Assessing Outside the Bubble: Performance Assessment for Common Core State Standards

May 30, 2011

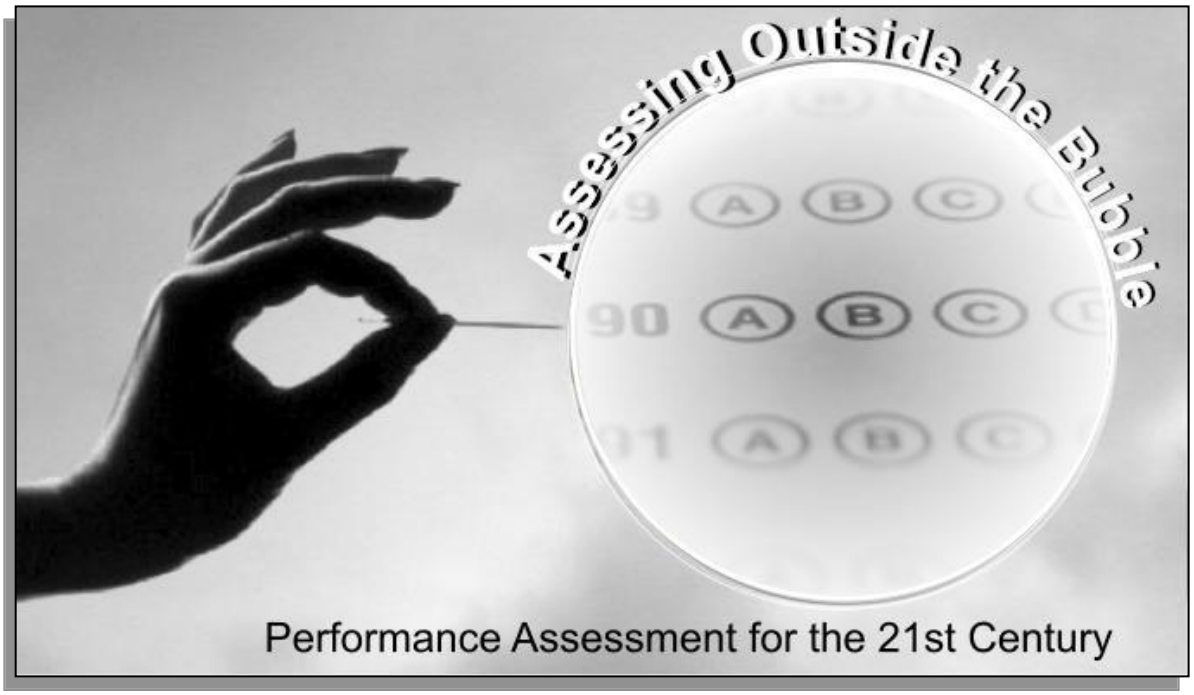
Jesica M. Bishop, Lora J. Bristow, Bryn P. Coriell, Mark E. Jensen, Leif E. Johnson, Sara R. Luring, Mary Ann Lyons-Tinsley, Megan M. Mefford, Gwen L. Neu, Emerson T. Samulski, Timothy D. Warner, Mathew F. White

Researched and written under the direction of Eric V. Van Duzer, Ph.D.

Editors: Eric V. Van Duzer & Gwen L. Neu

Abstract

The adoption of Common Core State Standards has increased the need for assessments capable of measuring more performance-based outcomes. This monograph brings together the current literature and resources for the development and implementation of performance assessment. The text was written as part of a project-based graduate course and has resulted in a clear, well researched and documented contribution that provides up to date information and references on Common Core State Standards and performance-based assessments.

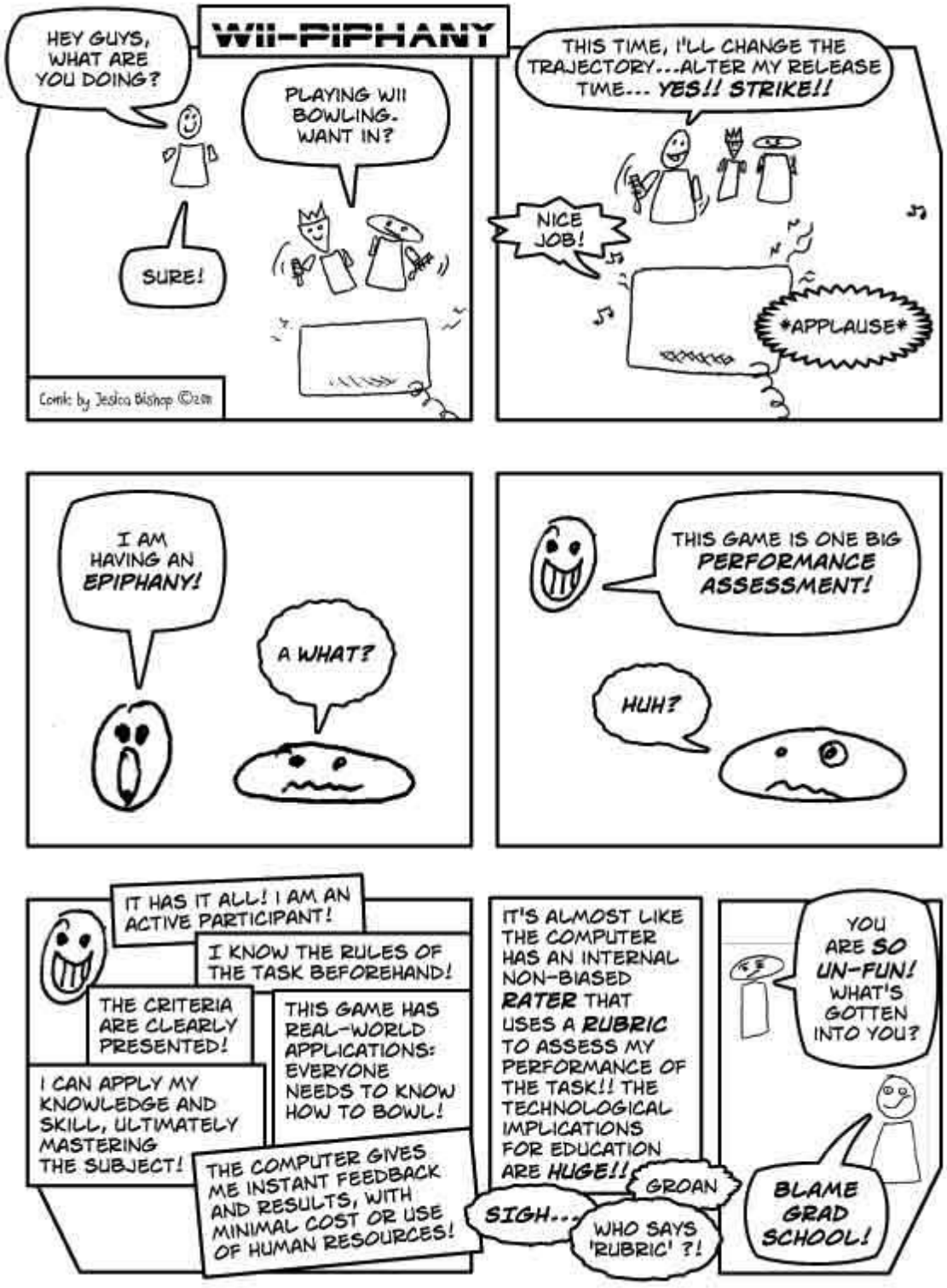


Jesica M. Bishop, Lora J. Bristow, Bryn P. Coriell, Mark E. Jensen, Leif E. Johnson, Sara R. Luring, Mary Ann Lyons-Tinsley, Megan M. Mefford, Gwen L. Neu, Emerson T. Samulski, Timothy D. Warner, Mathew F. White

Editors: Eric V. Van Duzer & Gwen L. Neu

With special thanks to Ann Diver-Stamnes

May 2011


WII-PIPHANY
HEY GUYS, WHAT ARE YOU DOING?
PLAYING WII BOWLING. WANT IN?
SURE!
Comic by Jessica Bishop ©2011
THIS TIME, I'LL CHANGE THE TRAJECTORY...ALTER MY RELEASE TIME... YES!! STRIKE!!
NICE JOB!
APPLAUSE
I AM HAVING AN EPIPHANY!
A WHAT?
THIS GAME IS ONE BIG PERFORMANCE ASSESSMENT!
HUH?
IT HAS IT ALL! I AM AN ACTIVE PARTICIPANT!
I KNOW THE RULES OF THE TASK BEFOREHAND!
THE CRITERIA ARE CLEARLY PRESENTED!
THIS GAME HAS REAL-WORLD APPLICATIONS: EVERYONE NEEDS TO KNOW HOW TO BOWL!
I CAN APPLY MY KNOWLEDGE AND SKILL, ULTIMATELY MASTERING THE SUBJECT!
THE COMPUTER GIVES ME INSTANT FEEDBACK AND RESULTS, WITH MINIMAL COST OR USE OF HUMAN RESOURCES!
IT'S ALMOST LIKE THE COMPUTER HAS AN INTERNAL NON-BIASED RATER THAT USES A RUBRIC TO ASSESS MY PERFORMANCE OF THE TASK!! THE TECHNOLOGICAL IMPLICATIONS FOR EDUCATION ARE HUGE!!
SIGH...
GROAN
WHO SAYS 'RUBRIC' ?!
YOU ARE SO UN-FUN! WHAT'S GOTTEN INTO YOU?
BLAME GRAD SCHOOL!

Contents

Chapter One

Introduction

Once you've made your plans for daily instruction and have your classroom management under control, how will you know if your students have learned what you taught them? How do you know if the teaching process was successful? A crucial component of the teaching process is determining its success. *Assessment,* the systematic collection, review, and use of information about educational programs, is necessary to verify what, and to what extent, objectives have been met in a lesson, a unit, or over the course of a school year. A*ssessment* is more than a way to inform instruction; it can also provide critical feedback to the learners making them more effective partners in meeting learning objectives. Improving students' efficacy in mastering content comes from the power of having detailed information about strengths and weaknesses (Pellegrino, Chudowsky, & Glaser, 2001). "One of the most important roles for assessment is the provision of timely and informative feedback to students during instruction and learning so that their practice of a skill and its subsequent acquisition will be effective and efficient" (Pellegrino, Chudowsky, & Glaser, 2001, p. 4). Teachers also need this feedback to inform their instructional strategies by assessing the effectiveness of various decisions such as pedagogical choices, use of materials, and the success of the teachers in differentiating instruction to meet students' varying needs.

http://www.assessment.uconn.edu/why/index.html

Knowing what kind of assessment to use under which circumstance is the trick. Today, the most common assessments in schools are the traditional pencil and paper tests, such as end of chapter tests, multiple choice tests, vocabulary tests, and the standardized tests administered at the end of each school year associated with No Child Left Behind and the Race to the Top initiatives. A different and often more complex form of learning assessment is *performance assessment*.

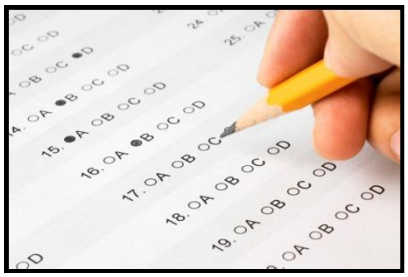
The goal of this handbook is to help you understand *performance assessments* – what they are, why you should use them, what forms they take, when you should employ one of those alternatives, and when it is more effective to utilize other means of evaluating your students.

Definition of Performance Assessment:

While there is no single, agreed upon definition for performance assessment in the academic community, some working definitions are:

- "procedures in which respondents are required to carry out tasks or processes in which they demonstrate their ability to apply knowledge and skills" (Arias, 2010, p. 85)
- "measures (of) students' cognitive thinking and reasoning skills and their ability to apply knowledge to solve realistic, meaningful problems," such as conducting a science investigation, constructing an original product, explaining a mathematics solution, writing a persuasive essay, or linking school activities to a real world experience (Lane, 2010, p.3)
- "assessments that capture the imaginations of students" (Baker, Aschbacher, Niemi, & Sato, 1992, p. 5)
- "assessment based on observation and judgment, like a driving test or an Olympic competition" (Arter, 1999, p. 30)

Why Use Performance Assessments?



pencil and paper assessment vs.

performance assessment

Performance assessments address the changing nature of educational goals, the relationship between assessment and the teaching/learning process, and the "limitations of the current methods of recording performance and credit" (Keyser & Howell, 2008, p. 7). As the American educational system evolves to produce globally competent workers, instructional practices need to evolve to include more complex skills in analyzing and applying knowledge to solve problems as well as mastering content. However, mastery of skills is difficult to assess through traditional paper and pencil, multiple choice tests. Few professionals today are evaluated with a test; rather, ongoing performance in achieving professional goals and meeting responsibilities is evaluated through an assessment of their actions. To bring education into the 21st century, we need to integrate the types of performances and assessments students will face when they enter the global economy. One change brought about through this evolution is the development of standards adopted across the country. These standards are part of "rethinking curricular and instructional efforts to promote quality and equality for all" (Clark, 2000, p. 202). A primary reason for adopting common standards across all fifty states is the issue of fairness in adequately preparing students for the realities they will face when they leave school. Clearly, this is in the best interest of both the students and society.

In 2010, a committee of experts representing the 50 states released what are now known as the Common Core State Standards. Common Core State Standards stress learning that requires more complex, higher-level-thinking than most state standards have to date. Californians expect the state Department of Education to phase in the new standards between 2012 and 2016. When teachers teach complex concepts and skills, assessment

needs to reflect that complexity. Performance assessment best allows for continuous classroom monitoring of student progress in mastering these complex tasks which positively impacts learning outcomes, student attitudes, and teacher success. An added bonus that comes from performance assessment is that students can solidify new skills while being tested and engaging in the performance.

Performance assessments act as vehicles that shape sound instructional practice by modeling for teachers what is important to teach and for students what is important to learn, resulting in student self-regulation through metacognition (Lane, 2010).

Performance assessments provide

- Authenticity
- Context for skills and knowledge being tested
- Cognitive complexity
- In-depth content coverage
- Examinee constructed response structure
- Reform to the educational system
- Information about what students do not know as well as well as what they do know

Performance assessments can overcome the shortcomings of standardized tests which ask students to choose from a set of pre-determined answers that are mainly focused on lower-level thinking skills such as information recall and rote memorization/application of algorithms. Performance assessments ask students to create their own responses to questions that primarily focus on *high-level thinking skills*, such as critical thinking, analysis, synthesis, and evaluation of the content. Because they are complex, performance assessment strategies are best utilized in concert with other forms of assessment. Similar to driver education or pilot certification, both factual knowledge and procedural knowledge are important components of a complete education, and performance assessment is better at measuring procedural knowledge.

Table 1.1 Test Format Continua

TEST FORMAT CONTINUUMS						
Work Samples	**Simulations**	**Projects**	**Essays**	**Short answer**	**Multiple-choice**	**True/False**
Most authentic		⟷				Least authentic
Cognitively most complex		⟷				Cognitively least complex
In-depth coverage		⟷				Coverage of content
Response structured by examinee		⟷				Response structured by test
Highest cost		⟷				Lowest cost

(Arias, 2010, p. 86)

The table above illustrates that traditional tests best assess simple learning objectives, such as language conventions, basic math algorithm, historical names and dates, and other rule driven activities. Performance assessments are necessary to evaluate complex learning objectives, such as writing to communicate, explaining how to solve mathematical problems, conducting a science experiment, thinking critically about historical issues, and demonstrating the ability to handle a motor vehicle.

Table 1.2 Types of Performance Assessments

To assess how well students can…	Provide this kind of material…	And ask students to…
Compare and contrast	Two texts, events, scenarios, concepts, characters or principles	• Identify elements in each • Organize the elements according to whether they are alike or different
Evaluate materials and methods for their intended purposes	A text, speech, policy, theory, experimental design, work of art	• Identify the purpose the author or designer was trying to accomplish • Identify elements in the work • Judge the value of those elements for accomplishing the intended purpose • Explain their reasoning
Assess their own work	A set of clear criteria and one or more examples of their own work	• Identify elements in their own work • Evaluate these elements against the criteria • Devise a plan to improve
Evaluate the credibility of a source	A scenario, speech, advertisement, Web site or other source of information	• Decide what portion of the information is believable, and explain their reasoning
Describe and evaluate multiple solutions strategies	A scenario or problem description	• Solve the problem in two or more ways • Prioritize solutions and explain their reasoning
Think creatively	A complex problem or task that requires either brainstorming new ideas or reorganizing existing ideas or a problem with no currently known solution	• Produce something original, OR • Organize materials in new ways, OR • Reframe a question or problem in a new way

(Brookhart, 2010)

Some performance assessments are considered *authentic assessments*. *Authentic assessments* are a form of assessment in which students are asked to perform real-world tasks that demonstrate meaningful application of essential knowledge and skills and often

involve a real audience or consequence rather than just classroom activities. Authentic measures are

> engaging and worthy problems or questions of importance, in which students must use knowledge to fashion performances effectively and creatively. The tasks are either replicas of or analogous to the kinds of problems faced by adult citizens and consumers or professionals in the field. (Wiggins, 1993)

Authentic tasks can range from analyzing a political cartoon to making observations of the natural world to computing the amount of paint needed to cover a particular room (Mueller, 2005).

It is sometimes difficult to distinguish between an authentic assessment and a performance assessment that looks authentic. Many performance assessments look authentic but are simply classroom tasks designed to assess complex thinking without some features of authentic assessment. The table below illustrates some of the differences between performance tasks and their authenticity.

Table 1.3 Degrees of Authenticity

Inauthentic	**Somewhat realistic**	**Authentic**
Explain a data set	Design a house using specific mathematical formulas and shapes	Design and build a model house that meets standards and client demands
Write a paper on laws	Write a persuasive essay on why a law should be changed	Write a proposal to present to appropriate legislators to change a current law
Read a teacher-chosen text	Read to class a self-selected text	Make a tape of a story for use in a library

(Wiggins, 1998, p. 28)

To determine if a performance task is authentic, compare your task with four characteristics of authentic assessments. Ask yourself, does my task:

1. Involve real-world problems that mimic the work of professionals?
2. Include open-ended inquiry, thinking skills, and meta-cognition?
3. Engage students in discourse and social learning?

4. Empower students through choice to direct their own learning? (Keyser, & Howell, 2008, p. 5)

If the answer to each question is "yes," then your task is most likely authentic.

Table 1.4 Difference between Traditional and Authentic Assessments

Typical tests	Authentic tasks	Indicators of authenticity
Require content responses only	Require quality product and/or performance, and justification	Assesses whether the student can explain, justify, apply, self-adjust, not just "correctness"
Unknown in advance, to ensure validity	Known in advance; involve excelling at predictable, demanding core tasks	Tasks, criteria, standards by which work will be judged are known
Disconnected from realistic context, constraints	Require real-world use of knowledge; student must "do" history, science, etc.	Task is a challenge and a set of constraints, likely to be encountered by professional, citizen, or consumer
Isolated items require use of recognition of known answers or skills	Integrated challenges in which knowledge and judgment must be innovatively used to fashion a quality product or performance	Task is multifaceted and non-routine, even if there is a "right" answer; requires problem clarification, trial and error, adjustments, etc.
Simplified for easy, reliable scoring	Involve complex and non-arbitrary tasks, criteria, standards	Task involves the important aspects of performance and/or core challenges in the field of study; not easily scored, but does not sacrifice validity for reliability
One shot	Iterative; recurring essential tasks	Designed to reveal whether the student has achieved real vs. pseudo-mastery, or understanding vs. familiarity
Depend on highly technical correlations	Provide direct evidence	The task is valid, fair; evokes student interest and persistence, is apt and challenging to students and teachers
Provide a score	Provide usable, diagnostic	Not designed merely to audit

feedback	performance, but to improve future performance; the student is the primary "customer" of information

(Wiggins, 1998, p.23)

When to use Performance Assessments

Because performance assessments are complex, you want to use them under certain circumstances and for specific purposes. You might use performance assessment to help diagnose what students know and what they do not know, to teach a skill while simultaneously assessing the skill, or to monitor progress toward a given objective.

Diagnostic Purposes

Performance assessments may be used for diagnostic purposes. What do students know about how to solve certain types of problems? Do they know how to control variables, use instruments, or evaluate findings? Information provided at the beginning of the course may help decide where to begin instruction or what topics need special attention.

Instructional Purposes

A good performance assessment often is indistinguishable from a learning activity, except for standardization and scoring. In this light, a performance task that simulates the authentic tasks of a scientist or mathematician may be used as either an instructional activity or an assessment activity or both. If the assessment task is used in such a way that the student would normally not know it is an assessment activity, it is called an *embedded* task.

Monitoring Purposes

The goal of a performance assessment is to judge the level of competence students have achieved in *doing* the science, design, mathematics, etc. Accordingly, performance assessment strategies are best used to monitor student process skills and problem solving approaches. The most effective performance assessments are authentic tasks that are open-ended with multiple-correct solution paths (Slater, n.d. p 3).

When not to use Performance Assessments

Performance assessments are meant to supplement, not supplant, other methods of evaluation. When you want to know if your students mastered the multiplication tables with fluency, a timed test of multiplication facts will do just fine. When you want to know that they can apply those facts to tasks (such as deciding how many of each of three products to purchase for the upcoming school fundraising booth) that require multiplication, judgment, and interpersonal negotiation, a multiple choice test might not be the best method.

By their very nature, performance assessments are expensive in resources and in time. For example, they require an initial investment of time to develop a quality *rubric* for scoring, people to validate scores, and extra time to administer the assessment, and they may involve special equipment such as a science lab. These limitations deserve careful consideration as you choose your learning objectives and think about how your students will demonstrate their mastery of them. If your objectives are simple and address knowledge at the lower end of Bloom's Taxonomy, you may best serve them with simple assessments, such as a chapter quiz or test from the textbook, a one or two sentence answer, or a brief oral interview with the student.

As you read more about performance assessments, think about the circumstances in which you might find yourself utilizing them. You will read more about different situations, and you will experience examples of suitable performance assessments throughout this handbook. Apply those ideas to your classroom and talk about them with your colleagues. The more you know about these powerful assessments, the more excited you will feel about them.

Chapter Two

Common Core Standards for 21st Century Education

You have probably heard about the new national standards and 21st Century Skills as the next steps in the ever-changing world of teaching and learning. This chapter will provide some background and information on these areas and connect them to performance assessment.

History of Academic Standards

1959 Pres. Eisenhower proposed national goals for education to support American competitiveness.

1965 Pres. Johnson signed the Elementary and Secondary Education Act (ESEA), as part of his "Great Society." The Act provides federal funding for Title I (high poverty schools) and bilingual education. It has been revised and reauthorized 5 times to date.

1983 *A Nation at Risk* report claimed higher standards were needed to reform education; lack of these had resulted in lowered achievement, placing United States at risk in competing globally.

1980s Measurement-driven models of education were prominent, and the use of standardized testing grew.

1988 Pres. Reagan and Sen. Edward Kennedy led revisions to the National Assessment of Educational Progress (NAEP) to allow for state-to-state comparisons of achievement.

1989 Pres. H. W. Bush proposed national standards, but the U.S. Senate rejected them. Pres. Bush and all 50 governors adopted National Education Goals for the year 2000 which proposed the development of national achievement standards.

1989 National Council of Teachers of Mathematics (NCTM) introduced standards; other organizations, including National Council of Teachers of English (NCTE) and the American Association for the Advancement of Science (AAAS), followed with standards in their fields throughout the 1990s. During the same period, many states developed their own standards.

1990s **I**n an effort to improve education by improving assessment, WYTIWYG (What you test is what you get) ideas encouraged movement towards development of performance assessments which resulted in large-scale use in some states (including California, for a brief few years).

1994 Pres. Clinton signed Goals 2000: Educate America Act which identified additional subject areas in which to develop standards.

2001 Pres. G. W. Bush signed a reauthorization of the ESEA, the No Child Left Behind Act (NCLB), which mandated all states, subject to sanction, to develop standards, and assess and report individual student performance on those standards in the high-stakes context of meeting targets. This demand staunched the continuation and development of large-scale performance assessment in part because it was difficult to get reliable individual scores.

2009 Federal Department of Education held the Race to the Top competition among states choosing to apply for large federal funds. The grant rules encouraged adoption of common core standards.

2010 Common Core State Standards (CCSS) were released by National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO). Reauthorization of the ESEA is overdue; President Obama's administration is working on proposals.

Timeline sources: Marzano &Kendall (1997), Sass (2011), Sloan (2011), Stecher (2010)

Standards-based reform: Historical pros and cons

The premise is simple: "In the ***standards-based reform*** movement, the primary goals of any school district are to establish appropriate standards for student performance and to build school cultures that ensure student success" (Clark, 2000, p. 201). In some countries, like Australia, and in higher education, this is called ***outcomes-based education (OBE),*** a name carried over from an earlier era of reform. *Outcomes-based education* is "an educational model in which curriculum and pedagogy and assessment are all focused on student learning outcomes" (Driscoll & Wood, 2007, p. 4), and these desired outcomes drive the curriculum. ***Learning outcomes*** are defined as "a stated expectation of what someone will have learned" (Driscoll & Wood, 2007, p. 5) and "descriptors of the intended results of educational activities" (Driscoll & Wood, 2007, p. 26). Outcomes are measured by assessment. Standards (or outcomes) are often divided into dimensions such as knowledge, skills, attitudes/values, and behaviors.

Some of the advantages claimed for standards are that they delineate what students should learn, not just what they should be taught; establish challenging norms; ensure teachers have common expectations; support accountability; and are part of "rethinking curricular and instructional efforts to promote quality and equality for all" (Clark, 2000, p. 202). When students know the goals, and get feedback on their performance towards them, higher achievement is supported; teachers know where students are and can better help them (Clark, 2000). Proponents also point out that learning theory and research connect the use of outcomes to "deep learning" (Driscoll & Wood, 2007, p. 8).

Challenges of standards-based education include: uneven and problematic implementation, due to different visions of learning; interpretations of what the language used in them actually means; overload from too many standards or demands; loss of local autonomy, as standards are created and passed down; and issues over how achievement is to be measured (Clark, 2000).

Some critics of standards argue that they are simply a revision to behavioral objectives associated with *behaviorist pedagogy*, back-to-basics rote learning, or a narrow view of measurement-based education. Even before NCLB, some saw

perils associated with rigid standards, narrow accountability, and tangible sanctions that can debase the motivations and performances of teachers and students. Teachers faced with reforms that stress such practices may become controlling, unresponsive to individual students, and alienated. Test- and sanction-focused students may lose intrinsic interest in subject matter, learn at only a superficial level, and fail to develop a desire for future learning. (Sheldon & Biddle, 1998, p. 1)

Standards and accountability may be seen as continuations of the factory model of education, and some studies supported these concerns (Sheldon & Biddle, 1998).

Table 2.1 Comparison of Traditional and Outcome Based Models

Traditional, or input-based model	Standards-Based or outcome model
• What is taught • Norm-referenced (Bell Curve) • It's normal and natural for some students to fail	• What is learned • Criterion-referenced • The goal is for all students to reach standards

Asking ourselves questions about why students need to achieve standards and how we will create instruction to support this can be a means of balancing the reality of mandated standards and our own beliefs about education.

The Common Core State Standards (CCSS)

Released in June, 2010, the ***Common Core State Standards (CCSS)*** are based on the goal of preparing students for ***College and Career Readiness (CCR).*** Their research-based development is documented in extensive reference lists. They are internationally benchmarked, meaning they have been compared to standards in high-performing countries, and built both upon review of current individual state standards and recommendations of professional organizations and post-secondary programs. As of March, 2011, all but seven states have adopted the CCSS and plan to implement them over the next few years. The CCSS are signposts of grade level achievement, with the explicit recognition that students

with various needs may have different learning rates and achievement levels (Common Core State Standards Initiative, 2010).

Here are some highlights:

English-Language Arts

- Four strands: reading, writing, speaking and listening, and language;
- CCR anchor standards are the reference points in each strand;
- organized by grade level K-8; by various grade combinations for high school (dependent on specific standard);
- K-5 foundational skills for concepts of print, the alphabetic principle, and other basic conventions;
- literacy in history/social studies, science, and technical subjects;
- responsibility for content-based literacy shared amongst the subject area teachers;
- literature and informational text, both classic and contemporary, encompass a broad range of cultures and genres;
- writing for a variety of purposes and in varied genres; use of technology to produce and publish;
- vocabulary acquisition practice threaded throughout the four strands;
- technology and varied media integrated across strands; used to gather and present information;
- expression, collaboration, integration, evaluation emphasized; skills learned in context; and
- sample reading texts, student writings, and translations of standards into performance tasks.

California and the CCSS

California has long been proud of its state standards, which have widely been hailed as some of the best in the country. In August, 2010, California decided to adopt the CCSS but invoked the option to add up to 15% to them, to keep some of its standards which it deems necessary and higher than the CCSS. These composite standards are called the California Common Core State Standards (CCCSS). A full version, showing the additions California has made to the CCSS, is available at http://www.scoe.net/castandards/index.html.

In a 2010 survey by the Gates Foundation and Scholastic, 64% of teachers in California said the state had too many standards (compared to 48% of teachers nationally), and 35% said they were too high (compared to about 15% nationally) (EdSource, 2010).

Mathematics

- grades K-8 organized by domain (grades K-5: operations & algebraic thinking; number & operations in base ten; number & operations—fractions; measurement & data; geometry;
- grades K-5 focus on whole numbers arithmetic (addition, subtraction, multiplication, and division) and development of a strong conceptual understanding, procedural skill with fractions, with some geometry and measurement;
- grades 6-7 extend work with fractions, develop concepts of rational numbers, proportional relationships;
- grades 6-8: ratios & proportional relationships; the number system; expressions & equations; statistics & probability; functions);
- students who master grade 7 standards move into 8th grade algebra I; or 8th grade continuation of preparation for algebra I
- high school standards are organized in two ways: by conceptual categories of number and quantity, algebra, functions, modeling, geometry, and statistics and probability; and in traditional "courses" like algebra 1 & 2, geometry, calculus, and advanced placement probability and statistics; and "integrated" (like the International Baccalaureate Model) and "accelerated" use;
- high school standards include a focus on modeling of mathematical situations;
- standards are designed to balance conceptual understandings with procedural skills, with an overall goal of focus and coherence; and for real world applications, construction of mathematical arguments, communication and precision in mathematical thinking (practice and content);
- particular care was taken to integrate research-based findings on practices in high-performing, high-poverty schools, as well as high-performing countries like Singapore, Hong Kong, and South Korea.

The CCSS do not specify a curriculum or pedagogical methods; they just work to "define what students should *understand* and be able to do" (Common Core State Standards Initiative, 2010). The new standards seem to follow a general progression, up through the grade levels, along Bloom's taxonomy, with the lower elementary grade

standards including more verbs related to knowledge and comprehension, and high school standards using those related to analysis, synthesis, and evaluation. They also progress from working with one source to multiple sources and increasingly prioritize deliberate consideration and use of information.

Although the CCSS documents are careful to explain that their origination was from the National Governors Association and Council of Chief State School Officers, not the federal government, U.S. Secretary of Education Duncan is a strong supporter.

Table 2.2 Arguments For and Against Common Core Standards

Arguments against the CCSS	Arguments in support of the CCSS
• Concern for federal involvement in education, which the Constitution implicitly delegates to the states	• Adoption is voluntary; states can choose to add up to 15% of their own standards to the CCSS • Allows comparability, collaboration between states; establishes a common language about learning expectations • Accountability measures (such as under NCLB) will be more fair
• Fear of rigidity, uniformity, lowest common denominator, one size fits all or fast-food education, and a few states feel their standards are higher	• Deemed better than 37 states in language arts, 39 states in math (Byrd Carmichael, Martino, Porter-Magee, & Wilson, 2010)
• Cost (to the states) of implementation	• Needed to ensure global competitiveness in the 21st century • Higher standards will prepare more students for post-secondary school, training, employment

Raised in the relevant literature are some specific objections to the CCSS:

The Pioneer Institute says the CCSS are not American-focused enough, don't explicitly focus on a back-to-basics approach, and are not traditional; so teaching will have to change (Stotsky & Wurman, 2010).

The Heritage Foundation stated, "The kind of comprehensive, comparable data that a

national test would supply is also a prerequisite for the liberal goal of creating an equal 'opportunity to learn' and achieve high standards through the equalization of resources among schools" (Burke & Marshall, 2010, p. 5) and would "undercut…the pockets of excellence that currently exist"(Burke & Marshall, 2010, pp. 9-10). The real problem is that teachers' unions work for their own interests, rather than supporting "student educational outcome objectives" (Burke & Marshall, 2010, p. 7).

Ronald Wolk of Big Picture Learning asserted, "To assume that these students ('disadvantaged') fail because of 'the soft bigotry of low expectations,' as President George W. Bush suggested in making the case for the No Child Left Behind Act, is preposterous. Their failure is due to the hard bigotry that generations of these kids have suffered. And high common standards won't rectify that. Indeed, they divert attention away from the real problem by creating the illusion that things will improve if students and teachers are held to even higher standards" (Wolk, 2011, pp. 1-2). He advocated instead for "disruptive innovation" and "personalized learning," with multiple educational pathways for students to choose, as alternatives to standardization (Wolk, 2011, p. 32).

What's Next with CCSS?

The federal government awarded $362 million to two consortia to develop common assessments of achievement based on the CCSS. California is part of the Partnership for the Assessment of Readiness for College and Careers (PARCC) consortium, along with 25 other states. The competitor is the SMARTER Balanced Assessment Consortium (SBAC), with 31 member states. Some states are members of both consortia. Assessment proposals from both consortia promise common features:

- a combination of summative (end of year) assessments for accountability purposes and formative and/or benchmark-based optional assessments for use throughout the school year
- computer-based administration

- data systems with quick turn-around, so information can be used by classroom teachers to inform and adjust instruction
    - a variety of test items, including "challenging performance tasks and innovative, computer-enhanced items that elicit complex demonstrations of learning and measure the full range of knowledge and skills necessary to succeed in college and 21st century careers" (Achieve, Inc., 2011) or "selected-response, constructed response, and technology-enhanced and performance tasks, which require application of knowledge and skills" (*A Summary of Core Components*, SBAC, 2010).

Differences between the efforts include that PARCC stresses the involvement of higher education leaders and faculty to develop high school assessments, with the intention of these becoming an indicator of students' level of preparation for entry-level postsecondary courses (*PARCC Assessment Design*, 2011), while SBAC claims there will be a large teacher role in developing and scoring assessments, with the explanation that teachers have the needed experience to apply cognitive development theory (Office of Superintendent of Public Instruction, State of Washington, n.d.).

A rush of new curriculum materials claiming alignment with CCSS will soon come forward, as for-profit publishers look to this market. The Common Core Curriculum Mapping Project already has free standards maps that include the entire CCSS English Language Arts (see www.commoncore.org for additional information). The maps, created by public school teachers, focus on the goal of a well-rounded liberal arts and science education. Cross-curricular units that focus on an essential question blend art, music, and media and include sample activities and assessments. High school maps include Socratic seminars, with a rubric for assessing student preparation and participation.

There is talk of core science and social science standards to come.

21st Century Skills---what are they?

One of the most common claims made in recent discussions about educational reform is the need for ***"21st Century Skills***." Although there are differences in the lists of these skills drawn up by various organizations, commonalities include:

- collaboration
- communication
- critical thinking and problem-solving
- creativity
- cross-cultural competence
- technology, information, and media literacy

These are not new ideas or capabilities, but they have taken on new facets and importance in today's context. For example, much information is now easily accessible and no longer the domain of certain individuals, or restricted by time and place; information literacy is not about storing information, but about finding the information needed to complete a task, including sifting out information of lesser applicability or reliability. It's about applying, synthesizing, and evaluating information, rather than memorizing.

Groups speaking from a business/economic perspective tend to add financial literacy, initiative or self-management, high productivity, and global competitiveness. Groups focused on technology add ideas like digital citizenship, the ability to transfer knowledge of one technology to another, and generally emphasize the intellectual capacities needed to work with technology, rather than proficiency with a particular technological tool. (Van Duzer, 1998). Groups interested in more of a whole person perspective add health and wellness awareness, civic and personal participation, and "habits of mind and heart" (Riordan, 2005, p. 4).

Some of the skills experts say are needed in today's world do not fit with how schools generally work: trouble shooting, risk taking, and autonomy on the part of students, for example. To teach these skills, our school cultures need to change.

Sheryl Nussbaum-Beach of the 21st Century Collaborative sees web 2.0 as "a revolution more profound than the shift from hunting to agriculture or the advent of printing

and mass literacy" (2011,p. 1), which necessitates a complete change in our practices of teaching to focus on metacognition and thinking, rather than content. Her proposal for 21st century learning has three prongs: face-to-face learning to build relationships, global communities of inquiry, and personal learning networks (Nussabaum-Beach, 2011).

How Performance Assessment fits with CCSS

In their focus on applied thinking skills, many parts of the CCSS may best be assessed using performance assessment. The CCSS themselves contain an appendix which provides sample performance tasks created from various English Language Arts (ELA) standards. Below are some examples. They also note that "each standard need not be a separate focus for instruction and assessment. Often, several standards can be addressed by a single rich task" (CCSS ELA, p. 5).

Table 2.3 Applying performance tasks for CCSS

Gr.	Strand	Standard	Performance Task
1st	Reading Literature: Craft and Structure	RL.1.4 Identify words and phrases in stories or poems that suggest feelings or appeal to the senses.	Students identify words and phrases within _(book)_ that appeal to the senses and suggest the feelings of __(emotion)_ experienced by __(character)__ (e.g., clapped, played, pouted).
5th	Reading Information: Key Ideas and Details	RI.5.1 Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text.	Students quote accurately and explicitly from _(text)__ to explain statements they make and ideas they infer regarding _(topic)__.

6-8th	Reading: Literacy in History/Social Studies 6-12: Integration of Knowledge and Ideas	RST.6.9 Analyze the relationship between a primary and secondary source on the same topic. RST.7.9 Compare and contrast treatments of the same topic in several primary and secondary sources. RST.8.9 Integrate information from diverse sources, both primary and secondary, into a coherent understanding of an idea or event, noticing discrepancies among sources.	Students construct a holistic picture of the history of Manhattan by comparing and contrasting the information gained from Donald Mackay's *The Building of Manhattan* with the multimedia sources available on the "Manhattan on the Web" portal hosted by the New York Public Library (http://legacy.www.nypl.org/branch/manhattan/index)
11-12th	Reading: Literacy in History/Social Studies 6-12: Key Ideas and Details	Rh.11-12.2 Determine the central ideas or information of a primary or secondary source; provide an accurate summary that makes clear the relationships among the key details and ideas.	Students determine the central ideas found in the Declaration of Sentiments by the Seneca Falls Conference, noting the parallels between it and the Declaration of Independence, and providing a summary that makes clear the relationships among the key details and ideas of each text and between the texts.

Source: CCSS ELA Standards and Appendix B

Will the CCSS bring together teaching, learning, assessment, and accountability suited to 21st century realities? Will these new standards provide an opportunity to build assessment, especially Performance Assessment (PA), as needed to measure results of complex instruction, focused on what is important in learning, inform further instruction, motivate and improve both student learning outcomes and teacher understanding?

Performance assessment, in being integrated into the learning context of the CCSS, has the capacity to move us forward in bringing together the twin aims of standards-based learning: equity in opportunities to learn, and excellence in meeting proficiency or readiness

targets. If the performance assessments used for external accountability measures are built on constructs that include not only content areas, but cognitive skills like having to explain, PA may be more able to serve as a measure of teachers' success in meeting their responsibility to their students.

An example provided by the Forum for Education and Democracy in a briefing to the U.S. Congress illustrates this potential, using information from Illinois' state test:

Table 2.4 Comparison of Performance Assessment and Multiple Choice Tests

Standard	Multiple Choice Test (current)	PA (possible)
Grade 8 Science 11B 'Technological design: Assess given test results on a prototype; analyze data and rebuild and retest prototype as necessary'	'What should Josh do if his first prototype sinks?' Desired answer: 'Change the design and retest his boat.'	Students are given clay, drinking straw, paper, and design a sailboat that will sail across a small body of water. Students can test and retest their designs. In the course of this assessment, students explore physics questions like displacement, and see real-world applications. They are likely to be more engaged, and learn while being "assessed." They use the scientific process and terminology, as well as a variety of cognitive skills to conduct hands-on inquiry.

Source: Wood, G., Darling-Hammond, L., Neill, M., & Roschewski, P. (2007). *Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills.* Forum for Education and Democracy: Brief to U.S. Congress.

Chapter Three

Performance Tasks

So, you are about to design and implement a performance assessment. Perhaps you have determined that implementing a performance assessment would be a worthwhile investment in your students and yourself as an educator. It may be that you are required to have your students complete a performance assessment. For whatever reason, there are critical decisions to make in choosing and designing the type of performance assessment to use in order to get the best results.

As mentioned in the introduction, you will want to use performance assessments under certain circumstances and for specific purposes. Some common uses for assessment are: helping you learn what students know and don't know, teaching a skill while simultaneously assessing students' understanding of the skill, and monitoring progress toward a given objective.

The primary focus in this chapter will be on the design of a quality **performance task** (see *Types of Performance Assessment*,.). A performance task is a "real or simulated situation that requires students to generate one or more products or performances in order to acquire mastery of identified learning outcomes (if the performance task is being administered for the purpose of instruction or formative assessment) or in order to demonstrate mastery of identified learning outcomes (if the performance task is being administered for the purpose of summative assessment)" (Gingrasso et al., 2009, p. 3). Typically PA is made up of a collection of performance tasks (Stecher, 2010, p. 3).

Experts in the field emphasize that any effective performance assessment task should have the following design features.

- Students should be active participants, not passive "selectors of the single right answer."
- Intended outcomes should be clearly identified and should guide the design of a performance task.
- Students should be expected to demonstrate mastery of those intended outcomes when responding to all facets of the task.

- Students must demonstrate their ability to apply their knowledge and skills to reality-based situations and scenarios.
- A clear, logical set of performance-based activities that students are expected to follow should be evident.
- A clearly presented set of criteria should be available to help judge the degree of proficiency in a student response (Maryland, 2011, p. 1).

Development of a task, in a collaborative setting with diverse experts interested in measuring a similar objective can provide an opportunity to share ideas and to design a task that measures an interdisciplinary skill set.

The next few pages will highlight the foundational, overlapping components of the task development structure. The development of performance assessment tasks can be broken down into three comprehensive steps: the development of clearly defined objectives and goals, the development or choice of performance activity and finally the development of clear performance and scoring criteria.

Developing a Performance Assessment Task

Step 1: Clearly define learning goals and objectives:

Goals and objectives provide a framework for learning and therefore provide a framework for the development of assessment. Goals and objectives can be defined as "broad statements of expected student outcomes" where "objectives divide the goals into observable behaviors" (Moskal, 2003, *Writing Goals and Objectives* section, paragraph. 1). Objectives lay the framework upon which a given goal is evaluated" (Moskal, 2003, ¶ 3). To be most effective in the classroom, teachers need to align the goals and objectives for both instruction and assessment.

The development of clear objectives and goals is part of the critical analysis of the concepts, skills, and knowledge needing assessment. According to Allen at the Wisconsin Education Association Council, "one should begin by identifying the types of knowledge and skills students are expected to learn and practice. These should be of high value, worth teaching to and worth learning" (Allen, n.d., para. 10). At this stage, developers needs to

consider what it is they want their students to know, how they want to assess the demonstration of knowledge, and in what ways they will determine or identify the appropriate display of skill.

To help developers identify the important skills, concepts, and knowledge to be learned and practiced and thereby capture the essential goals and objectives, consider the questions and examples below.

Educators need to reflect on the following five questions and examples as they identify what is to be learned or practiced by completing a performance task:

1. What important cognitive skills or attributes do I want my students to develop? (e.g., to communicate effectively in writing; to analyze issues using primary source and reference materials; to use algebra to solve everyday problems).
2. What social and affective skills or attributes do I want my students to develop? (e.g., to work independently, to work cooperatively with others, to have confidence in their abilities, to be conscientious).
3. What *metacognitive* skills do I want my students to develop? (e.g., to reflect on the writing process they use; to evaluate the effectiveness of their research strategies; to review their progress over time).
4. What types of problems do I want them to be able to solve? (e.g., to undertake research; to understand the types of practical problems that geometry will help them solve; to solve problems which have no single, correct answer)
5. What concepts and principles do I want my students to be able to apply? (e.g., to understand cause-and-effect relationships, to apply principles of ecology and conservation in everyday lives) (Herman, Aschbacher, & Winters,1992, pp.25-26, as cited in Allen, n.d., *Performance Assessment, Developing Performance* Tasks, para. 4)).

The development of clear learning and assessment objectives and goals will help guide you, the developer, in determining the appropriate type of assessment or performance to administer. For example, if you are interested in a student's memorization of the periodic table of elements, then a multiple choice or short response assessment may be more

appropriate. However, if you are interested in assessing higher order thinking skills and complex learning required for synthesizing and evaluating information, such as assessing a student's understanding of a chemical reaction, then a performance assessment may be most appropriate.

Step 2: Choose and develop performance task:

Once you identify that a performance assessment is the best form of evaluation, you need to choose and develop an appropriate performance task. You should ensure that the objectives and goals clearly align with measurable outcomes of the performance and that the task represents realistic or attainable goals for teaching and learning.

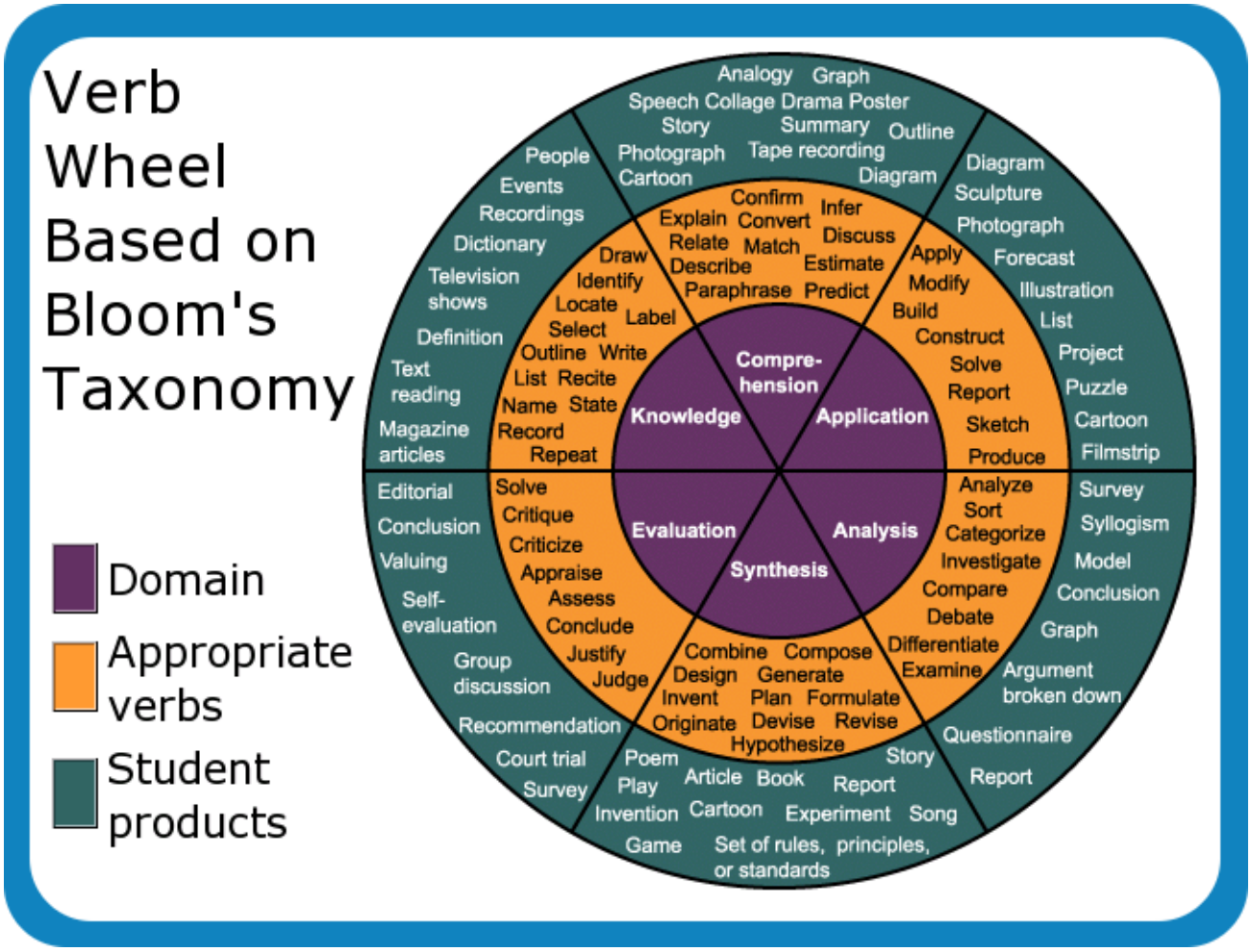
Some questions which may be helpful in guiding the process of developing performance tasks are:

1. How much time will it take to develop or acquire the skill or accomplishment?
2. How does the desired skill or accomplishment relate to other complex cognitive, social, and affective skills?
3. How does the desired skill or accomplishment relate to long-term school and curricular goals?
4. How does the desired skill relate to the school improvement plan?
5. What is the intrinsic importance of the desired skills or accomplishments?
6. Are the desired skills and accomplishments teachable and attainable for your students? (Herman, Aschbacher, & Winters, 1992, p. 31).

In summary, a task should not be overly complex, should hold the interest of the students, apply to a variety of situations, remain integral to long-range goals, teach to important and valuable skills, and be teachable and attainable (Herman et al., 1992). In addition, Brualdi (1998) found that attention to factors such as, "available resources, time constraints, and the amount of data required to make an adequate evaluation of the student's performance" must be considered (as cited in Wren, 2009, p. 1).

The development of an authentic performance assessment, geared towards the teaching of real-life skills, is often viewed as more valuable to the student as it has real life purpose and/or consequence. According to Wiggins (1990), "The best tests always teach students and teachers alike the kind of work that most matters; they are enabling and forward-looking, not just reflective of prior teaching" (as cited in Moskal, 2003, Developing Performance Assessment, paragraph 2). To provide a rich and valuable learning experience, it is important to develop tasks that align with long-term goals or to the 21st Century Skills needed to succeed beyond the schooling years. The incorporation of valuable 21st Century skills such as cross-cultural competence and technology, information, and media literacy will play a big part in the future development of performance assessment tasks.

For helpful guidance, the developer may want to revisit the critical domains of Bloom's taxonomy or consult the diagram below.



Source: *http://smu.edu/ir/assessment.htm 8/23/12*

Step 3: Develop clear performance and scoring criteria:

The final step in the development of a performance assessment is the articulation of clear performance and scoring criteria which align with the stated objective and goals. Clearly defined performance and scoring criteria should aim to help students identify the vital or necessary expectations of the final product or outcome and should ensure objective and consistent scoring over time and over different populations. Stiggins (1991) notes that "if a teacher fails to have a clear sense of the full dimensions of performance, ranging from poor or unacceptable to exemplary, he or she will not be able to teach students to perform at the highest levels or help students to evaluate their own performance" (as cited in Allen, Performance Assessment, Performance Criteria, paragraph 2). Performance and scoring criteria should include a complete description of the attributes being evaluated on a performance continuum or scale. The continuum model is necessary for providing valuable feedback to students on how to directly improve their performance score, not just assigning winners and losers.

Scoring *rubrics* are an important tool for providing a means to evaluate students' performance on a task. A rubric can be defined as "a set of criteria for grading assignments" (Rezaei, & Lovorn, 2010) and can be used as a "scoring instrument that specifies expectations for a given task by dividing it into its component parts and providing a detailed description of what constitutes various levels of performance for each of those parts" (Latimer, Bergee, & Cohen, 2010, p. 168).

The development of an effective rubric is as complex as the development of the performance assessment task itself; therefore, we have dedicated a separate, more specific chapter to the development and use of rubrics later in this handbook. For a quick preview, a review of Moskal's (2003) six general guidelines for the development of rubrics proves informative:

- The criteria set forth within a scoring rubric should be clearly aligned with the requirements of the task and the stated goals and objectives.

- The criteria set forth in scoring rubrics should be expressed in terms of observable behaviors or product characteristics.
- Scoring rubrics should be written in specific and clear language that the students understand.
- The number of points that are used in the scoring rubric should make sense.
- The separation between score levels should be clear.
    - The statement of the criteria should be fair and free from bias. (Moskal, B. 2003 http://pareonline.net/getvn.asp?v=8&n=14)

Technology: How can it help?

Every day, technology is playing a larger and more crucial role in education. "Advances in curriculum, instruction, assessment, and technology are likely to continue to move educational practice toward more individualized and mastery oriented approaches to learning. This evolution will occur across the K–16+ spectrum" (Pellegrino & Quellmalz, 2010, p.131).

Technology: A retail model for educational assessment

IIn the near future, our relationship to the information gleaned from educational assessments may more closely resemble a supermarket, or retail model of information processing. Just as barcodes, scanners and instant checkouts give retailers instant information about consumers, inventory and buying trends, so too may performance assessments (conducted using technology) give instant feedback on student progress, areas of improvement and mastery of skills. The days when schools were forced to interrupt normal instruction to administer mandated assessments would be a distant memory. (Pellegrino & Quellmalz, 2010, p. 131).

Currently, the Partnership for the Assessment of Readiness for College and Careers and SMARTER Balanced Assessment Consortium promise computer-based data systems with quick turn-around, computer-enhanced and technology- enhanced performance tasks (both of these federally funded consortia are involved in the development of assessments for Common Core State Standards.)

According to the Maryland Department of Staff Development,(n.d.), "Quality performance assessments such as simulations, projects and essays are the most authentic and cognitively complex of test formats." They are also the most expensive and require significant human resources to implement. (See Table 1 *Test Format Continuums* in the introduction chapter.)

[D]iagnostic assessments of individuals' learning, for example, must involve collecting, interpreting, and reporting significant amounts of information. No educator - whether a classroom teacher or other user of assessment data - could realistically be expected to handle the information flow, analysis demands, and decision-making burdens involved without technological support. Thus, technology removes some of the constraints that previously made high-quality formative assessment of complex performances difficult or impractical for a classroom teacher. (Pellegrino & Quellmalz, 2010, p.130)

According to Quellmalz, et al. (2009, p.77)

> Information and communications technologies such as Web browsers, word processors, editing, drawing, simulations, and multimedia programs support a variety of research, design, composition, and communication processes. These same tools can expand the cognitive skills that can be assessed, including the processes of planning, drafting, composing, and revising.

Recent examples illustrate the increasing and innovative use of technology in performance assessment:

- Katz writes (2007, p.9)

  > In 2006, Educational Testing Service (ETS) administers iSkills assessment tests to almost 6,400 students at sixty-three institutions. Students respond to fifteen interactive, performance-based tasks to assess Information and Communications Technology management skills. Each interactive task presents a real-world scenario, such as a class or work assignment, that frames the information problem. Students solve information-handling tasks in the context of simulated software (for example, e-mail, Web browser, library database) having the look and feel of typical applications.

- Pellegrino & Quellmalz (2010, p.121)

  > In 2006, the Programme for International Student Assessment (PISA) pilot tested a Computer-Based Assessment of Science to test knowledge and inquiry processes not assessed in the paper-based booklets. The assessment included such student explorations as the genetic breeding of plants.

Pellegrino & Quellmalz (2010) also discuss how technology is being integrated into the NAEP writing assessment, the new NAEP Technology and Engineering Literacy Framework, and Minnesota's use of simulations to test science skills.

In addition to online student assessments, web- based resources for educators will almost certainly play an increasing part in the development of quality performance tasks. The Assessment Design and Delivery System (ADDS) is one such design that is already in use. According to Vendlinski, Niemi, & Wang (n.d.), ADDS offers tools for educators in designing quality Performance Tasks. ADDS is a web-based assessment design tool developed by teachers and researchers at UCLA's Center for Research on Evaluation, Standards, and Student Testing (CRESST). The goal of ADDS is to improve the quality of classroom assessments. The tool is applicable across subject domains.

As part of the design process, teachers decide the attributes of an assessment as well as the context and type of responses the students will generate. ADDS has a database of various types of questions (essay, problem solving, concept mapping, simulations, and selected response) which teachers can choose from. Teachers are also free to upload their own questions. The tool allows teachers to deliver the assessments to their students either online or on paper

Other Considerations

There is a variety of information available to educators and other professionals wanting guidance on the creation and use of performance assessment tasks. There is no need to reinvent the wheel each time you want to use a performance assessment. However, having a solid understanding of the development process will help you identify the necessary and inevitable revisions that are required as conditions change.. In addition to the development steps, there are two critical considerations needing attention each time you administer any assessment; *test validity* and *inter-rater reliability* of scoring criteria. These will be addressed in detail in a later chapter, for now we will introduce some of the key ideas.

Validity can be defined as "the process and outcome of collecting and interpreting results of assessments or measurement so that inferences from findings are warranted by evidence" (Baker, Chung, & Delacruz, 2008, p. 596). Payne (2003) suggests a simple approach to understanding validity is to ensure that "a test does the job for which it is used" (as cited in Wren, 2009, p.5). Quality feedback on the teaching and learning process depends entirely on uncompromised validity of the interpretation of the results of the task.

Moskal (2003) provides recommendations important to task validity. First, she recommends that "the task should not examine extraneous or unintended variables" and that the task "should be fair and free from bias" (Developing Performance Assessment, paragraphs 5 & 6); for example when the language in word problems interfere with the student's ability to demonstrate their math knowledge. As stated by Brualdi (1999) one threat to test validity is the underrepresentation of the primary variables, meaning, "the tasks which are measured in the assessment fail to include important dimensions or facets of the construct" (p. 4). In addition to the threat of under representation, developers must also reevaluate tasks which end up measuring too many or irrelevant variables (Brualdi, 1999). To identify the measurement of unintended or arbitrary variables and biases, administer the task as a pilot study where feedback can be collected and used for any necessary revisions.

In addition to test validity, it is vital that the accompanying scoring rubric have a high degree of *reliability*. Herman et el. (1992) have found that a rater should feel confident "that the grade or judgment" is "a result of the actual performance, not some superficial

aspect of the product or scoring situation" (as cited in Wren, 2009, p.7). In addition to the confidence and consistency needed by an individual rater, the scoring should remain consistent across groups of raters. In essence, the search for consistency and inter-rater reliability is the ultimate purpose behind the design and implementation of the scoring rubric. Before administration, the rubric should be tested for reliability in a pilot study. In situations where a rubric is shared by a number of teachers, an additional step for ensuring reliable use of rubrics is to train raters. Provide exemplars for each level of performance, norm the process with test runs and comparative scoring, and check rating consistency (Wren, 2009).

Though the technical language associated with performance assessments may seem new, and technological advances may seem daunting, Allen (2011) reminds us "the concepts of authentic, or performance assessments are not new. Teachers always have assigned tasks which require their students to perform or develop products" (p. 1).

Now that you've read the criteria and primary steps for the development of quality performance assessment tasks, let's look at some examples. The following four sample tasks were excerpted from the website http://jfmueller.faculty.noctrl.edu/toolbox/index.htm *Authentic Assessment Toolbox*, created by Jon Mueller, author of *Assessing Critical Skills.*

Performance Tasks
Avoiding Critical Defects

Educational Testing Service (2011) notes, "Many things can go wrong with a task. Even if the content, in concept, is perfect for the job, in execution the task might have problems. For example, asking students to work in groups to build a mousetrap car to assess student understanding of mechanics, problem solving, scientific method, group process skills, and communication skills, might be a good idea—the content is sound. But, the actual instructions might not be clear (clarity), or the resources or time might not be equally available to all students (fairness), or it might be hard to judge individual skills in the context of group work (accuracy). A good task avoids these sources of bias and distortion" (p. 1).
(Educational Testing Service, 2011, p. 2)

Example: Persuasive Letter-Writing -Voice, Word Choice and Organization

Standards:

• Demonstrate focus and organization in written compositions.

• Write for a variety of purposes including description, information, explanation, persuasion and narration.

• Use describing words to enhance writing.

• Write with enthusiasm and personality.

• Writer speaks to the reader in an engaging way.

Task:

Think about something you really want your parents to let you do or something you want them to buy for you. Write a letter to your mom or dad persuading them to get this item for you. Remember the correct format of writing a letter: date, greeting, body, closing. Describe the item that you want. Remember to use adjectives and descriptive language to describe the item to your parents. Include three reasons why your parents should get you this item.

Table 3.1 Example Scoring Rubric

	3	**2**	**1**
Reasons	Has provided three different reasons that support the position.	Has provided two different reasons to support the position.	Has provided one or zero different reasons to support the position.
Descriptive Language	Consistently uses precise, fresh and original words which create vivid images.	Attempts to use descriptive words to create images.	Attempts new words but they don't always fit or uses general or ordinary words.
Letter format	Includes all four of the elements of a friendly letter.	Includes three of the four elements of a friendly letter.	Includes 0-2 elements of a friendly letter.
Voice	Point of view is clearly expressed, test elicits emotions, writer cares deeply about the topic and has clear sense of audience.	Attempts to express point of view, writing is expressive and shows perspective, writer conveys idea to the reader.	Expresses some predictable feelings, audience is fuzzy and reader has limited connection to writer.

(Mueller, 2011)

Portfolios

Portfolios are a common way to collect and demonstrate performance over time. There are two basic types of portfolios. First the formative portfolio preserves all of a student's work showing the evolution of their mastery. A "Best Pieces" portfolio requires students to judge which of their pieces of work to include and explain why each piece was selected.

According to Stecher (2010), "for portfolios to be useful as performance assessments, they must be standardized; that is, all students collect the same work products, and those work products are produced under similar conditions. In theory, it is easy to meet the former criterion but more difficult to achieve the latter" (p. 8).

Classroom Assessment Building (2011) defines a portfolio as a purposeful collection of artifacts that tells the story of a person and her/his skills, achievements, and/or growth, illustrated by a selection of her/his work. The selection of portfolio content and material should be based upon goals and standards, and it should include a broad range of accomplishments (including products, essays, quizzes, hobbies, etc.). Student portfolios can take many forms, including paper or electronic.

Electronic, multimedia portfolios have become increasingly popular. These digital portfolios combine textual, visual, and auditory components or artifacts (paper and electronic) that have been digitized for viewing on a computer or other digital viewing device. Multimedia portfolios can be in the form of videotapes or files on a disk, CD-ROM, DVD, or the World Wide Web. In some cases, the portfolios are created "by scratch" in a word processor or on a website; in other cases, a school might license a commercial portfolio program/manager. Electronic portfolios often include recitations, reports, presentations, or performances by the students (*Classroom Assessment Building*, 2011, p. 1).

In the past there have been several concerns regarding portfolios. First, students may differ in their ability or motivation to select the best entries in the portfolio. As the collection of materials is diverse, it is difficult to develop unambiguous and consistent scoring. Perhaps the single most common complaint is the time it takes to use portfolios effectively. When using portfolios to provide an ongoing record of student progress, ask the student to analyze

the materials and share their perspectives. Often a more tightly focused portfolio, say on essay writing, is more useful and produces better data than more diverse applications.

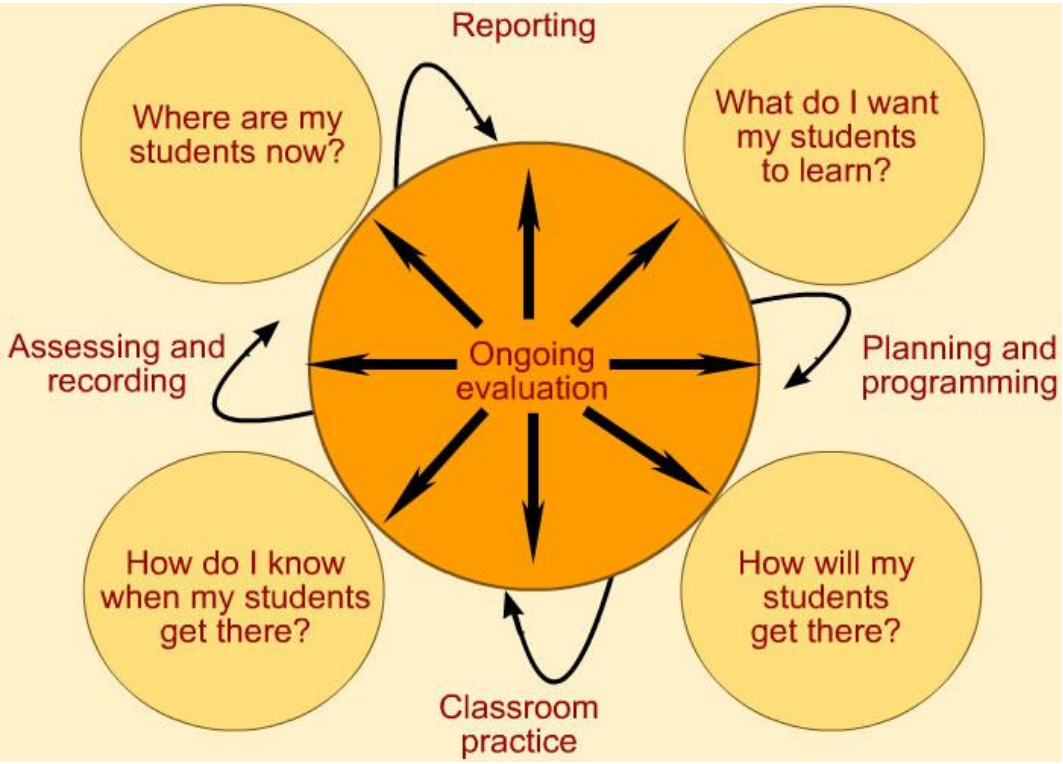
Chapter Four

Applications of Performance Assessments

How can performance assessment be most effective in the teaching and learning continuum?

The *teaching-learning continuum* is a dynamic relationship among students, teachers, knowledge, and skills, with evaluation at the center of all aspects of these interactions. The New South Wales Department of Education and Training materials state:

> The teaching and learning cycle represents the four stages that occur in the design and delivery of classroom tasks that incorporate an outcomes-based approach. The cycle has no start or end point, with each step informing the next. The process of gathering data and reflection dictates where in the cycle you need to be operating. (New South Wales Department of Education and Training, 2011, p. 1)



When applied throughout the learning continuum, performance assessment carries with it the potential to maximize student learning and improve the quality of instruction. Applying a performance assessment to a specific situation requires knowing which of two types to use: ***summative*** or ***formative***. ***Summative*** assessments, or assessment *of* learning, are often traditional pencil-paper measures of student work after the fact: standardized and customized tests, chapter reviews, quizzes, projects, or essays gathered at the end of a unit, semester, or course. ***Formative*** assessments (alternatively known as assessment *for* learning and embedded assessments) take place during the lesson, unit, or semester and inform both teacher and student of progress, understanding, and shortfalls in students' work. "When a comprehensive assessment program at the classroom level balances formative and summative student learning/achievement information, a clear picture emerges of where a student is relative to learning targets and standards" (Garrison, & Ehringhaus, 2007, p. 3).

New assessment systems are being developed which are both summative and formative; they include performance type responses, engage higher order thinking, and satisfy the sociopolitical desire for high-stakes testing (Way et al., 2011).

Learning Theories

Current research is inspiring a paradigm shift in regard to educational reform to meet the needs of the 21st century in which higher orders of thinking are becoming a necessity to succeed in the global community. Not only is this movement taking hold in the United States with the introduction of the Common Core State Standards, but the Education Minister of Singapore, Tharman Shanmugaratnam, stated recently:

> [We need] less dependence on rote learning, repetitive tests and a 'one size fits all' type of instruction, and more on engaged learning, discovery through experiences, differentiated teaching, the learning of life-long skills, and the building of character, so that students can… develop the attributes, mindsets, character and values for future success" (Darling-Hammond & Adamson, 2010, p.2).

Although the Minister of Education perfectly summarizes the need for the development of 21st century skills, he neglects to mention how those skills will be instructed or assessed. Learning theory has much to offer to guide construction of high-quality learning and testing products (Ertmer & Newby, 1993). "Careful examination of learning theory can yield linkages that ground design principles, provide reasoning for day-to-day design decisions, and can even offer assumptions useful for testing the viability of programs and products" (Zane, 2009, p.81).

What is a Theory

· A theory explains relationships between variables discovered through observations, with the goal of understanding, controlling, and predicting the phenomenon in the future. Theories evolve as new information, definitions, and forces come into play. The following are brief introductions to some of the commonly used theories in the field of education.

Behaviorism:

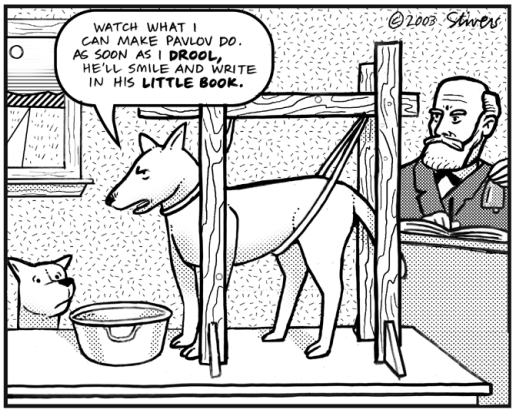
The behaviorist theory was popularized by Pavlov whose observations of the stimulus/response behavior of dogs led to the conclusion that when there is a response provoked by a stimulus, a conditioned behavior is created. The theory was later modified by Skinner who proposed that responses which are either rewarded or punished (providing a stimulus) create an atmosphere for repeatable learning (initiating a response). These observable behaviors are easy to quantify allowing subjects to be tested more easily.

Behaviorist learning allows:

- an exact score to be provided for precise testing results;
- simple standardization, and the focus on observable behaviors promotes the comparison of students with each other in terms of percentile ranks or grade equivalents;
- schools and school districts to be compared; and
- a single test which provides proof of achievement or lack thereof

(Ediger, 1993, p. 181).

The major flaw to this theory is that it treats the mind as a black box, focusing exclusively on input/output behaviors. The lack of interest of the internal world of the students results in less accurate predictions of complex behaviors and misses entirely the social, cognitive, and emotional precursors to observed behaviors.



Tasks requiring a low degree of processing (e.g., basic paired associations and rote memorization) seem to be facilitated by strategies most frequently associated with a behaviorism (e.g., stimulus-response and the connection to feedback) (Ertmer, & Newby, 1993).

(Stevens, retrieved 8/2/12, p.1)

Humanism

This theory represents a holistic approach of instructing a child in a fashion that is student centered with an emphasis on intrinsic motivation which establishes goals and fosters core values. In Maslow's Hierarchy of Needs, he suggests that teachers need to address the development of the whole child and through that process self-actualization will occur (Simons, 1987). According to this theory, once students realize what they were born to do (self-actualization), they can reach higher levels of thinking and understanding.

Cognitive Development

The term schema, coined by Jean Piaget in 1926, refers to the mental constructs that aid us in interpreting the world around us. Individuals search for further understanding in this world by creating and modifying schemas (Eggan & Kauchak, 2004). Without tying new information to established cognitive frameworks, students will have difficulties with learning in the future. Tasks that reflect the tenets of cognitive development require a higher level of processing (e.g., inquiring classifications, rule, or procedural executions) are primarily associated with strategies

having a stronger cognitive emphasis (e.g., schematic organization, analogical reasoning, algorithmic problem solving) (Ertmer & Newby, 1993).

Constructivism

Constructivists believe that "learners construct their own reality or at least interpret it based upon their perceptions of experiences, so an individual's knowledge is a function of one's prior experiences, mental structures, and beliefs that are used to interpret objects and events" (Jonessan, 1994, p. 182). What students know is a reflection of their social interactions and past experiences which are later processed to create an atmosphere for future learning. Tasks demanding high levels of processing (e.g., inquisitive problem solving, personal selection, and monitoring of cognitive strategies) are frequently learned with strategies that reflect the constructivist perspective (e.g., social and situated learning theories) (Ertmer & Newby, 1993).

Social Learning Theory

This theory suggests that the observation of others' attitudes, behaviors, and results of those behaviors generate participation in the learning process.

> Most human behavior is learned observationally through modeling: from observing others, one forms an idea of how new behaviors are performed, and on later occasions this coded information serves as a guide for action. (Bandura, 1977, p16).

Social learning theory explains the human learning process in terms of a continuous interaction among cognitive, behavioral, social, and environmental influences.

Situated Learning

This theory developed by Jean Lave argues that context is critical for learning and that most classroom learning activities that involve meaningless or abstract knowledge are ineffective because they exist without the rich connections possible when learning occurs in context. She argues that learning is a situational experience and that learning is embedded within an activity, the cultural context as well as the social interaction. It is also usually unintentional rather than deliberate (Lave, 1991).

Bloom's Taxonomy compared to a revised version

The following chart differentiates between the Original Bloom's Taxonomy (on the left) which was created in 1956 to the revised version (on the right) published in 2001 that was created by former Bloom student, Lorin Anderson and fellow Bloom researcher David Karthwohl. "The major differences in the updated version are in the more useful and comprehensive additions of how the taxonomy intersects and acts upon different types and levels of knowledge -- factual, conceptual, procedural and the metacognitive" (Wilson, 2005**).** This feature has the potential to make teacher assessment, teacher self-assessment, and student assessment easier or clearer as usage patterns emerge (Wilson, 2005).

Table 4.1 Taxonomies of the Cognitive Domain

Bloom's Taxonomy 1956	Anderson and Krathwohl's Taxonomy 2000
1. Knowledge: Remembering or retrieving previously learned material. Examples of verbs that relate to this function are: know, define, record identify, recall, name relate, memorize, recognize list, repeat, acquire	1. Remembering: Retrieving, recalling, or recognizing knowledge from memory. Remembering is when memory is used to produce definitions, facts, or lists, or recite or retrieve material.
2. Comprehension: The ability to grasp or construct meaning from material. Examples of verbs that relate to this function are: restate, identify, illustrate locate, discuss, interpret report, describe, draw recognize, review, represent explain, infer, differentiate express, conclude	2. Understanding: Constructing meaning from different types of functions be they written or graphic messages activities like i**nterpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.**
3. Application: The ability to use learned material, or to implement material in new and concrete situations. Examples of verbs that relate to this function are: apply, organize, practice relate, employ, calculate develop, restructure, show translate, interpret, exhibit use, demonstrate, dramatize	3. Applying: Carrying out or using a procedure through **executing, or implementing**. Applying related and refers to situations where learned material is used through products like models, presentations, interviews or simulations.

4. Analysis: The ability to break down or distinguish the parts of material into its components so that its organizational structure may be better understood. Examples of verbs that relate to this function are: analyze, differentiate, experiment compare, contrast, scrutinize probe, investigate, discover inquire, detect, inspect examine, survey, dissect contrast, classify, discriminate categorize, deduce, separate	4. Analyzing: Breaking material or concepts into parts, determining how the parts relate or interrelate to one another or to an overall structure or purpose. Mental actions included in this function are **differentiating, organizing, and attributing**, as well as **being able to distinguish between** the components or parts. When one is analyzing he/she can illustrate this mental function by creating spreadsheets, surveys, charts, or diagrams, or graphic representations.
5. Synthesis: The ability to put parts together to form a coherent or unique new whole. Examples of verbs that relate to this function are: compose, plan, propose produce, invent, develop design, formulate, arrange assemble, collect, construct create, set up, organize prepare, generalize, originate predict, document, derive modify, combine, write tell, relate, propose	5. Evaluating: Making judgments based on criteria and standards through **checking and critiquing**. Critiques, recommendations, and reports are some of the products that can be created to demonstrate the processes of evaluation. In the newer taxonomy evaluation comes before creating as it is often a necessary part of the precursory behavior before creating something. **Remember this one has now changed places with the last one on the other side.**
6. Evaluation: The ability to judge, check, and even critique the value of material for a given purpose. Examples of verbs that relate to this function are: judge, argue, validate assess, decide, consider compare, choose, appraise evaluate, rate, value conclude, select, criticize measure, estimate, infer deduce	6. Creating: Putting elements together to form a coherent or functional whole; **reorganizing** elements into a new pattern or structure through **generating, planning, or producing**. Creating requires users to put parts together in a new way or synthesize parts into something new and different a new form or product. This process is the most difficult mental function in the new taxonomy. **This one used to be #5 in Bloom's known as synthesis.**

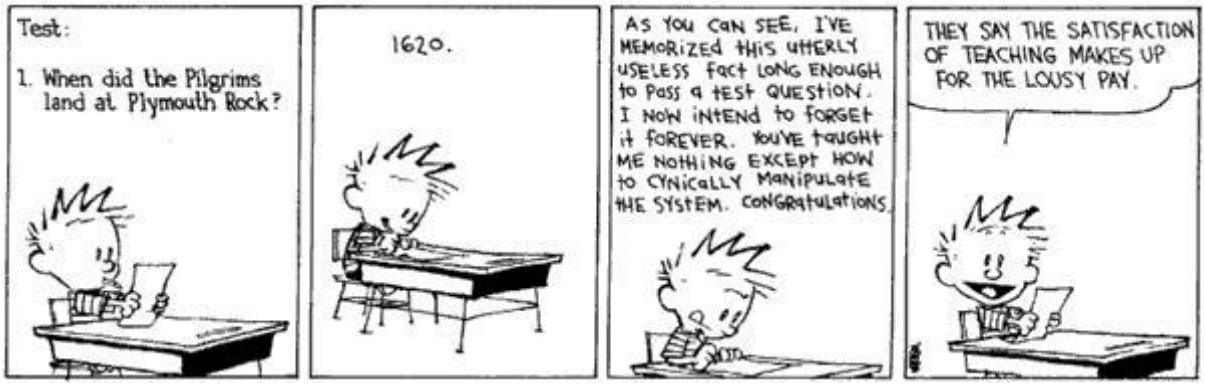
(Wilson, 2006, p. 1)

"Well-founded assessments provide an environment where students learn, struggle, produce, and then—by the way—are scored on their performance. Clearly, this sort of assessment is more complex than traditional end-of-unit mastery checking" (Zane, p. 86). These 21st Century assessment skills need to be acquired by the teacher to maximize the instructor's effectiveness.

Assessment Literacy

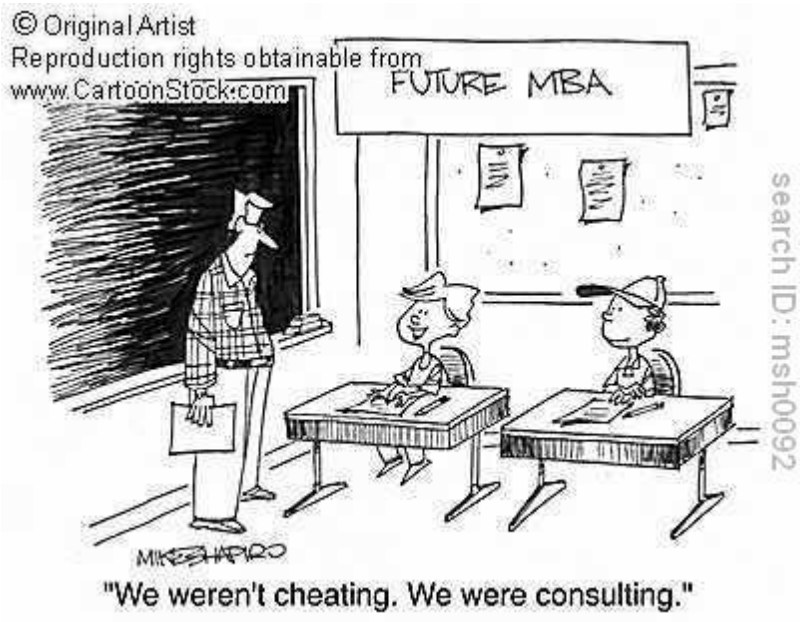
While a balanced assessment program can benefit students' educational performance and teachers' professional competencies, it is clear that most teachers do not have high *assessment literacy*. That is, they are often insufficiently trained to administer appropriate assessments in the classroom. "[E]ducators making … assessment-dependent decisions are doing so without a genuine understanding of educational assessment" (Popham, 2009, p. 1). Since evaluation ties together all aspects of the teaching-learning continuum, it is critical that teachers have a firm grasp of appropriate assessments for all phases of instruction.

Popham's experience as a former high school teacher and scholarship as a professor at UCLA provide a solid foundation for his views. He argues that "*Assessment literacy* is present when a person possesses the assessment-related knowledge and skills needed for the competent performance of that person's responsibilities" (Popham, 2009, p. 1). That is, teachers who decide which types of tests to give and when to give them have different responsibilities from their administrators. Students also need a certain amount of assessment literacy, especially when it comes to understanding the importance of the assessments their teachers assign throughout the year. Parents as well should know what tests their children are taking and why. All participants in the education process share the responsibility for assessment results.



Student and teacher roles in performance assessment

In the most literal sense, students can be given roles or jobs in order to most effectively complete a performance assessment: illustrator, reporter, civic planner, biologist, ship captain, taxi driver, etc. (Mc Tighe, 1998). If we are to provide our students 21st Century skills, imagining their possible jobs within the near future will help them contextualize open-ended problems designed at evaluating higher order thinking. After all, imagination is a requirement for creating answers based on unseen factors.



Students are often expected to collaborate with one another while staying on task when completing performance assessments. Active learning, questioning, research, and intellectual risk-taking will also lead to student successes (Office of Research,

1993). Teachers and students then enter into a relationship that reflects a cognitive apprenticeship, whereby learning partners explain and explore relationships between and among central concepts of content area, while discussing, reflecting, and creating new understanding of the material, relating it to personal experience and public issues (Newmann, Marks, & Gamoran, 1996).

Practical roles in the student-centered classroom, where the teacher acts as moderator and coach during content area lessons, include:

1. Having students think aloud when problem-solving;
2. Having students learn to label and understand the kinds of thinking they are using;
3. Asking students to probe each other's reasoning;
4. Offering students repeated opportunities to participate in developing sample test items that tap into reasoning;
5. Waiting for a response when asking questions that require reflection and thought;
6. Avoiding questions that call for a yes or no answer;
7. Using concept mapping;
8. Having students design essay and performance assessments that measure reasoning;
9. Setting up a classroom display on the kinds of reasoning valued (Stiggins, 1997).

One of the most crucial roles for teachers and students alike is that of rubric maker. When students are involved in creating *rubrics* (a process known as ***negotiable contracting)*** for their own performance assessments, a number of beneficial things occur.

> Given the appropriate direction by their teachers, youngsters are able to accurately evaluate their strengths and weaknesses and pinpoint where to focus their efforts to get the most out of what they're learning. As a result, students view assessment not as an arbitrary form of reward or humiliation (a common perception of middle school students), but as a positive tool for personal growth. (Stix, 1997, p. 1).

Students also begin using and accumulating academic language for both content area knowledge and assessment literacy as they participate in discussions during which rubrics are drawn up. This seems conducive to fostering students' awareness of metacognition and ultimately improving higher-level thinking skills on which educators need to focus. One

effective way to begin the process of rubric development with students is to provide them with two examples, an effective and an ineffective model, and to ask them to discuss the differences. This leads to criteria formulation.

Problem based learning

Problem based learning (PBL) is a teaching/learning approach rather than a theory which attempts to model and simulate the real world, creating conditions students observe or will encounter when they enter the workforce ("TIPS for Teachers Project," 2001). As a result, students collaborate with peers and teacher(s), investigate personal or public issues, and create the means to communicate appropriately those ideas to peers, and often, a wider community audience. Because of the dynamic and student-centered approach of PBL, various and ongoing assessments are necessary for both students and teachers to keep track of progress, facilitate understanding, and offer guidance for project completion or problem solution. Also, the open ended nature of this type of teaching/learning requires independence, trust among peers, and a sound work ethic to get the job done.

While PBL is a relatively new style of instruction, it holds the possibility of ensuring mastery of traditional skills and readiness for college or vocational classes. "If you use PBL in the classroom, you are not only teaching the stuff of school, you are supporting the social-emotional development of your students and getting them ready for college" (Markham, 2011, para. 7). PBL starts with a respectful classroom, ongoing learning, and assessment which ensure mastery, and "relevant, open-ended, student-centered question[s] that speak to a student's innate desire to know more about the world and how he or she fits into it" (Markham, 2011, para. 6).

Chapter Five

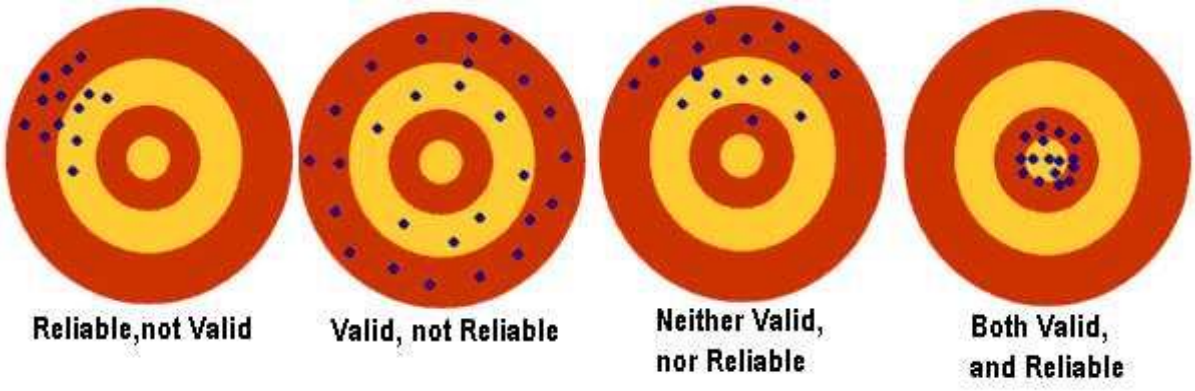
Validity & Reliability

Teaching and learning are a system of communication that entails passing knowledge from one person to another or from groups of people to other groups. To ensure that teaching is equally effective both locally and nationally, it is important to measure the effects of this teaching. By calibrating the teaching and learning process across the educational systems, educators can deliver more content with greater quality and effectiveness. Effective assessment requires an understanding and application of the concepts of validity and reliability. This chapter offers an exploration of validity and reliability. It defines the types of validity and reliability, threats to both, and the ways they interact with one another. It also explores their implications for performance assessment.

Validity and ***reliability*** are often discussed independently of one another. In fact, they are quite related. Reliability can be defined as the consistency of measurement. If you take the measurement over and over, do you get the same answer? Imagine a thermometer that gives radically different temperatures each time you use it. Without consistent measurement, you cannot make valid assessments of what your temperature really is. It is possible to get reliable measurements that are simply wrong or invalid. For example, if you adjust your scale at home so that your weight shows a couple of pounds lighter than it really is, you will get the same weight each time, but it will not be a valid measure of your weight. **A** reliable assessment that is not valid means you are measuring something that is quite repeatable but is not an accurate measure of what you are trying to measure. In contrast, you cannot have a valid assessment that is not reliable because without consistency you have no way of knowing what the results actually mean. This is why valid assessments require reliable measures; however, reliable measures are necessary but not sufficient for valid assessment. If there is one take-home message from the concern for demonstrating *validity* and *reliability* with your performance assessment, it is that you must have both. In order to be valid, the assessment must measure what you want it to measure, and the results must mean what you think they mean. For example, you may give a tenth grader a test of third grade math and get a reliable score, but to interpret the score of the assessment as showing

the student is clearly brilliant in math (after all he got 98%) would be an invalid interpretation.

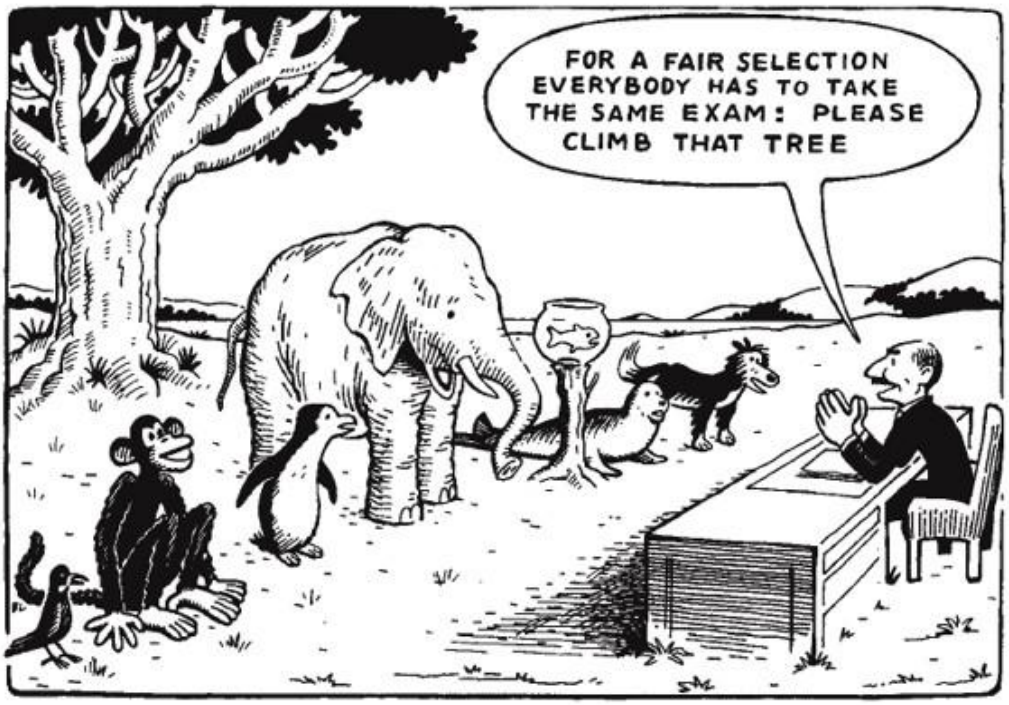
The following diagram is a good visual of the relationship between *validity* and *reliability*. As discussed above, you can have an assessment that is reliable but not valid. There will be consistency in the results, but it will be off-target from the results you were intending to measure. An assessment that is neither valid nor reliable is off-target and inconsistent. A high quality assessment that is a good measure of any theoretical construct or assessment of learning a lesson is valid and reliable, spot-on and consistent. For a quick reference, refer to the diagrams throughout this chapter.



http://www.aspiringminds.in/standardization.html

Validity

Validity is the extent to which an assessment measures what it was intended to measure. Validity indicates the degree of accuracy of either predictions or inferences based upon an assessment score. (CRESST.org, 2008). In essence, you can think of validity (as a stand-alone concept) as meaning that the data you collected mean what you think they mean.



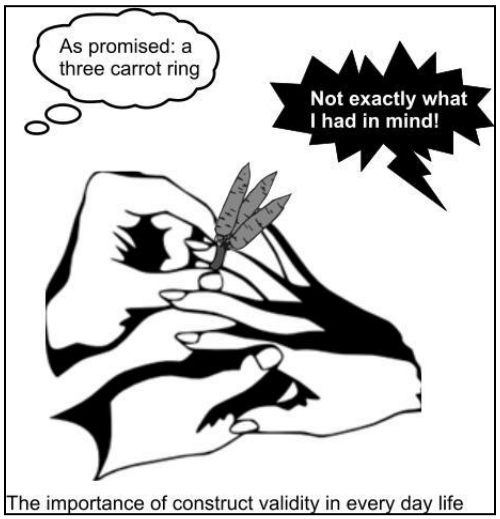
http://edusum.edublogs.org/

Validity is a key component of ensuring confidence in our testing methods. Traditional standardized assessment (such as True/False, multiple-choice, short answer, and essays) can have fairly high validity to the extent that the test measures what the authors intended due to the well-defined and tested protocols. "Among the advantages of conventional testing is its ability to get at specific concepts with relative ease" (Solomon, 1998, p. 111). Performance assessments for more complex types of activities (such as projects, simulations, and work samples) require greater effort to establish validity. These assessments have to account for a larger number of variables that interact to form a basis of experiential knowledge that can be more diverse, diffuse, and difficult to measure. Understanding validity and its role in effective assessment is important in creating meaningful assessments.

There are three common considerations used to establish validity: ***content validity***, ***construct validity***, and ***criterion validity***. A fourth, less commonly referenced consideration is ***consequence validity***.

"Content validity refers to the extent to which the test questions represent the skills being taught in the specified subject area" (Brualdi, 2000, p. 2). The content measured in the assessment should align with the content of the lesson both in terms of the range and intensity. For example, the relative number of items for each learning objective in an assessment should reflect the emphasis during instruction. One method of evaluating content validity involves having a group of subject matter experts (SMEs) rate items on an assessment as one of the following: essential, useful but not essential, or not necessary. If a preponderance of judges knowledgeable on the content that is taught agree that items on the assessment are essential, it's more likely that the assessment has content validity (Lawshe, 1975). A familiar example of assessment in which content validity is a crucial element are standardized state tests (Lissitz, 2007).

TEST Question: Who was Madonna's first husband?



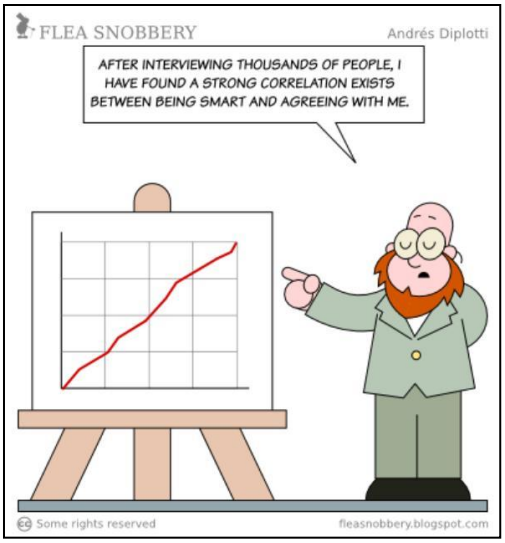
The importance of construct validity in every day life

Construct validity of an assessment means that the assessment measures what it claims to measure based on theory and that it does not measure something else; or if the theory is more complex, it can measure a number of factors that can be determined statistically. You can't see learning, so we have to use some performance (answering questions, applying the knowledge, writing an essay) to use as a proxy for what we really care about: learning. An assessment with significant construct validity will correlate with other measures which measure the same or similar construct. This is known as *convergent validity*. The same assessment will not correlate with measures with which it is not expected to correlate. This is known as discriminate validity. For example, if a math test using word problems correlates with other measures of math ability but less so with literacy assessments, it is said to demonstrate

convergent/divergent properties. On the other hand, if the test with word problems is correlated more strongly with the literacy and less so with the other math measures, then there is evidence that it is the ability to read and comprehend the word problems more than the ability to successfully do the math.

Construct validity is present when the assessment corresponds to the underlying theoretical construct of the material being taught. Is the construct appropriately represented? Constructs can be (a) uni-dimensional (e.g., measuring the length of bones in a skeleton), (b) bi-dimensional (e.g. measuring verbal and math skills of students as in the SAT test), and (c) multi-dimensional (e.g. student aptitude for successful school performance). Clarifying the theoretical construct with all of its relevant dimensions is a critical element in developing an assessment that provides valid evidence of that construct. For example, in the case of intelligence, the traditional construct which characterizes the intelligence on IQ tests as being bi-dimensional with intellect based on linguistic and logic operations would require a different assessment mechanism from one based on Gardner's theory of multiple intelligences. The full construct should be represented in the assessment. "*'Construct underrepresentation'* indicates that the tasks which are measured fail to include important dimensions or facets of the construct. Therefore, the test results are unlikely to reveal a student's true abilities within the intended construct.

Criterion validity seeks to demonstrate that the results of the assessment are correlated with one or more criteria. This is a measure of whether your assessment produces results that are comparable to other known measures of the material being assessed. For example, if students always raises their hand in class and give good, thoughtful answers demonstrating mastery of a topic, but then fail the quiz covering the same material, one must question whether the quiz is somehow measuring something other than what was intended. Do the results represent good



predictors? If people are measuring students' ability to succeed in college, then those who rate higher on assessments should be more successful overall than those whose assessment scores are lower. For example, SAT results are often used in college admission decisions. If SAT results were not correlated to some degree with student success, this lack of criterion validity would make the inferences from SAT scores an invalid interpretation.

Table 5.1 Validity Defined

Aspect of Validity	Definition	Example/Non-Example
Content	The extent to which the content of the test matches the instructional objectives.	A semester or quarter exam that only includes content covered during the last six weeks is not a valid measure of the course's overall objectives -- it has very low content validity.
Construct	The extent to which an assessment corresponds to other variables or assessments, as predicted by some rationale or theory.	If you can correctly hypothesize that ESOL students will perform differently on a reading test than English-speaking students (because of theory), the assessment may have construct validity.
Criterion	The extent to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) an external criterion.	If the end-of-year math tests in 4th grade correlate highly with the statewide math tests, they would have high concurrent validity.

Florida Center for Instructional Technology, (n.d.)

Authors often offer a range of terms for validity measures. Among these, *consequence validity* is uniquely focused not on measurement but on the consequences of the measurement. If a test consistently shows that women are better at personnel management skills than men, and this has the effect of generally excluding men from personnel jobs, then, is the measure valid given the results? *Consequence validity* involves

the argument that the consequences of the measurement may not reflect the capabilities of the people engaged in the measurement. To date, this is not a universally agreed-upon aspect of validity.

Threats to validity

Construct irrelevance refers to factors unrelated to the construct being measured that affect the results. For example, some have argued that high stakes testing creates a stressful environment that reduces test scores in ways that have nothing to do with representing what students know. Flaws in test design, such as having multiple choice questions in which the right answer can be ascertained by eliminating the other choices or that provide a clue to the correct answer by using a word from the answer in the question, are common threats to validity as the test results reflect the students' test-taking strategies rather than their mastery of the content. This is referred to as a problem with instrumentation.

Instrumentation is the design of the test and the test items as discussed above. Does the design of the test interfere with the respondents' ability to show what they know and are able to do? For example, if students have to flip back and forth between pages to answer test questions, this action may be distracting them from the content. Another example is visual interference, in which students are presented with a visually confusing range of information such as a large number of matching items instead of a smaller set.

Reliability

Reliability can most simply be defined as consistency. Types of reliability most commonly addressed are: ***inter-rater reliability, test-retest reliability, and internal consistency***. *Inter-rater reliability* refers to the degree that two (or more) people scoring an assessment produce the same or very similar scores. *Test-retest reliability* measures the degree to which the same assessment given twice produces similar

results. *Internal consistency* is a measure of the coherence of all the items in the assessment protocol. Assuming that the goal is to measure a specific construct, and the items are each focused on that task, then all of the items ought to be correlated with one another. A measure of internal consistency reveals the strength of those correlations.

Table 5.2 Measures of reliability

Type of Reliability	How to Measure
Inter-rater Reliability	Have two different people review the same assessment. Reliability is stated as the level of correlation between the two scores.
Test-Retest	Give the same assessment twice, separated by days, weeks, or months. Reliability is stated as correlation between the results of those two assessments.
Internal Consistency	A measure of internal consistency that measures the strength of the correlations between all items on a test.

Florida Center for Instructional Technology, (n.d.)

In terms of *inter-rater reliability,* there could be human error between two different raters. A number of techniques exist to strengthen *reliability*. The best way to maximize *inter-rater reliability* is to provide good training for raters. Providing anchor papers representing exemplars of each scoring level, doing a norming practice in which the same product is scored separately by raters and then differences are discussed and resolved, and performing ongoing checks on randomly selecting scored instruments to be re-scored by a second rate all improve the consistency of scores.

Test-retest reliability is rarely done with locally developed instruments but may become increasingly common as the accountability movement moves towards a value added approach. Two important considerations include the duration of time between tests (the longer the period the lower the correlation) and the process of administering the assessment.

Measures of internal consistency are based on two mutually supportive theories. First, if the test as a whole or subsets of the test are designed to test various aspects of the same construct, the results should be well correlated, and secondly, students who do well on one question are more likely to do well on other questions. Therefore, measures of internal consistency such as Chronbach's Alpha are used to measure the interblenal consistency (reliability) of instruments.

Validity and reliability are easier to quantify in traditional assessment. It becomes more difficult to quantify validity and reliability in performance assessment because performance assessment has more complexity, has greater diversity of approaches and products, and involves higher levels of thinking. When asked to apply an algorithm to solve a particular type of mathematical problem, test-takers find the process fairly standard, but when asked to design an environmentally friendly pathway through an urban forest, students' responses will be highly variable. Some of the variability can be reduced by restrictive instructions, but if those restrictions force students to adopt a particular approach, then the quality of the performance task and the ability to assess complex problem solving skills are reduced. "One of the challenges in establishing validity for performance assessments is lack of clarity about precisely what the assessments are intended to measure and what relationships ought to be found with other measures of related concepts" (Stecher, 2010, p. 23). This can be especially frustrating for parents who want to support the learning process of their children. One way to involve parents in the understanding of validity and reliability and their relation to performance assessment is to include the parents early in the discussion.

Tips for Parent/Student/Teacher Conferences

• **Involve** the parent and student early in your discussion. Have each of them discuss what they think are the student's strengths and weaknesses.

• Discuss **multiple** indicators of student performance including combinations of diagnostic, standardized, and performance assessments, formal and informal observations. No single assessment, grade, or performance can tell teachers or parents everything they need to know about a child. When possible, parents and students should be provided comparative information that enables them to know how other students are doing either on a local, statewide or a national basis.

• **Discuss student social skills** including the student's ability to work cooperatively and to assume leadership roles when appropriate. Teachers may wish to create a set of criteria and a checklist that assists them in assessing student social skills.

• **Show and discuss examples** of student work such as might be contained in a portfolio. Work should demonstrate a wide range of problem-solving skills as opposed to one or two single best pieces. Work that shows needed areas of improvement and comparative information with other students will be useful for both the student and parent to work towards improvement. As much as possible, the work should be the student's own, rather than group-work or work where parents or teachers have had considerable influence.

(CRESST.org, 2008, p. 12)

Understanding validity and reliability and the ways they fit with assessment is important for creating consistent, meaningful instruction and evaluating instruction and learning. Recognition of these relationships will make for stronger and more lasting teaching and learning events.

Chapter Six

Rubrics

What are rubrics?

Rubric originated from the Latin word rubrica which means red earth. It referred to the red ink used to highlight important components of writings in different contexts. There are remnants of this usage found in the modern St. James Bible. These highlighted passages were intonated in a different way than ordinary text in a given body of work. These rubrics helped achieve desired outcomes (Harper, 2010).

Grammarians conjecture that this transitioned to the red ink marks of correction a teacher would place on a student's paper. With the historical evolution to modern education, the term rubric has evolved to mean "tools that can help multiple instructors come to similar conclusions about construction of higher-level conceptual knowledge, performance skills, and attitudes" when assessing student work, evaluating their own instruction, and finding validity in the tools that they employ in the classroom (Bargainner, 2003, p. 1). The "rubric is an authentic assessment tool used for evaluating criteria that are subjective and complex" (Shepherd, & Mullane, 2008, p. 31).

Outside of the education field, we use rubrics in our everyday lives without realizing it. When we shop for a car, we set up a mental rubric based on safety, value, cost, and style. As we inventory the selections with which we are presented, we make our decision based on the overall scores we have given each of the sections of our mental rubric. Ideally, we look for a car that fits all our desires, wants, and needs. Realistically, we weigh our options, collaborate with family members and review our criteria to make the best possible choice.

clangnuts.blogspot.com

Just as we use mental rubrics in everyday life, changes in the way education is assessed led to a need to measure student outcomes as an integral part of a continuous

improvement process in classroom teaching. Assessments serve as comparisons as students move towards mastery. Ultimately, assessment supports the goal of becoming more effective educators with higher success rates in all the ways we value. This shift in ideology has led to some standardized approaches, but it is important to note that just as in buying a new car, as criteria change so do the measurements by which we judge. While there are now mandates and rules in regard to evaluating student progress overall, the use of a rubric is an invaluable tool in the educational process. It can be "time consuming for a teacher to calibrate a rubric, but its value as a learning tool, as well as a life skill, may be invaluable" (Shepherd & Mullane, 2008, p. 31).

In the educational setting, rubrics serve "as scoring guides, consisting of specific pre-established performance criteria, used in evaluating student work on performance assessments" (Mertler, 2001, Designing scoring rubrics section, para. 1). They can serve as a "scoring tool(s) that divides an assignment into its component parts and objectives, and provide(s) a detailed description of what constitutes acceptable and unacceptable levels of performance for each part" (Stevens & Levi, 2005).

When used appropriately and sparingly, they may guide the educational process in a variety of ways. They can inform the teachers of whether a new approach is producing desired results, whether the school is achieving national standards, and what improvements need to be made in different areas to create an overall success story for teachers and students.

Types of Rubrics

There are two main types of rubrics: analytical and holistic. Rubrics can also be classified as general rubrics and task-specific rubrics. General and task-specific rubrics can be designed as holistic, analytical, or combinations of the two main types.

A general rubric can be used to assess a performance that will be repeated with the same criteria being emphasized each time. For example, a teacher might use the same rubric for the successive drafts of a paper helping students pinpoint areas of weakness, provide clear direction to meet expectations, and improve their writing over time as reflected in the rubric scores. This repeated evaluation is known as formative assessment and can be described as iterative or repeatable in that it relates to the measurement of students'

development over the course of time and provides feedback to help the students improve their performance.

One flaw with general rubrics, such as a rubric used for all writing projects for the class, is that while they can be applied to different tasks and performances, the feedback produced might not be specific enough to be effective across a wide range of performances (Zimmaro, 2004).

Here is a very basic example of a general analytical rubric (breaks down the performance into four components each assessed separately) that might be used an entire term for a series of oral presentations. Note that this rubric can be used as a formative assessment with scores serving as the basis for feedback to the students as they progress.

Table 6.1 Oral Presentation Rubric

Criteria	never	sometimes	always
Makes eye contact	0	1	2
Volume is appropriate	0	1	2
Enthusiasm is evident	0	1	2
Summary is accurate	0	1	2

(Mueller, 2011, Step 4: Section, para. 1)

The following is an example of a general holistic rubric that could also be used for multiple performances. While the analytical rubric above provided detailed feedback, the following holistic rubric is best used as a summative assessment in which the score represents a grade for the end of an evaluation on a chapter, unit, or term of student work. In this instance, a homework assignment is the performance being assessed through the use of the general holistic rubric.

Table 6.2 Homework Problem Rubric

Homework Problem Rubric
Your homework problems will be graded according to this rubric. Each problem will be put into one of the following six categories and assigned the corresponding grade. Correct and clear (4) Answers in this category will give a completely correct solution to the problem and present it in a clear, logical way Correct and unclear (3) Answers in this category are correct, but presented in a way that is hard to follow or imprecise Clear, but has a small mistake or two (3) Answers in this category are almost correct, but have an algebra mistake or some small error. The answer is presented clearly. Partially correct (2) Answers in this category are along the right lines, but have some major flaws Demonstrated some understanding (1) Answers in this category demonstrate that the student understood what needed to be done to solve the problem and made some headway, but didn't get to a solution This ain't it (0) The student didn't try, or obviously doesn't understand what (s)he needed to do

(Muller 2011, Step 4: Section, para. 15)

Task-specific rubrics describe the criteria relevant to a specific performance. They take more time to create, but the specific feedback they generate can be of great assistance in informing students about their performance, as well as helping the teacher in assigning fair grades and improve instruction to better serve the learning goals. The language in a task-specific rubric should reflect the learning goals established and reinforced in the class. A disadvantage of task-specific rubrics is that they are not useful beyond the tasks or performance they are designed to measure. Here is an example of a *task-specific analytical rubric* (partial example shown – see the full example at: http://www.utexas.edu/academic/ctl/assessment/iar/students/report/rubrics-casestudy1.phpn).

Table 6.3 Analytic Rubric for a Case Study

	1 point	2 points	3 points	4 points
Introductory material	There is no introduction. The purpose is not identified.	The introduction is present. Identification of the purpose and central questions is sketchy.	The introduction provides an adequate context for the project. The purpose is identified through reference to one or more central questions.	The introduction provides a well-developed context for the project. The significance of central questions is illustrated by references to course materials.
Descriptions of the setting and data collection process	The narrative contains an incomplete or vague description of the setting, and no description of the data collection process.	The narrative contains an adequate description of the setting, but an incomplete description of the data collection process.	The narrative contains adequate descriptions of the case study setting and the data collection process.	The narrative contains well-developed descriptions of the setting and the data collection process (which is built upon concepts from current research, theory, and course materials).

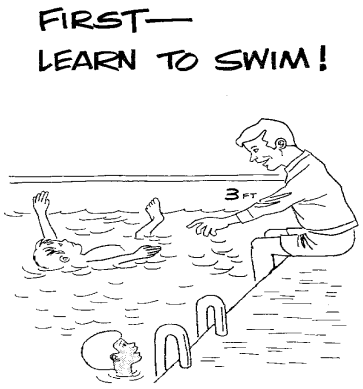
Record of observations	The narrative contains observations from only one perspective, or of a single type of data	The narrative contains observations from at least two sources.	The narrative contains observations from multiple sources or includes qualitative and quantitative data.	The narrative contains observations from multiple sources, includes qualitative and quantitative data, and makes references to models of appropriate practice that are supported by current research and theory.
Discussion, logic, and conclusions	The discussion is incomplete or illogical, and conclusions are missing or unrelated to the central questions.	The discussion is adequate, but conclusions--if present--do not match the central questions.	The discussion seems complete. Conclusions are logical and address the central questions.	The discussion seems complete. Conclusions are logical; they address the central questions, suggest possible strategies for addressing weaknesses, and are tied to the course work.

Presentation's clarity and style	At least three (3) of the following are true: The project contains multiple errors in grammar, spelling or mechanics. The page layout is cluttered. Navigation between sections is unclear. APA format is not used for in-text and bibliographical references to external resources.	Two (2) of the following are true: The project contains multiple errors in grammar, spelling or mechanics. The page layout is cluttered. Navigation between sections is unclear. APA format is not used for in-text and bibliographical references to external resources.	One (1) of the following is true: The project contains multiple or serious errors in grammar, spelling or mechanics. The page layout is cluttered. Navigation between sections is unclear. APA format is not used for in-text and bibliographical references to external resources	All of the following are true: The project contains no serious errors in grammar, spelling or mechanics. The page layout facilitates understanding of the narrative. ¨ Navigation between sections is clear. APA format is used for in-text and bibliographical references to external resources.

(University of Texas, 2010, Assess students: section, para. 1)

These are basic examples intended to begin familiarizing you with the form and function of rubrics. Now let's look a little closer at the two main types of rubrics: holistic and analytical.

Holistic Rubric



mattwardman.co/blog

Holistic rubrics are those in which the criteria of the authentic performance are included in the task(s) being evaluated. They are judged as a global or holistic indication of the overall success of the performance. In essence, the use of a holistic rubric is best when the teacher wants to make a brief but broad determination about a performance, and when the assessment is weighted lightly—for instance with a weekly homework assignment that emphasizes the cognitive habits of the student (L. Bristow, personal communication, March 1, 2011), or in tasks in which the performance criteria are not easily separated. An example of this might be found in the efficient swimming example below.

Efficient swimming requires both legs and arms to be in motion simultaneously. The swim coach might develop a holistic rubric that determines where the swimmer will be grouped with other similarly skilled swimmers based on an initial performance. If swimmers make it from one end of the pool to the other in a timely manner, they would be placed at a proficient stage of the scaling continuum. If the swim coach has to dive into the pool and rescue a beginning swimmer, that person would be placed at the most basic stage of the scaling continuum. Some swimmers might fit in the middle of the continuum, and designated groupings could be created for them. Separating the performance criteria of arm and leg movement for basic swimmers might be difficult as both skills would likely be underdeveloped, so evaluating the larger performance with a *holistic rubric* could help swim instructors understand the levels of swimmers with whom they are working.

Here is an example of a holistic rubric:

Fiction Writing Content Rubric – HOLISTIC

- 5pts. – The plot, setting, and characters are developed fully and organized well. The who, what, where, when, and why are explained using interesting language and sufficient detail.
- 4pts. – Most parts of the story mentioned in a score of 5 above are developed and

organized well. A couple of aspects may need to be more fully or more interestingly developed.

- 3pts. – Some aspects of the story are developed and organized well, but not as much detail or organization is expressed as in a score of 4.
- 2pts. – A few parts of the story are developed somewhat. Organization and language usage need improvement.
- 1pt. – Parts of the story are addressed without attention to detail or organization.

(Pearson Education, Inc., 2011, Para. 3)

A holistic rubric as an assessment tool is often too broad a measurement to elucidate and delineate a student's deficiencies in ways that improve instruction directly using its results (Arter, 2000). However, holistic rubrics do allow for the quick assessment of a performance and require minimal training on proper use. If rubric training equals time, and less time is required to acquire the skills to effectively utilize this assessment tool, it is likely that the holistic rubric will continue to be utilized if solely for its economic aspect (Johnson, Penny, & Gordon, 2009). In addition, one must ask, "Who is being taught*?"* Because there is less detail to analyze in the holistic rubric, younger students may be able to integrate it into their schema better than the analytic rubric" (Pearson Education, Inc., 2011, para. 2)

If you want to assess a product or performance quickly, broadly, or briefly, a holistic rubric is appropriate. Use a holistic rubric when you cannot easily separate the tasks involved in a performance as in the assessment of beginning swimming example used above. Is the whole body being used to propel itself in the performance of swimming? Would it be better to separately evaluate the arm stroke from the leg kick using distinct criteria in determining effective parts, or is it enough to tell that the parts are linked and therefore the swimmers are able to propel their bodies with some overall measurable degree of success? Would it be more useful to group the two traits and judge the result of the grouped traits (i.e., the body is propelled well based on multiple criteria)? There is no right answer; the decision regarding which type of rubric to use is based on the purpose of the assessment, the value of detailed feedback, and the time constraints the teacher faces.

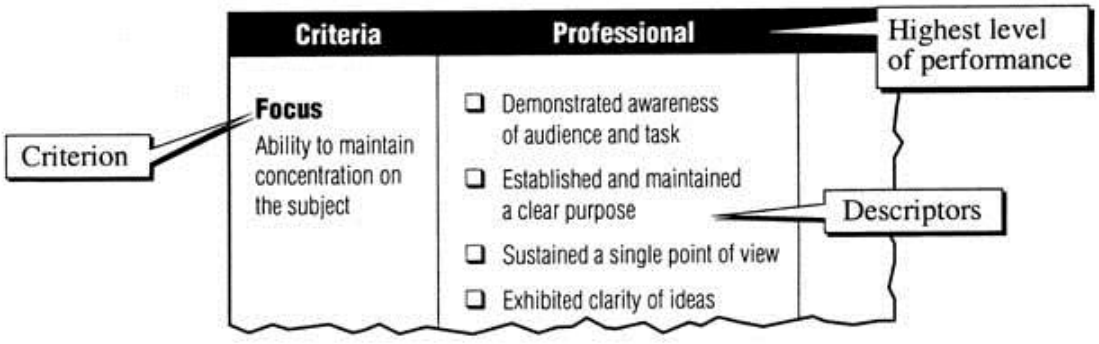
As a next step in the swimming analogy, the swim coach might create an *analytical rubric* to examine the specific skills each swimmer needs to improve. A score assigned to

the skill would rate the level of proficiency for each area or skill rather than as in a holistic rubric, which ranks the overall performance.

Analytical Rubric

"A rubric with two or more separate scales is called an analytical rubric, as it takes apart or breaks up the rating system for each trait" (Marcotte, 2006, Analytical or holistic section, paragraph 20). This type of rubric is particularly useful in terms of giving students teachers and parents feedback; when aspects of quality are being determined; and for designating standards in relation to "complicated skills, products and performances" (Arter, 2000, p. 5).

The example below shows one part of an analytical rubric and illustrates how an individual criterion (a task or skill) is established and presented, then subjected to a series of descriptors that are meant to detail the proficiency of the criterion being examined. The level of proficiency is determined by a sequential ranking on the scoring continuum. In this example, only one level is shown—the professional or highest ranking on this scale. Multiple criteria are shown to comprise the professional ranking, and a student would have met all of them to achieve this level of performance. In a complete analytical rubric, a full matrix with varying performance levels and their accompanying descriptors would be shown for each criterion being measured.



(Marketing & Business Administration Research and Curriculum Center, 2000, Analytical rubrics section, para. 3)

Next, let's take a look at a complete analytical rubric with a full range of performance levels and accompanying descriptors.

Table 6.4 Analytic Rubric Example

Criteria	Needs Improvement (1)	Developing (2)	Sufficient (3)	Above Average (4)
Clarity (Thesis supported by relevant information and ideas.)	The purpose of the student work is not well-defined. Central ideas are not focused to support the thesis. Thoughts appear disconnected.	The central purpose of the student work is identified. Ideas are generally focused in a way that supports the thesis.	The central purpose of the student work is clear and ideas are almost always focused in a way that supports the thesis. Relevant details illustrate the author's ideas.	The central purpose of the student work is clear and supporting ideas always are always well-focused. Details are relevant, enrich the work.
Organization (Sequencing of elements/ideas)	Information and ideas are poorly sequenced (the author jumps around). The audience has difficulty following the thread of thought.	Information and ideas are presented in an order that the audience can follow with minimum difficulty.	Information and ideas are presented in a logical sequence which is followed by the reader with little or no difficulty.	Information and ideas are presented in a logical sequence which flows naturally and is engaging to the audience.
Mechanics (Correctness of grammar and spelling)	There are five or more misspellings and/or systematic grammatical errors per page or 8 or more in the entire document. The readability of the work is seriously hampered by errors.	There are no more than four misspellings and/or systematic grammatical errors per page or six or more in the entire document. Errors distract from the work.	There are no more than three misspellings and/or grammatical errors per page and no more than five in the entire document. The readability of the work is minimally interrupted by errors.	There are no more than two misspelled words or grammatical errors in the document.

(DePaul University, 2011, Analytic Rubrics section, para. 9)

The example presented above is of an analytical rubric with the criteria being measured on the left vertical column and the descriptive details outlining the degree of mastery in ascending rank progressing along a continuum from left to right. By having an individualized criterion for specific tasks relating to the performance or product, students or

participants can get direct and detailed feedback about how the skill or task is ranked and what they need to work on to improve in the examined category. Similarly, teachers can use analytical rubrics to determine on what aspects of different lessons students were proficient and attempt to adjust lesson plans based on the feedback. Put another way, "when used as teaching tools, rubrics not only make the instructor's standards and resulting grading explicit, but they can give students a clear sense of what the expectations are for a high level of performance on a given assignment, and how they can be met" (Allen & Tanner, 2006, p. 203).

Descriptors in an analytical rubric must be clear and precise in assigning the meaning to the criterion at the appropriate level and designating what must be done to advance to the next level for that skill. There are two potential limitations to using the analytical rubric, the first being the length of time it might take to familiarize one-self or be trained by another to appropriately design an analytical rubric. This point was discussed previously in relation to the appeal holistic rubrics might have from an economic standpoint. A second point is the increased amount of time it takes for a scorer to evaluate a performance's criteria and appropriately rate them (Johnson, Penny, & Gordon, 2009). To effectively apply an analytic rubric to piece of work, the work typically has to be read multiple times so the rater can focus on a particular aspect. This can significantly increase grading time. However, students can apply rubrics to their own work or to peer' 'work. This reinforces the important features of the assignment while helping students develop self-regulating strategies.

Designing an Effective Rubric

In this section, we examine how to design an effective rubric. You should ask two questions at the outset of the design process: "What do I want students to know and be able to do?" and "How will I know when they know it and can do it well?" (Allen & Tanner, 2006, p. 198). Rubrics are assessment tools that vary in complexity and purpose, but which share some common features. Effective rubrics:

- focus on measuring a stated ***objective*** (performance, behavior, or quality);
- use a ***range*** to rate performance;
- contain specific performance characteristics arranged in levels indicating the **degree** to

which a standard has been met (Pickett & Dodge, 2007, Features section, para. 5).

The objective or the performance must be something that is observable and measurable in order for the rubric to be successfully implemented. We can understand the basis for an effective rubric by developing one. To begin conceptualizing the steps involved with creating an analytical rubric, let's pull up a blank template. The one below is from the SDSU Education website (Pickett & Dodge, 2007, Exercise section, para. 1).

Table 6.5 Rubric Template

(Describe here the task or performance that this rubric is designed to evaluate.)

Criteria	**Beginning 1**	**Developing 2**	**Accomplished 3**	**Exemplary 4**
Stated Objective or Performance	Description of identifiable performance characteristics reflecting a beginning level of performance.	Description of identifiable performance characteristics reflecting development and movement toward mastery of performance.	Description of identifiable performance characteristics reflecting mastery of performance.	Description of identifiable performance characteristics reflecting the highest level of performance.
Stated Objective or Performance	Description of identifiable performance characteristics reflecting a beginning level of performance.	Description of identifiable performance characteristics reflecting development and movement toward mastery of performance.	Description of identifiable performance characteristics reflecting mastery of performance.	Description of identifiable performance characteristics reflecting the highest level of performance.

Step 2: Identify the Measurable Criteria

The next design step in creating an analytical rubric is to identify the criteria of the performance you wish to examine. You can do this with your students if you wish. Examine the performance being assessed and separate it into the critical components that reflect the skills deemed to be the "essential aspects that define quality in a product or performance" (Arter, 2000, p. 13).

Step 3: Creating A Scale

Taking the next step in designing an effective analytical rubric entails the creation of the scale on which to rank the performance criteria. The example template above has a built-in scale. When determining the number of ranking categories, an instructor should consider what purpose the score will serve. Is the assessment formative? Is it meant to assist in the student's development, or is it summative with a score being assigned to the student's skill level (as at the end of a school term)?

Having multiple opportunities to improve a skill in subtle measurable ways would be more appropriate to a rubric having more scoring points on the continuum. Including more points on the scoring continuum requires more specificity in description between the levels of proficiency on the scale. However, it might also better inform the student as to what is needed to progress to the next level. There are advantages to both scale sizes as pointed out here: "While longer scales make it harder to get agreement among scorers (inter-rater reliability), extremely short scales make it difficult to identify small differences between students" (Marcotte, 2006, Dimensions and scales section, para. 19). The main point for this step is to consider different levels of quality as they relate to the performance criteria being examined and calibrate the scale to fit the purposes of the assessment.

Describing the Criteria

Arriving at the critical step of criteria description, the teacher must now examine exemplars of quality, benchmarks of excellence as in the use of anchor papers and standardized descriptions of best practices for the skill or criterion being assessed. Different authors suggest proceeding in a variety of ways, but however one proceeds, the level of description should be robust in detail in regard to what constitutes and delineates one level

of proficiency from another, illustrative of the learning goals established and reinforced in the instructional curriculum, and explanatory in a supportive manner as to what specific aspects must be improved upon by the student to progress to the next stage of mastery. Even a well-conceived analytical rubric can be subject to design flaws such as vague distinguishing descriptors of quality, inconsistent learning objectives, or descriptors that carry unsupportive judgments. Including students in the creation of a rubric empowers them to self-assess their own work habits, evaluate the quality of their projects, and reflect on what has been learned during the process (Brown, 2008).

There are a wide range of rubric templates online. Try these sites:
teacherplanet http://www.rubrics4teachers.com/
TeAch-nology http://www.teach-nology.com/web_tools/rubrics/
rubistar http://rubistar.4teachers.org/
SDSU Education website http://edweb.sdsu.edu/triton/july/rubrics/Rubric_Template.html .

Checking Your Rubric

Once you have a rubric, you need to perform some basic checks on it (Mueller, 2011). For example:

- Let a colleague review it.
- Let your students review it -- is it clear to them?
- Check if it aligns or matches up with your standards.
- Check if it is manageable.
- Consider imaginary student performance on the rubric.

As an analytical rubric, the concept of formative assessment as part of an iterative process again becomes relevant. The developer can determine the appropriateness and accuracy of design, based on the rubric's ability to measure the stated criteria of a performance or product. Because the learning related to the performance is ongoing, the rubric may be adjusted to increase accuracy of the intended goals of measurement and increase useable feedback. The following consistent elements of a rubric have been determined as indicators of an effective design: "(a)

measurements of the same skills or knowledge produce the same results, (b) every performance is evaluated by identical procedures, and (c) adjudicators share the same levels of expectation" (Latimer, Jr., Bergee, & Cohen, 2010, p. 169).

If we refer back to the validity chapter, we can recall that "performance assessments should require appropriate content, context, processes, and tools consistent with the defined domain and candidate characteristics" (Zane, 2009, p. 92). In essence, our rubric should measure what we claim we are attempting to measure in a clear, concise, and consistent manner that is supported by the curriculum and instructs future teaching and learning.

Arguments Against the Use of Rubrics

There is an ongoing and lively debate about the legitimacy of rubrics and their usefulness in performance assessment. Some feel that in attempting to move away from standardized tests and standardized curriculum, it is counterintuitive to then devise a strategy for standardizing performance assessment. Inter-rater reliability and consistent scoring as strength of rubrics are countered with the argument that students will write to the rubric, and evaluators will score from the rubric, limiting the exercise of independent judgment and the development of critical thinking skills by teacher and student alike. Objectivity in assessment is unlikely using a rubric, it is posited, as judgment still hinges on the interpretation of adjectives used to describe degrees of quality in a rubric. The associations of grades equivalent to the scores derived from assessment of a performance through the use of a rubric can succumb to a pitfall inherent to traditional assessment: children ultimately focus on how well they are doing instead of what they are doing (Kohn, 2006).

Conclusion

Rubrics can be effective assessment tools when evaluating student performances. When students become involved in the creation of a rubric used for evaluation of their work, they

1) are informed as to what criteria will be evaluated;
2) can observe what comprises and distinguishes different levels of proficiency;
3) can become empowered through their involvement in the processes of teaching,

learning, and self-assessment;

4) become invested in, and therefore take ownership of, the processes and products associated with a performance.

"Teachers clarify their goals, expectations, and focus, and even find that their paperwork is reduced because students are a part of the process of assessment development" (Pickett, & Dodge, 2007, Conclusion section, para. 8). While criticisms of rubrics may be valid in some instances, both sides of the rubric debate agree that using rubrics as an assessment tool should be done sparingly in an attempt to evaluate and create effective performance assessments.

Chapter Seven

Challenges and Opportunities

Problems with Performance Assessment

All teachers want to understand what their students are learning in their classes. If teachers are not inquisitive about this, they might be in the wrong profession. The question then is how exactly to measure the learning taking place in our schools. Answers to this question will vary depending on what one decides to measure. Because of the diversity of assessment practices, there is always a debate about which one is the best to use is in our schools for a particular purpose. As you have seen in the previous chapters, assessment can be formative or summative. Formative assessment informs a teacher throughout the year what their students are learning and what needs improving providing the basis for effective feedback to students. Summative assessment is typically an end of year, semester, or unit final examination when students' knowledge is assessed but without time allotted for improvement within the classroom. As we will see in this chapter, there is considerable criticism regarding the appropriateness of performance assessment (PA) as a summative measure. There is also concern over whether PA can be used economically on a large scale. PA is currently considered an alternative assessment not practiced in the mainstream accountability movement. The situation may be changing though as support for PA increases, and new standards in education call for 21st century skills more readily tested by PA.

As previous chapters have shown, the work students produce in performance assessment is diverse. It can range from a science project to an art design to an oral debate. Performance assessment stands in contrast to traditional assessment. All teachers are familiar with traditional tests as they are the most common type of classroom assessment and are mandatory across the nation's public schools in the form of standardized testing. The majority of these tests have students answering a combination of multiple choice and true/false questions as well as some writing samples. This form of assessment is favored for its efficiency to measure a large population of students' mastery of information. Some

teachers may prefer performance assessment to traditional tests which they see as too rigid and authoritarian. These teachers want more freedom in their classrooms and do not see standardized multiple choice tests as authentic assessment.

As the national debate over education is increasingly dominated by proponents of high stakes testing and traditional assessments, there is a backlash among some educators. Critics of policies like *No Child Left Behind* and *Race to the Top* propose more performance assessment in our schools as an alternative. In this chapter we will analyze this debate and look into the pros and cons of performance assessment's implementation in the classroom. This will include a presentation of common problems and challenges with performance assessment. So as not to leave our readers discouraged, we will also present guidelines for educators to effectively implement performance assessment in the classroom.

It is important to note that performance assessment and standardized testing are not mutually exclusive. Some proponents of performance assessment argue that PA, if done correctly, can be standardized and implemented across the nation to test students. Regardless of national and state policies, all teachers will have to find classroom assessments to evaluate the teaching and learning process.

Common Complaints Associated with Performance Assessment

- The financial cost of implementation is too high.
- Teacher training adds an extra burden.
- It takes up too much time.
- Parents are concerned with any radical departure from tradition.
- Students are not used to this form of assessment.
- It is not a good way to do summative assessment.
- It is not a valid or reliable test.
- It is too hard to implement on a large standardized scale.
- It suffers from rater bias and the halo effect.

Costs of Performance Assessment

Some of the faults critics find in performance assessment are practical in nature. One criticism leveled at proponents of this method is the increased cost in time and money

that it takes not only to prepare teachers and students but to administer and score these types of tests. Before these tests reach the students, though, tests must be designed. In the 1990s, the cost of performance assessments were prohibitively expensive, nearly twice as much as their multiple-choice counterparts or more. Some researchers studying science assessments found that compared to developing and validating a multiple choice type item it cost "80 times as much for an open ended item, 300 times as much for a content station, [i.e. science or computer lab] and 500 times as much for a full investigation item" (Lawrenz, Huffman, & Welch, 2000, p. 623). Fortunately, this issue may become a non-issue over time. "It is reasonable to assume that test development costs have declined as states and contractors learn from the past efforts" (Stecher, 2010, p. 27). In fact, as state governments collaborate with each other and with testing companies to improve performance assessment techniques, systems are being produced and implemented at comparative cost to multiple-choice tests. Currently, tests with multiple PA components are being developed and used in some states, such as with the New England Common Assessment Program. The states were able to pool resources and knowledge to develop a test that worked for their annual testing needs (Stecher, 2010). The hope is that advances in technology and testing knowledge will serve to bring the costs of performance assessment down to a reasonable level, and more states will be able to "take advantage of the economies of scale that will accompany states banding together in consortia, tapping the efficiencies of technology in administering tests and supporting scoring, and using teachers strategically in the scoring of performance items" (Adamson, Darling, & Hammond, 2010, p. 4).

Test scoring is another issue with performance assessment costs. Scoring costs of PAs tend to be at least twice the cost of multiple choice counterparts (Klein, 2008). However, uses of technology and progress in testing knowledge are helping to bring scoring costs down. In fact, computerized scoring has become a possibility with advances in scoring software. The hope is that computerized scoring programs will be able to replace the costly human scorers for some writing tests completely, and for other tests, replace at least one scorer (Klein, 2008).

Time is an additional cost in PA since it tends to be more time intensive than traditional methods of assessment. Teachers in some schools, districts, and states that have

adopted performance assessment techniques found that they had to jettison entire portions of their curriculum in order to fit this new approach. For example, some teachers felt like they had to drop portions of their curriculum to make time for portfolio writing and cooperative problem-solving exercises. (Kane, Khattri, & Reeve, 1998) Other teachers felt they were robbed of instructional time and had to drop particular skills and literature units from the curriculum (Kane, Khattri, & Reeve, 1991). In contrast, where districts adequately trained their teachers, and the assessment and curriculum were integrated and reinforced one another, teachers felt there was no negative change. These differences in perspective show that implementing a new assessment system without appropriate professional development can be difficult (Kane, Khattri, & Reeve, 1998). As state and district budgets become increasingly tighter, there is, unfortunately, little or no money allotted for the training and support of teachers using performance assessment. One can only hope that as administrators, legislators, and parents see the benefits in educational effectiveness that comes with performance assessments, they will be more likely to fund the training and implementation of performance assessment.

Teacher Issues in Performance Assessment

Lack of Teacher Training

Just as it is difficult to build a car engine without mechanic training, teachers with no background in performance assessment find themselves tinkering with pre-made tests and may be intimidated by the idea of creating a test or rubric from scratch. Teacher credentialing programs rarely include effective assessment instruction. Even when instruction is included, graduates are told that they should include assessment in their teaching, but they don't feel like they're adequately trained how to actually do this effectively. New teachers in Alberta, Canada, stated that their own university instructors didn't model sound assessment methods, and some prospective teachers in university classes "never heard the word 'rubric' used" (Scott, et al., 2011, p. 109). The study found that teaching programs should focus more on explicit instruction and emphasize the fact that student assessment is the responsibility of all educators in a school. The study also "included suggestions that university instructors needed to have school-based teaching

experience, to model good assessment approaches and strategies, to include alternative education assessment, and to explicitly teach differentiated instruction and assessment" (Scott, et al., 2011, p. 107). In other words, student teachers want more balance between theory and practice.

Need for More Professional Development

Related to the issue of teacher training is the issue of professional development. When teachers feel inadequate in assessing literacy, they are more likely to feel confusion, resistance, or anger when addressing assessment issues. Researchers in an Alberta, Canada study of teacher assessment literacy found that "educators rated their knowledge as high, (but) many interviewees were insecure about their assessment competency" (Scott, et al., 2011, p. 101). In order to use performance assessment in classrooms, teachers must be trained in "its purposes, format, pedagogical underpinnings, scoring procedures, and consequences" (Kane, Khattri, & Reeve, 1998, p. 75). As teachers learn to use performance assessments, they become more familiar with the methods of designing, implementing and scoring these tests. In daily use, teachers become more efficient and more comfortable with the idea of using a new method to assess their students.

In the Los Angeles Unified School District, UCLA's Center for Research on Evaluation, Standards and Student Testing (CRESST) found that part of the improvement in test scores and student outcomes was directly related to teacher training. As the district "worked to improve the quality and frequency of scoring training sessions, to collect additional training papers to create a larger library of student compositions for training purposes, and to assemble and report instructional 'best practices'" (Baker, Niemi, & Sylvester, 2007, p. 209), teachers felt more confident in using performance assessment in their classrooms and were able to evaluate assessment methods and curriculum with a critical eye for effectiveness and usefulness. Teachers also "continued to develop performance assessment prompts, rubrics, anchor papers, and training papers for the district's ongoing accountability programs." (Baker, Niemi, & Sylvester, 2007, p. 209). In another study of school districts that scaled up their use of performance assessments, teachers who were well trained began to see improvement in their instruction and curriculum. In fact, "teachers changed the types of questions they asked in class, and having assessment tasks helped teachers plan backwards and change their curriculum" (Tung, 2010, p. 42). These

improvements in teacher training served as the impetus to provide teachers with professional development that improved teaching methods in general, and specifically improved each teacher's use of performance assessment.

The following are recommendations for teacher professional development programs in the area of performance assessment:

- Use external partnerships to provide and facilitate professional development
- Assure that all new teachers become assessment literate
- Review the successes and challenges of scaled up performance assessment initiatives
- Include teachers in the design of and professional development for performance assessments
- Structure school days so that teachers have time to plan and debrief assignments and discuss student work, scoring, and performance assessment revision (Tung, 2010, p. 46-47).

Parent Resistance to Performance Assessment

Every educator knows that without parent support, the job of teaching is very difficult. This fact is no different in regard to performance assessment. A parent or guardian of a student might resist the idea of performance assessment for a variety of reasons. For example, changes to school curriculum have been opposed because of misinformation, lack of information, or preconceptions about the change (Konzal & Dodd, 1996). Most parents are not familiar with various types of assessments, so everything they know about the subject is from their own school experience. If a parent experienced multiple choice tests and letter grades as a student, then that parent might expect the modern student to be assessed in the same manner. Parents may have difficulty in understanding the assessment and therefore do not know how to express their ideas about the subject (Robinson, 1996). Because performance assessments may produce results that are more complicated than a letter grade, some parents may feel confused (Xue et al., 2000). In some places, curriculum reform is resisted and revoked, and in others, accepted. In Littleton, Colorado, where parents felt the adjustments to curriculum were enacted too quickly, resistance was the result. The consequences were disastrous, and changes were rescinded. In contrast, Vermont's movement toward performance assessment was readily accepted by parents.

Unlike Littleton, the state of Vermont made a large-scale effort to inform parents about the modifications before any changes were actually made. The difference in these two cases boiled down to timely dissemination of information (Khattri & Sweet, 1996).

When informing parents about new methodologies, they need to know what makes performance assessment different from other kinds of assessment, why it is useful, and how it helps the teacher and students. They should be informed about the ways in which curriculum and routines might change in the classroom. Including examples of this new form of assessment is also helpful for understanding. Studies show that "parents who learn the reasons, process and consequence of alternative assessments prior to and during the implementation can provide the support necessary to make this innovation a success" (Robinson, 1996, p. 300). Teachers should keep in regular contact with parents about what is happening in the classroom, and districts should offer in-service to teachers in how to answer questions from parents about assessment.

Here is a list of questions that parents might ask.

- Does the performance assessment cover important skills and knowledge?
- Are the test items varied to fairly test students having different experiences, backgrounds, and motivations?
- Does the assessment give my child worthwhile educational experiences?
- Does the assessment require my child to use higher level thinking and problem-solving skills rather than simply memorizing to determine the answer?
- Are teachers receiving training and assistance in designing and using performance assessments?
- How are assessment results going to be used? Are teachers using the results to evaluate their students' performance in their own classrooms and then tailoring instruction in areas of weakness, or are the results being compared to those in other classrooms and schools and for evaluating the teacher or school? ("What should parents," 2010, p. 2)
- Only through preliminary and ongoing communication about its processes and benefits will parents become the supportive partners needed for successful implementation of performance assessment.

Reliability of Performance Assessment

The reliability of performance assessment is often called into question when only a small amount of tasks are being used for measurement. To avoid this, tests like the S.A.T. use a large number of questions. Performance assessment most often uses a limited amount of tasks to be judged. With a small number of tasks or questions, chance can play a big part in the overall score. If there is only one performance at the end of the year, it will not be a reliable score with which to judge a student's comprehension. Thus, the reliability for each individual score in PA over the course of time is often considered unreliable. To enhance its reliability, PA has to measure a larger amount of tasks/performances.

The question of subjectivity also arises when discussing performance assessment. Critics of performance assessment argue that because of the diversity of assessments that can be deemed performance. there are not clear standards from which to rate students (Lissitz, 1997). Thus, the raters or graders of performance assessment are often accused of stamping their personal opinion on students' scores or credit. This is referred to as *rater bias*. This is not a problem for multiple choice and true/false tests completed on score sheets, since there are clear right and wrong answers, and the tests can be graded by a machine. However, even so-called objective tests are the result of decisions about what to test, in what way, and what counts as a demonstration of mastery. The reliability of performance assessment is called into question when the scores given to students are not consistent. An example of this would be a teacher giving very divergent scores to her students' performances without a clear rationale for the grading policy. For performance assessment to be reliable and effective, it must be made clear what the criteria for grading a performance will be.

One issue of reliability in performance assessment concerns *the halo effect*. This theory was first coined in 1920 by Edward Thorndike and has come to be widely discussed and applied to education. *The halo effect* can be defined as when a rater gives examinees inconsistent scores affected by what the teacher knows about the student (Bechger, Gunter, & Hsiao, 2010). This can have a negative or a positive effect on scores. For instance, the rater could like the first performance an examinee gives and pay less attention to the subsequent examinees, giving them the benefit of the doubt when it came to judgment.

Thus, raters might show greater leniency in grading students who have previously done excellent work. On the flip side of this is the possibility that a rater could dislike an initial performance and thus look poorly on all the following performances of an examinee. The halo effect can be assuaged by giving raters plenty of time to score performances, providing or creating clear rubrics with criteria for how a performance is scored, and writing clear instructions for how a performance analysis will take place (Bechger, Gunter, Hsiao, 2010). With most teachers lacking any rating assistance and being the sole raters of their students' performance, this issue can still be a problem. However, teachers with all the responsibility of scoring their students' work can still modify the halo effect and rater bias by increased use of rating criteria such as rubrics.

Inter-rater reliability is the term used to describe whether or not scores are consistent from one rater to the next. Again subjectivity is an issue when different educators with different standards are being asked to judge the same performance. "Scoring rubrics respond to this concern by formalizing the criteria at each score level. The descriptions of the score levels are used to guide the evaluation process. Although scoring rubrics do not completely eliminate variations between raters, a well-designed scoring rubric can reduce the occurrence of these discrepancies" (Moskal & Leydens, 2000, p. 3). Again, this problem can be reduced by using clear grading criteria provided and explained to each rater.

Other external factors can reduce reliability and reduce the consistency of a student's scores.

> For example, a rater may become fatigued with the scoring process and devote less attention to the analysis over time. Certain responses may receive different scores than they would have had they been scored earlier in the evaluation. A rater's mood on the given day or knowing who a respondent is may also impact the scoring process. A correct response from a failing student may be more critically analyzed than an identical response from a student who is known to perform well. (Moskal & Leydens, 2000, p. 5)

This remains one of the biggest complaints over performance assessment: it is too subjective.

Developing Clear Standards and Rubrics for Performance Assessment

All assessments are a way to test or measure what knowledge has been gained. When developing performance assessment, always consider what the desired outcome is. For the performance assessment to be valid and reliable, there has to be a decision as to the type or depth of knowledge the performance is actually measuring or demonstrating. A performance assessment is not always going to be the best way to assess knowledge. For teachers to determine when to use performance assessment, they must determine what kind of knowledge they want their students to demonstrate or perform. The following questions may "help in the identification of specific goals and objectives" (Moskall, 2003, p. 16):

- What do I hope to learn about my students' knowledge or skills?
- What content, skills and knowledge should the activity be designed to assess?
- What evidence do I need to evaluate the appropriate skills and knowledge?

(Moskall, 2003, p. 16)

One of the most common complaints about performance assessment is that it does not work well for summative and comparative assessment on a large scale. There is a desire in the world of education to collect large amounts of data from schools at the end of each year in order to compare and contrast how the schools are doing. This puts extra pressure on the advocates of performance assessment since the dominant form of assessment for national and state standardized testing is very traditional. Most states' standardized testing includes a multiple choice test at the end of the year. The information gathered from the tests is not available until the next year after students have moved to the next grade so it is not formative assessment at all. However, this information is useful to see what exactly the students learned during the year tested, and it is done in an efficient, reliable and valid manner on a large scale. Since it takes up a large amount of time, for performance assessment to become more widely accepted, educators must find a way to measure students' achievement on a large scale that is still reliable and practical.

Performance assessments often run into problems when they are used to measure student performance across a large geographical area. Critics of performance assessment argue that it cannot be used to track students across an entire state in a reliable and valid

manner. In this age of accountability, there is a huge desire to compare the data of different school districts to monitor their progress or lack thereof. This argument caused pilot programs in Vermont, Kentucky, and Maryland to discontinue use of a portfolio based performance assessment system in the mid 1990s. (Stetcher, 2010). These states were able to briefly experiment with performance assessment before they were pressured to abandon it as a state wide assessment measure. These states were also criticized for not being able to hold schools accountable for the learning taking place. (Stecher, 2010). Other critics argued bluntly, "If the state is going to require sufficient accuracy for individual student decisions while being concerned with the assessment time allocated by the school and the money spent for the testing effort, it is clear that performance assessment is not going to be the approach adopted" (Lissitz, 2007, p. 16). In this case, performance assessment was seen as an interesting educational device but not a way to produce data to measure students' development over time. (It can provide reliable scores for schools which aggregate a large number of individual scores into a single composite). On the other side of this debate are proponents of PA who see the higher order thinking skills in the new common core state standards as needing to be measured by PA in order to get a valid measure of student achievement (Stecher, 2010).

However, performance assessment does not have to be used on a large scale. Teachers can utilize PA for the beneficial formative assessment it provides throughout the year.

> While being critical of authentic performance testing for statewide application and as a sole approach to local testing, it is important to note that there are times when an assessment that combines group work, motivation, subject knowledge, manipulation of ideas, manner of expression etc. should be delivered as a package. Students need to see such assessment and be exposed to the demands of such an approach. (Lissitz, 2007 p. 17.

When used as formative assessment, PA has been shown to work well. Students' work improves greatly when teachers utilize formative assessment throughout the year instead of the traditional summative assessment wherein students are simply assigned a letter grade at the end of the semester. "Students respond better to give and take than to a letter grade. Assigning grades to student work had no positive effect. It didn't improve

students' performance. However, when teachers provided only descriptive feedback and no grade, student performance improved by 30 percent. When Butler looked at both assigning a grade and providing descriptive feedback, she found no positive effect - assigning grades negated the positive effects of the feedback" (American Federation of Teachers, 2010, p. 5).

A big obstacle to implementing performance assessment in most districts is the need to offer teachers instruction in this new type of assessment. Teachers might find it difficulty to adjust to a new form of teaching that integrates PA as part of a pedagogical approach. As it is, teachers do not generally have access to quality training in assessment techniques. "Findings indicate that, although there has been a lot of attention given to improving assessment, confusion remains for teachers about terminology, principles, and pragmatics that undermine teacher confidence about assessment and making sound judgments about students' work" (Scott, 2011, p. 98). This can be remedied through increased focus on assessment in teacher education and credentialing programs.

Making Accommodations for Performance Assessment

Issues of equity and fairness arise when implementing any type of assessment, including performance assessment in the classroom. All teachers have to investigate what accommodations they will have to make for students with special needs. Students are as diverse as the world they live in and, thus, the special needs of students will run the gauntlet from mild (minor case of attention deficit disorder) to the more serious (autism) as well as cultural (English language learners). Proponents of performance assessment often see PA as a more flexible way of assessing students than the traditional test. In this way, students with disabilities who need accommodations can be given that accommodation as each teacher sees fit.

> A national survey of elementary and secondary general education teachers found that traditional letter and number grades are typically preferred for general education students, whereas grading adaptations such as pass-fail, portfolios, and multiple grades were thought to be helpful for students with disabilities. (Mastergeorge &Martinez, 2010, p. 6).

Critics of this flexibility see a problem with grade inflation. Teachers more often than not grade their students with disabilities easier than their core population. Critics have argued

this does not actually help the student in need and is actually counterproductive (Mastergeorge & Martinez, 2010).

It is important for teachers to carefully consider what accommodations are being made as too many accommodations can have negative results. The validity of assessments is reduced when teachers provide too much help to their students during assessments. "If a test is designed to measure a student's reading skill, having the teacher read the content of a passage to a student could invalidate our results" (Neibling & Elliot, 2005, p. 3). Because PA can take many forms, the accommodations created will vary depending on what type of specific standards are formed for PA. Making accommodations is never easy nor should it be. It is an important part of creating equity in the school system. Making accommodations for students with special needs is not one of the larger obstacles in the way of performance assessment.

The research is mixed when it comes to performance assessment for English language learners (ELL). Some critics argue that PA complicates an already complicated situation. In this line of reasoning, ELL students are thrown into an assessment process that does not provide the proper scaffolding. Critics also worry that ELL students might not get the proper guidance if PA follows a more student-centered curriculum which leaves students to research on their own as opposed to a teacher-driven curriculum in which instruction always comes through the teacher. Much criticism has been directed toward the *No Child Left Behind* legislation for its high emphasis on accountability testing while leaving ELL students behind. "The law does little to address the most formidable obstacles to their achievement: resource inequities, critical shortages of teachers trained to serve ELLs, inadequate instructional materials, substandard school facilities, and poorly designed instructional programs" (Crawford, 2004, p. 2).

On the other hand, some argue that PA is a favorable tool for assessment of ELL students. The same argument that is used for PA in general is used for ELL students. "In language education, the value of language performance assessment is that it measures student's abilities to respond to real life language tasks. In other words unlike other types of tests, performance assessments can be used to approximate the conditions of a real task in a real life situation. As a result, performance assessments have value in that their scores can be used to predict student's abilities in future real life situations" (Norris, Brown,

Hudson, & Yoshioka, 2002, p. 5). These arguments in favor of PA have led to resurgence in the desire to see performance assessment as a part of education reform.

Conclusion

Performance assessment remains a vital tool for teachers to use in the classroom. Each form of assessment has its pros and cons. Teachers need to evaluate which type of assessment will work best for their classroom and goals. As we all know, teachers might have a great deal of control over their own classrooms, but they are not the absolute authority when it comes to choosing school-wide assessments. Schools mandate certain assessments that teachers must conduct. As stated previously, these assessments primarily consist of multiple choice tests that are used as summative assessments for how well each school and district are doing in comparison to others. Under *No Child Left Behind,* these tests have been used for accountability. With the dominant model of education in the United States being high stakes testing and standardization, proponents of performance assessment have felt downhearted. This does not have to be the case, though. There is a role for performance assessment to play in educational reform. This chapter outlined the common problems occurring with PA in order to address how they can be corrected. Proponents of PA must address the criticism of this assessment tool in order to overcome its faults. With the right justification and argument, PA can have a seat at the national educational table once again, and it appears that with the adoption of the new Common Core State Standards, PA will make a comeback. With these new national standards for education, it seems that educators and the public will take PA seriously as a necessary way to assess student learning.

References

Achieve, Inc. (2010) Partnership for the assessment of readiness for college and careers (PARCC). *Achieve, Inc.*, Retrieved April 5, 2011, from http://www.fldoe.org/parcc/pdf/prosum.pdf

Achieve, Inc. (2010). *Achieve, Inc. About Achieve.* Retrieved March 30, 2011, from http://www.achieve.org/about-achieve

Adamson, F., & Darling-Hammond, L. (2010). *Beyond Basic Skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Albert Shanker Institute | 'A Call for Common Content' Sign On Letter. (n.d.). *The Albert Shanker Institute*. Retrieved May 2, 2011, from http://shankerinstitute.org/curriculum.html

Allen, D, & Tanner, K. (2006). Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE—Life Sciences Education, 5*(201), 197-203.

Allen, R. (n.d.). Performance Assessment | Assessment & Testing | Professional Resources | Wisconsin Education Association Council. *Wisconsin Education Association Council*. Retrieved March 19, 2011, from http://www.weac.org/Professional_Resources/Testing/performance_assessment.aspx

American Federation of Teachers, AFL-CIO, (2010). *Quality Classroom Assessment Techniques,* Washington D.C.: retrieved March 31, 2011, from www.atf.org

Arias, R. M. (2010). Performance assessment. *Papeles del Psicólogo, 31*(1), Retrieved March 31, 2011, from http://www.cop.es/papeles

Arter, J. (1999). Teaching about performance assessment. *Educational Measurement: Issues and Practice, 18*(2). doi:10.1111/j.1745-3992.1999.tb00012

Arter, D.J. (2000). Rubrics, scoring guides, and performance criteria: classroom tools for assessing and improving student learning. *Proceedings of the American educational research association conference* New Orleans, La.

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). CRESST performance assessment model: assessing content area explanations. Retrieved March 31, 2011, from http://www.cse.ucla.edu/products/guidebooks/cmodels.pdf

Baker, E.L., Chung, G.K.W.K., & Delacruz, G.C. (2008). *Design and validation of technology-based performance assessments*. CRESST, UCLA, Los Angeles, CA.

Baker, E. L., Niemi, David, & Sylvester, Roxanne M. (2007). Scaling up, scaling down: seven years of performance assessment development in the nation's second largest school district. *Educational Assessment*, 12(3 & 4), 195-214.

Bandura, A. (1977). *Social learning theory*. New York: General Learning Press.

Bargainnier, S. (2003).Fundamentals of Rubrics. *Pacific Crest*,1-4.

Bechger, T., Gunter, M., & Hsiao, Y. (2010). Detecting halo effects in performance-based examinations . *Applied Psychological Measurement*, *34*(8), Retrieved from apm.sagepub.com

Beesley, A. (2009, April). *Measuring classroom assessment with a work sample.* Retrieved from ERIC database. (ED508465)

Birgin, O., & Baki, A. (2007). The use of portfolio to assess student's performance. *Journal of Turkish Education*, 4(2), 75-90. Retrieved from tused.org.

Brookhart, S. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, VA: ASCD.

Brown, C. A. (2008). Building Rubrics: A Step-By-Step Process. *Library Media Connection*, 16-18.

Brualdi, A. (1999). *Traditional and modern concepts of validity.* ERIC Digest. Retrieved from ERIC database. (ED435714)

Burke, L.M., & Marshall, J.A. (2010). Why national standards won't fix american education: misalignment of power and incentives. *Proceedings of The Heritage Foundation*, retrieved May 2, 2011, from http://www.heritage.org/Research/Reports/2010/05/Why-National-Standards-Won-t-Fix-American-Education-Misalignment-of-Power-and-Incentives

Byrd Carmichael, S., Martino, G., Porter-Magee, K., & Wilson, W. (2010). The state of state standards--and the common core--in 2010. *Proceedings of the Thomas B. Fordham Institute*, retrieved May 2, 2011, from http://www.edexcellence.net/publications-issues/publications/the-state-of-state.html

California Department of Education, Curriculum, Learning, and Accountability Branch. (2010). *Common core state standards.* California: Govt. Printing Office. Retrieved from http://www.cde.ca.gov/ci/cc/documents/generalccssoct2010.doc

Cizek, G. (2010). Translating standards into assessments: the opportunities and challenges of a common core. *Proceedings of the Brookings Institution Conference-Race to the Top Assessments: Common Core Standards and Their Impact on Student Testing*, http://www.brookings.edu/~/media/Files/events/2010/1028_race_to_the_top/1028_race_to_the_top_cizek_paper.pdf

Clark, D. C. (2000). Appropriate assessment strategies for young adolescents in an era of standards-based reform. *The Clearing House*, 73(4), 201-204.

Cole, H, Hulley, K, & Quarles, P. (2009). Does assessment have to drive the curriculum? *Forum on Public Policy Online, 2009*(1)

Common Core State Standards Initiative | About the Standards. (2010). *Common Core State Standards Initiative | Home*. Retrieved from http://www.corestandards.org/about-the-standards

Common Core State Standards Initiative | The Standards | English Language Arts Standards. (2010). *Common Core State Standards Initiative | Home*. Retrieved from http://www.corestandards.org/the-standards/english-language-arts-standards

Common Core State Standards Initiative | The Standards | Mathematics. (2010). *Common Core State Standards Initiative | Home*. Retrieved from http://www.corestandards.org/the-standards/mathematics

Crawford, J., (2004). *No child left behind: misguided approach to school accountability for English language learners*. National Association of Bilingual Education. Silver Spring, Maryland.

CRESST, Org., Report 738. (2008). *Providing validity evidence to improve the assessment of English language learners*, Center for the Study of Evaluation. University of California, Los Angeles. Retrieved March 31, 2011, from http://www.cse.ucla.edu/products/reports/R738.pdf

CRESST.Org. (2008). *Assessing the whole child*. Center for the Study of Evaluation, University of California, Los Angeles, California. Retrieved May 1, 2011 from http://www.cse.ucla.edu/products/guidebooks/wolekid.pdf

Darling-Hammond, L. & Adamson, F. (2010, p.2.) Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning. SCOPE retrieved 8/22/12 http://scale.stanford.edu/system/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning.pdf

Dede, C. (2010). Comparing frameworks for 21st century skills. In J. Bellanca & R. Brandt, Eds, *21st Century Skills*, pp. 51-76. Bloomington, IN: Solution Tree Press.

Dembo, M. (1994). *Applying educational psychology*. New York: Longman Publishing Group.

DePaul University. (2011). *Assessment criteria in a large-scale writing test: what do they really mean to the raters?*. Unpublished manuscript, Supplemental Instruction Teaching Commons, DePaul University, Chicago, IL. Retrieved from http://condor.depaul.edu/tla/Assessment/ConstructRubric.html#analytic

Driscoll, A., & Wood, S. (2007). *Developing outcomes-based assessment for learner-centered education: a faculty introduction.* Sterling,Virginia: Stylus Punlishing, LLC.

Ediger, M. (1993). Approaches to Measurement and Evaluation. *Studies in Educational Evaluation*. *19*(1), 41-50.

EdSource. (2010). *California and the "Common Core": will there be a new debate about K-12 standards?* Mountain View, California: EdSource, Inc.

Edwards, V. B., ed. (2006) Quality Counts at 10: A decade of standards-based education. *Education Week*, 25(17).

Educational Testing Service. (2011). *Classroom assessment for student learning, performance task rubric*.Retrieved March 26, 2011, from www.pfsd.com/uploads/Performance TaskRubric.pdf

Eggen, P., Kauchak, D. (2004). *Educational psychology windows on classroom*. Upper Saddle, New Jersey: Pearson Education, Inc.

Ertmer, P.A., Newby, T.J. (1993). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly*, 6 (4), 50-70.

Facione, P.A., & Facione, N.C. (2009). *The holistic critical thinking scoring rubric - hctsr a tool for developing and evaluating critical thinking*. Retrieved May 2, 2011, from http://www.insightassessment.com/pdf_files/Rubric%20HCTSR.pdf

Florida Center for Instructional Technology. (n.d.) *Classroom assessment: building student portfolios* University of South Florida, Tampa, Florida. Retrieved March 24, 2011, from http://fcit.usf.edu/assessment/performance/assessc.html

Garrison, C., & Ehringhaus, M. (2007). Formative and summative assessments in the classroom. National Middle School Association. Retrieved March 31, 2011, from http://www.nmsa.org/Publications/WebExclusive/Assessment/tabid/1120/Default.aspx

Gingrasso, S, Krause, T, Ploetz, P, Steinmetz, J , Terrell, P, & Warren, D. (2009). Performance tasks. *Proceedings of the 3rd annual critical thinking conference* (pp. 1-14). Stevens Point: The Center for Academic Excellence and Student Engagement at University of Wisconsin.

Guidelines for rubric development and establishing inter-rater reliability. (n.d.). Office of the Provost, Lesley University, Cambridge, MA. Retrieved May 2, 2011, from http://lesley.edu/provost/institutional_research/content/guidelines_for_rubric_development_and_establishing_irr.pdf

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan*, 80, 662-666.

Harper, D. (2010). *Online etymology dictionary*. Retrieved April 19, 2011, from http://www.etymonline.com/index.php?search=rubric&searchmode=none

Herman, Joan L., Aschbacher, Pamela R., and Winters, Lynn. (1992*). A Practical Guide to Alternative Assessment*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Introduction to Performance Assessment Tasks | AEA 267 Curriculum, Instruction & Assessment. (n.d.). *Area Education Agency 267*. Retrieved May 2, 2011, from http://www.aea267.k12.ia.us/cia/index.php?page=pat_intro

Johnson, R. L., Penny, J. A., Gordon, B. (2009). *Assessing performance: designing, scoring, and validating performance tasks.* New York, NY: The Guilford Press.

Jonassen, D. H. (1991) Objectivism versus constructivism: do we need a new philosophical paradigm? *Educational Technology Research and Development*, 39 (3), 5-14.

Kane, M, Khattri, N, & Reeve, A. (1998). *Principles and practices of performance assessment*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kane, M, Khattri, N, Reeve, A, & Adamson, R. U.S. Department of Education, Office of Educational Research and Improvement. (1997). *Assessment of student performance: studies of education reform* (RP 91-172004). Washington, DC: Government Printing Office. Retrieved from http://www2.ed.gov/pubs/SER/ASP/stude4-3.html

Karoly, J. C. & Franklin, C. (1996). Using portfolios to assess students' academic strengths: a case study. *Social Work in Education*, *18*(3), Retrieved March 31, 2011from http://ezproxy.humboldt.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=afh&AN=9609241989&site=ehost-live

Katz, I. (2007). Testing information literacy in digital environments: ETS's iSkills assessment. *Information Technology and Libraries*, *26*(3), 3-12.

Keyser, S, & Howell, S. (2008). *The state of authentic assessment*. (Informally published Manuscript) Department of Instructional Psychology and Technology, Brigham Young University. Retrieved from ERIC database. (ED503679)

Khattri, N., & Sweet, D. (1996) Assessment reform: promises and challenges. In M. Kane & R. Mitchell (Eds.*) Implementing performance assessment: promises, problems and challenges*, 1-21. New Jersey: Lawrence Erlbaum.

Klein, S. (2008). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. *Institute of Mathematical Statistics Collections, Probability and Statistics: Essays in Honor of David A. Freeman, 2,* 76–89.

Knoch, U, Read, J, & Randow von, J. (2007). Re-training writing raters online: how does it compare with face-to-face training? *Assessing Writing, 12*, 26-43.

Knoch, U. (2009). Diagnostic assessment of writing: a comparison of two rating scales. *Language Testing*, *26*(20), 275-304.

Kohn, A. (2006). The trouble with rubrics. *English Journal*, *95*(4). Retrieved from http://www.alfiekohn.org/teaching/rubrics.htm

Konzal, J. L., & Dodd, A. W. (1999). *Implementing higher standards and alternative assessment and grading policies in high schools: What do educators need to know about what parents think?* Paper presented at the AERA annual conference. Montreal, Canada.

Lane, S. (2010). *Performance assessment: The state of the art.* (SCOPE Student Performance Assessment Series). Stanford, CA: Stanford Center for Opportunity Policy in Education.

Retrieved March 31, 2011, from http://edpolicy.stanford.edu/pages/pubs/pub_docs/assessment/scope_pa_lane.pdf

Latimer,Jr., M.E, Bergee, M.J., & Cohen, M.L. (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment. *MENC: The National Association for Music Education, 58*(2), 168-183.

Lave, J., & Wenger, E. (1990). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press.

Lawrenz, F., Huffman, D., & Welch, W. (2000). Considerations based on a cost analysis of alternative test formats in large scale science assessments. *Journal of Research in Science Teaching, 37* (6), 615-626.

Learning Theories Knowledgebase. (2011). *Social Learning Theory (Bandura) at Learning-Theories.com*. Retrieved April 5th, 2011, from http://www.learning-theories.com/social-learning-theory-bandura.html

Lissitz, R. (1997) Statewide performance assessment; continuity, context, concerns., *Contemporary Education, 69(1), 15-19.*

Lombardi, Judy (2011). *Guide to performance assessment for California teachers*. Boston, MA: Pearson Education, Inc.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, *19*(3), 246-276.

Marcotte, M. (2006). *Building a better mousetrap: the rubric debate*. Unpublished manuscript, Education, Community College of Philadelphia, Philadelphia, PA. Retrieved May 2, 2011, from http://faculty.ccp.edu/dept/viewpoints/w06v7n2/rubrics1.htm

Markham, T. (2011, February 7). Want your students college ready? use pbl [Online Forum Comment]. Retrieved May 3, 2011, from http://www.edutopia.org/blog/project-based-learning-buck-thom-markham

Marketing & Business Administration Research and Curriculum Center. (2000). *Types of rubrics*. [Designer]. Retrieved March 10, 2011, from http://www.mark-ed.com/assessment/TypesofRubrics.htm

Maryland Department of Staff Development, (n.d.). *Developing performance assessment tasks.* Prince Georges County, Maryland: Retrieved May 2, 2011, from http://www.pgcps.org/~elc/developingtasks.html

Marzano, R., & Kendall, J. (1996). The fall and rise of standards-based education. *Proceedings of the A National Association of State Boards of Education (NASBE) Issues in Brief.* Retrieved May 3, 2011, from http://www.mcrel.org/PDF /Standards/ 5962IR_FallAndRise.pdf

Marzano, R., & Kendall, J. Ohio Department of Education, Instructional Management System. (1996). *What is the history of the standards movement?* Columbus Ohio: Retrieved May 2, 2011, from http://ims.ode.state.oh.us/ODE/IMS/SBE/FAQ/standards_history.asp

Marzano, R. J., Pickering, D. J., McTighe, J. (1993). *Assessing student outcomes: performance assessment using the dimensions of learning model*. Aurora, CO: McREL Institute.

Mastergeorge, A, and Martinez, J. (2010). Rating performance assessment of students with disabilities: a study of reliability and bias. *Journal of Psychoeducational Assessment*, *28*(536), 5-6.

McTighe, J. (Designer). (1998). *Possible student roles for performance tasks.* [Web]. Retrieved May 3, 2011, from http://www.aea267.k12.ia.us/cia/index.php?page=pat_roles

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). Retrieved April 8, 2011 from http://PAREonline.net/getvn.asp?v=7&n=25

Moskal, B.M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation*. *8*(14), Retrieved March 30, 2011, from E:\MA Education\EDUC 660 Assessment\Development of PA tasks, assessments with examples\Read\Recommendations for developing classroom performance assessments and scoring rubrics_Moskal, Barbara M.htm

Moskal, B.M. & Leydens, J. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10).

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation*, 8(14).

Mueller, J. (2005). The authentic assessment toolbox: enhancing student learning through online faculty development. *Journal of Online Learning and Teaching*, 1(1), 1-7.

Mueller, J. (2011). *Authentic assessment toolbox*. Retrieved March 27, 2011, from http://jfmueller.faculty.noctrl.edu/toolbox/index.htm

Muelle r, J. (2011). *Homework Rubrics*. Retrieved March 27, 2011, from http://cs.brown.edu/courses/cs016/courseinfo/rubric_info.pdf

Mueller, J. (2011). *Authentic assessment toolbox*. Psychology, North Central College, Naperville, IL. Retrieved March 27, 2011, from http://jfmueller.faculty.noctrl.edu/toolbox/howstep4.htm

Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic Pedagogy and Student Performance. *American Journal of Education*, *104*(4), 280-312. Retrieved April 3, 2011,

from http://www.jstor.org/stable/1085433

New South Wales Department of Education and Training. (Designer). (2009). *Teaching and learning cycle.* [Web]. Retrieved May 3, 2011, from http://www.curriculumsupport.education.nsw.gov.au/consistent_teacher/tlcycle.htm

Niebling B. & Elliot S., (2005) Testing accommodations and inclusive assessment practices. *Assessment for Effective Instruction,* 31(1), 1-6.

Norris, J., & Brown, J., & Hudson, T., & Bonk, W. (2002) Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing* 19, (395).

Nussbaum-Beach, S. (2011). *A futuristic vision for 21st century education*. Unpublished manuscript, Indiana Wesleyan University, Alexandria, Va. Retrieved May 3, 2011, from http://www.ascd.org/ascd-express/vol6/611-nussbaum-beach.aspx

PARCC Assessment Design (2011) http://www.parcconline.org/parcc-assessment-design

Payne, D.A. (2003). *Applied Educational Assessment* (2nd ed.). Belmont, CA: Wadsworth.

Pearson Education, Inc., Initials. (2011). *Analytic vs. holistic rubrics*. Retrieved May 3, 2011, from http://www.teachervision.fen.com/teaching-methods-and-management/rubrics/4524.html#ixzz1F7niAqEk

Pellegrino, J., Chudowsky, N., & Glaser, R. (Ed.). (2001). *Knowing what students know: the science and design of educational assessment*. Washington, D. C.: National Academy Press. Retrieved March 31, 2011, from http://www.nap.edu/openbook.php?isbn=0309072727

Pellegrino, J.W., & Quellmalz, E.S. (2010). Perspectives on the integration of technology and assessment. *Journal of Technology in Education*, 43(2), 119-134.

Pickett, N, & Dodge, B. (2007). *Rubrics for web lessons*. Dept. of Education, San Diego State University, San Diego, CA. Retrieved May 3, 2011, from http://webquest.sdsu.edu/rubrics/weblessons.htm

Popham, W. J. (1998). *Classroom assessment: what teachers need to know*. Boston, Massachesetts: Allyn and Baker.

Popham, W.J. (2009). *Is assessment literacy the "magic bullet"?*. Informally published manuscript, Graduate school of education, Harvard University, Cambridge, Mass. Retrieved May 3, 2011, from http://www.hepg.org/blog/19

Prus, J., & Johnson, R. (1994). Assessment & testing myths and realities. *New Directions for Community Colleges*. (88), 69 - 83.

Quellmalz, E.S. (2009). Technology and testing. *Science, 323*, 75-79.

Rezaei, A.R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*, 18-39.

Riordan, T. (2005). Education for the 21st century: teaching, learning, and assessment. *Change*, 37(1), 52-56. Retrieved March 7, 2011, from EBSCOhost

Robinson, J. (1996). Parents as allies for alternative assessment. In A. L. Goodwin (Ed.), *Assessment for Equity and inclusion: embracing all our children*, 297-303. New York: Routledge.

Rogers, G. & Sando, J. (1996). *Stepping Ahead: An Assessment Plan Development Guide*. Terra Haute, Indiana: Rose-Hulman Institute of Technology.

Rose, M. (2009). *21st century skills: education's new cliché*. Retrieved May 2, 2011, from http://www.truthdig.com/report/item/21st_century_skills_educations_new_cliche_20091208/

Rule, A. C. (2006). The components of authentic learning. *Journal of Authentic Learning,* 3(1),1–10.

Sada, A.G. (Designer). (2009). *Bloom's wheel*. [Web]. Retrieved May 2, 2011, from http://www.alline.org/euro/images/bloomwheel.png

Saettler, P. (1990). *The evolution of American educational technology*. Englewood, CO: Libraries Unlimited, Inc.

Sass, R. (2011). *American educational history: a hypertext timeline*. Informally published manuscript, Department of Education, College of Saint Benedict/Saint John's University, St. Joseph, Minn. Retrieved May 3, 2011, from http://www.cloudnet.com/~edrbsass/educationhistorytimeline.html

Schafer, W.D., Swanson, G., Bene, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education,* 14(2), 151-170.

SCOE | California State Academic Content Standards Commission. (2010). *SCOE | Sacramento County Office of Education*. Retrieved April 15, 2011, from http://www.scoe.net/castandards/index.html

Scott, S. (2011). Developing teachers' knowledge, beliefs, and expertise: findings from the Alberta student assessment study. *The Educational Forum, 75*, 96–113.

Scott, S., Webber, C. F., Aitken, N., & Lupart, J. (2011). Developing teachers' knowledge, beliefs, and expertise: findings from the Alberta student assessment study. *The Educational Forum, 75*, 96-113.

Sheldon, K. M., & Biddle, B. J. (1998). Standards, Accountability, and School Reform: Perils and Pitfalls. *Teachers College Record*, 164-180.

Shepherd, C.M., & Mullane, A.M. (2008). Rubrics the key to fairness in performance based assessments. *Journal of College Teaching and Learning*, *5*(9), 31.

Simons, J., Irwin, D. & Drinnien, B. (1987). *Psychology-The search for understanding*. New York, N.Y., West Publishing Company.

Slater, T. (n.d.). *Performance assessment*. Informally published manuscript, Department of Physics, Montana State University, Bozeman, Montana. Retrieved May 3, 2011, from http://www.flaguide.org/extra/download/cat/perfass/perfass.pdf

Slater, T. (n.d.). *Classroom assessment techniques performance assessment*. Unpublished manuscript, Department of Physics and Astronomy, University of New Mexico, Albuquerque, new Mexico. Retrieved May 3, 2011, from http://www.flaguide.org/cat/perfass/perfass1.php

Sloan, W. (2011). Coming to terms with common core standards. *ASCD Infobrief*, *16*(4).

Smarter Balanced Assessment Consortium, Initials. (2010). *A summary of core components*. Retrieved March 30,2011 from http://www.k12.wa.us/smarter/ pubdocs /SBACSummary2010.pdf

Soulsby, E. P. (n.d.). *Assessment*. Retrieved March 21, 2011, from http://www.assessment.uconn.edu/index.html

Stecher, B., (2010) "Performance Assessment in an Era of Standards-Based Educational Accountability." *Standford Center for Opportunity Policy in Education*. Standford Center for Opportunity Policy in Education, 2010. Retreived on April 17, 2011, from http://edpolicy.stanford.edu/pages/pubs/pub_docs/assessment/scope_pa_stecher.pdf

Stevens, D. D., & Levi, A. (2005). *Introduction to rubrics: An assessment tool to save time, convey effective feedback, and promote student learning.* Sterling, VA: Styus, Retrieved May 1, 2011, from http://www.introductiontorubrics.com/overview.html

Stevens. Classical Conditioning Cartoon. Retrieved August 23, 2012, from Google User content https://lh4.googleusercontent.com/NId7vhLFkD_TKj1-SVoNTfd3jPGOJC1lyX5tTPgX8zsSmA0YPaIlZaOksR0QrGAlyYsjhuNwFK-2lU9uZVm-yuD-VbE2ZzDptExvENx0WWPKVV6teZE

Stiggins, R. J. Assessment Literacy." *Phi Delta Kappan* 72(3), 534-539.

Stix, A. (1996). Creating rubrics through negotiable contracting and assessment. *Proceedings of the National Middle school Conference*. Retrieved from ERIC database. (ED411273)

Stotsky, S., & Wurman, Z. (2010). Common core standards still don't make the grade: why Massachussetts and California must regain control over their academic destinies. *A Pioneer Institute White Paper*, 65, 1-13. Retrieved May 3, 2011, from http://www.pioneerinstitute.org/pdf/common_core_standards.pdf

Sweet, D. U.S. Department of Education, Office of Research. (1993). *Performance assessment.* Washington, D.C.: Government Printing Office. Retrieved March 31, 2011, from http://www2.ed.gov/pubs/OR/ConsumerGuides/perfasse.html

Tinoca, L., Seung-Hyun, S., & Williams, L/. (2001). *Tips for teachers-assessing project-based learning*. Retrieved May 3, 2011, from http://www.edb.utexas.edu/minliu/pbl TIPS/assess.html

Tung, Rosann. (2010). *Including performance assessments in accountability systems: a review of scale-up efforts*. Boston, MA: Center for Collaborative Education.

University of Texas. (2010). *Assess students*. Unpublished manuscript, Instructional Assessment Resources, University of Texas, Austin, TX. Retrieved March 31, 2011, from http://www.utexas.edu/academic/ctl/assessment/iar/students/report/rubrics-casestudy1.php

Van Duzer, E. (1998). *Developing a Comprehensive View of General Technological Literacy*. ERIC Database. ED437033

Van Duzer, E. (2006). *Overcoming the limitations of the factory system of education*. ERIC Database. (ED490530)

Vendlinski, T, Niemi, D, & Wang, J. (n.d.). Learning assessment by designing assessments: an on-line formative assessment design tool. Manuscript submitted for publication, National Center for Research on Evaluation, Standards and Student Testing (CRESST) , UCLA, Los Angeles, USA. Retrieved May 1, 2011, from adds.cse.ucla.edu/reports/pdf/ learning_assessment.pdf

Way, W., McClarty, A.L., Murphy, D., Keng, L., & Furkhen, C. (2011). Through-course common core assessments in the united states: can summative assessment be formative? *Proceedings of the Annual Meeting of the American Educational Research Association*, Retrieved May 3, 2011, from http://www.pearsonassessments.com/hai/images /tmrs/AERA_Through_Course_Common_Core_Assessments_032411

Weigle, S.C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wiggins, G.(1990).. ERIC digest. Retrieved from ERIC database. (ED328611)

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco: Jossey-Bass, Inc.

Wilson, Owen. (2006). *Beyond bloom - a new version of the cognitive taxonomy*. Unpublished manuscript, Education, University of Wisconsin, Stevens Point, Wisconsin. Retrieved March 30, 2011, from http://www.uwsp.edu/education/lwilson/curric/newtaxonomy.htm

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, *7*(1), Retrieved March 15, 2011, from http://www.uncw.edu/cte/et/

Wolk, R. A. (2011, March). High Stakes of Standards-Based Accountability. *Education Week, 30*(23), 24-32. Retrieved May 1, 2011, from http://edweek.org/ew/…23volk eph.30.html?tk…

Wood, G., Darling-Hammond, L., Neill, M., & Roschewski, P. United States Congress, Forum for Education and Democracy. (2007). *Refocusing accountability: using local performance assessments to enhance teaching and learning for higher order skills*. Washington, D.C.: Government Printing.

Wren, D.G. Department of Research, Evaluation, and Assessment, Virginia Beach City Public Schools. (2009). *Performance assessment: a key component of a balanced assessment system* (2). Retrieved March 30, 2011, from http://www.vbschools.com/accountability /research_briefs/ResearchBriefPerfAssmtFinal.pdf

Xue, Y., Meisels, S. J., DiPrima Bickel, D., Nicholson, J., & Atkins-Burnett, S. (2000). An analysis of parents' attitudes towards authentic performance assessment. *Proceedings of the Aera 2000 annual meeting.* New Orleans, LA. Retrieved from ERIC database. (ED443868)

Zane, T. (2009). Performance assessment design principles gleaned from constructivist learning theory (Part 2). *TechTrends*, 53(3), 92.

Zane, T. (2009). Performance assessment design principles gleaned from constructivist learning theory (part 1). *TechTrends*, 53(1), 81-88.

Zimmaro, D.M. (2004). *Developing grading rubrics*. Informally published manuscript, Division of Instructional Innovation and Assessment, University of Texas, Austin, TX. Retrieved May 3, 2011, from http://www.utexas.edu/academic/mec/research/pdf/rubricshandout.pdf

Appendix A: Performance task examples

Podcast Project

Standards:

- Students will be able to identify and apply properties of triangles
- Students will be able to analyze congruent figures
- Students will be able to apply properties of points, lines, and planes

Task

We have been spending time studying congruent triangles and working on how to prove that two triangles are congruent. You and a partner will now create a podcast using a tablet computer. I will give you a proof that includes a diagram, given information, and what you are trying to prove. I would like you and your partner to complete the proof on the tablet computer, describing all of your steps as you write them. Be sure to describe how the properties of points, lines, angles, and triangles are used in your proof and help you to prove that the triangles are congruent. Both you are your partner must talk for part of your podcast. I will be grading you on the accuracy of your proof and the clarity of your explanation.

Total score: _____ / 20

	Poor	Good	Excellent
Proof is constructed correctly	There are 3 or more errors in the steps of the proof (1 pt)	Most steps in the proof are correct, with one or two errors (3 pts)	All steps in the proof are correct (5 pts)
Proof includes correct properties of points, lines, angles, and triangles	Students do not use the correct properties to justify their steps (1 pt)	Students use most of the correct properties of points, lines, angles, and triangles to justify their steps, but there are minor errors (3 pts)	Students use correct properties of points, lines, angles, and triangles in their proof to justify their steps (5 pts)

Students use correct congruence postulate	Students do not use the correct postulate to prove congruence (1pt)		Students use the correct postulate to prove congruence (2pts)
Podcast contains a complete and clear description of how to complete the proof	Podcast contains a description that is missing a lot of information and is difficult to understand (0 – 2 pts)	Podcast contains a description that is slightly unclear or incomplete (3 – 5 pts)	Podcast contains a complete and clear description of how to complete the proof (6 – 8 pts)

(Mueller, 2011)

Informative Speech

Standards:

- Students will demonstrate knowledge of the 50 states and capitals.
- Students will speak at a volume and pace appropriate for the situation.
- Students will communicate for different situations and audiences.
- Students will actively listen and give appropriate feedback.
- Students will recognize and appreciate different ideas.

Task**:**

Present a speech to the class to inform your classmates about your state. While you are a member of the audience, you will be giving your classmates useful feedback.

Process:

Your speech should be based on your paper. Be sure to include all the information so that the class knows as much about your state as you do! Take a look at the rubric before you start practicing since you will be graded on your speaking ability as well as the information you present. As you listen to each speech, you need to fill out a short "Presentation Evaluation Form" to give your classmates feedback about their presentation.

Rubric

Criteria	Excellent (3 points)	Acceptable (2 points)	Needs Work (1 point)
Information Presented x3	All information is included and detailed.	One or two pieces of information missing.	More than two pieces of information missing.
Pace of Speech	Speech is easy to understand, but doesn't feel slow.	Speech is either too fast or too slow.	Speech is hard to understand.
Volume of Speech	Speech is easily heard.		Speech is hard to understand.
Eye Contact	Eye contact is made throughout the presentation.	Eye contact is made more than half of the presentation.	Eye contact is not made during the presentation.
Feedback Given to Classmates x2	Two pieces of appropriate feedback for each classmate.		Less than two pieces of feedback for each student or not appropriate feedback.
Acceptance of Other Performances x2	Student is attentive and respectful during other performances.		Student is neither attentive nor respectful during other performances.

(Mueller, 2011)

Road Trip Directions

Standards:

- Students will locate places using map skills (scale, cardinal directions, latitude and longitude, etc.)
- Students will identify similarities and differences.
- Students will write useful directions.
- Students will use correct mechanics and spelling in well-organized writing.
- Students will solve problems using number facts and operations.
- Students will search for valuable and appropriate information on the Internet.
- Students will demonstrate word processing skills.

Task:

Plan a 500-mile, 3-day road trip in a region of the United States and write out step-by-step driving directions.

Process:

After deciding where you want your road trip to take place, you must write out detailed directions that include:

- Roads traveled and distance and direction traveled on each road
- Starting and stopping points each day (with total mileage traveled each day)
- Two daily activities (6 total)
- Total mileage
- Don't forget to plan for bathroom and food breaks!

When your directions are complete, you will compare and contrast your directions with a classmate's. Together, you will complete a Venn diagram.

Rubric:

Criteria	Excellent (5 points)	Acceptable (3 points)	Needs Work (0 points)
Detail of Directions	All roads, distances, and directions included	Some roads, distances, or directions missing	
Daily Mileage	Daily mileage included		Daily mileage missing
Daily Activities	Two activities each day	Less than two activities each day	
Total Mileage	Total mileage included		Total mileage missing
Realistic Plan	Plan is realistic and all necessary breaks included	Plan is realistic but not all necessary breaks included	Plan is not realistic
Venn Diagram	Complete (5 characteristics in each section)		Incomplete (less than 5 characteristics in each section)

(Mueller, 2011)

Appendix B: Creating Rubrics

Let's review the steps involved in creating a rubric (Mertler, 2001).

.Step 1:Re-examine the learning objectives to be addressed by the task. This allows you to match your scoring guide with your objectives and actual instruction.

Step 2: Identify specific observable attributes that you want to see (as well as those you don't want to see) your students demonstrate in their product, process, or performance. Specify the characteristics, skills, or behaviors that you will be looking for, as well as common mistakes you do not want to see.

Step 3: Brainstorm characteristics that describe each attribute. Identify ways to describe above average, average, and below average performance for each observable attribute identified in Step 2.

Step 4a: For holistic rubrics, write thorough narrative descriptions for excellent work and poor work incorporating each attribute into the description. Describe the highest and lowest levels of performance combining the descriptors for all attributes.

Step 4b: For analytic rubrics, write thorough narrative descriptions for excellent work and poor work for each individual tribute. Describe the highest and lowest levels of performance using the descriptors for each attribute separately.

Step 5a: For holistic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for the collective attributes. Write descriptions for all intermediate levels of performance.

Step 5b: For analytic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for each attribute. Write descriptions for all intermediate levels of performance for each attribute separately.

Step 6: Collect samples of student work that exemplify each level. These will help you score in the future by serving as benchmarks.

Step 7: Revise the rubric, as necessary. Be prepared to reflect on the effectiveness of the rubric and revise it prior to its next implementation

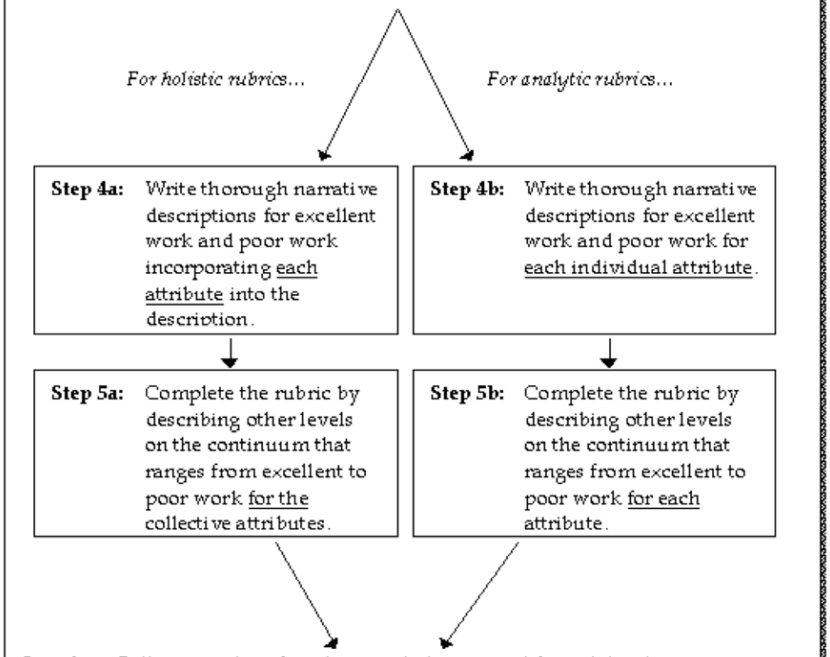
Designing Scoring Rubrics:

Step-by-Step Procedure

Step 1: Re-examine the learning objectives to be addressed by the task.

Step 2: Identify specific observable attributes that you want to see (as well as those you don't want to see) your students demonstrate in their product, process, or performance.

Step 3: Brainstorm characteristics that describe each attribute.



Step 6: Collect samples of student work that exemplify each level.

Step 7: Revise the rubric, as necessary.

Appendix C Glossary

21st Century Skills: Skills and habits of mind needed to succeed in today's world.

Assessment: The systematic collection, review, and use of information about educational programs. One goal of *assessment* is to provide feedback to both the learner and teacher.

Authentic assessments: A form of assessment in which students are asked to perform real-world tasks that demonstrate meaningful application of essential knowledge and skills. Authentic tasks can range from analyzing a political cartoon to making observations of the natural world to computing the amount of paint needed to cover a particular room.

Bell Curve: A symmetrical bell-shaped curve that represents the distribution of data around a central norm.

Bloom's Taxonomy: A hierarchy of cognitive domains that teachers use to guide their students through the learning process.

Cognitive: Conscious intellectual activity.

College and Career Readiness (CCR): Readiness to succeed, without remediation, in entry-level, credit-bearing academic college courses and in workforce training programs.

Common Core Standards (CCSS): K-12 standards, created by an association of state governmental and educational leaders, which support CCR, provide consistent educational standards across the country, are rigorous and include an emphasis on higher-order skills, compare well with standards in top-performing countries, and are evidence-based as well as built upon current state standards.

Consequence validity: Generalizing the results of an assessment to a larger population.

Construct validity: The assessment measures what it claims to measure based on theory and that it does not measure something else.

Construct underrepresentation: A threat to validity which indicates that the tasks measured in the assessment fail to include important dimensions or facets of the construct.

Content validity: The extent to which the test questions represent the skills being taught in the specified subject area.

Convergent validity: An assessment with significant construct validity will correlate with other measures with which it is expected to correlate.

Criterion validity: A measure of whether your assessment produces results that are comparable to other known measures of the material being taught.

Criterion-referenced: An assessment which rates how well a testee performed on the tested criteria, in comparison to the criteria.

Discriminate validity: An assessment will not correlate with measures which it is not expected to correlate.

Embedded task*:* An assessment task that is used in such a way that the student would normally not know it is an assessment activity.

Formative Assessment: A frequent assessment which informs the teacher throughout the year what the students are learning and what needs improving.

Halo Effect: A type of rater bias. When the rater of multiple performance assessments gives an examinee inconsistent scores.

Higher-level thinking skills: Include critical thinking, analysis and problem solving.

History effect: When an intervening event invalidates the content of the assessment.
Holistic: Educating the whole person.
Instrumentation: The design of the test.
Internal consistency: A measure of the coherence of all the items in the assessment protocol.
Inter-rater reliability: The term used to describe whether or not scores are consistent from one rater to the next.
Iterative: Repeatable in that it relates to the measurement of a student's development over the course of time.
Intra-rater reliability: The term used to describe the external factors that alter the consistency of a student's scores.
Inter-rater reliability: The possibility of human error between two different raters.
Intrinsic: Often used as internal motivation as it is essential.
Learning outcome(s): A statement that describes an essential goal for student learning, which the student should be able to demonstrate by the end of the course/etc.
Maturation: The idea that students change over time.
Metacognition: The awareness of how and what one knows, or one's knowledge concerning one's own cognitive processes or anything related to them.
Multimedia: Of or relating to the combined use of several media.
Norm-referenced: An assessment which rates the position of the testee, on the tested criteria, in comparison to a population.
Outcomes-based education: A student-centered education reform model which focuses on measuring student performance of outcomes, or what is learned, rather than the traditional focus on input, or what is taught.
Performance assessments: "Procedures in which respondents are required to carry out tasks or processes in which they demonstrate their ability to apply knowledge and skills" (Arias, 2010, p. 85).
Performance continuum: A succession or progression of steps or degrees.
Performance task**:** A "real or simulated situation that requires students to generate one or more products or performances in order to acquire mastery of identified learning outcomes."
Rater Bias: How the instructor or assessing professional's opinion, partiality, or preconceived notions affect reliability.
Reliability: The consistency or repeatability of ones measures.
Rubric: An authentic assessment tool used for evaluating criteria that are subjective and complex and that can help multiple instructors come to similar conclusions about construction of higher-level conceptual knowledge, performance skills, and attitudes.
Schema: Is the cognitive framework that assists the learner in the organization and interpretation of information.
Simulations: The technique of representing the real world by a computer program.
Standard(s): A statement of specific criteria for what students are expected to learn and be able to do; often written as content standards (what the student should know) and performance standards (what the student should be able to do).
Standardized test: A test that is administered and scored in a consistent, or "standard", manner. Standardized tests are designed in such a way that the questions, conditions for

administering, scoring procedures, and interpretations are consistent and are administered and scored in a predetermined, standard manner.

Standards-based reform: School reform movement, dominant since the 1990's, implemented differently in various states, but with common elements of: standards for what students should know and be able to do (and sometimes understand), assessment aligned to these standards used for accountability of schools, based on student achievement of proficiency.

Summative assessment: This kind of assessment is typically an end of year or semester final examination where student knowledge is assessed, but with little time allotted for improvement within the classroom.

Teaching-learning Continuum: In outcomes-based educational approach, the interrelationships between classroom practice, assessments, and planning revolving around continual evaluation.

Test-retest reliability: Measures the degree to which the same assessment given twice produces the same results.

Valid assessment: A measurement tool that measures what it was intended to measure.

Validity: The extent to which an assessment measures what it was intended to measure.