STATE BOARD OF EDUCATION

HEARING TYPE: __X__ INFORMATION/ACTION

DATE: AUGUST 23, 2006

SUBJECT: **COLLECTION OF EVIDENCE GUIDELINES AND PROTOCOLS**

SERVICE UNIT: Assessment and Research
Dr. Joe Willhoft, Assistant Superintendent

PRESENTERS: Dr. Joe Willhoft, Assistant Superintendent
Assessment and Research

Dr. Lesley Klenk, CAA Options Administrator

BACKGROUND:

Engrossed Substitute Senate Bill 6475 authorized the use of a Collection of Evidence (COE) as an option for meeting standards necessary to obtain a Certificate of Academic Achievement (CAA). Prior to implementation, the bill requires that the State Board of Education approve the guidelines, protocols, and scoring criteria for the collection. In making the approval decision, the board must find that the guidelines, protocols, and scoring criteria:

1) Meet professionally accepted standards for a valid and reliable measure of the Grade Level Expectations and the Essential Academic Learning Requirements; and
2) Are comparable to or exceed the rigor of the skills and knowledge that a student must demonstrate on the Washington Assessment of Student Learning (WASL).

The Office of Superintendent of Public Instruction (OSPI) submitted the draft guidelines and submission forms to the State Board for approval. The guidelines specify the number and types of work samples that must be submitted. The forms include the protocols that are proposed to ensure that the collection represents the student's work.

The draft guidelines and forms were distributed for review and comment on July 10, and were posted on the board's Web site. Individuals provided comments at the July 28 board meeting, and written comments also were received. A summary of the comments is attached.

OSPI is currently revising the guidelines and protocols based on: 1) the attached public comments, 2) comments from educators obtained in the OSPI survey, 3) recommendations received from the CAA Options Technical Committee, and 4) comments from State Board members. OSPI will prepare a new version of the guidelines and protocols that will be distributed to you prior to the August 23 meeting.

At the meeting, the board will be asked to consider and approve the revised guidelines and protocols.

Also, attached is a paper that includes professional standards for reliability and validity, which was reviewed by the OSPI National Technical Advisory Committee and the CAA Technical Advisory Committee. Recommendations from these two groups have been incorporated into this version of the standards. This document is intended to be used by the State Board as it makes its decisions regarding the COE scoring criteria, guidelines, and protocols.

Also attached is a paper written by Cathy Taylor that explains how the standards are applied to the COE. ***Pages 2-4 of this paper includes the proposed validity and reliability standards for the Collection of Evidence.***

August 9, 2006
State Board of Education

Comments Received on the Collection of Evidence Guidelines and Protocols

Summarized below are the comments that were received by the State Board of Education on the draft Collection of Evidence guidelines and protocols during the public hearing at the State Board meeting on July 28,2006, and that were received in writing prior to the August 7 comment deadline.

Seven individuals presented comments at the public hearing, including Mary Lindquist, WEA; Mary Kenfield, PTA; Nick Straley, Columbia Legal Services; Christie Perkins, WA State Special Education Coalition; Suzi Wright, Tulalip Tribe; Bonnie Bashaw, and Michael Tate, State Board for Community and Technical Colleges. Written comments were submitted by eight individuals, including Rebecca Venable; Michael Tate, State Board for Community and Technical Colleges; Charles Hasse, WEA; Nick Straley, Columbia Legal Services; Suzi Wright, Tulalip Tribe; Robert Allen; Rachel DeBellis; and Ann Lynch.

Summarized below are the major issues that were identified in the oral and written comments. To obtain a copy of the complete written comments, contact Laura Moore at lmoore@ospi.wednet.edu.

- The guidelines need to be clearer and less complex: Several individuals wrote or commented that it is crucial that students, teachers, and parents be able to clearly understand what is required, and that the guidelines and procedures need to be definite, concise, and not overly complex. Concerns were expressed that if the were too complex or unclear, students without support from parents or well-trained teachers and English language learners may not have access to the option.

Specific concerns were expressed regarding:

.. the requirement that teachers provide information on what assistance was provided;

- the mathematics requirement that work samples be representative of multiple strands; and

- the need for a parent or family-friendly set of guidelines.

- Steps must be taken to ensure equity of access: Concerns were expressed that all eligible students be given access to the alternative, which will require that parents and students be notified of the option, that teachers and other school officials have training in how to compile a collection and support students, and that funds be available to compensate schools for the added administrative burden. Also, as mentioned above, the guidelines need to be clear so that students and teachers will work together to complete sufficient and proficient collections. Specifically, the State Board was asked to be sensitive to students whose challenging personal circumstances, habits, and patterns of learning make it difficult to effectively access the alternative and navigate the system to put together a COE. Also, reviewers expressed concern regarding equal access and opportunity for special education students and English language learners.

One reviewer stated that school districts need to put more emphasis on supporting the creation and use of student learning plans, which can be a useful tool for supporting ELL students grappling with a new language and a new culture. The reviewer was aware of a number of schools that have failed to create plans for many or all of their students.

Also, WEA expressed concern that the system was only designed to support the compilation and scoring of 646 collections during the 2006-07 school year, which would likely also limit the number of students who would be eligible for the COE option.

- **Ample opportunities must be available for teacher training:** Several reviewers expressed that there must be ample opportunities for teachers and other individuals who are involved in assisting students to obtain information and training about the process.

- Compiling collections will result in a significant added workload for schools: The Washington Education Association and several other reviewers expressed concerns about the added workload on teachers and the capacity of building staff to assist students in compiling collections. The association was concerned that additional funds would not be available to support this work.

- The collection is not culturally appropriate for Native American students: A representative of the Tulalip Tribe and the Tribal Leader Congress on Education expressed concern that Native American teachers were not represented adequately in the scoring of the collections in the pilot and that larger scale projects cannot be used in collections. The reviewer also citied SB 6475, which allows the use of "performance tasks as well as written products," and expressed concern with a requirement in the guidelines that the work samples must be in writing. In addition, a concern was expressed that only expository and persuasive writing was permitted, and she questioned why narrative writing, which is more appropriate for Native American students, was prohibited. Lastly, she indicated that requiring students to analyze the author's purpose and point of view is considered arrogant and inappropriate in some cultures because it requires the questioning or interpretation of an authority.

- SPI should develop an electronic submission system: It was recommended that OSPI develop an electronic system to store submissions, archive documents, review records, enhance scoring opportunities, and communicate results.

- The time available to compile collections is too short: A concern was expressed that there would not be enough time for students to compile collections from the time they receive their WASL results (late October) and when they have to submit their collection for scoring (End of March). This will be the case for students who retake the WASL in the summer, although additional scoring opportunities will be available.

- The collections will be difficult to implement in Community College High School Completion programs: A representative of the State Board for Community and Technical Colleges stated that the process and guidelines will be unworkable in High School Completion programs since most all of their students will likely be compiling a

collection. It was suggested that the number of pieces of evidence and the paperwork requirements be reduced, and that school districts be responsible for getting COE materials to the college High School Completion programs.

- The implementation process should must comply with the Administrative Procedures Act and, as a result, be adopted by rule: Representatives from Columbia Legal Services stated that the Collection of Evidence implementation falls within the definition of a "rule" under the Administrative Procedures Act (APA), and that OSPI and/or the State Board of Education must engage in formal rule-making as set forth in the APA. In their view, rules are especially important for the scoring criteria, appeals process, and additional eligibility criteria.

- The current assessment score appeal process needs to better tailored to the COE: Columbia Legal Services representatives stated that the recently adopted score appeal process does not appear to adequately address the unique aspects of the COE assessment, and needs to be revisited.

- More time should be provided for feedback on the process: WEA asked that more time be provided to comment on the guidelines since most teachers were not available during the comment period.

- Broader concerns with the state standards and the WASL: A number of individuals raised concerns that were outside the authority of the State Board (e.g., the requirement that a student has to take the WASL twice and that ELL students be exempted from the graduation requirement) or other, broader concerns (e.g., the WASL and NCLB have absolutely ruined the teaching profession in Washington, the standards need to be improved).

The Application of Professionally Accepted Standards for Reliability and Validity to the Collection of Evidence

Prepared by C. Taylor and J. Willhoft

August 14, 2006

One of the three options that have been legislated as alternatives to performance on the Washington Assessment of Student Learning (WASL) as a means for students to earn a Certificate of Academic Achievement (CAA) is a collection of work samples, also referred to as the Collection of Evidence[1] (COE). Legislation requires that the guidelines and protocols for submission and the criteria used for scoring "meet professionally accepted standards for a valid and reliable measure of grade level expectations and the essential academic learning requirements." (SB 6475, Laws of 2006)

The process recommended by OSPI to the State Board of Education (SBE) is that the standards shown in Tables 1A and 1B, from the ***Standards for Reliability and Validity of Classroom-Based Assessments,*** be reviewed and approved by the National Technical Advisory Committee (NTAC). NTAC approval will assure the SBE that the criteria for reliability and validity against which the COE will be judged meet "professionally accepted standards". The review and approval of these reliability and validity standards will take place in two stages. First, the CAA Options Advisory Committee, composed of national and local educators and assessment experts (See Appendix A) will review, refine (as needed), and approve the standards. These standards will then be submitted to the NTAC for their approval in August of 2006. Once the NTAC adopts a set of reliability and validity standards for the COE, the design features of the COE will be submitted for their review. The NTAC will be asked to reach consensus on the

[1] Collections of Evidence are subject specific (i.e., reading, mathematics, and writing) collections of classroom-based assessments or work samples for individual students that demonstrate comparable curriculum standards as those assessed by WASL.

alignment of design features of the COE that address the standards. That work will be completed in mid-August, and will be presented to the SBE at its August meeting.

Table 1A: Validity Standards for Classroom-based Assessments

Validity Standard 1: **Representation and Fidelity**	Do the knowledge and skills required by the assessments represent the breadth of knowledge and skills defined in the standards?
Validity Standard 2: **Cognitive Demands**	Do the assessment tools and processes require students to demonstrate the targeted knowledge and skills at a cognitive level specified in the standards?
Validity Standard 3: **Consistency Across Assessments**	Do different assessments of the same knowledge and skills elicit comparable work?
Validity Standard 4: **Alignment with Instruction**	Does assessment align with the content taught and the instructional methods used?
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	Do the assessment tools and processes provide an equal opportunity for individuals, regardless of group or setting, to demonstrate the targeted knowledge and skills?
Validity Standard 6: **Consequences of the Interpretation and Use of Assessment Results**	Are there negative consequences for students that could be prevented if assessment tools, processes, events, or decisions had been more valid?

Table 1B: Reliability Standards for Classroom-based Assessments

Reliability Standard 1: **Generalizability**	Is the work typical of what the student knows and is able to do in relation to the learning targets?
Reliability Standard 2: **Sufficiency of Evidence**	Is there sufficient evidence so that one can make a dependable judgment about what each student knows and is able to do in relation to the learning targets?
Reliability Standard 3: **Clarity of Directions and Expectations**	Do the assessment directions provide clear, unambiguous expectations so that students can dependably demonstrate what they know and are able to do in relation to the learning targets?
Reliability Standard 4: **Quality of Scoring**	Are the scoring rules and scoring processes systematic enough to ensure consistent evaluation over time and across diverse samples of student work that demonstrate the same learning targets?

Two sources served as source materials for the attached *Standards for Reliability and Validity of Classroom-Based Assessments*. The first source was the *Standards for Educational and Psychological Testing* developed jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). The fourth edition of these standards was published in 1999. This document is widely accepted within the community of measurement professionals as encompassing the standards to be met for the development, evaluation, and use of tests that are commercially-developed or are used in large scale public assessment systems. The second source was Taylor and Nolen (1996, 2005), in which the authors adapted the *Standards* for application to the classroom assessment context. This latter work was used as the basis for the standards presented in the *Standards for Reliability and Validity of Classroom-Based Assessments*.

Considerations in Applying these Standards to Collections of Evidence

The Collections of Evidence (COE) to be used for the CAA involve the use of classroom-based assessments in a large-scale assessment context. The COE process requires students to collect work samples from classroom assignments and organize this evidence for a large scale purpose. In this case, not all standards for the validity and reliability of classroom-based assessments can be fully addressed by design features of a large scale assessment program. Three validity standards and one reliability standard for classroom-based assessments have limited applicability in this large scale context.

Validity Standard 4 (Alignment with Instruction) can best be evaluated by the classroom teacher or the students who know whether instruction has prepared the students to demonstrate the knowledge and/or skills required by the assessments.

Validity Standard 6 (Consequences of the Interpretation and Use of Assessment Results) requires ongoing research related to validity standards 1-5 and the consquences of the COE for students. Consequences related to students' self-concepts, their conceptions of school and the subject disciplines, and their academic choices as a results of their classroom-based assessment experiences are beyond the scope of the COE. However, consequences related to the COE should be examined. Positive or negative consequences that arise from decisions made based on the collections are relevant to validity ONLY if these consequences are due to problems related to validity standards 1 through 5.

In addition, although it is possible to Enhance Fairness and Minimize Bias (Validity Standard 5) through careful selection of collections to use for scorer training, it is difficult to thoroughly assess Validity Standard 5 without more information about the students. As with Validity Standard 4 (Alignment with Instruction), only the classroom teacher and the students know

whether the features of the assessment tools or events allow students to demonstrate what they know andare able to do. It is possible, however, to ensure that the COE provides opportunities for all qualified students, to demonstrate their knowledge and skills. The guidelines for the COE can be evaluated for the degree to which they enhance fairness and minimize bias.

Finally, for Reliability Standard 2 (Clarity of Expectations) the protocols for the COE, and any subsequent training materials and directions for teachers and students can be evaluated for clarity of expectations. The clarity of directions for assignments can be evaluated only if directions for assignments are provided along with students' work samples. Finally, if students include tests as part of their collections, test questions can be evaluated for clarity.

Above and beyond issues of reliability and validity, a separate standard has been recommended by the CAA Options Advisory Committee to answer the question: "Are there unintended consequences, for students, schools, and districts, of using the assessment system to make decisions about students?" This standard is important to consider when collections of evidence are used to judge students' proficiency in relation to the standards. Examples of unintended consequencies might include poor WASL performance due to the COE option (which would have implications for a school, district, or state AYP), a narrowing of the curriculum to a limited number of assessment tasks, repeated practice with a single task until the student prepares a proficient performance, or other unintended consequences. Studies should be planned to determine whether there are unintended negative consequences of the COE.

In Tables 2A through 2G of this document, the design features of the COE are more fully detailed. Tables 3A and 3B of this document present the approved links between the design features of the COE and the professionally accepted standards for reliability and validity from *Standards for Reliability and Validity of Classroom-Based Assessments*.

Table 2A:
Protocols – Directions to the COE users to indicate the types of evidence needed for each subject area

Writing Protocol
There are to be 5 to 8 written samples that together demonstrate proficiency in idea/development, organization, style, and the use of conventions. More work samples do not equate to a better score: Carefully selected work samples is a better indicator. Work samples should be written in blue or black ink or word processed. ➢ At least one expository or persuasive on-demand essay, timed and supervised in class ➢ At least two expository non-timed essays ➢ At least two persuasive non-timed essays ➢ 3 work samples (including the on-demand sample) may not include any adult assistance beyond setting the prompt and public expectations for an effective paper. ➢ Other work samples may include drafts read with teacher input and general comments (e.g., "You need to check for spelling errors." or "You need to rework your conclusion to wrap up your writing and give your reader something to think about.").
Reading Protocol
Work samples that cover all six strands that are assessed on the Reading WASL. ➢ A minimum of 8 and a maximum of 12 work samples from a classroom setting or a teacher-approved independent setting. Half of the work samples must represent responses to literary text and half of the samples must represent responses to informational text. ➢ All texts used in the work samples must meet high school expectations for rigor of reading material. The work samples must be comparable in rigor in skill and content to the High School Reading WASL. ➢ Work samples may feature work completed in other content areas—science, social studies, CTE coursework, etc. However, they must still address the literary or the informational strands listed above. ➢ One work sample must be a literary analysis paper of a significant piece of text—short story, narrative essay, novel, etc. that includes a demonstration of more than one literary strand. ➢ One work sample must be a research paper that includes at least two texts used for research purposes. Examples of this type of reading responses include: magazine or newspaper article analysis, analysis of historical events or scientific procedures, etc. The work sample should demonstrate more than one informational strand. ➢ One work sample that must be completed in an "on-demand" setting where students are provided an assignment to complete within a class period and without any teacher or peer assistance.
Mathematics Protocol
There must be 8 to 12 work samples. ➢ A variety of work samples such as projects, assignments, or exams ➢ Work samples of moderate or high complexity to ensure moderate or high level cognitive demands of the student ➢ At least two high school level work samples that and can be scored for an entire target from a strand of EALR 1: ➢ At least two high school level work samples can be scored for an entire target* from a strand of EALRs 2 through 5: ➢ Work samples that combine a content strand from EALR 1 and a process strand from EALRs 2 through 5. Work samples for EALRs 2 through 5 must be distributed across EALR 1 content strands. ➢ Work samples you select for EALR 1 should be representative of multiple High School WASL Mathematics Test Specifications ➢ Work samples you select must combine at least one content strand from EALR 1 and at least one process strand from EALRs 2–5. Work samples should be complex enough to demonstrate moderate to high level thinking skills.

Table 2B:

Sufficiency Review – Process used to determine that all of the WASL learning targets for a domain are included in the collection

Writing Protocol
In order to meet the sufficiency guidelines for successfully submitting a Writing Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Reading Protocol
In order to meet the sufficiency guidelines for successfully submitting a Reading Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Mathematics Protocol
In order to meet the sufficiency guidelines for successfully submitting a Mathematics Collection of Evidence, the student and teacher preparing the collection must comply with the following guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored

Table 2C:

Work Sample Documentation

Writing Protocol	*In the "Work Sample Documentation Form" teachers must provide documentation that the work sample demonstrates the state standards in writing. For each work sample, students must check one of the first three boxes on the form as well as the type of draft, process, and teacher-assisted for the work samples in the collection. The teacher must check that an "on-demand" essay is present in the collection. In the last box—teacher assistance—the student must describe what type of assistance he/she received beyond setting the prompt and the parameters of an effective paper.*
Reading Protocol	*In the "Work Sample Documentation Form" students and teachers must check all of the learning strands, both literary and informational. The student must provide of the titles of the texts must be provided to check the rigor of the readability of the texts. The student and the teacher must check each work sample to make sure that each sample addresses at least two strands. The student must identify which work sample is the short literary analysis paper and which is the short informational analysis paper. The teacher must check that an "on-demand" essay is present in the collection.*
Mathematics Protocol	*In the "Work Sample Documentation Form" students and teachers must check that all work samples address every high school content strand. Each work sample must address both a content strand and a process strand. Teachers must check that work samples meet the "rich problem" and high school level mathematics expectation. Students must check that each column and row have two entries. There must be an "on-demand" check*

Table 2D:

Scoring rules used to evaluate the collections – Performance criteria for the scoring rubrics used for each collection are given below along with an indication of the subject area EALRs and components within each EALR that are the focus of the performance criteria. Links to the EALRs are keys to authenticity validity.

Writing Criteria
Content, Organization & Style ➢ Has clear, focused main ideas or positions (EALR 1, Component 1) ➢ Elaborates by using reasons/arguments supported by well-chosen and specific details, examples, anecdotes, facts and/or statistics as evidence to support ideas or positions (EALR 1, Component 1) ➢ Includes information that is thoughtful and useful for the audience to know (EALR 1, Component 1) ➢ Organizes writing to make the best case to explain ideas or support positions (EALR 1, Component 2) ➢ Composes introductions that draw the reader into the main ideas or positions (EALR 1, Component 2) ➢ Writes conclusions that leave the reader with something to think about (EALR 1, Component 2) ➢ Organizes writing into effective, cohesive paragraphs (EALR 1, Component 2) ➢ Provides transitions which clearly serve to connect ideas (EALR 1, Component 2) ➢ Uses language effectively by exhibiting word choices that are effective and appropriate for intended audience, purpose, and form (EALR 1, Component 3) ➢ Writes (where appropriate) sentences or phrases that are varied in length and structure (EALR 1, Component 4) ➢ Provides the reader with a sense of the person behind the words (EALR 1, Component 5) Conventions ➢ Follows the rules of standard English [language] usage (EALR 1, Component 6) ➢ Spelling of commonly used words (EALR 1, Component 6) ➢ Capitalization (EALR 1, Component 6) ➢ Punctuation (EALR 1, Component 6) ➢ Exhibits the use of complete sentences except where purposeful phrases or clauses are used for effect (EALR 1, Component 6) ➢ Indicates paragraphs consistently (EALR 1, Component 6)
Reading Criteria
Comprehension of main ideas and details of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Identifies the main theme/main idea and uses evidence to demonstrate an overall understanding of the text (EALR 2, Component 1) ➢ Summarizes by providing an overarching statement about the text that connects to at least three events from the beginning, middle and end of text (EALR 2, Component 1) ➢ Infers and/or predicts about key elements of the text making connections with evidence (EALR 2, Component 1) ➢ Explains key vocabulary with both denotative and connotative definitions by linking them to the text (EALR 1, Component 2) Analysis, interpretation, & synthesis of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Applies knowledge of key literary/informational elements to enhance and expand understanding of text (EALR 2, Component 2) ➢ Compares and contrasts ideas to explain concepts within or between text (EALR 2, Component 3) ➢ Analyzes text to explain the relationship between cause(s) and effect(s) and links it back to the theme or main idea (EALR 2, Component 2) Thinks critically about literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Evaluate author's/ text's purpose and/or in order to judge effectiveness on intended audience ➢ Evaluates reasoning of ideas / themes within the text and makes connections with evidence Synthesizes information beyond the text by making generalizations, drawing conclusions, or applying information to evaluate a new text or context

Table 2D (Continued)

Mathematics Criteria
Uses high school content knowledge and procedures (EALR 1) with supporting work in: ➢ Number Sense (EALR 1, Component 1) ➢ Measurement (EALR 1, Component 2) ➢ Geometric Sense (EALR 1, Component 3) ➢ Probability & Statistics (EALR 1, Component 4) ➢ Algebraic Sense (EALR 1, Component 5) Solves Problems (EALR 2) ➢ Applies one or more strategies that lead to the answer (EALR 2, Component 2) ➢ Determines the answer to the problem (EALR 2, Component 3) Reasons Logically (EALR 3) ➢ Justifies conclusions, results, and/or answers by addressing the conditions and/or constraints in the problem Communicates Understanding (EALR 4) ➢ Gathers, represents, and/or shares mathematical information using clear mathematical language and organization Makes Connections (EALR 5) ➢ Uses and relates different mathematical models and representations of the same situation using clear mathematical language and organization (EALR 5, Components 1 and 2)

Table 2E

Range-Finding – The process of selecting exemplary collections to represent different performance levels

All Content Areas
Steps in the range-finding process ➢ Select a range of collections to serve as potential anchors for the rubrics during scoring training, practice collections to be used for practice during scoring training, and validity collections to be randomly inserted into scoring process to ensure adherence to scoring rubrics over time ➢ Ensure that all selected collections have met sufficiency criteria ➢ Discuss scoring rubrics ➢ Apply scoring rubrics to selected collections ➢ Discuss applied scores ➢ Adjust scoring rubrics and/or scores, if needed, based on collections ➢ Assign final scores to anchor collections ➢ Assign final scores to practice collections ➢ Assign final scores to validity collections

Table 2F

Scoring Training – The process of training scorers to apply scoring rubrics consistently using anchor collections to anchor rubrics

All Content Areas
Steps in the training process ➢ Review and discuss rubrics ➢ Review and discuss anchor collections ➢ Score practice collections ➢ Discuss assigned scores; work toward consensus with pre-assigned scores ➢ Score second practice collections ➢ Discuss assigned scores; work toward consensus with pre-assigned scores ➢ Scorers must qualify by meeting a criterion of exact agreement with pre-assigned scores

Table 2G

Table Scoring Process – The process of assigning scores to collections

All Content Areas
Steps in the scoring process ➢ Scorers assign scores ➢ Collections are randomly assigned to a second scorer (inter-rater agreement) ➢ Randomly selected collections are rescored by a table leader (supervisor) ➢ Validity collections are given to scorers randomly ➢ Scorers who drift from scoring rubrics are retrained as necessary

The next two tables, Tables 3A and 3B, link each of the Validity and Reliability standards COE design features.

Table 3A: Design Features of COE that Address Validity Standards

Validity Standard	Feature of COE Addressing Standard
Validity Standard 1: **Representation and Fidelity**	➢ Protocols for Reading, Writing, and Mathematics ➢ Sufficiency Review ➢ Scoring Rules ➢ Range-finding ➢ Scoring Training ➢ Scoring Process
Validity Standard 2: **Cognitive Demands**	➢ Protocols for Reading, Writing, and Mathematics
Validity Standard 3: **Consistency Across Assessments**	➢ Range-finding
Validity Standard 4: **Alignment with Instruction**	➢ Student self-report??
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	➢ Range-finding ➢ Scoring Training ➢ Scoring Process
Validity Standard 6: **Consequences of the Interpretation and Use of Assessment Results**	➢ Ongoing validity studies for the COE

Table 3B: Design Features of COE that Address Reliability Standards

Reliability Standard	Feature of COE Addressing Standard
Reliability Standard 1: **Generalizability**	➢ Protocols for Reading, Writing, and Mathematics
Reliability Standard 2: **Sufficiency of Evidence**	➢ Sufficiency Review ➢ Work Sample Documentation Form
Reliability Standard 3: **Clarity of Directions and Expectations**	➢ Protocols for Reading, Writing, and Mathematics ➢ Work Sample Documentation Directions ➢ Work Sample Sign-off Form
Reliability Standard 4: **Quality of Scoring**	➢ Scoring Rules ➢ Range-finding ➢ Scoring Training ➢ Scoring Process

References

Taylor, C. S. & Nolen, S. B. (1996). What does the psychometrician's classroom look like? *Educational Policy Analysis Archives*, v4, n17.

Taylor, C. S. & Nolen, S. B. (2005). *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms.* Columbus, OH: Pearson-Merrill-Prentice Hall.

Appendix A

Certificate of Academic Achievement (CAA) Options Advisory Committee Members

Linda Dobbs, Assistant Superintendent, ESD 189, Mt. Vernon, Washington

Deborah Gonzalez, Executive Director for Learning & Teaching, Puget Sound ESD, Highline, Washington

Gil Mendoza, Executive Director of Grants Management, Tacoma School District, Tacoma, Washington

Barbara Plake, Professor Emeritus, University of Nebraska-Lincoln

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

Edward Wiley, Professor, University of Colorado-Boulder

Appendix B

National Technical Advisory Committee for Assessment

Patricia Almond, University of Oregon, Eugene, Oregon

Peter Behuniac, University of Connecticut, Hartford, Connecticut

Richard Duran, Professor, California State University, Santa Barbara, California

George Englehard, Professor, Emory University, Atlanta, Georgia

Robert Linn, Professor Emeritus, University of Colorado-Boulder, UCLA-CRESST

William Mehrens, Professor Emeritus, Michigan State University, East Lansing, Michigan

Edys Quelmalz, Associate Director of the Center for Technology in Learning, Stanford Research Institute International, Palo Alto, California.

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

Standards for Reliability and Validity of Classroom-Based Assessment

Prepared by
Catherine S. Taylor[1]
University of Washington

In this document, we present standards for relilability and validity of classroom-based assessments. Two sources served to guide the development of the *Standards for Reliability and Validity of Classroom-Based Assessments* presented here. The first source was the *Standards for Educational and Psychological Testing, Fourth Edition* (1999), developed jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). The *Standards* were designed to establish professional expectations for the development, evaluation, and administration of individual tests as well as the interpretation and use of test scores.

Classroom-based assessment does not exactly align with the intent of the *Standards*. First, teachers use a wide range of assessment tools in classrooms including their observations, students' written papers, homework assignments, and other student work. Second, teachers' judgments about students are rarely made based on a single assessment event. When teachers give scores, grades, or written evaluations to students' work from a single assessment event (such as a test or a speech), the *Standards* may provide useful guidance. However, when all of the work from a school term (e.g., quarter, trimester, semester) is summarized into a grade or written summary, more guidance is needed. In this latter situation, standards are needed to address the range of different

[1] These standards are derived from a draft document by Catherine Taylor and Susan Nolen (2005) and include recommended clarifications and revisions received from the Washington State National Technical Advisory Committee and the Washington State Advisory Committee for the Certificate of Academic Achievement Options. We are grateful for the thoughtful review and excellent recommendations.

assessment tools used and the fact that students' knowledge and skills are likely to change because of instruction and experiences. In response to these differences, Taylor and Nolen (1996, 2005) designed framework for reliability and validity as they apply to the classroom context. Their work was the second source used to develop the standards presented here.

Tables 1A and 1B briefly outline six validity and four reliability standards relevant classroom-based assessment. In the text that follows, each standard is described for a classroom teacher audience in order to help guide their thinking about what they should consider when examining their own classroom assessments.

Table 1A: Validity Standards for Classroom-based Assessments

Validity Standard 1: **Representation and Fidelity**	Do the knowledge and skills required by the assessments represent the breadth of knowledge and skills defined in the standards?
Validity Standard 2: **Cognitive Demands**	Do the assessment tools and processes require students to demonstrate the targeted knowledge and skills at a cognitive level specified in the standards?
Validity Standard 3: **Consistency Across Assessments**	Do different assessments of the same knowledge and skills elicit comparable work?
Validity Standard 4: **Alignment with Instruction**	Does assessment align with the content taught and the instructional methods used?
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	Do the assessment tools and processes provide an equal opportunity for individuals, regardless of group or setting, to demonstrate the targeted knowledge and skills?
Validity Standard 6:	Are there negative consequences for students that

Consequences of the Interpretation and Use of Assessment Results	could be prevented if assessment tools, processes, events, or decisions had been more valid?

Table 1B: Reliability Standards for Classroom-based Assessments

Reliability Standard 1: **Generalizability**	Is the work typical of what the student knows and is able to do in relation to the learning targets?
Reliability Standard 2: **Sufficiency of Evidence**	Is there sufficient evidence so that one can make a dependable judgment about what each student knows and is able to do in relation to the learning targets?
Reliability Standard 3: **Clarity of Directions and Expectations**	Do the assessment directions provide clear, unambiguous expectations so that students can dependably demonstrate what they know and are able to do in relation to the learning targets?
Reliability Standard 4: **Quality of Scoring**	Are the scoring rules and scoring processes systematic enough to ensure consistent evaluation over time and across diverse samples of student work that demonstrate the same learning targets?

Before discussing the standards, however, it is necessary to clarify the meaning of the term *assessment*. Throughout the literature, *assessment* is used to describe assessment tools (e.g., individual test questions, entire tests or quizzes, directions for assignments, and scoring rubrics.), assessment processes (e.g., using a scoring rubric to assign points to students' essays or selecting the information to be used when giving course grades.), assessment decisions (e.g., giving course grades or placing students in special programs.) and assessment events (e.g., completing a test, writing a research paper, or doing a course project.). In the following discussion of standards for reliability and validity for classroom-based assessments, we attempt to identify these different aspects of assessment.

Validity Standards for Classroom-Based Assessments

"*Validity* is an integrative, evaluative judgment of the degree to which…evidence and [theory] support the…inferences and actions based on test scores and other modes of assessment (Messick, 1989)." In his treatise on validity theory, Messick outlined a 'half dozen or so' methods of gathering evidence for the validity of test scores:

> We can look at the content of the test in relation to the content of the domain of reference; we can probe the ways in which individuals respond to the items or tasks; we can examine the relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses; we can survey relationships of test scores with other measures and background variables, that is, the test's external structure; we can investigate differences in these test processes and structures over time, across groups and settings, and in response to . . . interventions such as instructional . . . treatment and manipulation of content, task requirements, or motivational conditions; finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. (p. 16)

To make valid inferences about students using classroom-based assessments, one needs to examine the validity evidence for individual assessment events and for judgments based on collections of student work from different assessment events. In the classroom, validity encompasses (a) whether the assessment tools actually require students to demonstrate the knowledge and/or skills[2] described in the learning targets, (b) whether instruction has prepared

[2] For this document ***knowledge*** includes ideas, principles, and facts as well as *understanding* of concepts, interrelationships among ideas, principles, and facts, and knowing how and when to use ideas, concepts, principles in relevant situations; ***skills*** include thinking and reasoning skills (e.g., making inferences, comparing and contrasting information, drawing conclusions), research skills (e.g., skill in

students for the assessed knowledge and skills *and* for the way the knowledge and skills are assessed, (c) whether the assessment tools or processes are biased in favor of or against individuals or groups, and (d) what occurs as a result of assessment processes, events and decisions including feedback, grading, and placement; students' self-concepts and academic behaviors; and students' understanding of the subject disciplines. Teachers must look at assessment tools, processes, events, and decisions for evidence of their validity. Teachers must consider alternate explanations of student performances (such as invalidity in assessments). Finally, teachers should consider the potential consequences of their assessment choices.

Validity Standard 1: Representation and Fidelity - Do the knowledge and skills required by the assessments represent the breadth of knowledge and skills defined in the standards? Before one can evaluate alignment to the standards, one must think clearly about the domains and/or disciplines that are the focus of education and define clear learning targets related to those domains and/or disciplines. Learning targets may include knowledge and skills; learning targets may also include valued performances that require application of knowledge and skills. Sometimes teachers define their own learning targets; sometimes learning targets are provided by schools, districts, or states. With clear learning targets, the first aspect of the validity of assessments can be evaluated: Whether the assessment tool is asking students to demonstrate valued knowledge and skills in a manner that is authentic to the domain and/or discipline.

Since assessment tools also include rules for assigning points or grades, validity standard 1 also has to do with the degree to which the scoring rules and processes used to assign points or grades are tied to the learning targets *and* whether these scoring rules and processes adequately

---

using the library and Internet to gather information), process skills (e.g., skill in conducting a scientific investigation, skill in using a process to go from initial ideas to a polished piece of writing), problem-solving skills, social skills, and communication skills.

represent the domain and/or discipline. For example, effective writing involves appropriate content, relevant ideas, logical organization, word choices, language usage, appropriate voice, and writing conventions (grammar, punctuation, spelling, and capitalization); therefore, if teachers evaluated students' writing *only* for writing conventions would make the assessment results less valid.

Validity Standard 2: Cognitive Demands – Do the assessment tools and processes require students to demonstrate the targeted knowledge and skills at a cognitive level specified in the standards? An important aspect of validity for classroom-based assessments is whether the assessments actually require students to use the targeted knowledge and/or skills to complete a test or other performance. For example, a student might get the right answer to a multiple-choice math question because she did the same problem for homework and she remembered the answer. Another student might get the right answer because three of the four answer choices were obviously wrong. A third student might get the right answer because he copied another student's answer. A fourth student might get the right answer because he worked out the answer during the test.

Standardized test makers use "tryouts" to find out how the test questions function *before* they use the questions on tests. Most teachers do not have the luxury to do this with their own assessments. Textbook assessments are rarely tried out with students before they are published. Therefore, teachers need to develop ways to find out whether test questions and performance directions actually tap into the concepts and skills they are intended to assess.

One way to do this is to ask students to explain their work or show their steps as they complete various assessments. When students explain their reasoning, their choices, and their solutions, a teacher may discover that an assessment isn't really tapping into the targeted

knowledge and skills. If this is the case, test questions and performance directions can be adjusted to ensure that students must use / demonstrate the targeted knowledge and skills in completing the assessment.

Validity Standard 3: Consistency - Do different assessments of the same knowledge and skills elicit comparable work? Another aspect of validity is whether students do similar work on different assessment tools that are intended to measure the same learning targets. One strategy for determining whether assessment tools and processes can be used to make valid inferences about examinees is to have students do more than one version of the same type of work. For example, a teacher might have 3-4 questions on a test to assess a particular science concept. She might have students do 2-3 science investigations to assess students understanding of investigative procedures. Multiple pieces of evidence provide information about whether students are performing similarly on assessments that are intended to measure the same thing. In short, for Validity Standard 3 teachers can review several sources of evidence to see whether examinees perform consistently across different assessments of the same knowledge and/or skills. If student performances on different tasks measuring the same knowledge or skill are very similar, the teacher can have more confidence that the test questions or performance tasks are measures of the same learning targets.

The grade book excerpt in Figure 1 shows students' performances on six essays, all of which were evaluated with the same two scoring rubrics – a five point rubric for content and a five point rubric for writing conventions. As can be seen, despite the fact that several students in the class consistently earn scores of 5 for the content of their essays, the highest score for Essay 4 was 3. This suggests that there may be a problem with the validity of scores for Essay 4.

Figure 1

An Example of Inconsistency of Assessment Scores in a Classroom as a Potential Threat to Validity

STUDENT SCORES ON 6 ESSAYS

	Essay 1		Essay 2		Essay 3		Essay 4		Essay 5		Essay 6	
Student	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**
Tanya	5	5	5	5	5	5	3	5	5	5	5	5
Mario	5	3	5	4	5	4	3	5	5	5	5	5
Emma	2	3	2	2	4	3	3	4	4	4	5	5
Juan	3	3	4	3	4	4	2	4	4	4	5	4
Geoff	2	3	3	3	3	3	3	3	3	4	4	4
Robin	4	5	4	5	4	5	3	5	5	5	5	5
Caitlyn	5	5	5	5	5	5	3	5	5	5	5	5
Points Possible	5	5	5	5	5	5	5	5	5	5	5	5

Cont. = Content
Conv. = Writing Conventions

Validity Standard 4: Alignment with Instruction - Does the assessment align with the content taught and the instructional methods used?One of the most fundamental validity questions a teacher should ask is whether the learning targets were actually taught, whether the method of assessment fits the *way* knowledge and skills were taught, and whether students had sufficient exposure to and practice with knowledge and skills to be successful on the assessments. For example, if students are asked to practice routine mathematical algorithms in class and for homework but are then asked to apply the algorithms in novel situations on a test, the assessment tool is not valid for the instructional context. A mismatch between what is taught and what is assessed can lead to frustration for teachers and students. It can also result in invalid grades for students.

Validity Standard 5: Enhancing Fairness and Minimizing Bias - Do the assessment tools and processes provide an equal opportunity for individuals, regardless of group or setting, to demonstrate the targeted knowledge and skills?alidity has to do with how well various assessment tools, processes, and/or events allow students to demonstrate their knowledge and skill. When an assessment favors some students over others, this is called bias. Bias occurs whenever students who have achieved the valued knowledge and skills do not or cannot demonstrate their achievements because of some aspect of the assessment tool or process.

Assessment decisions may be invalid when factors *within* the assessment tool prevent students from showing what they know and are able to do. In creating and selecting assessments, teachers must determine whether student work is influenced by factors irrelevant to the targeted learning objectives such as assessment context, format, response mode, cultural experiences, or other factors.

One factor that might affect students performance is when the *context* or *content* is unfamiliar to students *and* unrelated to the learning targets. For example, an assessment might require students to write on a topic about which they have little or no experience (e.g., "Write a story describing something that happened at Thanksgiving dinner."). Although students might be able to write effectively, the writing topic may prevent some students from demonstrating their writing skills. The context of the writing prompt is unrelated to what the teacher wants to know – whether students can write using important knowledge and skills related to the characteristics of effective writing (e.g., organization, word choices), the writing purpose (narrative), and writing conventions (e.g., grammar and spelling). When the context set for an assessment favors some students over others, the assessment tool is biased. Teachers are responsible for creating assessment contexts that allow *all* students to demonstrate their knowledge and skills. This may

mean that the contexts are different for different students. For example, the writing teacher could provide different writing prompts and allow students to select the one that works best for their backgrounds.

Bias also occurs when the *format* of the assessment tool prevents some students from demonstrating their knowledge and skills. Suppose a teacher wants to assess students' understanding of character development, plot development, theme, and setting in literary works. He assigns the same novel to all of the students in his class, and asks for a written essay. To demonstrate their literary analysis skills, students must read and write. Suppose, also, that some students are English language learners (ELL) who have good literary analysis skills but cannot read the novel because the text is too difficult and cannot write the essay because they are not yet skilled writers. A different assessment format may be required for these students (e.g., hearing a book on tape and giving an oral report) in order to make valid inferences about their literary analysis skills.

Teachers need to know whether differences in performance across students are because of true differences in students' knowledge and skills or whether differences are due to invalidity in the assessment tools, processes, or events. If the learned knowledge and skills *can* be demonstrated in a way other than through a specific assessment tool (without changing the target for what is assessed) and if some students can show their knowledge, conceptual understanding and skills through the alternate format, then a single format for the assessment tool is *biased* in favor of those who can perform in the chosen way and against those who cannot.

A third potential source of bias comes from the rules used to assign points to students work. To be valid scoring rules, the rules must focus *only* on the targeted knowledge and skills. For example, suppose a teacher evaluates literary analysis essays using scoring rules that award

points based on elegance of the writing or the creativity of presentation rather than the adequacy of literary interpretations. The assessment process will be biased in favor of students who are skilled or creative writers.

A final source of potential source of bias comes from the teacher. If the teacher 'colors' the process used to assign points to students' work with prior knowledge of students or attitudes toward students – rather than consistently applying scoring rules across all students' work – the resulting scores are unlikely to be valid reflections of students' knowledge and skills.

Validity standard 5 becomes increasingly critical as classrooms become more diverse and whole-group teaching becomes more difficult. Teachers must provide appropriate adaptations of assessment tools and processes while still obtaining valid evidence about student achievement related to the learning targets.

Validity Standard 6: Consequences of Interpretation and Use of Assessment Results - Are there negative consequences that could be prevented if assessment tools, processes, events, or decisions had been more valid? Assessments tools, processes, events, and decisions have effects on students. Tests, projects, teacher feedback, and grades can all influence student learning, self-concepts, motivation (Butler & Nisan, 1986; Covington & Omelich, 1984), and perceptions of the subject areas and disciplines being taught. Therefore, the final standard of validity for classroom-based assessments is related to how classroom assessments affect the students themselves.

If students develop a notion of the discipline of history as a collection of facts that are to be memorized, this consequence is *mis*-educative. If some students get poor grades whereas others get good grades *because of invalidity in standards 1 through 5,* then the consequences that arise from those grades (promotion to the next grade level, placement in special programs, access

to honors classes, etc.) are invalid consequences. Educators have an ethical responsibility to create and select valid assessment tools and to use valid processes so that consequences are fair, are based on appropriate information, and do not create misconceptions for students.

Reliability Standards for Classroom-Based Assessments

Reliability is the degree to which one can depend on the results of an assessment event to accurately reflect the students' proficiency on the knowledge and skills targeted by the assessment tool. The reliability standards presented are an elaboration on three topics found in the reliability literature: a) generalizability, b) standardization of directions, and c) objectivity of scoring (Cronbach, 1970). Reliability in classroom-based assessment refers to the degree to which one can rely on the results of assessment processes and events. Four standards of reliability are relevant to classroom based assessment tools and processes: (1) whether the examinee's performance on a given assessment tool is typical of the examinee's performance, (2) whether there is sufficient evidence available so that one can make dependable statements about what students have learned in relation to the learning targets; (3) whether students know exactly what is expected on tests, performances, and other assessment events so that they are likely to perform in a consistent way, regardless of the time in which the assessment is administered; and (4) whether the scoring rules and assessment processes are systematic enough to ensure that evaluators are consistent across students and over time.

Reliability Standard 1: Generalizability - Is the work typical of what the student knows and is able to do in relation to the learning targets? Assessment experts often talk about reliability as consistency in performance. Is a single throw of a basketball sufficient to make a dependable (reliable) statement about whether or not the student will make a basket during a game? Would the student perform in the same way a second time? To make a reliable

statement about what students know and are able to do, a teacher can ask them to do similar work several times (e.g., summarize the main ideas in a social studies text) and look to see whether their performance is consistent over time. Summative decisions made at the end of a grading period[3] can be much more reliable than the results of individual assessments.

Another aspect of generalizability is whether a student performs consistently on different measures of the same knowledge and skill. For example, during an instructional unit focused on algebraic problem-solving, the teacher can check to see whether the student applies algebraic strategies consistently across problems set in different real world contexts.

Reliability Standard 2: Sufficiency of Evidence - Is there sufficient evidence of student learning so that one can make a dependable judgment about what each student knows and is able to do relation to the learning targets? For summative decisions to be reliable, one must ensure that there is *sufficient, high-quality* assessment information from which to make trustworthy decisions about students. The reliability of summative decisions depends on the validity of the assessment tools and processes. If attention is given to validity standards one through five, then one can begin to ask whether there is sufficient information from which to make reliable decisions. Multiple, valid assessments are very likely to give reliable information about students. The more sources of valid assessment information teachers have at the end of a grading period, the more likely that their decisions will be ones that they and others can trust. Therefore, to address Reliability Standard 2, one must obtain as much valid information about students' achievement of the learning targets as possible. Classroom teachers can and should bring a wide range of information – observations, test scores, homework, class work, written papers, etc. – to bear on summative decisions such as course grades.

[3] A grading period is the time between report cards such as a quarter, trimester, or semester.

Reliability Standard 3: Clarity of Directions and Expectations - Do the assessment directions provide clear, unambiguous expectations so that students can demonstrate what they know and are able to do in relation to the learning targets? When students are not clear about what they are being asked to do, they are less likely to produce the expected response; they are more likely to respond in a way that is *inconsistent* with their own knowledge and skills or to respond differently depending on the context in which the assessment occurs. In contrast, when test items and performance directions are clear and explicit, students are more able to show what they know and are able to do consistently regardless of when or where an assessment event occurs. Expectations for student work are communicated to students in two ways – through directions for test items and assignments and through rules for assigning points (scores) or grades to students' work.

When the directions for tests and performances are clear, student responses are more likely to demonstrate their true knowledge and skills. For example, suppose a teacher created a multiple-choice test wherein students are expected to select the *best* conclusion for the results of a scientific investigation. All answer choices may present viable conclusions; however, only one provides the most thorough conclusion. Students must know, from the test directions, that all of the conclusions are possible and that they are to choose the *best* conclusion. Similarly, if students are expected to use both primary and secondary sources in a research study, the directions for the assignment should indicate this expectation.

When students know the criteria against which their performances are to be evaluated (scoring rules), they are more likely to demonstrate their knowledge and skills related to those expectations. For example, in a mathematics class, students need to know whether they will be evaluated on the *effectiveness* of their problem-solving processes as well as whether they

generate viable solutions to mathematics problems. For a research paper, students need to know whether they will be evaluated on how well they integrate information from several sources to write summaries of important ideas.

When directions are not clear and when students do not know the bases for evaluation of their work, they are less likely to provide a consistent performance from one assessment event to the next, even though the same knowledge and skills are being assessed during different events.

Reliability Standard 4: Quality of Scoring - Are the scoring rules and scoring processes systematic enough to ensure consistent evaluation over time and across diverse samples of student work that demonstrate the same learning targets?enerally three types of assessment tools that could be affected by the consistency of judgments about students' learning: short-answer and performance questions for tests; projects and performances; and different assignments for which a teacher has the same expectations. In these three situations, the consistency of judgments depends on whether the rules for scoring short-answer items, performance items, or extended performances are specific and clear enough that they can be applied consistently across students and whether the same rules for scoring are applied consistently across similar tasks and over time.

When teachers write rules for assigning points (scores) to students' responses to test questions and other assignments, these rules must help the teacher assign scores objectively. For example, a teacher wants students to write a paragraph describing the main character of a story. Figure 2A is a scoring rule that is so vague that the teacher would have difficulty applying it consistently across students. Figure 2B is a scoring rule that provides more specific guidance and is likely to result in more consistent evaluation across students.

Figure 2A: Vague Scoring Rubric for Character Description

4 points	The written work is a thorough and accurate description of the main character.
3 points	The written work is a mostly complete or mostly accurate description of the main character.
2 points	The written work is a partially complete or partially accurate description of the main character.
1 point	The written work is attempted with few details or is mostly inaccurate.
0 points	The written work shows no comprehension of the main character.

Figure 2B: More Specific Scoring Rubric for Character Description

6 points	The written work thoroughly describes the main character, including: • the main character's name • a physical description of the main character (age, sex, clothes, hair color, skin color, and what the character wears) • a reasonable statement about the main character's personality (e.g., friendly) or motives (e.g., wants to get rich) • at least <u>two</u> examples of the main character's actions or dialogue that show his/her personality or motives.
5 points	The written work addresses all four expectations but one or two required details are missing from the physical description
4 points	The written work completely addresses the first three expectations but gives only <u>one</u> example from the text to show personality or motives
3 points	The written work addresses all four expectations but many details are missing from the physical description OR no examples are given to show the character's personality or motives.
2 points	The written work addresses the first three expectations but many details are missing from the physical description OR the work completely addresses the first two expectations
1 point	The written work addresses the first two expectations but many details are missing OR the work gives a partial physical description of the character.
0 points	The written work gives only the name of the main character OR is illegible OR is off task OR shows no comprehension of the text.

In addition to clear scoring rules, teachers are likely to be more consistent if they have examples of previous students' work to show students what work at each performance level looks like.

When teachers ask students to do the same type of work at different times and in different

contexts or when teachers have the same expectations for different short-answer or essay questions on a test, reliability is enhanced when the same scoring rule is used each time. For example, suppose that, at the end of each science investigation, a science teacher asks students to write a paragraph in which they summarize the results of the investigation and explain how the results relate to the initial research question and the scientific theory from which the research question was drawn. The reliability of students' scores for 'drawing conclusions' depends on whether the teacher applies the same scoring rule consistently over time and across students.

Summary

The validity and reliability standards presented here are written for classroom teachers to help them evaluate their assessment tools, processes, events, and decisions. Research has shown that teachers spend from 30-70 percent of their time engaged in assessment processes (Stiggins, Faires-Conklin, & Bridgeford, 1986). The assessment processes, events and decisions are not neutral aspects of the classroom environment. They have important effects on students' self-concepts and their conceptions of school and the subjects taught in school. Therefore, it is essential that teachers learn how to evaluate published assessment tools and their own tools and processes so that they can make valid and reliable judgments about whether students have learned the targeted knowledge and skills. With high quality assessment tools and ethical assessment processes, teachers are more likely to make accurate decisions about how to support their students achievement.

References

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology, 78* (3), 210–216.Covington, M. V., & Omelich, C. L. (1979). Effort: the double-edged sword in school achievement. *Journal of Educational Psychology, 71* (1), 169–182.

Cronbach, L. (1970). *Essentials of Psychological Testing*, Third Edition. New York: Harper & Rowe, Publishers

Messick, S. (1989). Validity. In *Educational Measurement*, Robert Linn (Ed.). Washington, DC: American Council on Education.

Stiggins, R. J., Faires-Conklin, N., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5* (2), 5-17.

Taylor, C. S. & Nolen, S. B. (1996). What does the psychometrician's classroom look like? *Educational Policy Analysis Archives*, v4, n17.

Taylor, C. S. & Nolen, S. B. (2005). *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms.* Columbus, OH: Pearson-Merrill-Prentice Hall.

Appendix A
Certificate of Academic Achievement (CAA) Options Advisory Committee Members

Linda Dobbs, Assistant Superintendent, ESD 189, Mt. Vernon, Washington

Deborah Gonzalez, Executive Director for Learning & Teaching, Puget Sound ESD, Highline, Washington

Gil Mendoza, Executive Director of Grants Management, Tacoma School District, Tacoma, Washington

Barbara Plake, Professor Emeritus, University of Nebraska-Lincoln

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

Edward Wiley, Professor, University of Colorado-Boulder

Appendix B
National Technical Advisory Committee for Assessment

Patricia Almond, University of Oregon, Eugene, Oregon

Peter Behuniac, University of Connecticut, Hartford, Connecticut

Richard Duran, Professor, California State University, Santa Barbara, California

George Englehard, Professor, Emory University, Atlanta, Georgia

Robert Linn, Professor Emeritus, University of Colorado-Boulder, UCLA-CRESST

William Mehrens, Professor Emeritus, Michigan State University, East Lansing, Michigan

Edys Quelmalz, Associate Director of the Center for Technology in Learning, Stanford Research Institute International, Palo Alto, California.

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

Standards for Reliability and Validity of Classroom-Based Assessment

Prepared by
Catherine S. Taylor**1**
University of Washington

In this document, we present standards for relilability and validity of classroom-based assessments. Two sources served to guide the development of the *Standards for Reliability and Validity of Classroom-Based Assessments* presented here. The first source was the *Standards for Educational and Psychological Testing, Fourth Edition* (1999), developed jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME).

The classroom context does not exactly align with the intent of the *Standards*. The *Standards* were designed to establish professional expectations for the development, evaluation, and administration of individual tests as well as the interpretation and use of test scores. In contrast, teachers use a wide range of assessment tools in classrooms including their observations, students' written papers, homework assignments, etc. Second, teachers' judgments about students are rarely made based on a single assessment event. When teachers give scores, grades, or written evaluations to students' work on a single assignment (such as a test), the *Standards* can provide useful guidance. However, when all of the work from a school term (e.g., quarter, trimester, semester) is summarized into a grade or written summary, more guidance is needed. In this latter situation, standards are needed to address the range of different assessment tools used and the fact

[1] These standards are derived from a draft document by Catherine Taylor and Susan Nolen (2005) and include recommended clarifications and revisions received from the Washington State National Technical Advisory Committee and the Washington State Advisory Committee for the Certificate of Academic Achievement Options. We are grateful for the thoughtful review and excellent recommendations.

that students' knowledge and skills are likely to change during the course of a term because of instruction. In response to these differences, Taylor and Nolen (1996, 2005) designed frameworks for reliability and validity as they apply to the classroom context. Their work was the second source used to develop the standards presented here.

Tables 1A and 1B briefly outline six validity and four reliability standards relevant classroom-based assessment. In the text that follows, each standard is described for a classroom teacher audience in order to guide their thinking about what they should consider for their own classroom assessments.

Table 1A: Validity Standards for Classroom-based Assessments

Validity Standard 1: **Representation and Fidelity**	Do the knowledge and skills required by the assessments represent the breadth of knowledge and skills defined in the standards?
Validity Standard 2: **Cognitive Demands**	Do the assessment tools and processes require students to demonstrate the targeted knowledge and skills at a cognitive level specified in the standards?
Validity Standard 3: **Consistency Across Assessments**	Do different assessments of the same knowledge and skills elicit comparable work?
Validity Standard 4: **Alignment with Instruction**	Does assessment align with the content taught and the instructional methods used?
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	Do the assessment tools and processes provide an equal opportunity for individuals, regardless of group or setting, to demonstrate the targeted knowledge and skills?
Validity Standard 6: **Consequences of the Interpretation and Use of Assessment Results**	Are there negative consequences for students that could be prevented if assessment tools, processes, events, or decisions had been more valid?

Table 1B: Reliability Standards for Classroom-based Assessments

Reliability Standard 1: **Generalizability**	Is the work typical of what the student knows and is able to do in relation to the learning targets?
Reliability Standard 2: **Sufficiency of Evidence**	Is there sufficient evidence so that one can make a dependable judgment about what each student knows and is able to do in relation to the learning targets?
Reliability Standard 3: **Clarity of Directions and Expectations**	Do the assessment directions provide clear, unambiguous expectations so that students can dependably demonstrate what they know and are able to do in relation to the learning targets?
Reliability Standard 4: **Quality of Scoring**	Are the scoring rules and scoring processes systematic enough to ensure consistent evaluation over time and across diverse samples of student work that demonstrate the same learning targets?

Before discussing the standards, it is necessary to clarify the meaning of the term *assessment*. Throughout the literature, *assessment* is used to describe assessment tools (e.g., individual test questions, entire tests or quizzes, directions for assignments, and scoring rubrics.), assessment processes (e.g., using a scoring rubric to assign points to students' essays or selecting the information to be used when giving course grades.), assessment decisions (e.g., giving course grades or placing students in special programs.) and assessment events (e.g., completing a test, writing a research paper, or doing a course project.). In the following discussion of standards for reliability and validity for classroom-based assessments, we attempt to identify these different aspects of assessment.

Validity Standards for Classroom-Based Assessments

"*Validity* is an integrative, evaluative judgment of the degree to which…evidence and [theory] support the…inferences and actions based on test scores and other modes of assessment (Messick, 1989)." In his treatise on validity theory, Messick outlined a 'half dozen or so' methods of gathering evidence for the validity of test scores:

> We can look at the content of the test in relation to the content of the domain of reference; we can probe the ways in which individuals respond to the items or tasks; we can examine the relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses; we can survey relationships of test scores with other measures and background variables, that is, the test's external structure; we can investigate differences in these test processes and structures over time, across groups and settings, and in response to . . . interventions such as instructional . . . treatment and manipulation of content, task requirements, or motivational conditions; finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. (p. 16)

To make valid inferences about students using classroom-based assessments, one needs to examine the validity evidence for individual assessment events *and* for judgments based on collections of student work from different assessment events. In the classroom, validity encompasses (a) whether the assessment tools actually require students to demonstrate the targeted knowledge and/or skills2, (b) whether instruction has prepared students for the assessed

[2] For this document ***knowledge*** includes ideas, principles, and facts as well as *understanding* of concepts, interrelationships among ideas, principles, and facts, and knowing how and when to use ideas, concepts, principles in relevant situations; ***skills*** include thinking and reasoning skills (e.g., making inferences, comparing and contrasting information, drawing conclusions), research skills (e.g., skill in

knowledge and skills *and* for the way the knowledge and skills are assessed, (c) whether the assessment tools or processes are biased in favor of or against individuals or groups, and (d) what occurs as a result of assessment processes, events and decisions including feedback, grading, and placement; students' self-concepts and academic behaviors; and students' understanding of the subject disciplines. Teachers must look at assessment tools, processes, events, and decisions for evidence of their validity. Teachers must consider alternate explanations of student performances (such as invalidity in assessments). Finally, teachers should consider the potential consequences of their assessment choices. These issues are explained more fully in what follows.

Validity Standard 1: Representation and Fidelity - Do the knowledge and skills required by the assessments represent the breadth of knowledge and skills defined in the standards? Before one can evaluate alignment to the standards, one must know what the standards mean. Teachers may need to examine instructional materials and assessment tasks to clarify the meaning of state or district standards. For example, a standard such as, "The student will comprehend the main ideas and important details from text," is fairly straightforward. However, a standard such as "The student will make inferences and predictions from text," is less obvious. Are the students expected to make simple inferences (e.g., From the sentence, "She ran along the track until she reached the station," the reader can infer that the track and station are related to a railroad.)? Or, are the expected inferences more substantive. For example, inferences about a character's motives based on the character's actions, dialogue, and relationships with other characters require more thinking than simple inferences. Teachers must access documents and sources that help them clarify the meaning of standards before than can

using the library and Internet to gather information), process skills (e.g., skill in conducting a scientific investigation, skill in using a process to go from initial ideas to a polished piece of writing), problem-solving skills, social skills, and communication skills.

judge whether the assessments are aligned with the standards.

Usually standards are stated as learning targets related to those domains and/or disciplines taught in school. Standards may include knowledge and skills; standards may also include valued performances that require application of knowledge and skills. Teachers may define their own learning targets based on district or state standards or learning targets may be provided by schools, districts, or states. With clear learning targets, the first aspect of the validity of assessments can be evaluated: Whether the assessment tool is asking students to demonstrate valued knowledge and skills from the standards both in breadth and depth.

For example, suppose students are to achieve reading eight reading standards that range from comprehension to critical evaluation of text. The teacher should check to see whether their assessment tools represent all eight standards. Secondly, the teacher should check to see that the assessment tool represents the standards in a way that is authentic. If, for example, reading in the world beyond school involves more comprehension and interpretation of text than critical evaluation of text, then an assessment tool should have more items and tasks that assess comprehension and interpretation than for critical evaluation of text.

Since assessment tools also include rules for assigning points or grades, Representation and Fidelity (Validity Standard 1) also has to do with the degree to which the scoring rules and processes used to assign points or grades are tied to the learning targets *and* whether these scoring rules and processes adequately represent the domain and/or discipline. For example, effective writing involves appropriate content, relevant ideas, logical organization, word choices, language usage, appropriate voice, and writing conventions (grammar, punctuation, spelling, and capitalization); therefore, if teachers evaluated students' writing *only* for writing conventions would make the assessment results less valid.

Validity Standard 2: Cognitive Demands – Do the assessment tools and processes require students to demonstrate the targeted knowledge and skills at a cognitive level specified in the standards? An important aspect of validity for classroom-based assessments is whether the assessments actually require students to *use* the targeted knowledge and/or skills to complete a test or other performance. For example, a student might get the right answer to a multiple-choice math question because she did the same problem for homework and she remembered the answer. Another student might get the right answer because three of the four answer choices were obviously wrong. A third student might get the right answer because he copied another student's answer. A fourth student might get the right answer because he worked out the answer during the test.

Standardized test makers use "tryouts" to find out how the test questions function *before* they use the questions on tests; however, most teachers do not have the luxury to do this with their own assessments. Textbook assessments are rarely tried out with students before they are published. Therefore, teachers need to develop ways to find out whether test questions and performance directions actually tap into the concepts and skills they are intended to assess.

One way to do this is to ask students to explain their work or show their steps as they complete various assessments. When students explain their reasoning, their choices, and their solutions, a teacher may discover that an assessment isn't really tapping into the targeted knowledge and skills. If this is the case, test questions and performance directions can be adjusted to ensure that students must use or demonstrate the targeted knowledge and skills in completing the assessment.

Validity Standard 3: Consistency - Do different assessments of the same knowledge and skills elicit comparable work? Another aspect of validity is whether students do similar

work on different assessment tools that are intended to measure the same learning targets. One strategy for examining Consistency (Validity Standard 3) of assessment tools is to have students do more than one version of the same type of work. For example, a teacher might have 3-4 questions on a test to assess a particular science concept. She might have students do 2-3 science investigations to assess students' understanding of investigative procedures. Multiple pieces of evidence provide information about whether different assessments truly measure the same knowledge and skills. In short, for Validity Standard 3, teachers must review several sources of evidence to see whether examinees perform consistently across different assessments of the same knowledge and/or skills. If student performances on different tasks intended to measure the same knowledge or skill are very similar, the teacher can have more confidence that the test questions or performance tasks are measures of the same learning targets.

The grade book excerpt in Figure 1 shows students' performances on six essays, all of which were evaluated with the same two scoring rubrics – a five point rubric for content and a five point rubric for writing conventions. As can be seen, despite the fact that several students in the class consistently earn scores of 5 for the content of their essays, the highest score for Essay 4 was 3. This suggests that there may be a problem with the validity of scores for Essay 4. The teacher may wish to review the directions, his evaluation processes, and/or his instruction to determine whether scores from the fourth essay are valid.

Figure 1

An Example of Inconsistency of Assessment Scores in a Classroom as a Potential Threat to Validity

STUDENT SCORES ON 6 ESSAYS

	Essay 1		Essay 2		Essay 3		Essay 4		Essay 5		Essay 6	
Student	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**	**Cont.**	**Conv.**
Tanya	5	5	5	5	5	5	3	5	5	5	5	5
Mario	5	3	5	4	5	4	3	5	5	5	5	5
Emma	2	3	2	2	4	3	3	4	4	4	5	5
Juan	3	3	4	3	4	4	2	4	4	4	5	4
Geoff	2	3	3	3	3	3	3	3	3	4	4	4
Robin	4	5	4	5	4	5	3	5	5	5	5	5
Caitlyn	5	5	5	5	5	5	3	5	5	5	5	5
Points Possible	5	5	5	5	5	5	5	5	5	5	5	5

Cont. = Content
Conv. = Writing Conventions

Validity Standard 4: Alignment with Instruction - Does the assessment align with the content taught and the instructional methods used? One of the most fundamental validity questions a teacher should ask is whether the learning targets were actually taught, whether the method of assessment fits the *way* knowledge and skills were taught, and whether students had sufficient exposure to and practice with knowledge and skills to be successful on the assessments. For example, if students are asked to practice routine mathematical algorithms in class and for homework but are then asked to apply the algorithms in novel situations on a test, the assessment tool is not valid for the instructional context. A mismatch between what is taught and what is assessed can lead to frustration for teachers and students. It can also result in invalid grades for students.

Validity Standard 5: Enhancing Fairness and Minimizing Bias - Do the assessment tools and processes provide an equal opportunity for individuals, regardless of group or setting, to demonstrate the targeted knowledge and skills?

Another aspect of validity has to do with how well various assessment tools, processes, and/or events *allow* students to demonstrate their knowledge and skill. When an assessment favors some students over others, this is called bias. Bias occurs whenever students who have achieved the valued knowledge and skills do not or cannot demonstrate their achievements because of some aspect of the assessment tool or process. In creating and selecting assessments, teachers must determine whether student work is influenced by factors irrelevant to the targeted learning objectives such as assessment context, format, response mode, cultural experiences, or other factors.

Assessment decisions may be invalid when factors *within* the assessment tool prevent students from showing what they know and are able to do. One factor that might affect students' performance is when the *context* or *content* is unfamiliar to students *and* unrelated to the learning targets. For example, an assessment might require students to write on a topic about which they have little or no experience (e.g., "Write a story describing something that happened at Thanksgiving dinner."). Although students might be able to write effectively, the writing topic may prevent some students from demonstrating their writing skills. The context of the writing prompt is unrelated to what the teacher wants to know – whether students can write using important knowledge and skills related to the characteristics of effective writing (e.g., organization, word choices), the writing purpose (narrative), and writing conventions (e.g., grammar and spelling). When the context set for an assessment favors some students over others, the assessment tool is biased. Teachers are responsible for creating assessment contexts that

allow *all* students to demonstrate their knowledge and skills. This may mean that the contexts are different for different students. For example, the writing teacher could provide different writing prompts and allow students to select the one that works best for their backgrounds.

Bias also occurs when the *format* of the assessment tool prevents some students from demonstrating their knowledge and skills. Suppose a teacher wants to assess students' understanding of character development, plot development, theme, and setting in literary works. He assigns the same novel to all of the students in his class, and asks for a written essay. To demonstrate their literary analysis skills, students must read and write. Suppose, also, that some students are English language learners (ELL) who have good literary analysis skills but cannot read the novel because the text is too difficult and cannot write the essay because they are not yet skilled writers. A different assessment format may be required for these students (e.g., hearing a book on tape and giving an oral report) in order to make valid inferences about their literary analysis skills.

Teachers need to know whether differences in performance across students are because of true differences in students' knowledge and skills or whether differences are due to invalidity in the assessment tools, processes, or events. If the learned knowledge and skills *can* be demonstrated in a way other than through a specific assessment tool (without changing the target for what is assessed), and if some students can show their knowledge, conceptual understanding and skills through the alternate format, then a single format for the assessment tool is *biased* in favor of those who can perform in the chosen way and against those who cannot.

A third potential source of bias comes from the rules used to assign points to students' work. To be valid scoring rules, the rules must focus *only* on the targeted knowledge and skills. For example, suppose a teacher evaluates literary analysis essays using a scoring rubric that

awards points based on the elegance of the writing or the creativity of presentation in addition to the adequacy of literary interpretations. The scoring rules will be biased in favor of students who are skilled or creative writers.

A final source of potential source of bias comes from the teacher. If the teacher 'colors' the process used to assign points to students' work with prior knowledge of students or because of her/his attitudes toward students – rather than consistently applying scoring rules across all students' work – the resulting scores are unlikely to be valid reflections of students' knowledge and skills.

Validity standard 5 becomes increasingly critical as classrooms become more diverse and whole-group teaching becomes more difficult. Teachers must provide appropriate adaptations of assessment tools and processes while still obtaining valid evidence about student achievement related to the learning targets.

Validity Standard 6: Consequences of Interpretation and Use of Assessment Results - Are there negative consequences that could be prevented if assessment tools, processes, events, or decisions had been more valid? Assessments tools, processes, events, and decisions have effects on students. Tests, projects, teacher feedback, and grades can all influence student learning, self-concepts, motivation (Butler & Nisan, 1986; Covington & Omelich, 1984), and perceptions of the subject areas and disciplines being taught. Therefore, the final standard of validity for classroom-based assessments is related to how classroom assessments affect the students themselves.

If students develop a notion of the discipline of history as a collection of facts that are to be memorized, this consequence is *mis*-educative. If some students get poor grades whereas others get good grades *because of invalidity in standards 1 through 5*, then the consequences that

arise from those grades (promotion to the next grade level, placement in special programs, access to honors classes, etc.) are invalid consequences. Educators have an ethical responsibility to create and select valid assessment tools and to use valid processes so that consequences are fair, are based on appropriate information, and do not create misconceptions for students.

Reliability Standards for Classroom-Based Assessments

Reliability is the degree to which one can depend on the results of an assessment event to accurately reflect the students' proficiency on the knowledge and skills targeted by the assessment tool. The reliability standards presented are an elaboration on three topics found in the reliability literature: a) generalizability, b) standardization of directions, and c) objectivity of scoring (Cronbach, 1970). Reliability in classroom-based assessment refers to the degree to which one can rely on the results of assessment processes and events. Four standards of reliability are relevant to classroom based assessment tools and processes: (1) whether the examinee's performance on a given assessment tool is typical of the examinee's performance, (2) whether there is sufficient evidence available so that one can make dependable statements about what students have learned in relation to the learning targets; (3) whether students know exactly what is expected on tests, performances, and other assessment events so that they are likely to perform in a consistent way, regardless of the time in which the assessment is administered; and (4) whether the scoring rules and assessment processes are systematic enough to ensure that evaluators are consistent across students and over time.

Reliability Standard 1: Generalizability - Is the work typical of what the student knows and is able to do in relation to the learning targets? Assessment experts often talk about reliability as consistency in performance. Is a single throw of a basketball sufficient to make a dependable (reliable) statement about whether or not the student will make a basket

during a game? Would the student perform in the same way a second time? To make a reliable statement about what students know and are able to do, a teacher can ask them to do similar work several times (e.g., summarize the main ideas in a social studies text) and look to see whether their performance is consistent over time. Summative decisions made at the end of a grading period3 can be much more reliable than the results of individual assessments.

Another aspect of generalizability is whether a student performs consistently on different measures of the same knowledge and skill. For example, during an instructional unit focused on algebraic problem-solving, the teacher can check to see whether the student applies algebraic strategies consistently across problems set in different real world contexts.

Reliability Standard 2: Sufficiency of Evidence - Is there sufficient evidence of student learning so that one can make a dependable judgment about what each student knows and is able to do relation to the learning targets? For summative decisions to be reliable, one must ensure that there is *sufficient, high-quality* assessment information from which to make trustworthy decisions about students. The reliability of summative decisions depends on the validity of the assessment tools and processes. If attention is given to validity standards one through five, then one can begin to ask whether there is sufficient information from which to make reliable decisions. Multiple, valid assessments are very likely to give reliable information about students. The more sources of valid assessment information teachers have at the end of a grading period, the more likely that their decisions will be ones that they and others can trust. Therefore, to address Reliability Standard 2, one must obtain as much valid information about students' achievement of the learning targets as possible. Classroom teachers can and should bring a wide range of information – observations, test scores, homework, class work, written

[3] A grading period is the time between report cards such as a quarter, trimester, or semester.

papers, etc. – to bear on summative decisions such as course grades.

Reliability Standard 3: Clarity of Directions and Expectations - Do the assessment directions provide clear, unambiguous expectations so that students can demonstrate what they know and are able to do in relation to the learning targets? When students are not clear about what they are being asked to do, they are less likely to produce the expected response; they are more likely to respond in a way that is *inconsistent* with their own knowledge and skills or to respond differently depending on the context in which the assessment occurs. In contrast, when test items and performance directions are clear and explicit, students are more able to show what they know and are able to do consistently regardless of when or where an assessment event occurs. Expectations for student work are communicated to students in two ways – through directions for test items and assignments and through rules for assigning points (scores) or grades to students' work.

When the directions for tests and performances are clear, student responses are more likely to demonstrate their true knowledge and skills. For example, suppose a teacher created a multiple-choice test wherein students are expected to select the *best* conclusion for the results of a scientific investigation. All answer choices may present viable conclusions; however, only one provides the most thorough conclusion. Students must know, from the test directions, that all of the conclusions are possible and that they are to choose the *best* conclusion. Similarly, if students are expected to use both primary and secondary sources in a research study, the directions for the assignment should indicate this expectation.

When students know the criteria against which their performances are to be evaluated (scoring rules), they are more likely to demonstrate their knowledge and skills related to those expectations. For example, in a mathematics class, students need to know whether they will be

evaluated on the *effectiveness* of their problem-solving processes as well as whether they generate viable solutions to mathematics problems. For a research paper, students need to know whether they will be evaluated on how well they integrate information from several sources to write summaries of important ideas.

When directions are not clear and when students do not know the bases for evaluation of their work, they are less likely to provide a consistent performance from one assessment event to the next, even though the same knowledge and skills are being assessed during different events.

Reliability Standard 4: Quality of Scoring - Are the scoring rules and scoring processes systematic enough to ensure consistent evaluation over time and across diverse samples of student work that demonstrate the same learning targets? There are generally three types of assessment tools that could be affected by the consistency of judgments about students' learning: short-answer and performance questions for tests; projects and performances; and different assignments for which a teacher has the same expectations. In these three situations, the consistency of judgments depends on whether the rules for scoring short-answer items, performance items, or extended performances are specific and clear enough that they can be applied consistently across students and whether the same rules for scoring are applied consistently across similar tasks and over time.

When teachers write rules for assigning points (scores) to students' responses to test questions and other assignments, these rules must help the teacher assign scores objectively. For example, a teacher wants students to write a paragraph describing the main character of a story. Figure 2A is a scoring rule that is so vague that the teacher would have difficulty applying it consistently across students. Figure 2B is a scoring rule that provides more specific guidance and is likely to result in more consistent evaluation across students.

Figure 2A: Vague Scoring Rubric for Character Description

4 points	The written work is a thorough and accurate description of the main character.
3 points	The written work is a mostly complete or mostly accurate description of the main character.
2 points	The written work is a partially complete or partially accurate description of the main character.
1 point	The written work is attempted with few details or is mostly inaccurate.
0 points	The written work shows no comprehension of the main character.

Figure 2B: More Specific Scoring Rubric for Character Description

6 points	The written work thoroughly describes the main character, including: • the main character's name • a physical description of the main character (age, sex, clothes, hair color, skin color, and what the character wears) • a reasonable statement about the main character's personality (e.g., friendly) or motives (e.g., wants to get rich) • at least <u>two</u> examples of the main character's actions or dialogue that show his/her personality or motives.
5 points	The written work addresses all four expectations but one or two required details are missing from the physical description
4 points	The written work completely addresses the first three expectations but gives only <u>one</u> example from the text to show personality or motives
3 points	The written work addresses all four expectations but many details are missing from the physical description OR no examples are given to show the character's personality or motives.
2 points	The written work addresses the first three expectations but many details are missing from the physical description OR the work completely addresses the first two expectations
1 point	The written work addresses the first two expectations but many details are missing OR the work gives a partial physical description of the character.
0 points	The written work gives only the name of the main character OR is illegible OR is off task OR shows no comprehension of the text.

In addition to clear scoring rules, teachers are likely to be more consistent if they have examples of previous students' work to show students what work at each performance level looks like.

When teachers ask students to do the same type of work at different times and in different contexts or when teachers have the same expectations for different short-answer or essay questions on a test, reliability is enhanced when the same scoring rule is used each time. For example, suppose that, at the end of each science investigation, a science teacher asks students to write a paragraph in which they summarize the results of the investigation and explain how the results relate to the initial research question and the scientific theory from which the research question was drawn. The reliability of students' scores for 'drawing conclusions' depends on whether the teacher applies the same scoring rule consistently over time and across students.

Summary

The validity and reliability standards presented here are written for classroom teachers to help them evaluate their assessment tools, processes, events, and decisions. Research has shown that teachers spend from 30-70 percent of their time engaged in assessment processes (Stiggins, Faires-Conklin, & Bridgeford, 1986). The assessment processes, events and decisions are not neutral aspects of the classroom environment. They have important effects on students' self-concepts and their conceptions of school and the subjects taught in school. Therefore, it is essential that teachers learn how to evaluate published assessment tools and their own tools and processes so that they can make valid and reliable judgments about whether students have learned the targeted knowledge and skills. With high quality assessment tools and ethical assessment processes, teachers are more likely to make accurate decisions about how to support their students achievement.

References

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology, 78* (3), 210–216.

Covington, M. V., & Omelich, C. L. (1979). Effort: the double-edged sword in school achievement. *Journal of Educational Psychology, 71* (1), 169–182.

Cronbach, L. (1970). *Essentials of Psychological Testing*, Third Edition. New York: Harper & Rowe, Publishers

Messick, S. (1989). Validity. In *Educational Measurement*, Robert Linn (Ed.). Washington, DC: American Council on Education.

Stiggins, R. J., Faires-Conklin, N., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5* (2), 5-17.

Taylor, C. S. & Nolen, S. B. (1996). What does the psychometrician's classroom look like? *Educational Policy Analysis Archives*, v4, n17.

Taylor, C. S. & Nolen, S. B. (2005). *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms.* Columbus, OH: Pearson-Merrill-Prentice Hall.

Appendix A
Certificate of Academic Achievement (CAA) Options Advisory Committee Members

Linda Dobbs, Assistant Superintendent, ESD 189, Mt. Vernon, Washington

Deborah Gonzalez, Executive Director for Learning & Teaching, Puget Sound ESD, Highline, Washington

Gil Mendoza, Executive Director of Grants Management, Tacoma School District, Tacoma, Washington

Barbara Plake, Professor Emeritus, University of Nebraska-Lincoln

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

Edward Wiley, Professor, University of Colorado-Boulder

Appendix B
National Technical Advisory Committee for Assessment

Patricia Almond, University of Oregon, Eugene, Oregon

Peter Behuniac, University of Connecticut, Hartford, Connecticut

Richard Duran, Professor, California State University, Santa Barbara, California

George Englehard, Professor, Emory University, Atlanta, Georgia

Robert Linn, Professor Emeritus, University of Colorado-Boulder, UCLA-CRESST

William Mehrens, Professor Emeritus, Michigan State University, East Lansing, Michigan

Edys Quelmalz, Associate Director of the Center for Technology in Learning, Stanford Research Institute International, Palo Alto, California.

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

The Application of Professionally Accepted Standards for Reliability and Validity to the Collection of Evidence

Prepared by C. Taylor and J. Willhoft

August 14, 2006

One of the three options that have been legislated as alternatives to performance on the Washington Assessment of Student Learning (WASL) as a means for students to earn a Certificate of Academic Achievement (CAA) is a collection of work samples, also referred to as the Collection of Evidence[1] (COE). Legislation requires that the guidelines and protocols for submission and the criteria used for scoring "meet professionally accepted standards for a valid and reliable measure of grade level expectations and the essential academic learning requirements." (citation)

The process recommended by OSPI to the State Board of Education (SBE) is that the standards shown in Tables 1A and 1B, from the ***Standards for Reliability and Validity of Classroom-Based Assessments,*** be reviewed and approved by the National Technical Advisory Committee (NTAC). NTAC approval will assure the SBE that the criteria for reliability and validity against which the COE will be judged meet "professionally accepted standards". The review and approval of these reliability and validity standards will take place in two stages. First, the CAA Options Advisory Committee, composed of national and local educators and assessment experts (See Appendix A) will review, refine (as needed), and approve the standards. These standards will then be submitted to the NTAC for their approval in August of 2006. Once the NTAC adopts a set of reliability and validity standards for the COE, the design features of the COE will be submitted for their review. The NTAC will be asked to reach consensus on the

[1] Collections of Evidence are subject specific (i.e., reading, mathematics, and writing) collections of classroom-based assessments or work samples for individual students that demonstrate comparable curriculum standards as those assessed by WASL.

alignment of design features of the COE that address the standards. That work will be completed in mid-August, and will be presented to the SBE at its August meeting.

Table 1A: Validity Standards for Classroom-based Assessments

Validity Standard 1: **Representation and Fidelity**	Do the knowledge and skills required by the assessments represent the breadth of knowledge and skills defined in the standards?
Validity Standard 2: **Cognitive Demands**	Do the assessment tools and processes require students to demonstrate the targeted knowledge and skills at a cognitive level specified in the standards?
Validity Standard 3: **Consistency Across Assessments**	Do different assessments of the same knowledge and skills elicit comparable work?
Validity Standard 4: **Alignment with Instruction**	Does assessment align with the content taught and the instructional methods used?
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	Do the assessment tools and processes provide an equal opportunity for individuals, regardless of group or setting, to demonstrate the targeted knowledge and skills?
Validity Standard 6: **Consequences of the Interpretation and Use of Assessment Results**	Are there negative consequences for students that could be prevented if assessment tools, processes, events, or decisions had been more valid?

Table 1B: Reliability Standards for Classroom-based Assessments

Reliability Standard 1: **Generalizability**	Is the work typical of what the student knows and is able to do in relation to the learning targets?
Reliability Standard 2: **Sufficiency of Evidence**	Is there sufficient evidence so that one can make a dependable judgment about what each student knows and is able to do in relation to the learning targets?
Reliability Standard 3: **Clarity of Directions and Expectations**	Do the assessment directions provide clear, unambiguous expectations so that students can dependably demonstrate what they know and are able to do in relation to the learning targets?
Reliability Standard 4: **Quality of Scoring**	Are the scoring rules and scoring processes systematic enough to ensure consistent evaluation over time and across diverse samples of student work that demonstrate the same learning targets?

Two sources served as source materials for the attached *Standards for Reliability and Validity of Classroom-Based Assessments*. The first source was the *Standards for Educational and Psychological Testing* developed jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). The fourth edition of these standards was published in 1999. This document is widely accepted within the community of measurement professionals as encompassing the standards to be met for the development, evaluation, and use of tests that are commercially-developed or are used in large scale public assessment systems. The second source was Taylor and Nolen (1996, 2005), in which the authors adapted the *Standards* for application to the classroom assessment context. This latter work was used as the basis for the standards presented in the *Standards for Reliability and Validity of Classroom-Based Assessments*.

Considerations in Applying these Standards to Collections of Evidence

The Collections of Evidence (COE) to be used for the CAA involve the use of classroom-based assessments in a large-scale assessment context. The COE process requires students to collect work samples from classroom assignments and organize this evidence for a large scale purpose. In this case, not all standards for the validity and reliability of classroom-based assessments can be fully addressed by design features of a large scale assessment program. Three validity standards and one reliability standard for classroom-based assessments have limited applicability in this large scale context.

Validity Standard 4 (Alignment with Instruction) can best be evaluated by the classroom teacher or the students who know whether instruction has prepared the students to demonstrate the knowledge and/or skills required by the assessments.

Validity Standard 6 (Consequences of the Interpretation and Use of Assessment Results) requires ongoing research related to validity standards 1-5 and the consquences of the COE for students. Consequences related to students' self-concepts, their conceptions of school and the subject disciplines, and their academic choices as a results of their classroom-based assessment experiences are beyond the scope of the COE. However, consequences related to the COE should be examined. Positive or negative consequences that arise from decisions made based on the collections are relevant to validity ONLY if these consequences are due to problems related to validity standards 1 through 5.

In addition, although it is possible to Enhance Fairness and Minimize Bias (Validity Standard 5) through careful selection of collections to use for scorer training, it is difficult to thoroughly assess Validity Standard 5 without more information about the students. As with Validity Standard 4 (Alignment with Instruction), only the classroom teacher and the students know

whether the features of the assessment tools or events allow students to demonstrate what they know andare able to do. It is possible, however, to ensure that the COE provides opportunities for all qualified students, to demonstrate their knowledge and skills. The guidelines for the COE can be evaluated for the degree to which they enhance fairness and minimize bias.

Finally, for Reliability Standard 2 (Clarity of Expectations) the protocols for the COE, and any subsequent training materials and directions for teachers and students can be evaluated for clarity of expectations. The clarity of directions for assignments can be evaluated only if directions for assignments are provided along with students' work samples. Finally, if students include tests as part of their collections, test questions can be evaluated for clarity.

Above and beyond issues of reliability and validity, a separate standard has been recommended by the CAA Options Advisory Committee to answer the question: "Are there unintended consequences, for students, schools, and districts, of using the assessment system to make decisions about students?" This standard is important to consider when collections of evidence are used to judge students' proficiency in relation to the standards. Examples of unintended consequencies might include poor WASL performance due to the COE option (which would have implications for a school, district, or state AYP), a narrowing of the curriculum to a limited number of assessment tasks, repeated practice with a single task until the student prepares a proficient performance, or other unintended consequences. Studies should be planned to determine whether there are unintended negative consequences of the COE.

In Tables 2A through 2G of this document, the design features of the COE are more fully detailed. Tables 3A and 3B of this document present the approved links between the design features of the COE and the professionally accepted standards for reliability and validity from *Standards for Reliability and Validity of Classroom-Based Assessments*.

Table 2A:
Protocols – Directions to the COE users to indicate the types of evidence needed for each subject area

Writing Protocol

There are to be 5 to 8 written samples that together demonstrate proficiency in idea/development, organization, style, and the use of conventions. More work samples do not equate to a better score: Carefully selected work samples is a better indicator. Work samples should be written in blue or black ink or word processed.

- At least one expository or persuasive on-demand essay, timed and supervised in class
- At least two expository non-timed essays
- At least two persuasive non-timed essays
- 3 work samples (including the on-demand sample) may not include any adult assistance beyond setting the prompt and public expectations for an effective paper.
- Other work samples may include drafts read with teacher input and general comments (e.g., "You need to check for spelling errors." or "You need to rework your conclusion to wrap up your writing and give your reader something to think about.").

Reading Protocol

Work samples that cover all six strands that are assessed on the Reading WASL.

- A minimum of 8 and a maximum of 12 work samples from a classroom setting or a teacher-approved independent setting. Half of the work samples must represent responses to literary text and half of the samples must represent responses to informational text.
- All texts used in the work samples must meet high school expectations for rigor of reading material. The work samples must be comparable in rigor in skill and content to the High School Reading WASL.
- Work samples may feature work completed in other content areas—science, social studies, CTE coursework, etc. However, they must still address the literary or the informational strands listed above.
- One work sample must be a literary analysis paper of a significant piece of text—short story, narrative essay, novel, etc. that includes a demonstration of more than one literary strand.
- One work sample must be a research paper that includes at least two texts used for research purposes. Examples of this type of reading responses include: magazine or newspaper article analysis, analysis of historical events or scientific procedures, etc. The work sample should demonstrate more than one informational strand.
- One work sample that must be completed in an "on-demand" setting where students are provided an assignment to complete within a class period and without any teacher or peer assistance.

Mathematics Protocol

There must be 8 to 12 work samples.

- A variety of work samples such as projects, assignments, or exams
- Work samples of moderate or high complexity to ensure moderate or high level cognitive demands of the student
- At least two high school level work samples that and can be scored for an entire target from a strand of EALR 1:
- At least two high school level work samples can be scored for an entire target* from a strand of EALRs 2 through 5:
- Work samples that combine a content strand from EALR 1 and a process strand from EALRs 2 through 5. Work samples for EALRs 2 through 5 must be distributed across EALR 1 content strands.
- Work samples you select for EALR 1 should be representative of multiple High School WASL Mathematics Test Specifications
- Work samples you select must combine at least one content strand from EALR 1 and at least one process strand from EALRs 2–5.

Work samples should be complex enough to demonstrate moderate to high level thinking skills.

Table 2B:

Sufficiency Review – Process used to determine that all of the WASL learning targets for a domain are included in the collection

Writing Protocol
In order to meet the sufficiency guidelines for successfully submitting a Writing Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Reading Protocol
In order to meet the sufficiency guidelines for successfully submitting a Reading Collection of Evidence, the student and teacher preparing the collection must comply with the COE guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored.
Mathematics Protocol
In order to meet the sufficiency guidelines for successfully submitting a Mathematics Collection of Evidence, the student and teacher preparing the collection must comply with the following guidelines. If the collection does not meet these guidelines in any capacity, the collection will not be scored

Table 2C:

Work Sample Documentation

Writing Protocol	*In the "Work Sample Documentation Form" teachers must provide documentation that the work sample demonstrates the state standards in writing. For each work sample, students must check one of the first three boxes on the form as well as the type of draft, process, and teacher-assisted for the work samples in the collection. The teacher must check that an "on-demand" essay is present in the collection. In the last box—teacher assistance—the student must describe what type of assistance he/she received beyond setting the prompt and the parameters of an effective paper.*
Reading Protocol	*In the "Work Sample Documentation Form" students and teachers must check all of the learning strands, both literary and informational. The student must provide of the titles of the texts must be provided to check the rigor of the readability of the texts. The student and the teacher must check each work sample to make sure that each sample addresses at least two strands. The student must identify which work sample is the short literary analysis paper and which is the short informational analysis paper. The teacher must check that an "on-demand" essay is present in the collection.*
Mathematics Protocol	*In the "Work Sample Documentation Form" students and teachers must check that all work samples address every high school content strand. Each work sample must address both a content strand and a process strand. Teachers must check that work samples meet the "rich problem" and high school level mathematics expectation. Students must check that each column and row have two entries. There must be an "on-demand" check*

Table 2D:

Scoring rules used to evaluate the collections – Performance criteria for the scoring rubrics used for each collection are given below along with an indication of the subject area EALRs and components within each EALR that are the focus of the performance criteria. Links to the EALRs are keys to authenticity validity.

Writing Criteria
Content, Organization & Style ➢ Has clear, focused main ideas or positions (EALR 1, Component 1) ➢ Elaborates by using reasons/arguments supported by well-chosen and specific details, examples, anecdotes, facts and/or statistics as evidence to support ideas or positions (EALR 1, Component 1) ➢ Includes information that is thoughtful and useful for the audience to know (EALR 1, Component 1) ➢ Organizes writing to make the best case to explain ideas or support positions (EALR 1, Component 2) ➢ Composes introductions that draw the reader into the main ideas or positions (EALR 1, Component 2) ➢ Writes conclusions that leave the reader with something to think about (EALR 1, Component 2) ➢ Organizes writing into effective, cohesive paragraphs (EALR 1, Component 2) ➢ Provides transitions which clearly serve to connect ideas (EALR 1, Component 2) ➢ Uses language effectively by exhibiting word choices that are effective and appropriate for intended audience, purpose, and form (EALR 1, Component 3) ➢ Writes (where appropriate) sentences or phrases that are varied in length and structure (EALR 1, Component 4) ➢ Provides the reader with a sense of the person behind the words (EALR 1, Component 5) Conventions ➢ Follows the rules of standard English [language] usage (EALR 1, Component 6) ➢ Spelling of commonly used words (EALR 1, Component 6) ➢ Capitalization (EALR 1, Component 6) ➢ Punctuation (EALR 1, Component 6) ➢ Exhibits the use of complete sentences except where purposeful phrases or clauses are used for effect (EALR 1, Component 6) ➢ Indicates paragraphs consistently (EALR 1, Component 6)
Reading Criteria
Comprehension of main ideas and details of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Identifies the main theme/main idea and uses evidence to demonstrate an overall understanding of the text (EALR 2, Component 1) ➢ Summarizes by providing an overarching statement about the text that connects to at least three events from the beginning, middle and end of text (EALR 2, Component 1) ➢ Infers and/or predicts about key elements of the text making connections with evidence (EALR 2, Component 1) ➢ Explains key vocabulary with both denotative and connotative definitions by linking them to the text (EALR 1, Component 2) Analysis, interpretation, & synthesis of literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Applies knowledge of key literary/informational elements to enhance and expand understanding of text (EALR 2, Component 2) ➢ Compares and contrasts ideas to explain concepts within or between text (EALR 2, Component 3) ➢ Analyzes text to explain the relationship between cause(s) and effect(s) and links it back to the theme or main idea (EALR 2, Component 2) Thinks critically about literary (EALR 3, Component 4) or informational (EALR 3, Component 1) text ➢ Evaluate author's/ text's purpose and/or in order to judge effectiveness on intended audience ➢ Evaluates reasoning of ideas / themes within the text and makes connections with evidence Synthesizes information beyond the text by making generalizations, drawing conclusions, or applying information to evaluate a new text or context

Table 2D (Continued)

Mathematics Criteria
Uses high school content knowledge and procedures (EALR 1) with supporting work in: ➢ Number Sense (EALR 1, Component 1) ➢ Measurement (EALR 1, Component 2) ➢ Geometric Sense (EALR 1, Component 3) ➢ Probability & Statistics (EALR 1, Component 4) ➢ Algebraic Sense (EALR 1, Component 5) Solves Problems (EALR 2) ➢ Applies one or more strategies that lead to the answer (EALR 2, Component 2) ➢ Determines the answer to the problem (EALR 2, Component 3) Reasons Logically (EALR 3) ➢ Justifies conclusions, results, and/or answers by addressing the conditions and/or constraints in the problem Communicates Understanding (EALR 4) ➢ Gathers, represents, and/or shares mathematical information using clear mathematical language and organization Makes Connections (EALR 5) ➢ Uses and relates different mathematical models and representations of the same situation using clear mathematical language and organization (EALR 5, Components 1 and 2)

Table 2E

Range-Finding – The process of selecting exemplary collections to represent different performance levels

All Content Areas
Steps in the range-finding process ➢ Select a range of collections to serve as potential anchors for the rubrics during scoring training, practice collections to be used for practice during scoring training, and validity collections to be randomly inserted into scoring process to ensure adherence to scoring rubrics over time ➢ Ensure that all selected collections have met sufficiency criteria ➢ Discuss scoring rubrics ➢ Apply scoring rubrics to selected collections ➢ Discuss applied scores ➢ Adjust scoring rubrics and/or scores, if needed, based on collections ➢ Assign final scores to anchor collections ➢ Assign final scores to practice collections ➢ Assign final scores to validity collections

Table 2F

Scoring Training – The process of training scorers to apply scoring rubrics consistently using anchor collections to anchor rubrics

All Content Areas
Steps in the training process ➢ Review and discuss rubrics ➢ Review and discuss anchor collections ➢ Score practice collections ➢ Discuss assigned scores; work toward consensus with pre-assigned scores ➢ Score second practice collections ➢ Discuss assigned scores; work toward consensus with pre-assigned scores ➢ Scorers must qualify by meeting a criterion of exact agreement with pre-assigned scores

Table 2G

Table Scoring Process – The process of assigning scores to collections

All Content Areas
Steps in the scoring process ➢ Scorers assign scores ➢ Collections are randomly assigned to a second scorer (inter-rater agreement) ➢ Randomly selected collections are rescored by a table leader (supervisor) ➢ Validity collections are given to scorers randomly ➢ Scorers who drift from scoring rubrics are retrained as necessary

The next two tables, Tables 3A and 3B, link each of the Validity and Reliability standards COE design features.

Table 3A: Design Features of COE that Address Validity Standards

Validity Standard	Feature of COE Addressing Standard
Validity Standard 1: **Representation and Fidelity**	➢ Protocols for Reading, Writing, and Mathematics ➢ Sufficiency Review ➢ Scoring Rules ➢ Range-finding ➢ Scoring Training ➢ Scoring Process
Validity Standard 2: **Cognitive Demands**	➢ Protocols for Reading, Writing, and Mathematics
Validity Standard 3: **Consistency Across Assessments**	➢ Range-finding
Validity Standard 4: **Alignment with Instruction**	➢ Student self-report??
Validity Standard 5: **Enhancing Fairness and Minimizing Bias**	➢ Range-finding ➢ Scoring Training ➢ Scoring Process
Validity Standard 6: **Consequences of the Interpretation and Use of Assessment Results**	➢ Ongoing validity studies for the COE

Table 3B: Design Features of COE that Address Reliability Standards

Reliability Standard	Feature of COE Addressing Standard
Reliability Standard 1: **Generalizability**	➢ Protocols for Reading, Writing, and Mathematics
Reliability Standard 2: **Sufficiency of Evidence**	➢ Sufficiency Review ➢ Work Sample Documentation Form
Reliability Standard 3: **Clarity of Directions and Expectations**	➢ Protocols for Reading, Writing, and Mathematics ➢ Work Sample Documentation Directions ➢ Work Sample Sign-off Form
Reliability Standard 4: **Quality of Scoring**	➢ Scoring Rules ➢ Range-finding ➢ Scoring Training ➢ Scoring Process

References

Taylor, C. S. & Nolen, S. B. (1996). What does the psychometrician's classroom look like? *Educational Policy Analysis Archives*, v4, n17.

Taylor, C. S. & Nolen, S. B. (2005). *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms.* Columbus, OH: Pearson-Merrill-Prentice Hall.

Appendix A

Certificate of Academic Achievement (CAA) Options Advisory Committee Members

Linda Dobbs, Assistant Superintendent, ESD 189, Mt. Vernon, Washington

Deborah Gonzalez, Executive Director for Learning & Teaching, Puget Sound ESD, Highline, Washington

Gil Mendoza, Executive Director of Grants Management, Tacoma School District, Tacoma, Washington

Barbara Plake, Professor Emeritus, University of Nebraska-Lincoln

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington

Edward Wiley, Professor, University of Colorado-Boulder

Appendix B

National Technical Advisory Committee for Assessment

Patricia Almond, University of Oregon, Eugene, Oregon

Peter Behuniac, University of Connecticut, Hartford, Connecticut

Richard Duran, Professor, California State University, Santa Barbara, California

George Englehard, Professor, Emory University, Atlanta, Georgia

Robert Linn, Professor Emeritus, University of Colorado-Boulder, UCLA-CRESST

William Mehrens, Professor Emeritus, Michigan State University, East Lansing, Michigan

Edys Quelmalz, Associate Director of the Center for Technology in Learning, Stanford Research Institute International, Palo Alto, California.

Joseph Ryan, Professor Emeritus, Arizona State University – West

Catherine Taylor, Associate Professor, University of Washington