National Alternate
Assessment Center

Scoring Alternate Assessments Based on Alternate Achievement Standards:

A Proposed Typology of AA-AAS Scoring Practices

Mari Quenemoen
Marianne Perie
Jacqui Kearns

Executive Summary

As states develop AA-AAS programs, they should consider how the scoring design may articulate certain values and assumptions about the ways in which students with significant cognitive disabilities learn and how best to observe and measure their learning. This report touches on some implications of each scoring design for assessment validity, as well as the reasons why states may choose to incorporate various criteria into the scoring design. The typology of scoring designs proposed in this report, and the discussion of implications and policy values, should represent a starting point for a conversation on AA-AAS scoring designs that employs shared language and shared understandings about how best to measure what students with significant disabilities know and can do.

The scoring design types proposed in this report are *accuracy approaches*, *multi-dimensional matrices*, and *scaffolded scales*.

1. *Accuracy* approaches either simply score each item as correct/incorrect, or assign a one-dimensional score that can include partial credit for responses that are partially correct. Many item-based tests, and some portfolios and rating scales, are scored with accuracy approaches.

2. *Multi-dimensional matrices* are scoring designs that assign separate points (beyond correct/incorrect) to two or more dimensions. Many states have developed AA-AAS portfolios to be scored with multi-dimensional matrices.

3. *Scaffolded scales* combine two dimensions into a uni-linear scale according to a scripted protocol of item administration that states often call "scaffolding." The scaffolding protocol usually instructs the test administrator to provide increasing levels of prompting or support until the student produces a correct answer or engages in the process. States tend to use scaffolded scales to score item-based assessments.

Accuracy approaches score only one dimension of proficiency: the quality or accuracy of a student's performance. Multi-dimensional matrices may assign separate maximum points to two or more dimensions, and scaffolded scales score two or three dimensions combined into a uni-linear scale matched to a scripted administration protocol.

This report also analyzes the kinds of criteria that are embedded into states' scoring designs, focusing specifically on the criteria in multi-dimensional matrices. *Student criteria* score how a student performs on each task/item; *item criteria* score for the quality of the task/item; *generalization criteria* score for the student's ability to perform a skill across multiple tasks or settings; and *system criteria* are designed to ensure that the educational program is providing the opportunities and supports the student needs to engage in his/her own learning. All 50 states' scoring designs include student criteria, and some scripted items on item-based tests may be pre-coded with item-criteria. Multi-dimensional matrices may include any combination of student and other criteria.

Quenemoen, M., Perie, M., & Kearns, J. (2010). *Scoring Alternate Assessments Based on Alternate Achievement Standards: A Proposed Typology of AA-AAS Scoring Practices.* Lexington, KY: National Alternate Assessment Center.

Acknowledgements

The authors would like to thank Rachel Quenemoen of the National Center on Educational Outcomes and Charlie Depascale and Brian Gong of the Center for Assessment for their expert review of this report and their helpful observations and suggestions. Thank you also to the many state staff members who provided helpful feedback and additional information. Any remaining factual or interpretive errors are solely the responsibility of the authors.

Table of Contents

List of Tables

List of Figures

Scoring Alternate Assessments Based on Alternate Achievement Standards:

A Proposed Typology of AA-AAS Scoring Practices

Mari Quenemoen, Marianne Perie, & Jacqui Kearns

Introduction

Since the reauthorization of the Individuals with Disabilities Education Act (IDEA) mandated more inclusionary testing for students with disabilities in 1997, states have been developing alternate assessments for students who previously might have been excluded from federal accountability systems. Federal regulations in 2003 provided more specific guidance on the responsibility of states to assess and report on the achievement of students with the most significant cognitive disabilities using alternate assessments based on alternate achievement standards (AA-AAS). Today, states administer a range of AA-AAS approaches and report the results to meet accountability requirements under the Elementary and Secondary Education Act (ESEA). However, there has been little consensus on how to categorize the different types of AA-AAS or which scoring designs best suit each assessment approach.

National surveys of state AA-AAS practices over the years have largely failed to capture the range of assessment approaches and scoring designs used by the 50 states. The National Center on Educational Outcomes (NCEO) has been monitoring states' alternate assessment practices since 1997, but warned in recent reports that the categories used to describe test approaches in the survey may be of limited utility. The 2005 survey stated that, "It may be that the traditional way of describing alternate assessment approaches is no longer the best because there is considerable overlap across approaches that states take" (Thompson, Johnstone, Thurlow, & Altman, 2005, p. 11), and the 2009 report maintained that approaches "defied easy categorization" (Altman et al., 2010, p. 20). In response to the confusion in the field about how to talk about AA-AAS approaches, researchers from the National Alternate Assessment Center (NAAC) developed a typology of test approaches based on a comprehensive review of states' AA-AAS materials. The three proposed AA-AAS approach types are 1) portfolios; 2) rating scales; and 3) item-based tests comprised of performance tasks, writing prompts, constructed-response items, or multiple-choice items (Quenemoen, Quenemoen, Kearns, & Kennedy, 2010). This proposed typology is intended to help clarify discussions of AA-AAS approaches by standardizing the vocabulary that states, stakeholders, and test developers use to describe and understand these assessments.

Even states that use the same AA-AAS approach, however, often use different methods to score student performance, which may include measures of accuracy, independence, or progress, as well as measures of item quality or system criteria. To date, the field has no widely agreed-upon typology of AA-AAS scoring approaches. Quenemoen, Thompson, and Thurlow (2003) analyzed the scoring criteria used by five states used to score AA-AAS, and examined the values and assumptions embedded in those criteria. They proposed drawing a distinction between criteria that evaluate the student, such as skill and progress, and those that evaluate the student's educational system, such as staff support and variety of settings (Quenemoen et al., 2003). The report provided examples for how five states used these criteria in various ways to score very different AA-AAS approaches, but did not propose a systematic way to name and understand each state's scoring design in relation to others.

Two recent national surveys also attempted to capture features of states' scoring programs. In 2009, SRI and Policy Associates International released the National Profile on Alternate Assessments based on Alternate Achievement Standards (NSAA) derived from a survey administered during the 2006-07 school year. The NSAA asked states which "elements of student performance" each state scores: 1) accuracy of student response; 2) ability to generalize across settings; 3) amount of independence; or 4) amount of progress. The NSAA report does not, however, attempt to describe how these scoring elements fit into states' AA-AAS scoring designs, i.e. how those elements are scored. The 2009 NCEO survey does attempt to capture states' scoring designs, asking states to choose from four designs to characterize how their state's AA-AAS is scored: 1) rubric; 2) points assigned on a rating scale; 3) number of items correct; or 4) reading rate or accuracy (Altman et al., 2010). The NCEO report also captures information about which elements, or dimensions, are scored on states' rubrics. However, the NCEO survey's scoring design options may have been too limited to fully capture the range of scoring designs across the fifty states, and the report did not show how scoring designs interact with assessment approaches.

Like the AA-AAS approach typology report (Quenemoen et al., 2010), this report proposes a typology of scoring approaches based on observed characteristics of the 50 states' AA-AAS, including the scoring design and the dimensions that are scored. The three main scoring design types proposed in this paper, *accuracy approaches, multi-dimensional matrices*, and *scaffolded scales*, should help policymakers, test developers, teachers, and parents communicate better about student performance on the AA-AAS by providing a shared language and shared understandings about scoring designs. This report characterizes the ways that points are assigned to items, student materials, or observed student performance, and it touches on some ways in which these points are aggregated into total assessment scores. The report does not analyze the policy implications of different methods of aggregating scores, or what those scores mean in terms of student proficiency. In the discussion, we do suggest possible implications of each scoring design for assessment validity. As states continue to improve assessment methods for this population of students, they should think carefully about how assessment approach interacts with scoring design to reflect values and assumptions about student learning and student performance.

Methods

Two researchers from NAAC, a federally funded research center, identified and reviewed states' online AA-AAS materials, including administration and technical manuals, training materials, and sample test items using both a manual search on states' Department of Education websites and keyword searches on Google. Broad features of scoring design and scored dimensions for each state were recorded and labeled. The researchers compared results, and, when inconsistent, re-analyzed the data until interrater agreement reached 100%.

After determining the scoring approach from each state, one researcher verified this information with an AA-AAS staff person from that state via email or telephone communication. To verify scoring approaches, the researcher asked states to respond to a short narrative describing the scoring design, including, where appropriate, the specific scaffolding protocol or the points assigned to matrix dimensions. All but three states verified their information or submitted corrections.[1] Then, the researchers qualitatively analyzed the scoring approaches and proposed a typology for categorizing them as described in this paper. Researchers collected data on scoring designs simultaneously with data on AA-AAS approaches for Quenemoen et al., 2010.

AA-AAS Scoring Designs: A Proposed Typology

Assessments for the general population of students are usually scored for one dimension only: the accuracy or quality of student responses, either in terms of correct/incorrect or with options for partial credit. About a third of states' AA-AAS are scored only for student accuracy or performance, including approaches such as mastery scales or options for partial credit. Under the proposed typology, scoring approaches that focus solely on correct/incorrect responses are called *accuracy approaches*. Unlike general assessments, however, many AA-AAS are also scored for a number of additional dimensions. *Multi-dimensional matrices* assign points independently to two or more dimensions, which may include student performance but also a number of other criteria such as independence of performance or complexity of task. *Scaffolded scales* also combine two or three dimensions, but employ a scripted protocol of item administration along a uni-linear scale that results in one score that can no longer be separated into distinct dimensions. The proposed three primary scoring types are as follows:

1) *Accuracy approaches* either simply score each item as correct/incorrect or assign a one-dimensional score that can include partial points for task completion. Many item-based assessments are scored with accuracy approaches, but a few portfolio and rating scale assessments are also scored using some variation of this one-dimensional accuracy approach.

2) *Multi-dimensional matrices* are scoring designs that assign separate points (beyond correct/incorrect) to two or more dimensions. Many portfolios, and some item-based assessments, are scored with multi-dimensional matrices.

[1] States that did not confirm their information were Montana, Nebraska, and New Mexico.

3) *Scaffolded scales* combine two or three dimensions into a uni-linear scale according to a scripted protocol of item administration that states often call "scaffolding." The scaffolding protocol usually instructs the test administrator to provide increasing levels of prompting or support until the student produces a correct answer or engages in the process. States tend to use scaffolded scales to score item-based assessments.

See Table 1 for generic examples of each scoring design. See Figure 1 for the number of states' AA-AAS that use each scoring design.[2]

Table 1. Examples of each scoring design

Example test question:

A verb is an action word. Which of the following words is a verb?

a. ball
b. run
c. shoe

Accuracy Approach:

The correct answer is "b. run." Score correct or incorrect.

Multi-dimensional Matrix:

Allow student to work on five tasks for each standard over the course of two months. For each standard, use the following matrix:

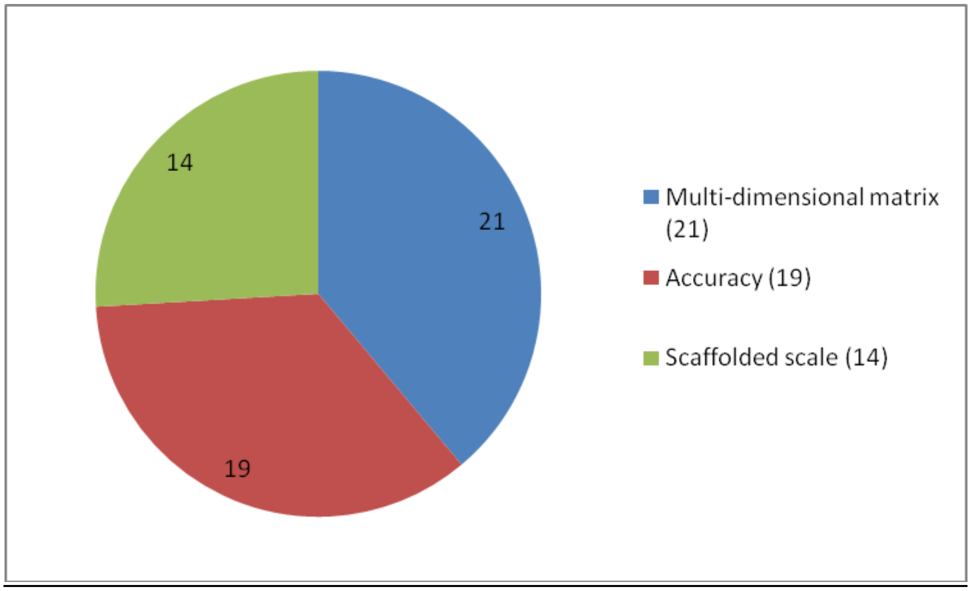
Accuracy	4: 100% accuracy across 5 tasks 3: 80% accuracy (4 correct tasks) 2: 60% accuracy (3 correct tasks) 1: 40% accuracy or less (2 or fewer correct tasks)
Level of Independence	4: Performs tasks with full independence 3: Requires assistance or prompting on some tasks 2: Requires assistance or prompting on most tasks 1: Cannot complete tasks without assistance or prompting

Scaffolded Scale:

If the student answers "b. run," score a "3" and continue to the next question.
If the student answers incorrectly, remove the incorrect answer and allow the student to try again.
If the student answers "b. run" the second time, score a "2" and continue to the next question.
If the student answers incorrectly, score a "1." If the student is unengaged, score a "0."

[2] The total exceeds 50 states because some states have more than one AA-AAS or use different scoring procedures within a single assessment, e.g. one to score multiple-choice items and another to score constructed-response items.

Figure 1. Number of AA-AAS scored with each approach



Accuracy Approaches

Nineteen states use a scoring design that scores student performance or accuracy only, using methods such as correct/incorrect, credit for partial accuracy or partial task completion, or mastery scales. Twelve state alternate assessments[3] score multiple-choice or constructed-response items only for accuracy or number of items correct. Several other assessments allow for partial accuracy, which may mean that a student's answer on a constructed-response item is partially correct, that a student partially completes a task, or that a certain percentage of trials were correct.

As for assessments for the general population of students, an accurate response also implies an independent response. Unless otherwise stated in the state's administration manual or scoring guide, accuracy approaches are predicated on the assumption that the student is performing without assistance (beyond allowable accommodations which, when implemented correctly, do not detract from independence). This distinguishes the accuracy approach from scaffolded scales, which guide the test administrator through scripted levels of prompting or support, and multidimensional matrices, which may or may not assign separate points for level of assistance. A violation of this assumption results in scores that may be difficult to interpret.

Different AA-AAS approach types also require scoring designs to be implemented differently. An item-based test may be scored with an accuracy approach consisting of a relatively simple determination of correct/incorrect or partial accuracy/partial completion. Rating scales, on the other hand, require the test administrator to assess the level of a student's performance based on observed behavior on the test or in the classroom, sometimes in addition to more structured

[3] This includes Michigan's highest level test, but not its two lower level tests. This number also includes tests that score some items only for accuracy, but may score other items with a scaffolded scale.

trials. Additionally, a few portfolios use the accuracy approach by scoring only for accuracy or using a uni-dimensional performance scale. Two portfolios score for the percentage of accurate trials, and one uses a uni-dimensional scale to judge the degree to which a student's materials demonstrate skill and knowledge. Accuracy approaches that stretch the boundaries of the typology somewhat, like the uni-dimensional scale, are discussed in greater depth below. See Table 2 for examples of accuracy approaches applied to multiple AA-AAS approach types.

Table 2. Examples of accuracy approaches

State	AA-AAS Approach	Scoring details
IA	Rating scale	The teacher scores for percent accuracy on the most recent trial for each skill, or marks the skills that were "already mastered," "not taught," or "fully prompted."
KS	Portfolio	1-5 scale according to degree of accuracy across the five trials.
LA	Item-based (multiple choice and performance tasks)	Each performance task is scored on a 0 to 2 point or a 0 to 1 point scale, according to an item-specific rubric. Two-point tasks allow the possibility of a partially correct response.
MD	Portfolio	Students must perform under 50% at baseline, and must perform with 80-100% accuracy for an objective to be scored "mastered."
NE	Item-based (multiple choice)	Scored right/wrong.

Multi-dimensional Matrices

While accuracy approaches only measure one dimension, multi-dimensional matrices always assign points to at least two dimensions. Twenty-one states use multi-dimensional matrices to score their assessments. Multi-dimensional matrices are usually used to score portfolios, but can also apply to item-based tests.[4]

Across the 21 AA-AAS matrices analyzed, 15 different scoring dimensions were identified (see Figure 2). States' matrices include from 2 to 6 dimensions, with an average of 4. Seventeen of these multi-dimensional matrices include "accuracy" or "performance" dimensions, and 12 include "level of independence" or "level of assistance." See Figure 2 for the distribution of dimensions across states' AA-AAS scoring matrices. Certain dimensions can be pre-requisites

[4] See the Appendix for more information about states' multi-dimensional matrices.

for scoring (e.g. alignment to standards), others are included in a matrix with specific point values assigned, and still others are aggregate dimensions that include several elements, as specified in a state’s scoring guide (e.g. choice, self-evaluation, and interactions could be combined as one dimension called “self-determination”). See Table 3 for an example of a multi-dimensional matrix that makes a dimension a pre-requisite for scoring. In this example, Alabama’s matrix requires that the materials submitted for each standard be aligned to that standard. Materials that are not aligned will not be considered for scoring. See Table 4 for an example of a matrix that combines several elements into an aggregate dimension. In this example, Delaware’s matrix combines progress, appropriateness, and supports into a dimension called “Activity,” and self-determination, self-evaluation, and choice into a dimension called “Self-Determination.” Delaware’s scoring guide explains how to interpret each dimension.

Figure 2. Matrix dimensions

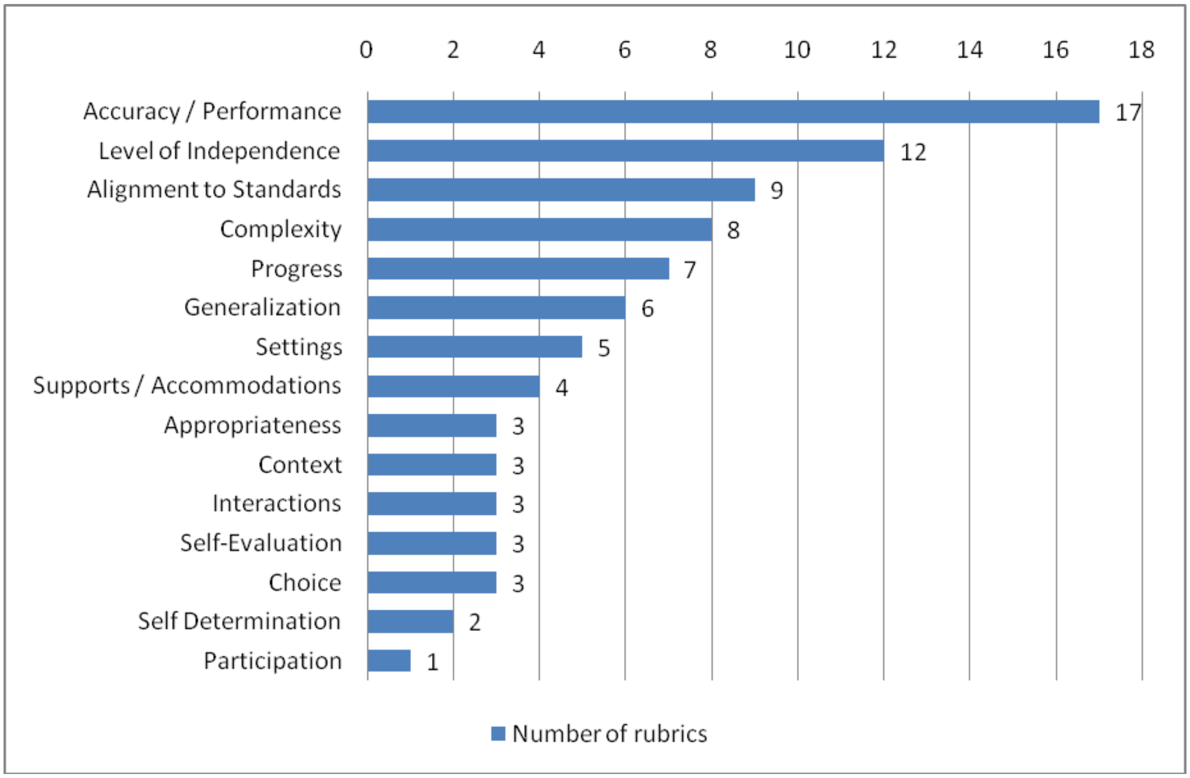


Table 3. Example of a multi-dimensional matrix that makes one dimension a pre-requisite for scoring[5]

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice
AL	Unstructured Portfolio	Alignment to the extended content standard (non-alignment makes the evidence unscorable), complexity (4), level of assistance (3), and mastery of content (3).	3	3		pre-req	4										

Table 4. Example of a multi-dimensional matrix that combines several elements into each dimension

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice
DE	Structured Portfolio	The "activity" element (5) combines age appropriateness, using a schedule, using supports, and the inclusion of a progress update. "Self-determination" (5) combines choice, planning, self-monitoring, and feedback. Other elements are settings (5) and interactions (5).			Activity (5)			Activity (5)			5	5	Self-Determination (5)	Self-Determination (5)	Activity (5)		Self-Determination (5)

Four Categories of Matrix Criteria

To clarify what kinds of criteria are embedded in various multi-dimensional matrices, and to understand how they interact in various combinations, we have grouped all of the dimensions used by states into four categories: student, item, generalization, and system criteria. *Student criteria* score how a student performs on each task/item; *item criteria* score for the quality of the task/item; *generalization criteria* score for the student's ability to perform a skill across multiple tasks or settings; and *system criteria* are designed to ensure that the educational program is providing the opportunities and supports the student needs to engage in his/her own learning.[6] Table 5 shows which dimensions fall under each criterion category.

[5] Numbers in each cell represent the number of points assigned to each dimension on the matrix. "Pre-req" is short for pre-requisite and indicates that the dimension determines whether the material is eligible to be scored.

[6] This categorization schema builds on the work of Quenemoen, Thompson, & Thurlow (2003).

Table 5. Four categories of matrix criteria

Student Criteria: Performance / Accuracy Level of Independence[7] Progress[8]	Item Criteria: Alignment to Standards Complexity Appropriateness[9] Context[10]
Generalization Criteria: Generalization Settings	System Criteria: Self-Determination[11] Interactions[12] Self-Evaluation Participation Choice Supports/Accommodations

Most matrices include dimensions from at least two of these categories, and all of them score at least one student criterion. Of the 21 state scoring matrices analyzed, 18 include both student and item criteria, and of those, 9 are comprised *only* of student and item criteria (see Figure 3). An example of a matrix with student and item criteria is one that measures accuracy/performance (student), level of independence (student), and alignment to standards (item); another is one that measures progress (student) and complexity (item). The second most frequent combination of criteria is a matrix with dimensions from all four categories: student, item, generalization, and system criteria. See the Appendix for more information about the combination of dimensions on each state's multi-dimensional matrix. In the Appendix and in the matrix tables in this report, the four categories of matrix criteria are color-coded: student criteria in white, item criteria in green, generalization criteria in blue, and system criteria in pink.

[7] Level of independence can also be called level of assistance, and measures how much teacher support (beyond accommodations) the student needs to produce a response, including additional prompting, refocusing, or even hand-over-hand assistance. Each state may define the levels of independence/assistance differently.

[8] Most states define progress as change in performance after established baseline.

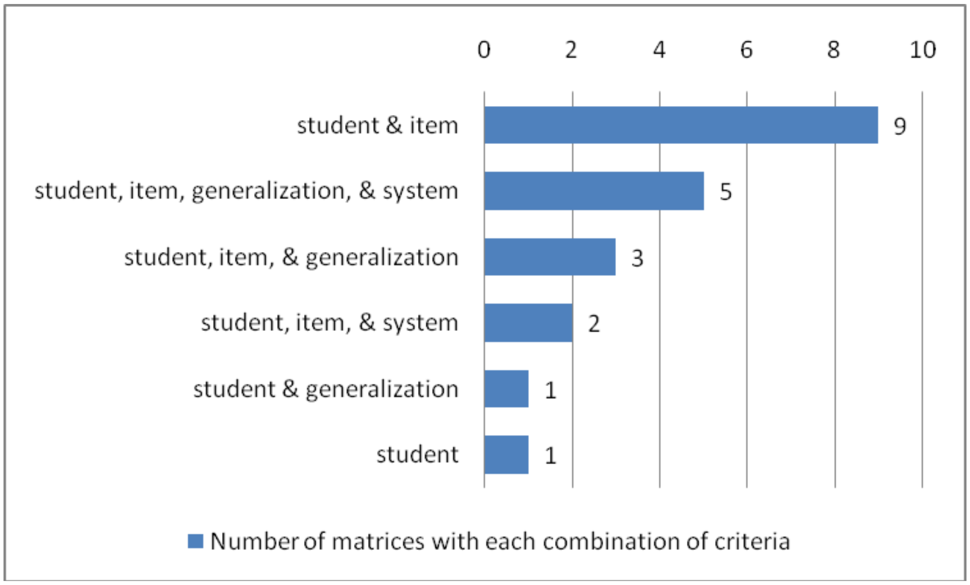
[9] Appropriateness may mean age-appropriateness or appropriate challenge.

[10] Context refers to the use of age-appropriate and meaningful tasks/materials.

[11] Self-determination can mean a variety of things, but often implies other factors such as self-evaluation, participation, and choice.

[12] Interactions may mean interactions with peers or with others in the community.

Figure 3. Number of multi-dimensional matrices with each combination of criteria



Every state's multi-dimensional matrix includes at least one student criterion. While accuracy approaches always assume independent student performance (unless otherwise stated), the performance/accuracy dimension on multi-dimensional matrices is often paired with the level of independence dimension, implying that a student's performance must be understood in the context of the level of assistance provided by the test administrator. When the matrix does not include level of independence, we can assume that the student must perform independently with or without approved accommodations. Whether paired with independence or not, the performance/accuracy dimension either refers to accuracy on individual tasks or across structured trials, or level of performance on a one-dimensional scale. Some matrices use a progress dimension instead of performance/accuracy, which usually implies a change in student performance after established baseline.

Item criteria may help add standardization to the assessment by controlling for some of the variance in teacher-designed tasks. Since teachers often design or modify the tasks that comprise portfolio assessments, this approach relies on a high degree of teacher judgment over what constitutes an appropriately challenging task as well as strong knowledge of the subject matter. Such reliance presents a measurement challenge, since a high score on a task that is insufficiently challenging to a student provides little useful information about that student's learning. Item-level dimensions can control somewhat for the quality of the task and allow for greater inferences from the student's score. For instance, Alabama uses a relatively unstructured portfolio approach, but its scoring matrix makes "alignment to standard" a pre-requisite for each teacher-designed task, and it assigns a relatively higher number of points to the "complexity" of the task. Not only must the task be aligned to the standard in order to be scored, but teachers have an incentive to design a task that is challenging and academically rigorous enough to score points for complexity.

Generalization criteria are meant to show that a student can perform a skill in multiple settings or across multiple tasks. Academic instruction for students with significant cognitive disabilities may include rote learning and repeated drills on discrete skills, and many special education experts believe that these students may benefit from dedicated instruction on generalizing those skills to other tasks or other settings (Browder & Cooper-Duffy, 2003). Generalization criteria also provide an incentive to teachers to vary the ways in which students practice and perform new skills.

System-level criteria are designed not to measure student performance but to hold schools accountable for providing appropriate access and supports to allow a student to engage in the learning process. These dimensions are usually weighted less heavily in comparison with student and item criteria, and can act as a "reminder" or a "check" for teachers. Some AA-AAS now collect these data for internal program monitoring and improvement purposes, but do not count system criteria points toward a student's final performance score. The 2009 NCEO survey of state practices shows that, compared with previous years, system criteria are being used with declining frequency (Altman et al., 2010).

Relative Weight

Even though many states use more than one dimension, they do not necessarily give them equal weight. Some states weight dimensions differently in the design by allocating a greater number of points to certain dimensions. Others give each dimension the same point value, but weight them after the fact in determining a total score (e.g., total score = 2 * accuracy + complexity). Oklahoma follows the former approach with a matrix that assigns different maximum point values to each dimension: level of independence (8), progress (5), accuracy (4), participation (1), alignment to standards (1), and age-appropriateness (1). In this case, a student who demonstrates a high level of independence but completes tasks with only partial accuracy may earn more points than a student who performs with complete accuracy but requires a high level of assistance from the teacher. However, the student must also demonstrate some measure of progress (as defined in Oklahoma's scoring guide) to achieve the highest possible score. Maine's item-based test allocates points as follows: appropriateness (pre-requisite), complexity (8), accuracy (3), and level of independence (3). In this case, each item must be age-appropriate in order for the item to be scored. Assuming that the task is age-appropriate, a student who completes a complex task with partial accuracy and independence may receive more points than a student who completes a very simple task with a high level of accuracy and independence. As this should make clear, the relative weight of each dimension can create various administrative and instructional incentives for teachers, and must be considered carefully. See Table 6 for representations of Oklahoma and Maine's matrices.

Table 6. Examples of multi-dimensional matrices with weighted dimensions prior to final performance score

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice
ME	Item-based (performance task, constructed-response, and multiple-choice items)	A task that is not age appropriate is unscorable. Level of complexity (8), accuracy (3), and level of assistance (3).	3	3			8	pre-req									
OK	Structured Portfolio	Level of independence (8), progress (5), accuracy (4), participation (1), connection to standard (1), and age-appropriateness (1).	4	8	5	1		1								1	

Some states use an extrinsic weighting method when calculating a final performance score. Arkansas assigns the same number of points (4) to each of its the three matrix dimensions, but in deriving the final performance score, the performance points are weighted by 4, context by 2 (context for this purpose indicates age-appropriateness), and level of assistance by 1, and settings points apply only once per content area (see Table 7). For example, if a student performs with partial accuracy (2 of 4 - weighted x4), on an age-appropriate task (4 of 4 - weighted x2), with full independence (4 of 4 - weighted x1), that student would receive a final performance score of 8 + 8 + 4 = 20. On the other hand, if a student performs with complete accuracy (4 of 4 - weighted x4), on an age-appropriate task (4 of 4 - weighted x2), with a relatively low level of independence (2 of 4 - weighted x1), then the student would receive a final performance score of 16 + 8 + 2 = 26 (settings points would be applied once per academic content area, rather than once per portfolio entry). Thus, a student who performs a task accurately with a high level of teacher assistance may earn a higher final score than a student who performs a task with less accuracy but fully independently.

Likewise, Mississippi scores its portfolio using a matrix with two dimensions that are each worth up to 4 points. To calculate a total score for each task, they double the performance score and triple the complexity score before summing them together. (Total score = 3*Complexity + 2*Performance.) Full independence is required and stated clearly in the administration and scoring guidelines. Therefore, a portfolio will earn more points if it contains more complex tasks that a student performs with only partial accuracy than if it contains relatively easy tasks that a student performs with complete accuracy. If teachers understand that certain dimensions will be weighted for the final performance score, this method may provide the same kind of administrative and instructional incentives as weighting dimensions prior to calculating the final performance score.

Table 7. Example of a multi-dimensional matrix that weights the final performance score

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice
AR	Unstructured Portfolio	Each dimension can be scored up to 4 points, but is weighted as follows: performance x 4, context x 2, level of assistance x 1. Settings is scored only once for each content area.	4 (x4)	4 (x1)					4 (x2)		4 (scored once)						

Scaffolded Scales

Twelve states use a scale that combines two or three dimensions into a uni-linear scale. Rather than assign separate points to each dimension and then combine them into a total score, as a multi-dimensional matrix does, scaffolded scales employ a scripted system of test administration that results in a single score. This scripted system of administration is often called "scaffolding" in AA-AAS test administration manuals. The term "scaffolding" as it is used for scoring AA-AAS should not be confused with scaffolding for instructional purposes. Teachers use instructional scaffolding to teach skills and concepts by gradually removing levels of support until a student can perform the skill or demonstrate the concept independently. In terms of scoring, teachers *add* levels of support according to a scripted scaffolding protocol.

Most current scaffolded scales use a scripted protocol for increased levels of prompting or support, resulting in a score that combines a student's performance/accuracy with the level of support the administrator provides for each item. For example, if a student does not answer a multiple-choice item correctly, the test administrator may be instructed to remove the incorrect option the student selected and allow the student to try again. A correct answer after this distracter is removed may receive a "3" instead of a perfect "4," for example. Some scoring systems grant at least one point if a student answers correctly after the administrator has told him/her the right answer or delivered hand-over-hand assistance, and others grant a point if the student answers incorrectly every time but engages in the process. Table 8 provides examples of different scaffolded scales from Minnesota and Colorado.

Table 8. Examples of scaffolded scales using different protocols

Minnesota Test of Academic Skills Scoring Guide[13]

3	Correct Response	The student responds correctly without assistance.
2	Correct Response with Additional Support	The student responds correctly to the task after the test administrator has provided additional support as indicated in the task script
1	Incorrect Response	The student responds incorrectly to the task after the test administrator has provided additional support as indicated in the script.
0	No Attempt or No Response	The student does not respond to the task or the student's response is unrelated to the task.

Colorado Scoring Rubric for Constructed-Response Item Types

Total Score	Content Score	Level of Independence
6	Correct	Level 4: INDEPENDENT - Performs task without assistance
5	Partially Correct/Some Error	Level 4: INDEPENDENT - Performs task without assistance
4	Correct	Level 3: PARTIAL - Partial physical, verbal, or gestural prompt
3	Partially Correct/Some Error	Level 3: PARTIAL - Partial physical, verbal, or gestural prompt
2	Correct	Level 2: LIMITED - Full physical prompt
1	Partially Correct/Some Error	Level 2: LIMITED - Full physical prompt

Further coded:*
4 – Independent and incorrect
3 – Partial and incorrect
2 – Limited and incorrect
0 - Incorrect or No response

* Incorrect answers are all scored as zero, but coded as follows.

As the Minnesota and Colorado examples show, scaffolded scales encompass varied approaches to student performance and engagement. Most scaffolded scales, like Minnesota's, allow the test administrator to provide additional prompting and support even after a student has answered incorrectly. The protocol may allow a student to "try again" until the student produces a correct answer or the protocol instructs the administrator to stop. In contrast, Colorado's scale assigns a "zero" for any incorrect answer and allows for additional prompting and support only if the student is unengaged or unresponsive.

Nearly all scaffolded scales combine accuracy and level of assistance, but Florida uses a variation of a scaffolded scale that combines accuracy with level of assistance and level of complexity. The test administrator begins with a multiple-choice item of lowest complexity. If the student responds incorrectly, the administrator removes the incorrect answer and allows the

[13] Provided in *Task Administration Manual: 2010 Minnesota Test of Academic Skills (MTAS),* retrieved August 12, 2010 from http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/Alternate/Alternate_Manuals_Directions/index.html.

student to respond again. If the student responds incorrectly again, the administrator removes the second incorrect answer and allows the student to respond again. A correct answer after the second distractor is removed is scored a 1, and an incorrect answer or no response a 0. On the other hand, if the student answers the first lowest level item correctly, the test administrator then presents a more challenging item. If the student responds incorrectly to the second item, the administrator scores a 3. But if the student responds correctly to the second item, the administrator administers yet a more complex item. An incorrect answer to this item is scored a 6, and a correct answer is scored a 9. See Table 9 for details about Florida's AA-AAS.

Table 9. Florida's scaffolded scale combining accuracy, level of assistance, and complexity

FL	Item-based (multiple choice). Each question is written at 3 levels of complexity.	Students progress through three levels of complexity per item in a grade level content based assessment (starting at Participatory). Possible item scores are 0, 1, 2, 3, 6, or 9 based on the highest level of complexity (for 3, 6, or 9 points) or level of support (for 2 or 1 point) at which a student provides an accurate response.

Variations and Gray Areas

Most states only use one design to score their AA-AAS. A few states that use multiple types of items also use multiple scoring procedures. For instance, Arizona uses a scaffolded scale to score its performance tasks and constructed-response items, but scores multiple-choice items only for accuracy. Other states use different scoring approaches for students at different "levels," where "level" typically refers to level of communication. For instance, Mississippi uses a performance dimension for students who communicate symbolically and a progress dimension for students who do not use symbol language, in addition to a complexity dimension for all students. Alaska does not sort students into pre-test levels, but uses a mechanism to shift students into lower-level items that are scored differently. If a student scores zero on three consecutive regular items, the test administrator begins to administer Expanded Level of Support (ELOS) items, which are rated only for level of assistance. These examples represent various ways scoring designs can be modified, but they do not change the typology proposed in this paper.

A few other states' scoring designs stretch the bounds of the typology slightly, particularly the accuracy approach. Rating scales are all scored with some version of an accuracy approach, but four of them represent unique variations. Indiana's rating scale requires teachers to rate student performance along a continuum of skills, or a "performance thread." For each standard strand, teachers must identify how a student performs relative to a series of skills, from least to most complex. See Table 10 for an example of a performance thread on Indiana's AA-AAS. South Dakota's and Connecticut's rating scales use performance scales that include level of independence as a factor within performance levels, but do not use a scaffolded scale. These are the only examples of accuracy approaches that do not assume independent student performance. Hawaii's rating scale and Virginia's portfolio both measure only student performance, but use uni-dimensional performance scales based on observation or evidence or student performance.

For the purposes of this typology, all of these scoring designs are characterized as variations of accuracy approaches because of their primary focus on the quality of student performance.

Two other assessments categorized as using accuracy approaches also account for other factors to a much lesser extent. North Dakota uses an unusual system of scoring multiple-choice and constructed-response items for accuracy, but also providing a set of "secondary indicators," including settings, choice, planning, and supports that can add a small number of additional points to the student's final score. Utah also scores primarily for accuracy, but to achieve the top score, a student must also demonstrate a level of generalization (see Table 11). Because the focus on accuracy significantly outweighs the focus on other dimensions, this report categorizes these scoring approaches as accuracy approaches, though they do occupy a "gray area" between approaches. See Table 11 for more information about these scoring approaches.

Table 10. Example of Indiana's "performance thread": Grade 3 – 5 "number sense"[14]

Least complex

- → demonstrates awareness of the presence of objects
- → identifies more
- → uses numbers to compare
- → names and orders quantities
- → describes relationships between numbers and quantity
- → identifies numbers and quantity to 100
- → identifies numbers and quantity to 1000
- → compares numbers on a number line
- → compares parts and whole

Most complex

Table 11. Gray Areas

HI	Rating Scale	Each standard is rated by the teacher and a second rater as "non-existent," "emerging," "progressing," or "mastered."
ND	Item-based (multiple choice and constructed response)	Items are scored primarily for accuracy, but a set of "secondary indicators" can add additional points, including for settings, choice, planning, supports, and self-monitoring.
SD	Rating scale	A five point rating scale combines accuracy and level of assistance for each item. Additionally, teachers select one item from each indicator in reading and each content strand in math and science, and collect student evidence of performance on

[14] Retrieved February 5, 2010 from https://ican.doe.state.in.us/beta/tm.htm.

		those items, documenting at least 3 trials of each skill.
UT	Item-based (performance tasks)	Each task can be scored up to 4 levels: level 1 is minimal (no correct trials), level 2 is partial (1 correct trial), level 3 is sufficient (2 correct trials), and level 4 is substantial (3 correct trials, but also 3 activities/objects, 3 people, and 3 settings).
VA	Portfolio	Performance on each standard is scored 0 – 4 points according to evidence of student skill and knowledge.

Scoring Procedures and Reliability

The analysis of scoring design presented here examines the dimensions measured and the approach to scoring, including the number of points to be given to each dimension and weights applied after initial scoring. A second, equally vital aspect of the scoring process is the implementation of that design, including intended approach, reliability, and procedural fidelity. The validity of the interpretation of the test results is dependent not only on the method used for scoring but how well that method is applied. Particularly for scoring open-ended items and performance tasks, it is important to select qualified scorers and train them well to ensure both scoring accuracy on individual assessments and reliability across many assessments. Likewise, recalibrating the scorers during the process is important to ensure that scorers do not begin to "drift," that is, begin to score more rigorously or leniently over time. Furthermore, strong assessment programs often have two individuals score each assessment, or they use a second, independent scorer to rescore a certain percentage of assessments that have been randomly selected to calculate the reliability of the scoring process. These and other methods allow test developers to determine the accuracy with which each assessment is scored and calculate an overall reliability estimate of the scored results.

The AA-AAS faces the same issues of identifying and training potential scorers and monitoring scoring sessions for scorer drift, reliability, and accuracy. Scoring procedures for AA-AAS present unique challenges however, because scoring decisions are often embedded into the test administration. The test administrator, who is usually the student's teacher, may impact the score from the start simply by the selection of tasks. Scoring designs that include a "level of independence/assistance" dimension depend heavily on the judgment of the teacher both to provide distinct levels of support and to report that support accurately in terms of a score. Even scores that only score right/wrong assume independence and rely on the teacher to administer the test as intended. Evidence of procedural fidelity is desirable. For AA-AAS scored with a scaffolded scale, it is important to monitor the fidelity with which the scaffolding protocol is administered, and see that the assessment protocol procedures as well as the student responses are recorded accurately. Procedural fidelity for scoring both portfolios and item-based assessments can be monitored by direct observation by an external observer or through submission of videotapes. Recalibrating scorers using this model may present challenges, since procedural fidelity data are generally analyzed during field testing or after the assessment administration.

Finally, for states that use a rating scale approach based on direct classroom observation, collecting procedural fidelity and inter-rater reliability measures may be problematic. Connecticut requires that each student maintain a folder of supporting evidence that could verify the student's response on the assessment items. A random sample of student folders is collected as an auditing mechanism to verify that those individual samples were scored correctly and to make inferences about the reliability of all of the scores. In short, states must think carefully about setting up scoring implementation procedures that maximize scoring accuracy and reliability. A detailed discussion of implementation falls outside the scope of this paper, but must be considered an important aspect of technical quality.

Conclusion and Policy Implications

In summary, the three main approaches to AA-AAS scoring proposed in this report are (1) accuracy approaches, (2) scaffolded scales that combine performance/accuracy and level of assistance dimensions into a uni-linear scale through scripted scaffolding protocols, and (3) multi-dimensional matrices that independently assign points to at least two different dimensions. See Table 12 for an overview of the proposed typology.

Table 12. Proposed AA-AAS scoring typology

Scoring Design	Accuracy Approaches	Multi-dimensional Matrices	Scaffolded Scales
Criteria	Student criteria (accuracy only)	Student criteria Item criteria Generalization criteria System criteria	Student criteria (accuracy and level of independence)

The four scoring criteria categories proposed in this report (student, item, generalization, and system criteria) are clearly demonstrated in states' multi-dimensional matrices, while the other two approaches usually include only student criteria. However, other dimensions may be pre-coded during item development. For example, pre-scripted multiple-choice items may be pre-coded for alignment or complexity. Scaffolded scales combine two student criteria, namely student performance and level of assistance, but, again, item criteria may or may not be pre-coded for other dimensions. For teacher-designed tasks in portfolios, item criteria may need to be included on a scoring matrix to control for the quality of the task.

Scoring designs present technical and validity considerations, but they also communicate a state's policy values. The inclusion of item criteria such as complexity, alignment, or appropriateness on a multi-dimensional matrix may communicate a policy message that students in this population should work on tasks that are academically rigorous and challenging. System criteria may be used as a reminder to educators about the importance of student self-evaluation and choice. The use of generalization criteria may imply that policymakers want students to show that they are able to apply the skills they learn beyond a singular test setting. More research is needed to determine whether or not these scoring criteria actually produce the intended policy outcomes, as well as the validity challenges they pose.

Scoring approaches also reflect assumptions about how best to measure student achievement. Accuracy designs, though they may use different approaches, all reflect a belief that we can understand student achievement on these assessments solely by measuring accuracy of student performance. By contrast, many states' AA-AAS scoring approaches reflect an assumption that information on student performance or accuracy is not enough to understand what students in this particular population know and can do. Scaffolded scales represent a belief that some students in this population may not be able to show what they know and can do without levels of teacher intervention that are normally prohibited in general assessment practices. Some scaffolded scales embody the belief that a student must answer an item correctly in order to earn points toward proficiency, but may require extra prompting to engage with the item, while others reflect the belief that even after a student produces an incorrect response, the student should earn points if s/he can produce a correct response with additional support. What those points mean in terms of proficiency depends entirely on how the state sets its proficiency levels. But as Gong and Marion (2006) point out, providing different levels of support for each student may in fact change the "construct similarity" (p.12) of each item, making even a multiple-choice assessment less standardized and turning a single item into multiple items for some students but not all. Finally, multi-dimensional matrices also reflect an assumption that measuring accuracy for students who participate in AA-AAS provides insufficient data about these students' learning and that other dimensions of proficiency including supports and opportunity to learn are also important.

A strong AA-AAS should reflect the best available research about how students with significant cognitive disabilities learn and demonstrate knowledge and skill, but states can also emphasize policy values in the development of scoring approaches. The typology presented here is a starting point for developing shared understandings about AA-AAS scoring approaches to facilitate better communication between states, test developers, teachers, and parents about the policy values and validity challenges posed by each approach. Communicating and sharing the policy rationale behind scoring approaches prior to administering the assessment may help state policymakers convey to educators and parents what is expected from the student, and how best to measure the his/her learning.

References

Altman, J., Lazarus, S.S., Quenemoen, R.F., Kearns, J., Quenemoen, M. & Thurlow, M.L. (2010). *2009 Survey of the states: Accomplishments and new issues at the end of a decade of change.* Minneapolis MN: University of Minnesota, National Center on Educational Outcomes.

Browder, D., & Cooper-Duffy, K. (2003). Evidence-based practices for students with severe disabilities and the requirement for accountability in “No Child Left Behind.” *The Journal of Special Education*, *37*(3), 157-163.

Cameto, R., Knokey, A.-M., Nagle, K., Sanford, C., Blackorby, J., Sinclair, B., & Riley, D. (2009). *National profile on alternate assessments based on alternate achievement standards. A report from the national study on alternate assessments* (NCSER 2009-3014). Menlo Park, CA: SRI International.

Gong, B., & Marion, S. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities* (Synthesis Report 60). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kearns, J., Towles-Reeves, E., Kleinert, H., Kleinert, J., & Thomas, M. (in press). Characteristics and implications for students participating in alternate assessments based on alternate academic achievement standards. *The Journal of Special Education*, XX(X) xx-xx.

Quenemoen, M., Quenemoen, R., Kearns, J., & Kennedy, S. (in press). *A Proposed Typology for Characterizing States’ AA-AAS: Developing a Common Language to Describe These Assessments.* Lexington, KY: National Alternate Assessment Center.

Quenemoen, R., Thompson, S., & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria* (Synthesis Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., Thurlow, M. L., & Altman, J. R. (2005). *2005 State special education outcomes: Steps forward in a decade of change*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Appendix

1. States that use a multi-dimensional matrix[15]

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice	Combination of criteria
AL	Unstructured Portfolio	Alignment to the extended content standard (non-alignment makes the evidence unscorable), complexity (4), level of assistance (3), and mastery of content (3).	3	3		pre-req	4											student & item
AR	Unstructured Portfolio	Each dimension can be scored up to 4 points, but is weighted as follows: performance x 4, context x 2, level of assistance x 1. Settings is scored only once for each content area.	4 (x4)	4 (x1)					4 (x2)		4 (scored once)							student, item, & generalization
DE	Structured Portfolio	The "activity" element (5) combines age appropriateness, using a schedule, using supports, and the inclusion of a progress update. "Self-determination" (5) combines choice, planning, self-monitoring, and feedback. Other elements are settings (5) and interactions (5).			Activity (5)			Activity (5)			5	5	Self-Determination (5)	Self-Determination (5)	Activity (5)		Self-Determination (5)	student, item, generalization, & system

[15] Numbers in each cell represent the number of points possible for each matrix dimension per scored objective/strand, mirroring the scoring details text for each assessment.

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice	Combination of criteria
GA	Structured Portfolio	Context (4), achievement/progress (4), generalization (4), and fidelity to standard (3). "Generalization" includes settings and interactions.			4	3			4	Generalization (4)	Generalization (4)	Generalization (4)						student, item, generalization, & system
ID	Unstructured Portfolio	Accuracy (4), level of independence (4), and complexity/alignment (4).	4	4			4											student & item
KY	Portfolio	Complexity (4), supports (4), and performance accuracy (4) or progress (3). Students who communicate at the pre-symbolic level can score up to 3 points for a gain of at least 40 points over baseline performance. Dimension A students can score up to 4 points for 90 – 100% accuracy (taken from the highest scoring probe submitted for each assessment target along with one student work sample).	4 (Dimension A)		3 (Dimension B)		4								4			student, item, & system
ME	Item-based (performance task, constructed-response, and multiple-choice items)	A task that is not age appropriate is unscorable. Level of complexity (8), accuracy (3), and level of assistance (3).	3	3			8	pre-req										student & item

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice	Combination of criteria
MA	Structured Portfolio	Level of complexity (5), accuracy (4), independence (4), self-evaluation (4), and generalization (3).	4	4			5			3				4				student, item, generalization, & system
MS	Structured Portfolio	Students who are symbolic language learners use the "attainment" scoring guide, which measures accuracy on a 5 point scale. Pre-symbolic language users use the "progress" scoring guide, which measures progress on a 5 point scale. Each item is also assigned up to 5 points for level of complexity. All students must score below 40% at baseline.	5 (Attainment)		5 (Progress)		5 (both)											student & item
MO	Structured Portfolio	Skill performance (4), level of independence (4), and connection to standards (3). Performance and independence are scored 6 times throughout two collection periods for each indicator, resulting in an average score for each.	4	4		3												student & item
NH	Structured Portfolio	Evidence of progress (4), connection to general curriculum (4), supports (4), generalization (4), and self-determination (4). "Generalization" includes settings, and "self determination" includes evidence of choice, self-monitoring, planning, and self-evaluation.			4	4				generalization (4)	generalization (4)		self-determination (4)	self-determination (4)	4		self-determination (4)	student, item, generalization, & system

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice	Combination of criteria
NJ	Structured Portfolio	Complexity/link to indicator (4), level of independence (4), and accuracy (4). The student must perform each skill with less than 40% accuracy at baseline, and final accuracy score is based on the final activity.	4	4			4											student & item
NY	Structured Portfolio	Accuracy (4) and level of independence (4).	4	4														student
OH	Unstructured Portfolio	Instructional context (4) and performance/accuracy (3). "Instructional context" implies age-appropriateness. The student achievement score is calculated by multiplying performance by instructional context. The matrix also measures level of independence (4), and settings and interactions (4), but these do not count toward the achievement score.	3						4									student & item
OK	Structured Portfolio	Level of independence (8), progress (5), accuracy (4), participation (1), connection to standard (1), and age-appropriateness (1).	4	8	5	1		1								1		student, item, & system
RI	Structured Portfolio	Connection to the content-strand (8), evidence of progress (8), accuracy (4), and level of independence (4).	4	4	8	8												student & item

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice	Combination of criteria
TN	Structured Portfolio	Connection to content standards (50), documentation of progress (50), evidence of choice (20), supports (10), settings (10), and interactions (10). For students with excessive absences, another matrix applies: connection to content standards (30), documentation of progress (30), choice (12), supports (6), settings (6), and interactions (6). A third matrix applies to homebound students: connection to standards (30), documentation of progress (30), and choice (12).			50/30/30	50/30/30					10/6/-	10/6/-			10/6/-		20/12/12	student, item, generalization, & system
TX	Item-based (performance task)	Accuracy (2), but level three tasks are weighted by 1.5, level two tasks are weighted by 1.2, and level one tasks are weighted by 1. Level of support (2) according to standardized levels of prompting. For a level 2 or 3 complexity level task, the student can earn an additional generalization point for each predetermined criterium that is performed without prompting. For generalization, the same task can be performed with a change in personnel, materials, or environment.	2 (x 1, 1.2, or 1.5)	2						1 (for level 2 or 3 task)								student & generalization

State	Test/item format(s)	Scoring Details	Accuracy / Performance	Level of Independence	Progress	Alignment to Standards	Complexity	Appropriateness	Context	Generalization	Settings	Interactions	Self Determination	Self-Evaluation	Supports/Accommodations	Participation	Choice	Combination of criteria
VT	Structured Portfolio	After demonstrating an appropriate baseline and evidence of instruction, products are rated "strong," "sufficient," or "insufficient" for alignment depth/ breadth and accuracy.	strong/sufficient/insufficient			strong/sufficient/insufficient												student & item
WA	Structured Portfolio	Alignment with grade level expectations (4), alignment to targeted skill (4), level of performance (4), and generalization/contexts (4).	4			4 + 4				4								student, item, & generalization
WY	Unstructured Portfolio	Performance (4), level of independence (4), and generalization across contexts (4). Level of complexity is scored 1 - 4, but is then weighted (1 = 1, 2 = 3, 3 = 5, and 4 = 8).	4	4			4 (weighted up to 8)			4								student, item, & generalization

2. States that use a scaffolded scale

Accuracy and Level of Assistance

State	Test / item formats	Scoring details
AZ	Item-based (performance tasks, constructed response, and multiple choice)	"Performance Tasks" (both performance tasks and constructed-response items) are scored on a 0-2 scale of no response, modeled responses to independent responses. "Rater Items" (both performance tasks and constructed-response items) are scored up to 4 points for a combination of accuracy and level of independence (scaffolded scoring). Multiple choice items are correct/incorrect.
CA	Item-based (performance tasks)	For level II – V tests, tasks are scored up to 4 points for mastery and level of task completion (partially completes task, minimally completes task). For level I tests, tasks are scored up to 5 for a combination of accuracy and level of prompting/assistance (scaffolded scoring).
CO	Item-based (constructed response and multiple choice)	Constructed response items are scored up to 6 points, combining accuracy and level of independence, and allowing for partially correct answers. Multiple choice items are scored 1-3 for a correct response according to level of independence.
IL	Item-based (multiple choice)	Scaffolded scoring up to 4 points combining accuracy and level of assistance. A correct answer after the administrator provides the answer is scored 2, and an incorrect answer is scored 1.
MI	Supported Independence and Participation tests: Item-based (performance tasks and constructed response)	For Participation and Supported Independence tests, items are scored up to three (P) or two (SI) for a combination of accuracy and independence (scaffolded scoring).
MN	Item-based (multiple choice)	Scaffolded scoring up to 3 points. An incorrect answer after two levels of assistance is scored a 1. No answer is scored 0.
MT	Item-based (performance tasks, constructed response, and multiple choice)	Scaffolded scoring up to 4 points.
NM	Item-based (performance tasks)	Each task is scored up to 2 or 3 combining accuracy and level of assistance (scaffolded scoring).
PA	Item-based (performance tasks, constructed response, and multiple choice)	Scaffolded scoring up to 5 points.
SC	Item-based (multiple choice)	Maximum points vary for each item, and points are scaffolded, combining accuracy and level of assistance. A small number of tasks are rated only for level of engagement.

WV	Item-based (constructed response and multiple choice)	CR items are scaffolded up to 6 points, and MC items are scaffolded up to 3 points.
WY	Item-based (portfolio and constructed response)	Items use a scaffolded scoring system, combining accuracy and level of assistance, up to 4 points. Hand-over-hand assistance results in score of one point. Zero points are earned if the student refuses to complete the task.

Accuracy/Level of Assistance and Accuracy/Complexity

State	Test details	Scoring details
FL	Item-based (multiple choice). Each question is written at 3 levels of complexity.	Students progress through three levels of complexity per item in a grade level content based assessment (starting at Participatory). Possible item scores are 0, 1, 2, 3, 6, or 9 based on the highest level of complexity (3, 6, or 9) or level of support (2, 1, or 0) at which a student provides an accurate response.

3. States that use accuracy approaches

AK	Item-based (multiple choice, constructed response, and performance tasks)	Regular test items are scored for accuracy and can include partial credit for partial accuracy. If a student scores a zero on three consecutive items in three consecutive tasks for a content area, the assessor administers the Expanded Level of Support (ELOS) items, which represent prerequisite skills. ELOS items are scored 1 – 4 according to level of assistance.
AZ	Item-based (multiple choice, constructed response, and performance tasks)	“Performance Tasks” (both PT and CR) are scored on a 0-2 scale of no response, modeled responses to independent responses. “Rater Items” (both PT and CR) are scored up to 4 points for a combination of accuracy and level of independence. Multiple choice items are correct/incorrect.
CA	Item-based (performance tasks)	For level II – V tests, tasks are scored up to 4 points for accuracy and level of task completion (partially completes task, minimally completes task). For level I tests, tasks are scored up to 5 for a combination of accuracy and level of prompting/assistance.
CT	Rating Scale	Teachers rate students on skills that are downward extensions of each essence statement for each performance standard. Each skill is rated as mastery/independent, developing/supported or does not demonstrate. Mastery indicates

		accuracy and independence at least 80% of the time.
HI[16]	Rating Scale	Each standard is rated by the teacher and a second rater as "non-existent," "emerging," "progressing," or "mastered."
IN	Rating Scale	Each student is rated through the use of a matrix of advancing approximations (a performance thread) to the content standard of the grade level band in which the student is enrolled.
IA	Rating Scale	The teacher scores for percent accuracy on the most recent trial for each skill, or marks the skills that were “already mastered,” “not taught,” or “fully prompted.”
KS	Structured Portfolio	1-5 scale according to degree of accuracy across the five trials.
LA	Item-based (multiple choice and performance tasks)	Each performance task is scored on a 0 to 2 point or a 0 to 1 point scale, according to an item-specific rubric. Two-point tasks allow the possibility of a partially correct response.
MD	Structured Portfolio	Students must perform under 50% at baseline, and must perform with 80-100% accuracy for an objective to be scored "mastered."
MI	Functional Independence tests: Item-based (multiple choice and writing prompts)	On the Functional Independence test, multiple choice items are scored for accuracy only.
NE	Item-based (multiple choice)	Scored for accuracy.
NV	Item-based (multiple choice and writing prompt)	Scored for accuracy, or flagged as "guided response," which indicates that the student could not answer without teacher intervention (which renders a 0 score).
NC	Item-based (multiple choice and writing prompt)	Scored for accuracy. Writing prompts scored for factual content.

[16] Hawaii will use a new assessment format in 2010-11.

ND	Item-based (multiple choice and constructed response)	Items are scored primarily for accuracy, but a set of “secondary indicators” can add additional points, including for settings, choice, planning, supports, and self-monitoring.
OR	Item-based (performance tasks)	2 (correct), a 1 (partially correct), a 0 (incorrect), a D (teacher did not administer item because it was deemed too difficult for the student), or I (inappropriate)
UT	Item-based (performance tasks)	Each task can be scored up to 4 levels: level 1 is minimal (no correct trials), level 2 is partial (1 correct trial), level 3 is sufficient (2 correct trials), and level 4 is substantial (3 correct trials, but also 3 activities/objects, 3 people, and 3 settings).
SD	Rating scale	A five point rating scale combines accuracy and level of assistance for each item. Additionally, teachers select one item from each indicator in reading and each content strand in math and science, and collect student evidence of performance on those items, documenting at least 3 trials of each skill.
VA	Structured Portfolio	Performance on each standard is scored 0 – 4 points according to evidence of student skill and knowledge.
WI	Item-based (multiple choice, constructed response, and writing prompt)	Multiple-choice items are scored for accuracy, and constructed-response items may be awarded partial credit for a partially correct answer.