From **Gal, I. & Garfield, J. B. (editors). The Assessment Challenge in Statistics Education**. IOS Press, 1997 (on behalf of the ISI). Pages 27-36. ISBN 90 5199 333 1. Copyright holder: International Statistical Institute. Permission granted by ISI to the IASE to make the chapters from this book freely available on the Internet. The References section is a separate chapter and is available on the Internet at http://www.stat.auckland.ac.nz/~iase/publications/assessbkref.

3. Authentic Assessment Models for Statistics Education

Shirley Colvin
Kenneth E. Vos

Purpose

Authentic assessment is an emerging field within assessment models. It claims to measure by direct means the student performance on tasks that are relevant to the student outside of the school setting. Most educators will agree with the need to assess learning within the context of applications. This chapter will address the following issues:

- A vision of an effective assessment system must be articulated. What are the standards (visions) for an effective assessment system?
- A well-thought-out plan for designing an effective program must be constructed. What are the components of a process for designing an effective authentic assessment program?
- Classroom teacher readiness to change assessment plans is crucial to any program of assessment. How do you determine the degree of readiness of classroom teachers for a new assessment plan?
- Promises abound in the assessment field but limitations can strangle an assessment program at conception. What are the promises and limitations of recent assessment reforms?

A crucial aspect of teaching and learning is knowing what and how much is learned. Assessment should be the source of this information. This chapter will give a glimpse of how to design an authentic assessment plan in statistics education.

INTRODUCTION

At the forefront of reform of education and its many inherent issues is how to assess the product of schooling, which certainly includes statistics education. The recent literature points to the need for assessment measures that (a) require the student to transfer learning to new situations/view situations from different perspectives for problem-solving purposes (National

Council of Teachers of Mathematics, 1989), and (b) have instructional utility (NCTM, 1995; Webb, 1993; Stenmark, 1991). Therefore, the field of education is actively searching for “alternative” assessment programs. These alternatives can be of many forms but most commonly they are given the label of performance, authentic or real-world application. This chapter will focus on authentic assessment models but obviously it is not possible to present an authentic assessment model without incorporating performance or real-world applications. The three terms associated with assessment, performance, real-world applications and authentic, are sometimes used interchangeably to describe the same type of assessment tasks. However, in this chapter the term, authentic assessment, is used to describe the means of measuring student performance on tasks that are relevant to the student outside of the school setting. The term, performance assessment, relates to the direct, systematic observation of student actions related to a performance task. The term, real-world, is sometimes overused in education to describe any aspect of life outside of school. In order to be consistent and reflect the belief for the need to communicate clearly, this chapter will use the term authentic assessment. At the present time, a concise definition of authentic assessment is being formulated. An acceptable description could be the direct examination of student performance on significant tasks that are relevant to life outside of school (e.g., Herman, Aschbacker, &Winters; 1992). The use of real-world situations in assessment settings is promoted by the NCTM Assessment Standards (NCTM, 1993) and other reform documents (NCTM, 1989; Mathematical Sciences Education Board, 1993). In some ways, the most important aspect of assessment is not the struggle with the correct assessment term to describe the phenomenon but rather the struggle to construct an effective, efficient and meaningful assessment system in an educational setting. The term, system, is used to describe both an individual teacher designing an authentic assessment program for classroom instruction and a more elaborate program of assessment designed by school districts or states. Therefore, a system can be fairly complex but it does not necessarily need to be if a teacher wants to begin a small assessment program in a single classroom setting.

Authentic assessment is a viable vehicle to assess a student’s understanding of statistics education. However, authentic assessment is a fairly recent phenomenon in the assessment field. Many issues need to be resolved or clarified before authentic assessment will impact traditional testing situations in American schools. Classroom teachers are excited about the potential of moving from traditional multiple-choice tests to an authentic assessment model. However, the excitement wanes when the reality of this enormous undertaking emerges. Issues of appropriate item development, technical measures of reliability and validity, effective scoring designs, and acceptance by the public to new reporting schema quickly appear. These issues should be viewed as challenges within the field of assessment and not barriers. A clear and concise linear model of dealing with these issues may not be possible but a process of thinking about how to effectively design an assessment system can be developed. This chapter examines issues involved in creating and using authentic assessment tasks as part of a broader system of assessment in statistics education. The first section, Task Issues, focuses on the nature of appropriate assessment items and on scoring of students’ performance. The second section, System Issues, focuses on principles for designing an effective authentic assessment system and reiterates known limitations and barriers to success. The last section, Implications, discusses the resulting challenges and the potential for changing and improving how assessment occurs in various statistics education settings.

TASK ISSUES

Authentic assessment tasks must involve activities appropriate to a student's life outside of school. One way to judge if a task is authentic is to apply the Reality Principle (see Chapter 6). If the student believes the situation within the task could really happen in "real life," it is more likely the student will attempt a possible solution. A significant challenge for anyone considering the possibility of writing an authentic assessment task is the need to capture the interest of the student, make the situation believable to the student, incorporate sufficient real data to lend credibility and, most importantly, focus the process toward important statistical concepts. If the authentic assessment task does not engage the student, a major part of the goal of the assessment task is not satisfied. If the student doubts the situation occurs in "real life" outside of school, a reluctance to continue to explore in depth may occur. If the situation incorporates real data but it seems to be too simple in form or accuracy, a student will quickly question its authenticity. If the assessment task is engaging, believable, incorporates sufficient real data but obviously lacks the need to use any important concept in statistical education, then the assessment task should be redesigned or eliminated from the assessment program.

The following examples can be considered authentic assessment tasks. Each example lists an appropriate age range but with adaptation, considering the development level of the student, the example could be extended to older or younger students. A few examples of authentic assessment tasks will be given in a brief outline format which includes a general comment, grade level(s), important ideas, and some techniques needed. A more detailed example adapted from *Measuring Up* (1993) completes the list of examples and serves as a context for elaborating on scoring issues in authentic assessment.

Who Fits My Rule?

"Who Fits My Rule" is classification guessing game in which players try to figure out the common characteristics, or attributes, of a set of animals or other living objects.

Grade levels: primary--kindergarten through third grade

Important ideas: collecting, organizing and interpreting categorical data

Some techniques needed: classification schema (e. g., Venn diagrams, visual displays, charts)

Fit-all Mitten Company

A small group of students (3 or 4) constitute a project group responsible for an action plan answering at least the following questions posed by the management council of the Fit-all Mitten Company.

What size mittens will reflect the company's name "Fit-all?"

What color or colors should the Fit-all mitten be if limited yarn colors are available?

What style (closed or glove) Fit-all mitten should be produced?

Grade levels: middle, junior high, early high school

Important ideas: collecting information from a limited sample, organizing the data in visual displays, making decisions based on acquired information

Some techniques needed: designing a method of obtaining information, displaying information, determining hand size by an approximation method

Heart Disease

While the cause of heart disease is not known there are many risk factors associated with heart disease. Have either individual or groups of students choose a factor or combination of factors to investigate. Use information currently available from a variety of sources.

Grade levels: upper high school, introductory college

Important ideas: significant issue to society, pose complex questions using data from external sources (surveys, experiments)

Some techniques needed: multiple representations, causation and relationship factors

Heavy Bears

The core of the example comes from *Measuring Up* (1993) pp. 125-132.

Grade levels: advanced 3rd, 4th grade and above

Important ideas: use real-world context for data, organizing unordered data in a meaningful way, compare data sets

Some techniques needed: drawing a line plot, analyzing sets of data, choosing appropriate representations

The data in Table 1 give the weights of some grizzly bears and black bears living in the Rocky Mountains in Montana. Some of the employees of the Montana agency responsible for natural resources in the state believed a study needed to be done to verify the cubs (young bears) of different types of bears maintained the relationship of weight and type of bear found in older more mature bears.

1. Organize these data in a way that would help you find which kind of bear is heavier—grizzly or black bear?

(More space allotted on student response sheet)

2. Write down three things that you can tell about the weights of the bears.

(More space allotted on student response sheet)

3. Based on these data, how much heavier is a typical bear of one kind than a typical bear of the other kind?

Show how you figured out your answer.

(More space allotted on student response sheet)

Table 1: Weights of Bears

GRIZZLY			BLACK		
Bob	male	220 lbs	Blackberry	female	230 lbs
Rocky	male	170 lbs	Greta	female	150 lbs
Sue	female	210 lbs	Freddie	male	140 lbs
Linda	female	330 lbs	Harry	male	230 lbs
Wilma	female	190 lbs	Ken	male	170 lbs
Ed	male	180 lbs	Hilda	female	220 lbs
Glenda	female	290 lbs	Grumpy	male	160 lbs
Bill	male	230 lbs	Blackfoot	female	150 lbs
			Marcy	female	170 lbs

	Grempod	male	200 lbs

Quality standards

The data set must be viewed as a whole in order to allow comparison of the two groups of bears. Since these data are unordered there must be evidence of displaying these data to show the overall shape and features of the data set. The data sets must be described using three different statements. A summary of these data must show the comparison of the two data sets, not just comparing the heaviest and lightest bears.

Scoring rubric

A general scoring rubric as well as a specific scoring rubric are shown in Figure 1. The general scoring rubric is applicable to many similar situations while the specific scoring rubric applies only to this assessment task. Therefore, the general scoring rubric is a model for future specific scoring rubrics of other assessment tasks.

General Scoring Rubric	Specific Scoring Rubric
3 points • used reasonable strategy(ies) to reach a conclusion • gave clear explanation • complete explanation or only a minor error	3 points [HIGH] • clear, accurate graph or plot • description includes range and distribution comments • grizzly bears are heavier than black bears • value reported is difference between central values of grizzly/black bear data sets
2 points • used reasonable strategy(ies) but did not finish or reach a conclusion • gave unclear explanation • some deficiencies • incomplete	2 points [MEDIUM] • inaccurate display of data • description does not include range and center of data sets • concluded all grizzly bears are heavier than all black bears • incomplete explanation of typical bear comparisons
1 point • started a description but unsuitable statements • major errors • inappropriate approach	1 point [LOW] • graph or plot has major flaws • description focuses on individual bears rather than features of the whole data set • no typical data compared
0 points • blank or unreadable • incorrect with no logical explanation	0 points • no explanation • random statements about bears

Figure 1: Scoring Rubrics

In *Measuring Up* (1993, pp. 130-32, permission granted for use) the results of three different 4th grade students are reported. These results are reproduced in Figure 1. The results were scored using the Specific Scoring Rubric with High being 3 points, Medium is equivalent to 2 points and Low is 1 point. The example of the High score included an accurate line plot with both medians shown. The display is designed to assist in supporting the conclusion that young grizzly bears weigh more than young black bears. This response met or exceeded most of the High (3 points) criteria. The example of the Medium score shows the young grizzly bears weigh more than the young black bears but the horizontal axis is not meaningful in this situation. It assumes the bears are numbered or ordered from least heavy to most heavy in weight. This display does not include a description of the centers of the data sets. The example of the Low score shows a display of all the weights of the bears without distinguishing between grizzly and black bears. It is a display of the information but has major flaws since it focuses on individual bears rather than on the comparison between the types of bears.

After reviewing a few tasks it is appropriate to step back and consider an entire assessment program within a classroom, department, or school district. The word chosen to describe this type of assessment program is an assessment system. The next section gives a brief introduction to standards for an effective assessment system, appropriate development phases for assessment task construction and some impediments to assessment reform.

SYSTEM ISSUES

Standards for an effective assessment system

An assessment system can only be considered effective if it is set in a context of a vision of what constitutes sound practice in assessment. The word, standard, is gaining stature and importance in the current curriculum reform movement. Unfortunately, with its vaulted stature has come multiple meanings. A standard can be a vision or goal or benchmark to reach in the near future or it can be a hurdle to jump before considered successful. In this chapter a standard is a vision or benchmark used to measure the appropriateness of the assessment system. It is not a hurdle to jump over or slide under. These standards follow the general format and substance of other compilations of standards (NCTM, 1995; MSEB, 1993). Each standard can be viewed separately but the impact of each standard is greater when all six standards are considered together. These six standards are general principles for beginning the process of designing an assessment model. This general principles section is followed by development phases which describe the more practical steps taken to design an assessment model for the classroom, department or school district. Both the general principles and the development phases lead into a section on impediments to assessment reform.

General principles

1. Important Statistics Education Content. The assessment should reflect the statistics content that is most important for the students to learn. Important statistics content should be stated as completely as possible. Agreement obviously may be difficult to attain. However, there exists at least a core of big ideas agreed to among statistics educators.

2. Enhanced Learning Of Statistics. Assessment should intensify the learning process by a student. In other words, the assessment should add to the learning of statistics and not be viewed as a necessary evil, e. g. test to confuse. Ideally learning and assessment should be a seamless process.

3. Development Levels Of Learners. The development levels of learners should be reflected in the assessment. Some areas of mathematics education such as geometry and number theory have stated the specific stages of development by learners. Statistics learning does not possess a precise schema. However, general knowledge of the development of learning from the early years through adult should be used to match the appropriate assessment task with the development level of the learner.

4. Criteria For Performance. The assessment plan should include a process to determine how the criteria of performance are developed. Experience working with classroom teachers has shown that a very difficult aspect of an assessment system is how to establish sound criteria of performance. There must be stated a flexible process in determining who sets the criteria, how the criteria are set, who implements the criteria, how are the criteria revised, and how are the criteria reported.

5. Multiple Sources Of Information. Assessment should include multiple situations and models. An assessment system must incorporate more than a single source of information. Authentic assessment tasks alone are not rich enough to give a complete picture of individual students or groups of students. There needs to be a diversity of assessment sources such as multiple-choice test, authentic assessment tasks, and portfolios to name just a few. Recognizing the importance of multiple sources of information is most crucial to the success of an assessment system.

6. Openness. All aspects of the assessment process should be open to review and scrutiny by the public, teachers, learners, and administrators. The use of the results should be clearly understood by everyone involved. Openness in assessment can be very elusive. One can believe in openness for an assessment system but practice just the opposite. This is a common pitfall for many assessment systems. Openness is necessary for support from the various groups who have a stake in the assessment results. Many assessment systems neglect to communicate the process or clearly state the results in a readable manner.

Development phases

On a practical level, the process of designing an authentic assessment system is not linear, concise, or prescriptive. However, there are phases which should be considered in the design. A group of instructors or an individual teacher could follow these phases to begin to understand how authentic assessment could impact the classroom. One can start at the first phase and continue in a linear fashion or begin at different entry points and skip around in the process. The general principles are a net over the development phases. The general principles do not match one-to-one with the phases but rather are just that: principles. The following process has proven to be successful with many classroom teachers over the last few years.

Identify the important ideas in statistics education	*Note*: the number of important ideas should be small (6-10) in order to focus the content and assessment.
Identify knowledge, skills, and techniques needed to understand the important ideas	*Note*: avoid too finely detailed statements, rather keep the framework as broad as possible.
Explore how the important ideas in statistics are used in society/business/education	*Note*: "authentic" means related to life outside of school, not just look like real-world.
Design a limited number of authentic assessment situations at different grade levels for pilot use	*Note*: suggested grade level ranges--K-4, 5-8, 9-12, college which matches NCTM *Standards* ranges.
Create quality standards for each authentic assessment situation	*Note*: a quality standard is a statement that describes a successful solution to the task. Also it can be a description of features to consider in a successful solution.
Using the quality standards develop a specific scoring rubric for each authentic assessment task	Note: experience suggests the scoring rubric be a scale with an odd number (excluding zero) of values, not exceeding six values.
Devise a method of recording and reporting learner results in an accurate manner	*Note*: authentic assessment scoring schema are usually not easy to quantify. A concise plan of documentation must be developed.
Conduct a quality review of the entire process	*Note*: quality review means every decision is eligible for revision, nothing is a given or obvious.
More questions, more ideas	*Note*: similar to the process of data analysis, there should be more questions after going through the process than before beginning the process.

Applying the development phases to yield authentic assessment tasks is both a process and obviously a product. An authentic assessment task should be forthcoming at the completion of the phases. The next section includes a few barriers to success in implementing an assessment system.

Some impediments faced by assessment reform

Designing an authentic assessment program for a classroom, department, or school district should be a straightforward process. However, experience and anecdotal evidence does not support this assumption. Many barriers and impediments are inherent in schooling and seem to slow down any changes in assessment programs. These barriers can be grouped into two broad categories: conditions for readiness by instructors and promises/liabilities. The conditions for readiness give a non-exhaustive checklist to use with teachers to gauge their acceptance of a new assessment plan.

Conditions for readiness

1. Desire for better assessment information. Change is not born out of contentment. If teachers are satisfied with the present assessment activities, they will resist any suggestion of moving toward authentic assessment or any other assessment model.

2. Staff openness to innovations. New assessment models can be threatening to many teachers--by the way, not only "deadwood" teachers can react in this manner.

3. Conceptual clarity about assessment. There is a certain amount of fuzziness currently associated with assessment. Teachers must be convinced about the conceptual underpinnings of any assessment model available.

4. Assessment literacy of staff. Few teachers consider themselves to be assessment experts but they must have at least a working knowledge of assessment. Fulfilling this condition may be the first priority of any assessment plan.

5. Clarity about learning goals. Assessment is not possible with non-existent or out-of-focus learning goals. Before any classroom, department or school district can develop assessment tasks, it must develop or articulate learning goals for statistics education. Failure to fulfill this condition will sink the entire assessment plan. The common learning goals described in Chapter 1 together with age-specific learning goals discussed elsewhere (e.g., NCTM, 1989) could serve as a starting point for a local discussion of learning goals for students in a particular group or level.

6. Openness of community/parents to new methods of assessment. The obsession with "single number" success ratings will be difficult to modify. Many parents are comfortable with the traditional assessment system and may resist any change which seemingly gives less information.

Promises and potential liabilities

An authentic assessment program holds out promises as well as liabilities. Any assessment model should be reviewed for potential liabilities and promises. A balance of promises and potential liabilities should be the goal of an assessment program that hopes to make an impact in the classroom, department or school district. The following chart summarizes some of the promises and liabilities of assessment reform plans.

Promises	**Potential Liabilities**
Emphasis on critical thinking	Authentic assessment is not equated with complex, higher order problem solving ability
Written communication of mathematical and statistical information highly regarded	Emphasis on a specific mode of communication assumes a certain learning style
High stakes testing more closely aligned to statistics education curriculum	"Opportunity to learn" statistics becomes paramount
New challenges in technical quality e. g. validity, reliability	Conventional quality measures for testing may not be feasible
Classroom teacher an integral key to the assessment reform	Heavy dependence on teachers for item design, field testing of items, and scoring

IMPLICATIONS

An authentic assessment model is a natural for assessing the learning of statistics. Statistics is making sense of data drawn from authentic settings. To fully understand the power of statistics a student must be able to "make sense" of data. An individual, school or college may use the standards of an effective assessment system, but should always be aware of barriers that may impede the assessment models used.

Authentic assessment holds an alluring promise for many important features of statistics education: critical thinking emphasized; written communication highly regarded; close alignment to statistics curriculum; ability to make decisions based on data. However, to successfully implement an authentic assessment system there must be articulation of the vision of an effective assessment system, building on this vision with a process of designing an assessment system, but always with a keen awareness of the barriers to success in implementing this assessment system.

The success of any assessment system depends on a carefully constructed plan from conception to implementation. It is possible to make authentic assessment be a success in our schools. Success in statistics assessment must be measured by the increase in learning and understanding the power of statistics. Only time will tell if statistics understanding of our students will increase within the authentic assessment era.

With a renewed interest in a balance among curriculum reform effort, new instructional strategies, and alternative assessment techniques, educators are faced with a difficult task of teaching and learning in the 1990s and in the next millennium. Navigating through this maze of contradictions and unfilled promises, educators need to at least keep this focus: It is better to try at least one different assessment technique than be paralyzed with all this information and refuse to try any new assessment program. Experience has shown that trying at least one new assessment program will open a new dialogue of understanding between a student and the classroom teacher. If this goal of opening a dialogue is met with attempting authentic assessment, we would call the effort a tremendous success!