RESEARCH BRIEF
DEPARTMENT OF TEACHING AND LEARNING
DEPARTMENT OF PLANNING, INNOVATION, AND ACCOUNTABILITY
OFFICE OF RESEARCH AND EVALUATION – May 22, 2017

VIRGINIA BEACH CITY PUBLIC SCHOOLS
CHARTING THE COURSE

Performance Assessment: A Key Component of a Balanced Assessment System

Author: Douglas G. Wren, Ed.D., Educational Measurement and Assessment Specialist
Other Contact Persons: Amy E. Cashwell, Ed.D., Chief Academic Officer
Donald E. Robertson, Jr., Ph.D., Chief Strategy and Innovation Officer

Performance assessment is about performing with knowledge in a context faithful to more realistic adult performance situations, as opposed to out of context, in a school exercise.

Grant Wiggins (2006)

ABSTRACT

Performance assessment is recommended for evaluating students' higher-order thinking and the acquisition of knowledge, concepts, and skills required for success in the 21st century. Relevant literature on performance assessment was reviewed for this report, which is an updated version of a 2009 research brief by the same author.[1] The current brief addresses the following topics: clarification of the term "performance assessment," a comparison of traditional and performance assessments, the procedures involved in developing performance assessments and rubrics, and the role performance assessment plays in a balanced assessment system. Links to various online sources are included in the reference list.

KEY TOPICS:

INTRODUCTION: A STRATEGIC FRAMEWORK

On December 16, 2014, the School Board of Virginia Beach adopted a new strategic plan for Virginia Beach City Public Schools (VBCPS). *Compass to 2020: Charting the Course* included four goals, including "Goal 1 - High Academic Expectations." One strategy under this goal stated VBCPS will "continue to implement a balanced assessment system with an emphasis on standards-based assessments and performance-based assessments" (VBCPS, 2015, p. 2).

What exactly are performance-based assessments? While some professionals tend to use jargon freely and assume everyone is familiar with the terms and meanings, this report defines and gives examples of performance-based assessments. In addition, the process for developing performance assessments as well as how this type of assessment fits in a balanced assessment system is explained.

PERFORMANCE-BASED ASSESSMENT: DEFINITIONS

The term "performance-based assessment" is frequently referred to as performance assessment, or by its acronym, PBA.[2] Herrington and Herrington (1998) noted that the terms "performance assessment" and "authentic assessment" are also used interchangeably; occasionally they are used in tandem. For example, recent legislation by the Virginia General Assembly alluded to "age-appropriate, authentic performance assessments" (Virginia Department of Education, 2014, p. 1). Performance assessment and PBA are simply shortened versions of performance-based assessment, but there is one notable

[1] Many of the sources retained from the original brief are seminal works by renowned authorities in educational measurement and assessment (e.g., Darling-Hammond, Herman, Marzano, McTighe, Popham, Stiggins, and Wiggins).

[2] Performance assessments are also called performance tasks (PTs). Although some might argue that the terms refer to different processes, the term performance assessment is used throughout this report to describe what could be considered PTs.

difference between authentic assessment and performance-based assessment. Gulikers, Bastiaens, and Kirschner (2004) cited earlier literature to explain the distinction between performance assessment and authentic assessment:

> Some see authentic assessment as a synonym to performance assessment (Hart, 1994; Torrance, 1995), while others argue that authentic assessment puts a special emphasis on the realistic value of the task and the context (Herrington & Herrington, 1998). Reeves and Okey (1996) point out that the crucial difference between performance assessment and authentic assessment is the degree of *fidelity* of the task and the conditions under which the performance would normally occur. Authentic assessment focuses on high fidelity, whereas this is not as important an issue in performance assessment. These distinctions between performance and authentic assessment indicate that every authentic assessment is performance assessment, but not vice versa (Meyer, 1992). (p. 4)

There has been a considerable amount of information written on performance assessment during the past 30 years. Consequently, there are numerous definitions available. Palm (2008) observed that some of the definitions were extremely broad, others were quite restrictive, and that most of the definitions of performance assessment were either response-centered (i.e., focused on the response format of the assessment) or simulation-centered (i.e., focused on the student performance observed during the assessment).

A publication from the Office of Educational Research and Improvement of the U.S. Department of Education (1993) provided an example of a response-centered definition of performance assessment:

> Performance assessment … is a form of testing that requires students to perform a task rather than select an answer from a ready-made list. For example, a student may be asked to explain historical events, generate scientific hypotheses, solve math problems, converse in a foreign language, or conduct research on an assigned topic. (para. 1)

A simulation-centered approach with reference to "real-life contexts" appears in another 1990s definition of performance assessment. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCTM], 1999) defined the performance assessment as "Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied" (p. 186).

It is important to note that the word "emulate" is used in the Standards definition.[3] Educators should keep in mind that performance assessments are more meaningful when they imitate real-life situations. Furthermore, Wiggins (1992) suggested that well-designed performance assessments are "enticing" to students, which seems plausible. In all probability, more students would rather participate in one of the activities listed below (i.e., examples of performance assessments) than take a paper-and-pencil test:

- Design and construct a model
- Develop, conduct, and report the results of a survey
- Perform a science experiment
- Write a mock letter to the editor of a newspaper

[3]The definition of performance assessment was revised in the most recent version of the *Standards for Educational and Psychological Testing*: "Assessments for which the test taker actually demonstrates the skills the test is intended to measure by doing tasks that require those skills" (AERA, APA, & NCTM, 2014, p. 221).

TRADITIONAL ASSESSMENTS VS. PERFORMANCE ASSESSMENTS

The popularity of performance assessments during the late 1980s and 1990s came about in part because of dissatisfaction with traditional, multiple-choice tests (Kahl, 2008). By the end of the 20th century, performance assessment had moved "from a trendy innovation to an accepted element of good teaching and learning" (Brandt, 1998, p. v). Soon afterward, large-scale performance assessments at the state and national levels nearly disappeared due to the testing requirements of the *No Child Left Behind Act* (Stecher, 2010). Recently, the pendulum has been swinging back, as educators, politicians, and the public have once again realized the limitations of standardized, multiple-choice tests (Darling-Hammond, 2014).

Table 1 illustrates the advantages and disadvantages of performance assessment and traditional assessment methods. Performance assessment is inherently student-centered and caters more to higher-order thinking compared to traditional assessment. Sophisticated multiple-choice tests—such as the *College and Work Readiness Assessment* (CWRA+; Council for Aid to Education, 2007), which includes a document-based, selected-response section in addition to a performance task—can be developed but the work is complex and time-consuming. Regardless, performance assessment "is viewed as having better possibilities to measure complex skills and communication, which are considered important competencies and disciplinary knowledge needed in today's society" (Palm, 2008, p. 3). In short, performance assessments are better suited for measuring students' attainment of 21st century skills than are traditional assessments.

Table 1: Attributes of Traditional Assessments and Performance Assessments

Attribute	Traditional Assessment	Performance Assessment
Assessment Activity	Selecting a response	Performing a task
Nature of Activity	Contrived activity	Activity emulates real life
Cognitive Processes	Remembering/Understanding	Creating/evaluating/analyzing/applying
Development of Solution	Teacher-structured	Student-structured
Objectivity of Scoring	Easily achieved	Difficult to achieve
Evidence of Mastery	Indirect evidence	Direct evidence

Sources: Liskin-Gasparro (1997), Mueller (2008), and Pohl (2000).

Advocates have also emphasized performance assessment is more in line with good teaching than traditional assessment (Palm, 2008). Popham (2001) and other testing experts (e.g., Haladyna, Nolen, & Haas, 1991; Mehrens, 1991) stated that teaching to the test, which Popham refers to as item-teaching, is highly unethical in preparation for traditional assessments, but teaching to the test is actually encouraged when it comes to performance assessments (Mueller, 2008). With performance assessment, students have access to scoring rubrics in advance so they will know how their performance (e.g., oral response, written response, presentation, or journal) will be evaluated. Additionally, teachers should allow their students to preview examples of high-quality and poor-quality performance products to use as models, provided the products cannot be mimicked.

Critics of performance assessment routinely call attention to the fact that the scoring of performance assessments can be highly subjective. Even though the creation of functional scoring rubrics or other standards for evaluating performance assessments is achievable, applying one standard or a set of standards consistently across a group of written responses, research projects, oral performances, or portfolios can be difficult. Scoring becomes a Herculean task when the group includes every student in a particular grade level across a large school district. However, the outcomes often justify the means: "Two decades of research has found that when teachers use, score, and discuss the results of high-quality performance assessments over time, both teaching and learning improve" (Darling-Hammond, 2014, p. 11).

DEVELOPING PERFORMANCE ASSESSMENTS

A number of authors have described the general process for developing performance assessments (Allen, 1996; Brualdi, 1998; Herman, Aschbacher, & Winters, 1992; Moskal, 2003). There are three basic steps in this process: defining the purpose, choosing the activity, and developing the scoring criteria. These will be explained in the next few sections.

Defining the Purpose

The first step in developing performance assessments involves determining which concepts, knowledge, and/or skills should be assessed. The developer needs to know what type of decisions will be made with the information garnered from the assessment. Herman et al. (1992) suggested that teachers ask themselves five questions as they narrow down the myriad of possible learning objectives under consideration:

1. What important cognitive skills do I want my students to develop? [e.g., communicate effectively in writing, employ algebra to solve real-life problems]
2. What social and affective skills do I want my students to develop? [e.g., work independently, appreciate individual differences]
3. What metacognitive skills do I want my students to develop? [e.g., reflect on the writing process, self-monitor progress while working on an independent project]
4. What types of problems do I want them to be able to solve? [e.g., perform research, predict consequences]
5. What concepts and principles do I want my students to be able to apply? [e.g., use principles of ecology and conservation, understand cause-and-effect relationships] (pp. 25-26)

The initial step in developing performance assessments is analogous to the first stage in the "backward design" model espoused by Wiggins and McTighe (2005) in *Understanding by Design*. The questions posed in Stage 1 (Identify Desired Results) include these: "What should students know, understand, and be able to do? What content is worthy of understanding? What 'enduring' understandings are desired? What essential questions will be explored?" (McTighe, n.d., p. 2). For both backward design and performance assessment, the priority in the first step is establishing a clear focus for both instruction and assessment in terms of measurable objectives.

Choosing the Activity

The next step in the development of a performance assessment is to select the performance activity. Brualdi (1998) reminded teachers they should first consider several factors, including available resources, time constraints, and the amount of data required to make an adequate evaluation of the student's performance. In her synthesis of the literature on developing classroom performance assessments, Moskal (2003) made these recommendations:

- *The selected performance should reflect a valued activity*. [An example would be a real-life situation.]
- *The completion of performance assessments should provide a valuable learning experience*. [Since performance assessments typically require a considerable investment of classroom time, there should be a comparable payoff for students in terms of acquired knowledge and for teachers in their understanding of the students' knowledge.]
- *The statement of goals and objectives should be clearly aligned with the measurable outcomes of the performance activity*. [The elements of the activity must correspond with the objectives specified when the purpose was defined in the previous step.]

- *The task should not examine extraneous or unintended variables*. [Students should not be required to possess knowledge or skills that are not relevant to the activity's purpose.]
- *Performance assessments should be fair and free from bias*. [Activities that give some students an unfair advantage over other students should not be selected. Moskal gave an example of an activity that included baseball statistics and could penalize students who are not knowledgeable about baseball.] (p. 2)

The five recommendations above are inherently related to the validity of the performance assessment. Validity is defined as "the extent to which a test does the job for which it is used" (Payne, 2003, p. 579). Validity is the most important attribute of a good assessment (Lyman, 1998). Concerning performance assessments, Randhawa and Hunter (2001) stated, "It is precisely because of their potential benefits that validity issues need to be systematically addressed through multiple lines of inquiry" (p. 22).

Publishers of nationally-normed, standardized tests go to great lengths to acquire different types of validity evidence for their products. If there is no evidence of validity for an assessment, then interpreting and using the assessment's results cannot be done with confidence. The most common type of validity evidence is content validity, which can be obtained by having an assessment reviewed by qualified content experts. A content expert is "someone who knows enough about what is to be measured to be a competent judge" (Fraenkel & Wallen, 1996, p. 156). Each content expert is tasked with determining if the performance activity matches the learning objective(s) it was intended to measure. The rubrics that are designed to score performance tasks and products should also be reviewed for content validity. Rubric development for performance assessments will be discussed in the next section.

Developing the Scoring Criteria

The last step in constructing a performance assessment is developing the scoring criteria. While traditional assessments are comprised mostly of items for which the answer is either right or wrong, the difference is not as clear-cut with performance assessments (Brualdi, 1998). To evaluate a presentation, project, or composition, a rubric is generally used (Popham, 1997). Wiggins and McTighe (2005) provided this definition of a rubric: "a criterion-based scoring guide consisting of a fixed measurement (4 points, 6 points, or whatever is appropriate) and descriptions of the characteristics for each score point. Rubrics describe degrees of quality, proficiency, or understanding along a continuum" (p. 173).

Two types of rubrics are used to evaluate performance assessments. "Analytic scoring rubrics divide a performance into separate facets and each facet is evaluated using a separate scale. Holistic scoring rubrics use a single scale to evaluate the larger process" (Moskal, 2003). Moskal's six general guidelines for developing either type of rubric are the following:

- 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. (pp. 2-3)

When creating analytic scoring rubrics, McTighe (1996) noted that teachers could allow students to assist, "based on their growing knowledge of the topic" (p. 9). There are other practical suggestions to consider when developing rubrics. Stix (1997) recommended using neutral words (e.g., novice, apprentice, proficient, distinguished; attempted, acceptable, admirable, awesome) instead of numbers for each score level to avoid the perceived implications of good or bad that come with numerical scores. Another suggestion from Stix was to use an even number of score levels to avoid "the natural temptation of instructors—as well as students—to award a middle ranking" (p. 3). For analytic rubrics, it may be necessary to assign different weights to certain components depending on their importance relative to the overall score. Whenever different weighting is used on a rubric, the rationale for this must be made clear to all stakeholders (Moskal, 2003).

Gathering evidence of content validity is critical for both performance assessments and rubrics, but it is also imperative that rubrics are written is such a way to maximize interrater and intrarater reliability. Without reliability, the interpretation of the performance assessment results cannot be valid. Interrater reliability refers to the scoring agreement and consistency of multiple raters, while intrarater reliability refers to the consistency of scores assigned by one rater at different points of time (Moskal, 2000). When more than one person is involved in scoring a performance assessment, acceptable interrater reliability can only be achieved if all of the scorers have the same interpretation of the rubric and consistently apply this shared understanding each time they score. A high degree of intrarater reliability occurs when an individual interprets and applies the rubric consistently during all scoring activities, regardless of when and where the scoring activity takes place (e.g., in the morning over coffee, in the afternoon during planning, in the evening after dinner, or the following week just before a deadline).

Herman et al. (1992) emphasized the importance of having "confidence that the grade or judgment was a result of the actual performance, not some superficial aspect of the product or scoring situation" (p. 80). Key to the scoring situation is the quality of training that teachers and other scorers receive. Researchers have agreed the degree of interrater reliability is correlated with training (Herman et al., 1992; Oakleaf, 2009; Weigle, 1999). Appropriate training allows teachers to come to a common understanding of the rubric and consensually define important components of student performance.

In order to avoid what Herman et al. (1992) referred to as "capricious subjectivity" (p. 81) among a large group of scorers (e.g., schoolwide or districtwide), compulsory, comprehensive training should be conducted. These experts provided the elements of and other considerations for a training session, which have been paraphrased as follows:

- Performance assessment orientation – Scorers are given a complete overview and are actually administered the assessment.
- Scoring criteria clarification – Scorers discuss all aspects of the rubric/scoring guide and view a range of samples, which facilitates further dialogue on score assignment.
- Scoring practice – Herman et al. (1992) called this "the heart of the rater training process" (p. 84). As scorers become more proficient, clear-cut samples give way to more challenging or borderline samples. Open discussion about each scored sample follows.
- Revision of protocols – When scorers encounter unexpected responses that are not clearly defined in the rubric/scoring guide, they will need to devise guidelines for scoring these.
- Recording scores – Scorers are trained on specific procedures for documenting scores.
- Reliability documentation – Scorers may begin working independently after they have attained an acceptable level of agreement. In general, scores must be within one rubric point of each other (i.e., adjacent reliability). Reliability checks should be ongoing.

- Scheduling – The amount of time it will take to train scorers adequately depends on a number of factors:
  - The experience level of the scorers – have they done this before?
  - The scorers' familiarity with the scoring criteria – have they seen the rubric yet?
  - The time it takes for scorers to come to consensus about the meaning of the criteria – is the rubric clearly written with well-defined criteria?
  - The complexity of the scoring criteria in relation to the quality of the work to be scored – how many of the "borderline" performance assessment products are there to be scored?

Despite the fact that developing rubrics and training scorers can be a complicated process, the ensuing rewards are worth the effort. Perhaps the greatest value of rubrics is that "they provide information to teachers, parents, and others interested in what students know and can do … [and] *promote* learning by offering clear performance targets to students for agreed-upon standards" (Marzano, Pickering, & McTighe, 1993, p. 29).

Other Considerations

Performance assessments should always be field-tested before they are fully implemented in schools. As Wiggins warned, "Unpiloted, one-event testing in the performance area is even more dangerous than one-shot multiple-choice testing" (Brandt, 1992, p. 36). Invaluable feedback from the persons who administer and score the assessments, as well as from the students themselves, can be obtained in pilot studies. Field tests can provide information to determine if the assessment contains bias or is measuring any unintended variables. Roeber (1996) mentioned that writing the directions for performance assessments is difficult, but it is more easily facilitated after field-testing the assessment. Test administrators should note "areas of students [*sic*] confusion, responses that students provided which are vague or incomplete, and ways in which some or all of [the] students responded that were not anticipated" (p. 2).

Performance assessments have been used, revised, and reused by educators for decades, and the ensuing paper trail is extensive. There is a seemingly endless supply of PBAs, performance tasks, and rubrics available commercially or at no cost from numerous online sources. A helpful resource for evaluating rubrics is *A Rubric for Rubrics* (Mullinax, 2003), which can be accessed at http://www.tltgroup.org/resources/rubrics/A_Rubric_for_Rubrics.htm. There is no need to reinvent the wheel when it comes to performance assessments; however, the processes required to ensure valid and reliable results from performance assessments involve a great deal of time and attention to detail.

A BALANCED ASSESSMENT SYSTEM

Testing authority Rick Stiggins (2008a) made this prognosis in the not-too-distant past:

> We have come to a tipping point in American education when we must change our assessment beliefs and act accordingly, or we must abandon hope that all students will meet standards or that the chronic achievement gap will close. The troubling fact is that, if all students don't meet standards—that is, if the gap doesn't close between those who meet and don't meet those standards—our society will be unable to continue to evolve productively in either a social or an economic sense. Yet, paradoxically, assessment as conceived, conducted, and calcified over the past has done as much to perpetuate the gap as it has to narrow it. This must change now and it can. As it turns out (again paradoxically), assessment may be the most powerful tool available to us for ensuring universal student mastery of essential standards. (p. 1)

In other words, assessment is not only part of the problem; it is also an important part of the solution. High-quality performance assessments have the potential to play a major role in the betterment of K-12 education. As Darling-Hammond (2014) recently noted, "These activities [performance assessments] not only engage students in more intellectually challenging work that reflects 21st century skills, they also serve as learning opportunities for teachers, when they are involved in using the assessments and scoring them together" (p. 11).

During the address quoted above, Stiggins (2008a) also said Americans "have invested literally all of our resources in once-a-year testing for decades" (p. 2). Previously, the National Education Association (NEA, 2003) reported that most assessment systems in the U.S. were "out of balance, with standardized tests dominating" (p. 1). Consequently, school districts began re-evaluating their assessments and implementing what they called balanced assessment systems.

A balanced assessment system is comprised of different types of formative and summative assessments administered on both a large scale and at the classroom level. In this context, "balanced" does not refer to assessments of equal weight (Redfield, Roeber, & Stiggins, 2008). The NEA (2003) asserted both traditional and performance assessments that yield reliable, valid results in a timely manner have a place in a balanced assessment system. Stiggins (2008b) also explained, "Truly productive assessment systems within schools and districts serve the information needs of a wide variety of assessment users" (p. 4). In keeping with this, Gong (2010) stated that a balanced assessment system should provide the following information:

- Coherently informs different actors responsible for different levels of the educational system, including state/national, district, school, classroom/individual
- Provides "vertical" information integrating summative, interim, and formative assessments
- Is comprehensive enough to inform different purposes, including accountability, program improvement, and instruction
- Inclusive of students – provides appropriate assessment for all students
- Inclusive of valued content & skills
- Provides both diagnosis and prescription information of "what is" and "what should be done" to improve (p. 9)

Because most parents have many preconceived ideas about assessment—usually based on their own experiences with testing—educating them about how performance assessment fits in a balanced assessment system is critical (Meisels, Xue, Bickel, Nicholson, & Atkins-Burnett, 2001). Meisels et al. cited evidence from previous research and their own study of parental support for performance assessment. They further advised, "If it [performance assessment] is ever to become more generally accepted by parents and policy makers, it is essential that parents' reactions be taken into account and shaped through positive and informative interactions with teachers and other educators" (p. 16).

CONCLUSION

Performance assessments have been defined as "Assessments for which the test taker actually demonstrates the skills the test is intended to measure by doing tasks that require those skills" (AERA, APA, & NCTM, 2014, p. 221). Not only does performance assessment allow students to demonstrate their abilities in more genuine contexts than other types of assessment, performance assessment has other advantages over the traditional assessments that are still prevalent in most schools. Students will recognize real-life connections and therefore find high-quality performance assessments more engaging. Measurable outcomes necessary for students to achieve success in the 21st century are more easily evaluated with performance assessments.

It takes a great deal of time and effort to develop high-quality performance assessments and scoring rubrics. Additional time should be invested to ensure that the assessments, rubrics, and the scoring methods yield valid and reliable results. Still more time is required to communicate to students, parents, and other stakeholders the information they need to know regarding performance assessments. What are the benefits of these collective efforts? With the effective utilization of performance assessments within the framework of a balanced assessment system, significant improvements can be made to both teaching and learning.

REFERENCES

Allen, R. (1996). *Performance Assessment.* Wisconsin Education Association Council. Retrieved from http://weac.org/articles/performance-assessment

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Brandt, R. (1992). On Performance Assessment: A Conversation with Grant Wiggins. *Educational Leadership, 49*(8), 35-37. Retrieved from http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_199205_brandt2.pdf

Brandt, R. (1998). Foreword. In G. Wiggins and J. McTighe, *Understanding by Design* (pp. v-vi). Alexandria, VA: Association for Supervision and Curriculum Development.

Brualdi, A. (1998). Implementing Performance Assessment in the Classroom. *Practical Assessment, Research & Evaluation, 6*(2). Retrieved from http://PAREonline.net/getvn.asp?v=6&n=2

Council for Aid to Education. (2007). *College and Work Readiness Assessment* [Measurement instrument], New York, NY: Council for Aid to Education.

Darling-Hammond, L. (2014). Testing to, and Beyond the Common Core. *Principal, 93*(3), 8-12. Retrieved from http://www.naesp.org/sites/default/files/Darling-Hammond_JF14.pdf

Fraenkel, J. R., & Wallen, N. E. (1996). *How to Design and Evaluate Research in Education* (3rd ed.). New York, NY: McGraw-Hill.

Gong, B. (2010). *Using Balanced Assessment Systems to Improve Student Learning and School Capacity: An Introduction*. Paper commissioned by the Council of Chief State School Officers and the Research and Development Consortium sponsored by Renaissance Learning. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/Balanced%20Assessment%20Systems%20GONG.pdf

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). Perceptions of Authentic Assessment: Five Dimensions of Authenticity. Paper presented at the Second Biannual Joint Northumbria/EARLI SIG Assessment Conference, Bergen, Norway. Retrieved from http://www.ou.nl/Docs/Expertise/OTEC/Publicaties/judith%20gullikers/paper%20SIG%202004%20Bergen.pdf

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution. *Educational Researcher, 20*(5), 2-7.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). A Practical Guide to Alternative Assessment. Alexandria, VA: Association for Supervision and Curriculum Development. Retrieved from http://files.eric.ed.gov/fulltext/ED352389.pdf

Herrington, J., & Herrington, A. (1998). Authentic Assessment and Multimedia: How University Students Respond to a Model of Authentic Assessment. *Higher Education Research and Development, 17*(3), 305-322. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0729436980170304#.VcunbflViko

Kahl, S. (2008). *The Assessment of 21st Century Skills: Something Old, Something New, Something Borrowed*. Paper presented at the Council of Chief State School Officers 38th National Conference on Student Assessment, Orlando, FL.

Liskin-Gasparro, J. (1997). *Comparing Traditional and Performance-Based Assessment*. Paper presented at the Symposium on Spanish Second Language Acquisition, Austin, TX.

Lyman, H. B. (1998). *Test Scores and What They Mean* (6th ed.). Boston: Allyn and Bacon.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model.* Alexandria, VA: Association for Supervision and Curriculum Development. Retrieved from http://files.eric.ed.gov/fulltext/ED461665.pdf

McTighe, J. (n.d.). *Understanding by Design* [White paper]. New York, NY: McGraw-Hill Education Global Holdings, LLC.

McTighe, J. (1996). What Happens Between Assessments? *Educational Leadership, 54*(4), 6-12. Retrieved from http://jaymctighe.com/wordpress/wp-content/uploads/2011/04/What-Happens-Between-Assessments.pdf

Mehrens, W. A. (1991). *Defensible/Indefensible Instructional Preparation for High Stakes Achievement Tests: An Exploratory Trialogue*. Paper presented at the Annual Meetings of the Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Meisels, S. J., Xue, Y., Bickel, D. D., Nicholson, J., & Atkins-Burnett, S. (2001). *Parental Reactions to Authentic Performance Assessment.* Ann Arbor, MI: University of Michigan, Center for the Improvement of Early Reading Achievement. Retrieved from http://www.ciera.org/library/archive/2001-06/0106prmx.pdf

Moskal, B. M. (2000). Scoring Rubrics: What, When, and How? *Practical Assessment Research and Evaluation, 7*(3). Retrieved from http://pareonline.net/getvn.asp?v=7&n=3

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation, 8*(14). Retrieved from http://PAREonline.net/getvn.asp?v=8&n=14

Mueller, J. (2016). Authentic Assessment Toolbox. Retrieved from http://jonathan.mueller.faculty.noctrl.edu/toolbox/whatisit.htm

Mullinax, B. B. (2002). *A Rubric for Rubrics.* TLT Group. Retrieved from http://www.tltgroup.org/resources/rubrics/A_Rubric_for_Rubrics.htm

National Education Association. (2003). *Balanced Assessment: The Key to Accountability and Improved Student Learning.* Washington, DC: Author. Retrieved from http://www.ferris.edu/HTMLS/administration/academicaffairs/charterschools/resources/docs/balanced.pdf

Oakleaf, M. (2009). Using Rubrics to Assess Information Literacy: An Examination of Methodology and Interrater Reliability. *Journal of the American Society for Information Science and Technology, 60*(5), 969–983. Retrieved from http://meganoakleaf.info/oakleafjasist2009.pdf

Palm, T. (2008). Performance Assessment and Authentic Assessment: A Conceptual Analysis of the Literature. *Practical Assessment, Research & Evaluation, 13*(4), 1-11. Retrieved from http://pareonline.net/pdf/v13n4.pdf

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

Pohl, M. (2000). *Learning to Think, Thinking to Learn: Models and Strategies to Develop a Classroom Culture of Thinking.* Melbourne, VIC: Hawker Brownlow Education.

Popham, W. J. (1997). What's Wrong—and What's Right—with Rubrics. *Educational Leadership, 55*(2), 72-75. Retrieved from http://www.ascd.org/publications/educational-leadership/oct97/vol55/num02/What's-Wrong%E2%80%94and-What's-Right%E2%80%94with-Rubrics.aspx

Popham, W. J. (2001). Teaching to the Test? *Educational Leadership, 58*(6), 16-20. Retrieved from http://www.ascd.org/publications/educational-leadership/mar01/vol58/num06/Teaching-to-the-Test%C2%A2.aspx

Randhawa, B. S., & Hunter, D. M. (2001). Validity of Performance Assessment in Mathematics for Early Adolescents. *Canadian Journal of Behavioural Science, 33*(1), 14-24. Retrieved from http://www.researchgate.net/profile/Darryl_Hunter/publications https://www.researchgate.net/publication/232432585_Validity_of_performance_assessment_in_mathematics_for_early_adolescents

Redfield, D., Roeber, E., & Stiggins, R. (2008). *Building Balanced Assessment Systems to Guide Educational Improvement.* Paper presented at the Council of Chief State School Officers 38th National Conference on Student Assessment, Orlando, FL. Retrieved from http://browenhorst.tie.wikispaces.net/file/view/2_Article_AssessmentSystem.pdf

Roeber, E. D. (1996). Guidelines for the Development and Management of Performance Assessments. *Practical Assessment, Research & Evaluation, 5*(7). Retrieved from http://pareonline.net/getvn.asp?v=5&n=7

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. Retrieved from https://scale.stanford.edu/system/files/performance-assessment-era-standards-based-educational-accountability.pdf

Stiggins, R. J. (2008a). *Assessment FOR Learning, the Achievement Gap, and Truly Effective Schools*. Presentation given at the Educational Testing Service and College Board Conference, Educational Testing in America: State Assessments, Achievement Gaps, National Policy and Innovations, Washington, DC. Retrieved from http://www.ets.org/Media/Conferences_and_Events/pdf/stiggins.pdf

Stiggins, R. J. (2008b). *Assessment Manifesto: A Call for the Development of Balance Assessment Systems*. Portland, OR: Educational Testing Service, Assessment Training Institute. Retrieved from http://www.nyscoss.org/img/uploads/file/Assessment Manifesto Article - Rick Stiggins.pdf

Stix, A. (1997). *Empowering Students Through Negotiable Contracting*. Paper presented at the National Middle School Initiative Conference, Long Island, NY. Retrieved from http://files.eric.ed.gov/fulltext/ED411274.pdf

U.S. Department of Education, Office of Educational Research and Improvement. (1993). *Consumer Guide: Performance Assessment* (ED/OERI 92-38). Retrieved from http://www2.ed.gov/pubs/OR/ConsumerGuides/perfasse.html

Virginia Beach City Public Schools. (2015). *Compass to 2020: Charting the Course - The Strategic Framework of Virginia Beach City Public Schools.* Retrieved from http://www.vbschools.com/compass/2020/ourstrategicframework.asp

Virginia Department of Education. (2014). *Guidelines for Local Alternative Assessments for 2014-2015, Developed in Response to 2014 Acts of Assembly*. Retrieved from http://www.doe.virginia.gov/testing/local_assessments/index.shtml

Weigle, S. C. (1999). Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing, 6*(2), 145-178. Retrieved from https://eric.ed.gov/?id=EJ617674

Wiggins, G. (1992). Creating Tests Worth Taking. *Educational Leadership, 49*(8), 26-33. Retrieved from http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_199205_wiggins.pdf

Wiggins, G., & McTighe, J. (2005). *Understanding by Design* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.