anderson and Krathwohl (2001), places emphasis on the highest category of cognition—that is, ‘create’. For visual display projects, students use high-level cognitive skills within this category to create their visual displays. Tufte (1990) assert that “to document and explain a process, to make verbs visual, is at the heart of information design” (p. 55). Wainer (2005b) reports, “an efficacious way to add context to statistical facts is by embedding them in a graphic” (p. 86). Examples of the visual display projects effectively communicate the context of the data collected, within the framework of an authentic student assessment (Dickinson & Hall, 2006).

Baker (2005) contends that student assessments have to be scored, or otherwise judged, to determine the level of performance. Rubrics were developed for each of these assessments. Letter grades are associated with each rubric level and are used to represent the relative quality of student achievement (Johnson & Johnson, 2002).

Statistics Notebook

In the graduate-level statistics course, this performance assessment involved students analyzing data provided by the instructor for each statistical technique taught in the course.
Students then complete a “statistics notebook” detailing their analysis and interpretation of the provided data (Onwuegbuzie, 2003; Onwuegbuzie & Leech, 2003). Scoring rubrics were developed for this assessment by the instructor (Wilson & Onwuegbuzie, 1999). In addition, the instructor provides detailed editorial feedback for each notebook—preferably copyediting the students’ write-up—such that students can build on each previous notebook report. This notebook is extremely flexible because it can be used in introductory (e.g., correlation, *t*-tests), intermediate (e.g., multiple analysis of variance, discriminant analysis), and advanced (e.g., structural equation model, confirmatory factor analysis, hierarchical linear modeling) statistics courses.

EDUCATIONAL IMPLICATIONS

As statistics instructors, we endeavor to expand our efforts to incorporate in our classes assessments that are innovative and that more accurately measure students’ ability to apply statistical knowledge and skills to solving meaningful problems. The three assessment protocols described above are examples of the type of assessments that are designed to present students with tasks situated in contexts that are similar to ones that they would encounter in the real world. For tasks of this nature, students’ skills are not measured in isolation. Rather, students are required to construct responses and through their performance, the instructor is provided some indication of their levels of understanding of statistical concepts and their ability to apply this knowledge to real-life situations. Both performance and authentic assessments allow for statistics instructors to promote and assess students’ higher-order thinking and problem solving skills. Indeed, they are examples of what Derry, Levin, and Schauble (1995) termed *statistical* *authencity*. Assessments also influence both our methods of instruction and the content of instruction. According to Onwuegbuzie and Leech (2003), assessment should be treated as an "essential component of statistics classes that influences and is influenced by the context, content and pedagogical style" *of the instructor* (p. 124).

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (Eds.) (2001). *A taxonomy of learning, teaching, and assessment: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Baker, E. L. (2005). Technology and effective assessment systems. In the 104th Yearbook of the National Society for the Study of Education (pp. 358-378). Malden, MA: Blackwell Publishing.

Bermant, M. (2007). Retrieved from March 31, 2007http://www.PlasticSurgery4U.com/.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3). Retrieved November 29, 2006, from http://www.amstat.org/publications/jse/v10n3/chance.html

Derry S., Levin, J. R., & Schauble, L. (1995). Stimulating statistical thinking through situated simulations, *Teaching of Psychology, 22*, 51-57.

Dickinson, W. B. (2001). Escaping Flatland: Chernoff's faces revisited. *Proceedings of the SAS Users Group International Conference: SUGI 26, SAS Institute, Cary: NC.*

Dickinson, W. B., & Hall, B. W. (2006). Developing pictorial space: Instructional strategies for statistical graphics. *Proceedings of the Seventh International Conference on Teaching Statistics*. Salvador, Brazil.

Elliott, S.H. (1995). *Creating meaningful performance assessments*. ERIC Digest E531. (ERIC document reproduction service No. ED 381 985)

Haladyna, T. M. (2002). *Essentials of Standardized Achievement Testing: Validity and Accountability*. Boston: Allyn and Bacon.

Hanna, G. S., & Dettmer, P.A. (2004). *Assessment for effective teaching: Using context-adaptive planning.* Boston: Pearson Education.

Johnson, D. W., & Johnson, R. T. (2002) *Meaningful assessment: A manageable and cooperative process*. Boston, MA: Allyn and Bacon.

McMillan, J.H. (2004). *Classroom Assessment: Principles and Practice for Effective Instruction* (3rd edition). Boston: Pearson Education.

Newmann, F.M. and Archbald, D.A. (1992). "The Nature of Authentic Academic Achievement" (in *Toward a New Science of Educational Testing and Assessment,* A.R. Tom, Editor.). Albany, NY: State University of New York.

Nitko, A.J. and Brookhart, S.M. (2007). Educational Assessment of Students (5th Edition). Upper Saddle River ,NJ: Pearson Education.

Onwuegbuzie, A.J. (2003). Teaching Statistics Courses: Some important considerations. *Academic Exchange Quarterly, 7,* 319-325.

Onwuegbuzie, A.J. & Leech, N. L. (2003). Assessment in Statistics Courses: More than a tool for evaluation. *Assessment and Evaluation in Higher Education, 28*, 115-128.

Ryan, C.D. (1994). Authentic Assessment. Westminster, CA: Teacher Created Materials, Inc.

Stiggins, R.J. (1994*). Student-Centered Classroom Assessment.* New York: Macmillan College Publishing.

Tufte, E. (1990). *Envisioning Information*. Connecticut: Graphics Press.

Wainer, H. (2005a). Old Mother Hubbard and the United Nations: An adventure in exploratory data analysis. *Chance, 18(3),* 38-45.

Wainer, H. (2005b). Graphic Discovery: A Trout in the milk and other visual adventures. Princeton, NJ: Princeton University Press.

Wiggins, G. (1990). *The case for authentic assessment*. ERIC Digest (ERIC Document Reproduction Service No. ED328611)

APPENDIX A

Sumo Wrestling – scatterplot graph

Sumo is “ a Japanese style of wrestling and Japan's national sport. It originated in ancient times as a performance to entertain the Shinto gods. Many rituals with religious background are still followed today. The basic rules of sumo are simple: The wrestler who either first touches the floor with something else than his sole or leaves the ring before his opponent, loses. The fights themselves usually last only a few seconds and in rare cases up to one minute or longer”. [Source: http://www.japan-guide.com/e/e2080.htm]

Sumo Wrestlers often bulk up for their profession. During wrestling, “the chest and stomach remains exposed revealing the typical fat distribution of men with massive weight”. [Source: © 1996-2006 Michael Bermant, MD] Six tournaments are held every year, each one lasting 15 days. Three of the tournaments are held in Tokyo (January, May, September), and one each in Osaka (March), Nagoya (July), and Fukuoka (November).

Left: Archival image of sumo wrestler

Images courtesy of Michael Bermant, M.D.
http://www.PlasticSurgery4U.com/

Below: photograph of contemporary Sumo wrestlers

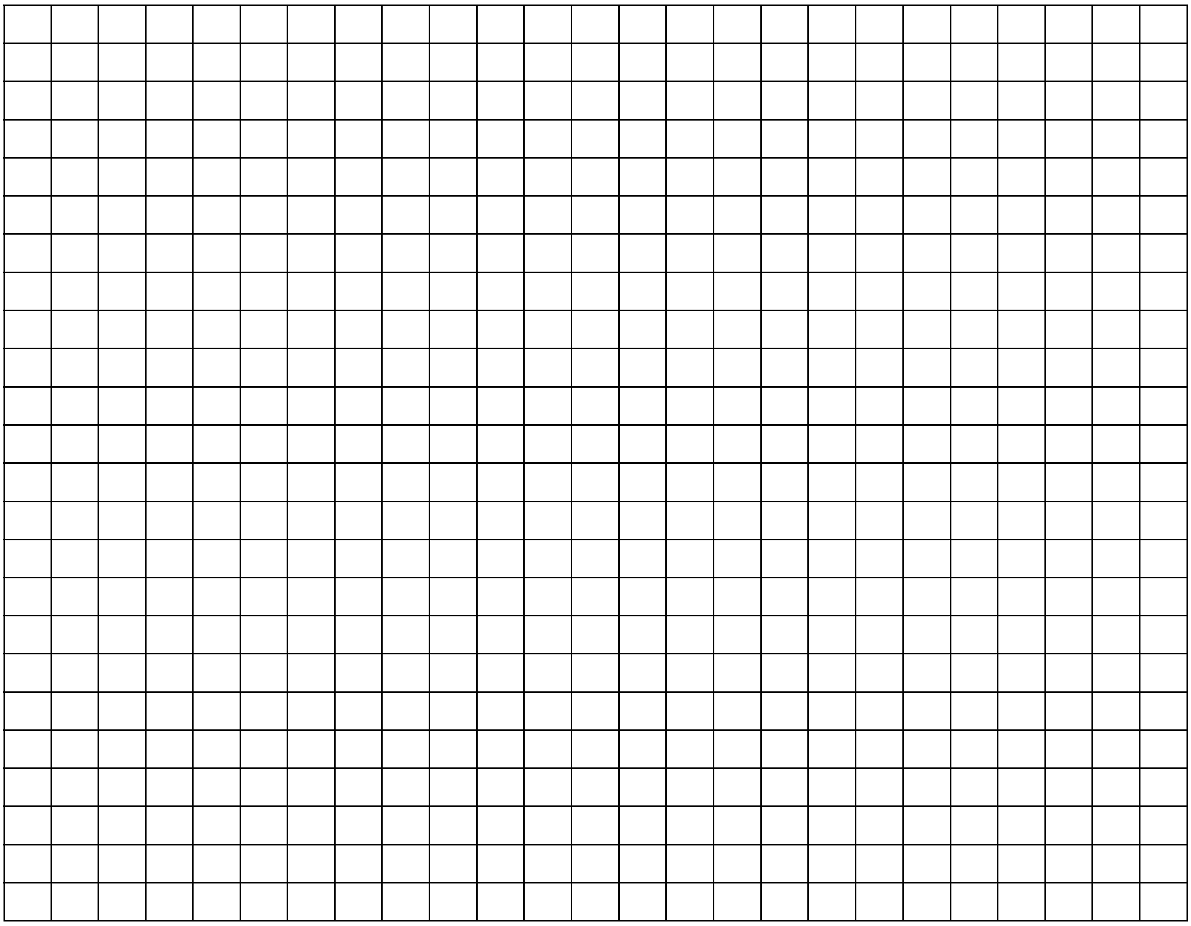
Directions:
Shown below is a dataset of sumo wrestler body weight (kilograms) and the number of wrestling victories (wins).

Sumo wrestler data: ordered pairs

body weight = X-coordinate
number of wins = Y- coordinate

Sumo wrestler	**X** **Body weight (kilograms)**	**Y** **Number of wins**
A	51	16
B	18	3
C	40	10
D	45	12
E	30	6
F	15	2
G	19	4
H	25	5
I	55	20
J	48	13
K	49	14
L	60	22
M	50	15
N	35	8

1. Using ordered pairs (X,Y), graph each ordered pair on the coordinate plane below.
2. Label the X and Y axes.



3. Which sumo wrestler had the most victories (wins)? ________________

4. Which sumo wrestler had the least victories? _________________

5. Based on your graph of the data, explain the relationship between the sumo wrestlers' weight and number of wins.

____

____

____

____

APPENDIX B

Visual Information Display Project

1. Collect and record a multivariate dataset of your choice.

•Identify each variable and its type.
•Construct a graphical display by hand of your dataset.
•Construct a graphical display using computer software of your choice.
NO PIE CHARTS unless part of a series.
Provide labels and legend as needed.

2. Write a paper (approximately 5 pages total) discussing your project.
Written Paper must include the following elements:

- Identify the type of display: data map, time-series, space-time-narrative, or relational graphics.
- Identify each variable and its type.
- Copy of the dataset used for graph
- Construct a graphical display by hand of your dataset.
- Construct a graphical display using computer software of your choice.
  Provide labels and legend as needed.

•Discuss the use of color, line, form, and composition in the graphical display.

•Discuss any problems you encountered during the project.

3. Present your project to the class (10- 15 minutes) in oral-visual-written format.

Bring a color transparency of your final graph to share with the class.

4. Turn in to Instructor:

Written paper discussion (4-5 pages, typed)
Copy of the dataset
Copy of hand-drawn graph
Copy of computer-aided graphical display)

APPENDIX C
Statistics Notebook Assignment

Directions for the Statistical Assignments:

1. Each research question is to be regarded as a separate study and written up as a separate study.
2. Each research question is to have a title page with a title that appropriately reflects the independent variable(s), the dependent variable(s), and the population. The running head should be a 50-character or less summation of the essence of the study.
3. The beginning on Page 2 should be a repetition of the title from page one, followed by the research question, and then with the results section.
4. The results section should begin with the statistical analysis you performed in relation to your specific research question. Then you should indicate that you check the assumptions associated with the underlying test (e.g., normality) and a brief discussion of any values computed to assess each assumption.
5. After the assumption checks have been addressed, then examine such indices as the main effect(s), variables/item coefficients, and *p*-values, using appropriate statistical terminology and language. If post-hoc tests were conducted, then indicate the type and findings.
6. Effect sizes are to be reported at the end of the results section. These effect sizes should be interpreted for all statistically significant findings.
7. Reference(s) page is next, whereby each citation appearing in the text must appear in a consistent form in the reference list.
8. Following the reference(s) page are the tables, with each table having its own separate page. Each table should be appropriately titled and fully descriptive for your reader, as well as conforming to the American Psychological Association (APA) style.
9. Attached in a combined section at the end of the research questions for each section should be your statistical output.

Statistical Notebook Assignment: Regression

1. Regression data set "Spring 2007.RegressionData" is the name of the data set you need to use to address these research questions.

2. Variables in this data set are:
   a. Group
   b. Grade
   c. Age
   d. Math1 through Math 10
   e. Math Total
   f. Reading 1 through Reading 10
   g. Reading Total
   h. Pmath 1 through Pmath 8; Pmath 10
   i. Pmath total
   j. Preading 1 through Preading 8; Preading 10
   k. Preading total

3. Research questions you are to address are:
   a. What student-generated reading variables are predictive of parents' rating of their interest toward reading?
   b. What student-generated math variables are predictive of parents' rating of their interest toward mathematics?
   c. Which reading items are predictive of students' mathematics attitude total scores?
   d. Which parent-generated reading items are predictive of students' overall reading attitude total scores?

4. Report, in table form, the relevant statistical information for each regression procedure performed for each of the research questions above. Your tables should be developed in accordance with each specific research question.

5. You are to write, in APA style, a paper for each research question above. Include the title page, results, and relevant tables. Statistical output should be attached at the end of each paper.