# FLVS Assessment Guidelines

**Updated March 2015**

## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of Contents</td>
<td>1</td>
</tr>
<tr>
<td>Definitions and Types of Assessment</td>
<td>3</td>
</tr>
<tr>
<td> • Mastery</td>
<td>3</td>
</tr>
<tr>
<td> • Diagnostic Prescriptive Assessments</td>
<td>3</td>
</tr>
<tr>
<td> • Formative assessments</td>
<td>3</td>
</tr>
<tr>
<td> • Summative assessments</td>
<td>3</td>
</tr>
<tr>
<td> • Authentic assessment</td>
<td>3</td>
</tr>
<tr>
<td> • Performance assessment</td>
<td>3</td>
</tr>
<tr>
<td>Authenticity, Integrity and Rigor</td>
<td>3</td>
</tr>
<tr>
<td> Authenticity</td>
<td>4</td>
</tr>
<tr>
<td> Academic Integrity</td>
<td>4</td>
</tr>
<tr>
<td> Rigor</td>
<td>5</td>
</tr>
<tr>
<td>Assessment Design</td>
<td>6</td>
</tr>
<tr>
<td> Validity and Reliability</td>
<td>6</td>
</tr>
<tr>
<td> Assessment Bias</td>
<td>6</td>
</tr>
<tr>
<td> Test Blueprints</td>
<td>6</td>
</tr>
<tr>
<td> Test Development Process</td>
<td>7</td>
</tr>
<tr>
<td>Test Items</td>
<td>8</td>
</tr>
<tr>
<td> Scoring open-ended Test Items - Rubrics</td>
<td>8</td>
</tr>
<tr>
<td> Cognitive Complexity</td>
<td>8</td>
</tr>
<tr>
<td> Tagging Test Items</td>
<td>9</td>
</tr>
<tr>
<td> Item Difficulty</td>
<td>9</td>
</tr>
<tr>
<td> Cognitive Complexity vs. Item Difficulty</td>
<td>12</td>
</tr>
<tr>
<td> Item Difficulty and Grouping Items on Assessments</td>
<td>13</td>
</tr>
<tr>
<td> How to Build Groups</td>
<td>14</td>
</tr>
<tr>
<td> Item Analysis</td>
<td>15</td>
</tr>
</tbody>
</table>
Question Banks and Academic Integrity..............................................................................17
Additional Guidelines from FLVS Global.............................................................................19
Exams..................................................................................................................................20
General Information ...........................................................................................................20
Test Banks ............................................................................................................................20
When to Begin Item Writing and Purpose of the Exams .....................................................20
Testing a Subset of Standards ..............................................................................................20
Test Difficulty .......................................................................................................................21
Module Pretests/Diagnostic Assessments .........................................................................21
Lesson Quizzes ....................................................................................................................21
Module Assessments ............................................................................................................22
Honors or Advanced Assessments .......................................................................................22
Segment Exams ...................................................................................................................22
Test Time Limits and Multiple Parts to a Test ....................................................................23
Points on a Test and Weighting Parts of Tests ...................................................................23
Recommendations for Number of Items on a test ...............................................................24
Naming conventions for exams divided into multiple parts ..............................................24
Turnitin ..................................................................................................................................24
Discussion-Based Assessments ...........................................................................................24
Pre- and Post-tests for Efficacy Studies ............................................................................25
Advanced Placement ...........................................................................................................25
AP Redesign .........................................................................................................................25
AP Item Types Summary .....................................................................................................25
Multiple Choice (MC) ...........................................................................................................25
Free Response Questions (FRQ) ..........................................................................................25
Reusing Items from Pre to Post ..........................................................................................26
Document-Based Questions ...............................................................................................26
AP Published Scoring and Practice Resources .................................................................26
Other FLVS AP Guidelines .................................................................................................26
Appendix A Cognitive Complexity – Depth of Knowledge (DOK) .....................................28
Appendix B Test Blueprint Template .................................................................................29
Appendix C Sample Exemption Guide .............................................................................30
Definitions and Types of Assessment

FLVS curriculum is designed to engage students in a variety of activities and formative assessments throughout their learning experience. Guided and independent learning experiences are designed to meet the needs of 21st century learners.

Assessments are methods of measuring student learning and/or student achievement against a desired outcome or standard.

- **Mastery** – An important goal in FLVS curriculum development is to create a path for students that is the shortest road to mastery. The goal of mastery learning approaches is to have all students learn instructional material at roughly equivalent, high levels... In their review of mastery learning programs, Kulik, Kulik, and Bangert-Drows (1990) cite Bloom’s (1976) formulation as the classic approach. In Bloom’s model, students receive individualized instruction as necessary so that they all master course material. The basic approach reduces variation in final student performance through instruction suited to all students’ needs. [http://technologysource.org/extra/407/definition/1/](http://technologysource.org/extra/407/definition/1/)
- **Diagnostic Prescriptive Assessments** are pretests designed to identify a student’s strengths, weaknesses, knowledge and skills prior to instruction. Results provide an important snapshot of information to help individualize instruction.
- **Formative Assessments** are on-going assessments, reviews, and observations of student performance used to modify instructional methods and provide self-assessment.
- **Summative Assessments** make a judgment of student competency after an instructional phase is complete.
- **Authentic Assessment** refers to assessment tasks that resemble reading and writing in the real world and in school (Hiebert, Valencia & Afflerbach, 1994; Wiggins, 1993).
- **Performance Assessment** requires students to demonstrate their knowledge, skills, and strategies by creating a response or a product (Rudner & Boston, 1994; Wiggins, 1989).

Authenticity, Integrity and Rigor

The focus of FLVS assessments submitted to the instructor is authenticity, integrity and rigor. Approximately 40% of assessments submitted to the instructor are authentic. Academic integrity is monitored by using test banks and requiring students to participate in periodic discussion-based assessments facilitated by instructors. Rigor is promoted by ensuring that module and segment exams are written with 30–60% moderate- to high-level complexity as defined using the Depth of Knowledge model [see Appendix A].
**AUTHENTICITY**

Authentic Assessment occurs when the student is producing work that demonstrates meaningful application of knowledge and skills analogous to work performed in the real world. The framework document for the Partnership for 21st Century Skills elaborates: “Working on authentic tasks is a useful, engaging activity in itself; it becomes an "episode of learning" for the student (Wolf, 1989). From the instructional perspective, designing such tasks guarantees that we are concentrating on worthwhile skills and strategies (Wiggins, 1989).” Students are learning and practicing how to apply important knowledge and skills for authentic purposes. Authentic assessment promotes content mastery and high-quality performance. It shows what the student is learning, how the student is learning it, and the quality of the understanding over time.

People in the 21st century live in a technology- and media-driven environment, marked by access to an abundance of information, rapid changes in technology tools, and the ability to collaborate and make individual contributions on an unprecedented scale. The ability to navigate the complex life and work environments in the globally competitive information age requires students to pay rigorous attention to developing learning and innovation skills as well as life and career skills. To be effective in the 21st century, citizens and workers must be able to exhibit a range of functional and critical thinking skills. [Partnership for 21st Century Skills http://www.p21.org/our-work/p21-framework.

Authentic assessments may take the form of one or more of the following:

- Portfolios
- Peer review and/or self-reflection using rubrics
- Exhibition, presentations, Web development, narrations of a process
- Analyzing data, drawing conclusions
- Student-created content using 21st century tools and frameworks
- Literacy abilities in real-life contexts
- Projects
- Labs, simulations, and virtual field trips
- Collaborations

**ACADEMIC INTEGRITY**

For academic integrity purposes, each test bank includes at least three times as many items as will be administered to a single examinee for each traditional, auto-graded assessment. For example, for item one on the test; three items may be grouped together in the item bank. The computer randomly assigns one of these three items to the student. The same process occurs for each item on the test. Consequently, each of our students is administered a nearly unique test form.

For AI purposes, the order that “groups” of items are administered to students on a test is routinely scrambled.
At predefined points in courses, students participate in discussion based assessments. Students are required to initiate a discussion about the course content that is facilitated by the instructor. The student is rated by the instructor on how well they understand the course content, how clearly they make connections to other relevant concepts and ideas, how well-prepared they are for the discussion and how fully they participate.

FLVS instructors utilize technologies to check for authenticity.

A student may be randomly chosen to take any segment exam in a proctored environment. Additionally, teachers may request a proctored exam at any time.

Complete information on Academic Integrity is available in the FLVS document entitled Academic Integrity: The FLVS Non-Negotiable at http://www.flvs.net/areas/flvscourses/Documents/AcademicIntegrity.pdf

RIGOR

True rigor is creating an environment in which each student is expected to learn at high levels, each student is supported so he or she can learn at high levels, and each student demonstrates learning at high levels (Blackburn, 2008).

The Florida Standards, the Common Core State Standards and other national standards provide guidance on what is expected of students to ensure they are learning at high levels. Each course is written to ensure students have opportunity to master the standards at a high level and to demonstrate mastery of learning in meaningful ways. Courses also work toward building necessary background knowledge and scaffolding within lessons to provide support for students as they learn and attempt to fill in learning gaps. Instructors are available to provide additional support through live lessons and individualized attention.

As a part of the curriculum design process, guiding questions will be asked to ensure a focus on authenticity, integrity and rigor:

- How would you describe the level of authenticity in each assessment?
- Are discussion-based assessments scheduled appropriately?
- Do all traditional assessments have test banks according to FLVS guidelines?
- Are students expected to participate in project- and problem-based activities that promote individualized responses?
- Are there opportunities for choice in the way a student experiences the learning and shows mastery in assessments?
- Are all traditional assessments written with appropriate moderate- and high-level complexity as defined using the Depth of Knowledge model (See Appendix A)?
Assessment Design

Validity and Reliability

The two most important qualities of a well-constructed test are validity and reliability. Validity is the accuracy of measurement in relation to the intended use of the scores. It is not the assessment itself that is valid or invalid, but rather the inferences made from the scores. In creating standards based classroom assessments, the most important type of validity is content validity. Content validity requires that the test measures a complete and representative sample of the content and skills to be learned at appropriate levels of cognitive complexity. Sources of evidence in ensuring content validity include blueprinting tests to ensure appropriate alignment to the standards and using subject matter experts to evaluate the test items and test forms.

Most of the standards in any course are subject specific; however, English Language Arts (ELA) standards are also present in most all subject areas. The way in which these are taught and assessed may vary from subject to subject, but to ensure content validity, at least these three things should be true when the ELA standards are assessed as part of the course:

- Students should have received instruction on these standards within the course
- The standards are listed as standards for instruction within each of the pertinent modules
- Students are aware that these are important skills to be learned during module instruction

Reliability is the degree to which a test provides consistent results. Creating test forms based on the blueprint helps ensure reliability of scores by specifying that each test form is equal in length and item type and that it measures the same standards at the same level of cognitive complexity. Test items that are “grouped” together should be of similar difficulty. Additionally, using rubrics to score open-ended items promotes better inter-rater reliability.

Assessment Bias

W. James Popham, in Assessment Bias, How to Banish It, provides a concise definition of assessment bias: “assessment bias is present whenever one or more items on a test offend or unfairly penalize students because of those students’ personal characteristics such as race, gender, socioeconomic status, or religion.” Test items at FLVS are carefully inspected by subject matter experts to prevent assessment bias.

Test Blueprints

A test blueprint is a detailed written plan for a test. Blueprints are created by the Curriculum Specialist and are often reviewed by the Psychometrician for module and segment exams. They are created in Excel using a template that is available in Box (see Appendix D for a screenshot of the template). Consult the SOP (Standard Operating Procedure) for blueprinting a test for complete information. The SOP is also located in Box. All blueprints for a course should be created in a single Excel document using separate tabs for each individual test.
Test blueprints are uploaded to the project site and include the following documentation:

- Common Core Standard and/or state and national benchmark for each item, including the identifying code and text
- Number of items that will be written to measure each benchmark with the specified item type and complexity level of each item
- If a benchmark is not measured on a traditional blueprinted test (such as for a lab or performance requirement), an indication on the blueprint of where in the course that benchmark will be assessed
- Total number of items on the overall test
- Total number of items from each complexity level, percentage of items from each complexity level, and rationale for these percentages (see Appendix A – the rationale is whether the course is a core or elective course or other specialty course such as AP or whether it has a state required EOC)
- *Core courses that focus on teaching content tested on high stakes exams such as state developed EOC’s, FCAT 2.0, AIR, PARCC, Smarter Balanced and AP will use the Core requirements listed in Appendix A as a minimum and adjust according to available item specifications. Any adjustments to the core percentages in Appendix A should be approved by the manager for the course.
- If an exam is to be delivered in parts, include details for each part of the test and be sure the overall test meets the complexity level guidelines
- Date the test was blueprinted
- Caution: Do not weight multiple choice items on a single test by cognitive complexity level. For example, do not award five points for a high-complexity multiple choice item but only two points for a low-complexity multiple choice item.
- When blueprinting a test, identify the intended scoring weight for each item and incorporate the weight for each item into tallying the final percentages of cognitive complexity.

**Test Development Process**

FLVS follows best practices in current classroom test development as follows:

- Define the test purpose and objectives (based on relevant state standards, Common Core, and/or national standards depending on the subject area)
- Blueprint tests for module and segment exams to ensure the relevant standards are tested and they are tested at appropriate levels of cognitive complexity
- Write test items based on the blueprint and according to the “FLVS Item Writing Guidelines” (see Appendix G).
- Review, edit and validate the test items (work done by subject matter experts)
- Assemble and validate the tests per the test blueprint
- Administer and score the tests (some test items may be revised based on item performance)
**Test Items**

**Scoring open-ended Test Items - Rubrics**

Rubrics with carefully developed descriptors for each score point should be provided and used by teachers to score open-ended items on tests. This is especially important on module and segment exams. The use of a generalized rubric is often acceptable and sometimes preferable to creating a new rubric for each assignment. Teachers and students can become familiar with the generalized rubric as they use it repeatedly. Model papers illustrating each score point are highly desirable in helping teachers produce reliable scores.

Providing a “best answer paper” does not provide enough information to teachers because the “what is important” factor remains undefined. For example, is it the content, length, writing style, number of grammatical errors, spelling, sentence structure or some other quality that makes it a “best paper?”

Examples of generalized rubrics that can be useful in scoring test items are displayed in Appendix F.

**Cognitive Complexity**

Cognitive complexity refers to the level of cognitive activity required by an examinee to successfully complete a test item. We specifically design our tests to measure appropriate levels of complexity by using a complexity schema called Depth of Knowledge (DOK) and by blueprinting tests to require specific percentages of items from each complexity level. Brief definitions of the DOK levels are located in Appendix A. DOK is a challenging schema to understand and master. Training resources, including an Elluminate session by the school psychometrician are available in the Development Handbook in the “Cognitive Complexity Resources” folder at [https://my.floridavirtualschool.net/Departments/CurriculumServices/CD/Development%20Handbook/Forms/Checked%20out.aspx](https://my.floridavirtualschool.net/Departments/CurriculumServices/CD/Development%20Handbook/Forms/Checked%20out.aspx).

Norman L. Webb, the developer of DOK specified four levels of complexity: 1) Recall, 2) Basic Application of a Skill or Concept, 3) Strategic Thinking and 4) Extended Thinking. Until late 2012, the FLDOE used a slightly modified version of DOK. They collapsed the four levels into three categories - low complexity (LC), moderate complexity (MC), and high complexity (HC) — forming an ordered description of the demands an item may make on a student. During the summer of 2010, FLVS adopted the same three-level modification.

In late 2012, the Florida Department of Education adopted the full Depth of Knowledge model. They are now categorizing items by Levels 1, 2, 3, and 4. As they implement the common core based Mathematics Florida Standards (MAFS) and Language Arts Florida Standards (LAFS) and the new Florida Standards Assessments (FSA’s), they may introduce some Level 4 items into high-stakes testing. For more information go to [http://www.cpalms.org/cpalms/dok.aspx](http://www.cpalms.org/cpalms/dok.aspx) - click the link entitled “Content Complexity Common Definitions.”

The state’s definition document provides helpful insights into guidelines for cognitive complexity requirements in test development:

"Because the DOK model of content complexity was designed primarily as a framework for aligning content standards and assessments, it is important to distinguish between the DOK rating for a given standard and the
possible DOK ratings for assessment items designed to address the standard. The DOK level for an individual content standard is intended to represent the typical performance level of cognitive complexity that a learning activity or assessment item associated with that standard might entail. This is particularly important for assessment purposes, since 50% or more of assessment items associated with a given standard should meet or exceed the DOK level of the standard.”

This statement also implies that 50% of items on state sponsored tests can be written to cognitive complexity levels that build toward but do not meet or exceed the DOK level of the standard. This is consistent with the FLVS complexity guidelines and with findings from cognitive psychology:

“The cognitive processes that are most esteemed — logical thinking, problem solving, and the like — are intertwined with knowledge” (Willingham, 2009).

While FLVS is ultimately interested in how well students can research, apply, create, and share information with a community of learners, we recognize and emphasize that students cannot effectively perform at higher levels of cognitive complexity without mastering important background and factual knowledge. In developing assessments, we create a mix of items at all levels of cognitive complexity, as appropriate (see Appendix A.)

**Tagging Test Items**

As test items are developed, we need to keep important information such as which standard the item measures and the complexity level of the item with each item. To that end, we tag each item with relevant information as shown in the example below. Each assessment item should reference the corresponding lesson, the appropriate benchmark code, and the complexity of the item [LC = low complexity, MC = medium complexity, HC = high complexity] as illustrated in the sample question:

**Question 4** (Worth 2 points)

(01.02 LC)

SC.912.N.1.3: 

A farmer is trying to determine why some weeds sprayed with …..

**Item Difficulty**

Item difficulty is defined in two different situations:

- Before a test is administered – it is the percent of students likely to answer an item correctly as estimated by experienced educators
- After a test is administered- it is the percent of students that answer an item correctly on a test
For example, if only 40% of students answer an item correctly or are estimated to do so, it is considered a much more difficult item than if 85% answer it correctly or are estimated to answer it correctly. After a test is administered, the item difficulty level is also referred to as an item p-value. If 40% of students get the item correct, the p-value is .40; if 85% get it correct the p-value is .85.

FLVS develops classroom tests as a publisher of curriculum and corresponding assessments using the guidelines outlined earlier in the section entitled “Test Development Process.” We are not a publisher of high stakes tests comparable to those developed at the state or national level for accountability purposes. During the development process at FLVS, we do not have access to item responses from students that allow us to do an item analysis to obtain item statistics such as p-values or item discrimination values. Even so, we need to understand principles of test development related to p-values and item analysis because it impacts the quality and difficulty level of our assessments. There are some things we can do as content experts to help control for difficulty before a test is administered, and we sometimes have limited resources in Educator or from Efficacy Studies after tests are administered to use item p-values and other item statistics.

The overall average p-value of the items on a test is directly related to the average score students will achieve on the test. For example, if 100 students* take a five item test with difficulty values as shown in the chart, then the average score on the test for comparable examinees is expected to be about 62%.

<table>
<thead>
<tr>
<th>Item #</th>
<th>Difficulty or p-value (percent of correct answers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.70</td>
</tr>
<tr>
<td>2</td>
<td>.90</td>
</tr>
<tr>
<td>3</td>
<td>.60</td>
</tr>
<tr>
<td>4</td>
<td>.50</td>
</tr>
<tr>
<td>5</td>
<td>.40</td>
</tr>
</tbody>
</table>

*Note that difficulty values can change depending on the ability level of the group of students that respond to a test item or because the item has been compromised or “over-exposed.”

Depending on what cut score an organization uses to interpret the score on this particular test, 62% may show high performance or low performance or in-between.

- FLVS and most school districts use the 90=A, 80=B, 70=C, 60=D scale which means that a 62% is a very low average test grade even if the test is very difficult.

Consider the following quote from the Algebra 1 End-of-Course Assessment Test Item Specifications, Version 2, from the FLDOE:

The difficulty of FCAT 2.0 and EOC items is initially estimated by committees of educators participating in Item Content Review meetings each year. As each test item is reviewed, committee members make a prediction of difficulty based upon their knowledge of student
performance at the given grade level. The classification scheme used for this prediction of item difficulty is based on the following:

Easy - More than 70 percent of the students are likely to respond correctly.

Average - Between 40 percent and 70 percent of the students are likely to respond correctly.

Challenging - Fewer than 40 percent of the students are likely to respond correctly.

After an item appears on a test, item difficulty refers to the actual percentage of students who chose the correct answer.

Since the FLDOE sets cut scores for their EOC’s, they can freely allow for this range of difficulty values on their tests. In fact, the best items on their tests will generally be average difficulty items because those items best discriminate between high and low performers.

However, if we want to ensure that our module and segment exams result in an average score of 70% or higher, then our average p-value over all items on a test must be .70 or higher. What does this mean for us?

- We can have some items with average difficulty values (.40 to .70)
- We have very little tolerance for very difficult items (<.40)
- We need a large number of items that most of our students will be able to answer correctly after instruction (.70 - .90+)

- This is one reason that we allow for a fairly high percentage of low complexity (recall) items on our tests. Note: keep in mind during item reviews that not all low complexity items are “easy.” The nature of a recall item is that you either know the answer or you do not. In general, low complexity items at FLVS should not be difficult for most students when assessment and instruction are properly aligned.
- We know from cognitive psychology that students cannot effectively perform at higher levels of cognitive complexity without mastering important background and factual knowledge. We need to test this important background knowledge with our low complexity items.
- Also note that to show growth from pretest to posttest, the most desirable items are those that students cannot answer correctly before instruction but can answer correctly after instruction.
- Even though our test development process is different from high states state exams, it is good and informative practice to run a correlation analysis between student EOC scores and segment exam scores to confirm a positive relationship.
- Also consider that a module or segment exam should not be so easy that most all students are scoring close to the top of the scale. There should be some variation in scores on these tests.

What are some reasons that p-values obtained on a test after instruction is completed would be low?
The test item may be a good item but many students have not yet mastered the increased cognitive complexity levels required by Common Core and The Florida Standards.

- The item is not closely aligned to the standards and/or the instruction and content in the course.
- The item is keyed incorrectly or has more than one correct answer.

Since we use the 90=A, 80=B, 70=C, 60=D scale, we need to do our best to ensure that item difficulty levels and average test scores for a course work as well as possible given the scale we use. What strategies can we use to accomplish this?

- Blueprint the test according to the guidelines and follow the test blueprint during item writing.
- Before course release, review and revise test items to ensure they follow best practices in item writing.
- As part of the item reviews during the development process, ask SME’s to estimate each item’s difficulty level and use the estimates to estimate the overall test difficulty level. Use the information to revise items that may be too difficult.
- After a test is administered, check Educator for item analysis information that may help you identify and improve test items that are too difficult (see the “Item Analysis” section of this document – Note: changes to items after administration can only be done during field study, product review or work on enhancements. Check with your manager before using this type of analysis).
- Check that items are “grouped” as appropriately as possible during test development.

**Cognitive Complexity vs. Item Difficulty**

Cognitive Complexity is NOT the same as item difficulty. You might think that low complexity items will be easy and high complexity items will be hard. However, it is fairly easy to write difficult low complexity items though it is not our intention to do this. Consider two test items where Item 1 is low complexity and Item 2 is high complexity. Notice the percent of students that selected each answer choice (in red) and the percent who answered each question correctly:

**Item 1 - (LC) SC.L.18.7**

What ultimately happens to the light energy captured during photosynthesis?

- 44% x-It is used to produce carbohydrate molecules.
- 15% It is turned into more chlorophyll for the plant.
- 13% It is used to release the energy from H₂O.
- 29% It is put into ATP molecules for long-term storage.
Item 2 - (HC) SC.H.1.3.1 (a released FCAT item)

Scientists once thought there was no life on the deep-sea floor due to the lack of sunlight. However, in the 1970s, complete ecosystems were discovered on the deep-sea floor that depend on energy from chemical reactions rather than energy from sunlight. Which conclusion can be drawn from this discovery?

19% A. Organisms are able to survive under any conditions.
19% B. Observations will always lead to the discovery of new organisms.
55% C. Organisms may be discovered in remote areas and extreme conditions.
7% D. Observations collected on existing organisms can apply to any organism.

Only 44% of students got the low complexity Item 1 correct while 55% got the high complexity item 2 correct. This means that the low complexity item is quite difficult and even slightly more difficult than the high complexity item. At FLVS, we need to ensure that our low complexity items test important information presented clearly in the course that our students will master if they are working conscientiously. Lead teachers reviewing Item 1 recommended a change to answer choice D after seeing the item statistics and realizing it had an element of truth that made it an especially strong a distractor. If they could have changed the course content they may have opted to clarify the explanation in the lesson instead. Since we usually do not have item statistics to guide our test development, we have to be extra diligent to check during item reviews that the items are not too difficult given the course content and especially that the low complexity items are not too difficult.

**Item Difficulty and Grouping Items on Assessments**

When a student takes one of our tests, the set of items she is administered is randomly assigned to her from a “group” of similar items. For Academic Integrity purposes, each test bank includes at least three times as many items as will be administered to a single examinee for each traditional, auto-graded assessment. For example, for item 1 on the test, three items may be grouped together. The computer randomly assigns one of these three items to the student. The same process occurs for each item on the test. Consequently, each of our students is administered a nearly unique test form. We do this to promote academic integrity. To be fair to students, each test form must be as close to same level of difficulty as possible.

To accomplish this, we must evaluate our “groups” of test items. Items in a group are designed to share a set of common characteristics; each item in a group should

- measures the same benchmark
- be written to the same level of cognitive complexity
- have approximately the same difficulty value

For example, consider the groups of items in the chart below – groups 10, 24 and 26. The information is taken from an item analysis from an efficacy study. Though we do not have item statistics to do this level of analysis during test development, it is helpful to understand why it is important to have groups made up of items with similar difficulty values.
The average difficulty value over the nine items in three groups is 66% or 0.66. If, hypothetically, our test consisted only of these items, the expected average score on the test would be 66%. However, any student receiving item 71 may be at a disadvantage because the item is so much more difficult than items 70 or 72. Grouping items that have similar levels of difficulty helps ensure that each student will receive a test that is similar in overall difficulty to the tests other students receive.

**How to Build Groups**

<table>
<thead>
<tr>
<th>Item #</th>
<th>Benchmark</th>
<th>DOK</th>
<th>Difficulty</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>SC.912.L.15.6</td>
<td>MC</td>
<td>76%</td>
<td>10</td>
</tr>
<tr>
<td>29</td>
<td>SC.912.L.15.6</td>
<td>MC</td>
<td>70%</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>SC.912.L.15.6</td>
<td>MC</td>
<td>66%</td>
<td>10</td>
</tr>
<tr>
<td>70</td>
<td>SC.912.L.15.10</td>
<td>LC</td>
<td>92%</td>
<td>24</td>
</tr>
<tr>
<td>71</td>
<td>SC.912.L.15.10</td>
<td>LC</td>
<td>23%</td>
<td>24</td>
</tr>
<tr>
<td>72</td>
<td>SC.912.L.15.10</td>
<td>LC</td>
<td>84%</td>
<td>24</td>
</tr>
<tr>
<td>76</td>
<td>SC.912.L.15.13</td>
<td>HC</td>
<td>70%</td>
<td>26</td>
</tr>
<tr>
<td>77</td>
<td>SC.912.L.15.13</td>
<td>HC</td>
<td>49%</td>
<td>26</td>
</tr>
<tr>
<td>78</td>
<td>SC.912.L.15.13</td>
<td>HC</td>
<td>62%</td>
<td>26</td>
</tr>
</tbody>
</table>

Group 10 items are of similar difficulty. They measure same benchmark at same DOK.

Item 71 is much too difficult for Group 24.

Group 26 items are borderline. Item 77 is somewhat difficult for this group.

For example, form groups of three items where one item will be selected; do not put six items together in a single group and choose two of them or nine items in a group and choose three of them. Why this requirement?

- It is straight forward to revise a group of three items, choose one. For example – To improve Group 24 we need only replace item 71
- We could even duplicate item 70, if necessary, and use it as a replacement for item 71 without risking a student getting the same item twice on a test (as can happen if there were 6 items choose 2)
- It is easier to find 3 items with approximately the same difficulty level for a group than to find 6 or 9 items with the same difficulty level
ITEM ANALYSIS

After a test is administered to students, Educator offers some limited but helpful item analysis at the teacher shell level for students who are active in a course.

IMPORTANT: Item Analysis information to make changes to items can be used ONLY when a course is in field study or product review or when working on a course enhancement. Even then, manager approval should be obtained before making changes to active items.

Accessing information from a few teacher shells until you have at least 30 item responses representative of the students you want to study can give you data on how the students are performing on those items (100 responses is much better but 30 can give you an idea). To access the information, you must enter Educator with manager or teacher access. Once you are in the teacher shell click ASSESSMENTS>EXAM and select the exam you want to analyze. Click ACTIONS>GRADEBOOK. Then scroll to the bottom of the page. You will see two buttons and some summary statistics as illustrated next:

Number of submissions: 74
Standard Deviation: 7.53
Min: 19
Max: 50
Range: 31
Mode: 42
Mean: 38.01

To see the item analysis, click ADVANCED QUESTION STATISTICS. You will see an exam with questions as displayed next:
In a few cases, FLVS Marketing has done efficacy studies for a course that included a full item analysis with complete information on how each item on a test is performing. Here is a link to these studies:

http://www.flvs.net/areas/aboutus/Pages/Research.aspx
**Question Banks and Academic Integrity**

In creating item banks for Academic Integrity purposes (i.e., writing three items that measure the same standard so students do not all receive the same items on their individual tests), it is NOT the intent to ask the same question three different ways. The intent is to have three different questions available that address the same standard at the same level of difficulty and cognitive complexity. In some subject areas this can be accomplished by cloning items. Consider the following examples (and non-example):

Example A (non-example and example)

<table>
<thead>
<tr>
<th><strong>Non-example</strong> (Do NOT write items that ask the same question using a different structure):</th>
<th><strong>Example:</strong> Cloning Items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Which step in the scientific method follows immediately after testing the hypothesis?</strong></td>
<td><strong>Which step in the scientific method follows immediately after testing the hypothesis?</strong></td>
</tr>
<tr>
<td>State the problem.</td>
<td>State the problem.</td>
</tr>
<tr>
<td>x-Analyze data.</td>
<td>x-Analyze data.</td>
</tr>
<tr>
<td>State the conclusion.</td>
<td>State the conclusion.</td>
</tr>
<tr>
<td>Gather information.</td>
<td>Gather information.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>The step of analyzing data occurs immediately after which step in the scientific method?</strong></th>
<th><strong>Which step in the scientific method follows immediately after stating the problem?</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>State the problem.</td>
<td>Test the hypothesis.</td>
</tr>
<tr>
<td>x-Test the hypothesis.</td>
<td>State the conclusion.</td>
</tr>
<tr>
<td>State the conclusion.</td>
<td>Analyze data.</td>
</tr>
<tr>
<td>Gather information.</td>
<td>x- Gather information</td>
</tr>
</tbody>
</table>

| **This is the same question in two different forms. The same information is required to answer each question.** | **These questions test two different concepts. Additionally, the first item could be considered a parent and the second a “clone” because the second item is written using the first item as a template.** |
Example B: Math items are often ideal for writing clones:

**Question 2** (Worth 5 points)
01.02 MACC.912.F-BF.1.1B:

Given that \( f(x) = 2x + 1 \) and \( g(x) = \frac{3x-1}{2} \), solve for \( g(f(3)) \).

- 5
- 7
- 9
- 10

This is a correct answer

**Question 3** (Worth 5 points)
01.02 MACC.912.F-BF.1.1B:

Given that \( f(x) = 3x + 1 \) and \( g(x) = \frac{4x+2}{3} \), solve for \( g(f(0)) \).

- 2

This is a correct answer

- 6
- 8
- 9
Example C: Not all questions can be written as clones. In such cases, write different questions that measure the same benchmark at the same level of difficulty and cognitive complexity.

The following questions measure SC.912.N.1.4: *Identify sources of information and assess their reliability according to the strict standards of scientific investigation.*

**Question 11** (Worth 2 points)  
(SC.912.N.1.4)

Claims about new scientific breakthroughs often appear in the media. If such a claim is made in each of the following, which media source should be considered most reliable?

- Peer-reviewed journals  
  *This is a correct answer*
- Popular magazines
- Internet websites
- Newspaper advertisements

**Question 12** (Worth 2 points)  
(SC.912.N.1.4)

Scientists review the work of other scientists to

- ensure the results are in agreement with accepted scientific beliefs
- guarantee federal funding by supporting each other's work
- keep the scientific community better informed
- validate the results and check the evidence’s reliability  
  *This is a correct answer*

**Additional Guidelines from FLVS Global**

Be sure to incorporate additional guidelines from the FLVS Global document entitled *FLVS Global Guidelines for Content Development*. This document provides information that will ensure that content and assessments will work in other LMS’s and with our varied client needs. It includes an entire section on Quizzes, Exams and Final Exams and another section on Answer Keys. It specifies which item types are allowed and explains some
limitations that are necessary regarding group selection and item types. The link to this document in Box is https://flvs.box.com/s/mbyk2uv8t8c3eeyg540h

Exams

GENERAL INFORMATION

Test Banks
A test bank or item bank is a database of questions from which a subset of test items is selected for administration to an individual student. At least three times as many items as will be administered to a single examinee are written for each traditional, auto-graded assessment for most types of exams. See the section entitled “Item Difficulty and Grouping Items on Assessments” for information on why the item banks are necessary and how they are used.

When to Begin Item Writing and Purpose of the Exams
Before writing traditional type test items such as multiple choice, the initial blueprint should be completed and course content must be available for item writers to reference as they write items. Having content available to item writers is important to ensure consistency and coherence between content and assessments. As an example, mathematical notation used in test items should be consistent with what was taught in a course and terminology should also be consistent (i.e., avoid focusing only on “central ideas” in the course but asking about “main ideas” on an assessment). Assessment may be unreasonably difficult or may require additional teacher explanations if alignment is not achieved. Even so, it is not the intent to create questions that are overly specific to certain types of content given the intended purpose of module and segment exams, which is to measure mastery of the standards. For example, if a lesson in a course presented a story about lions it would not be appropriate to ask on the module or segment exam, “What was the name of the main character in the story about the lions?”

Testing a Subset of Standards
Classical test theory assumes that the items on any single test are a sample from an infinite domain of possible items. Tests are blueprinted and items are developed with appropriate representation of the domain. Test scores are used to infer mastery of that domain. While longer tests generally increase the reliability of scores, due to constraints on administration time and examinee stamina, we must make choices about how many items the test can include. Consequently, it may not be possible to test every individual standard on a module exam or especially on a segment exam. Some standards must be tested together or tested on other assessments. The standard(s) an item assesses should be identified on the test blueprint. If multiple standards are tested by a single item, list all the appropriate standards on the blueprint. The FLDOE illustrates this idea in their draft release of the Algebra I EOC Item Specifications where they list a standard that will be assessed and group it with other standards that are “also assessed” by the same item(s). The PARCC math item prototypes illustrate this use with their most “relevant standards” idea. As FLVS develops courses, teams decide for each standard whether to “tap it, nail it, or smash it.” This analysis should also be
useful in deciding which standards are most important to test on the module and segment exams. Decisions about which Common Core or Florida Standards should be tested individually and which with other standards should be made in consultation with the respective curriculum manager and other FLVS experts on Common Core and the Florida Standards.

**Test Difficulty**

As test items are written and reviewed, keep in mind that students in FLVS courses are graded using the 90/80/70/60 scale and not by cut scores such as states use for PARCC, Smarter Balanced, tests of Florida Standards and EOC’s. We must ensure the exams are of an appropriate difficulty level given the scale we will use for scoring. For additional information, see the section of this document entitled “Item Difficulty.”

**Module Pretests/Diagnostic Assessments**

The primary objectives of the module pretests are to assess prior knowledge and to act as one tool used to create a learning path appropriate for the learner. Sometimes a course is selected for an efficacy study. In such cases the pretest may also be used as a baseline to measure growth.

Teachers use pretest results, discussion based assessments, and other student information to create a path for each student. For courses that have a prescriptive learning focus, it is especially important that students are informed about the importance of the pretest in determining what they already know and what they still need to learn.

Instructors will be provided with an exemption guide to support their decisions regarding individualized learning paths for students [Appendix C]. Exemption guides are created by the Curriculum Specialist in collaboration with the writer and members of the instructional team.

Pretest items are written to module benchmarks and aligned to learning objects or lessons in order to support the prescriptive learning goal.

When students take the pretest, they respond to approximately 2-5 items per benchmark. Due to the quantity of benchmarks in some subject areas and the complexity of those benchmarks, the project team will make some collaborative decisions regarding number of items, benchmarks to be assessed, and complexity of the test items during test blueprint development. In some cases, such as courses with EOC exams, the state (Florida) item specifications document may provide helpful information by specifying benchmarks that the state plans to test together, test by a particular item type, or not test at all. The pretest is diagnostic and will have little to no impact on the student’s grade in the course. Even so, the importance of completing the pretest needs to be conveyed to the student.

At the module level, it is permissible to reuse up to 30% of module exam items on a module level pretest.

**Lesson Quizzes**

Lesson quizzes may or may not be used in a course. For subjects that use quizzes [mini-module assessments], test banks are not required.

Lesson-level quizzes should begin preparing students for success on the module- and segment-level assessments. While they may focus at times on the factual, procedural, and background knowledge that is prerequisite to higher levels of cognitive complexity, they must also begin to form a bridge between lower and higher levels of cognitive complexity. In general, the scores students receive on a group of lesson-level
assessments should bear some relationship to scores they might later receive on module and segment exams. It is at the lesson level that practice, feedback, and scaffolding during assessment can help students make the difficult transition from lower to higher levels of cognitive complexity in their thinking and experience.

**Module Assessments**

The objective of the module assessments is to assess student mastery of the standards and benchmarks for a given module. Module assessments may take many forms. FLVS does not mandate a traditional module assessment. Project teams may choose to create a traditional module assessment, a non-traditional module assessment, or a mix of assessment components.

For traditional module exams, individual students will respond to approximately 1-3 items per standard/benchmark. For academic integrity purposes, a test bank is required. Module assessments should help prepare students for success on the segment exams. Because there are fewer benchmarks to test at the module level than at the segment level, module assessments offer the chance to test benchmarks using a larger quantity of items from each level of cognitive complexity than will appear on a segment exam. They should provide students an opportunity to be familiar with and successful on test items written to the same levels of difficulty and cognitive complexity they will encounter on the segment exam. In general, the scores students receive on the module assessments should show sound correlation to their scores on segment exams.

At the module level, it is permissible to reuse up to 30% of module exam items on a module level pretest.

**Honors or Advanced Assessments**

Honors or advanced students will have learning opportunities throughout the course that require deeper critical thinking and, in some subject areas, additional content knowledge acquisition. There may be additional lessons and/or extensions to regular lessons. It is not a requirement that the honors or advanced lessons and assessments are evenly spaced throughout the course. Instead, honors and advanced students are provided appropriate learning opportunities in a way that meets the standards and is in the best interest of the students.

Historically, honors or advanced students have had a unique segment exam in each FLVS course. To prepare students for that experience, honors or advanced students have often also had unique module assessments where applicable. Project teams may choose to reuse questions from the regular exam with additional or different questions that address additional content and/or higher complexity levels. Completing a summative project to test comprehension is an alternate path that has been implemented in some recent high school math courses.

**Segment Exams**

The objective of the segment exam is to assess student mastery of standards and benchmarks for a given segment. Since many students take only one semester with FLVS, segment exams should cover material taught during the current segment and should not be cumulative across the entire course, even in cases where a student must prepare for a cumulative exam outside of FLVS such as an EOC or AP exam.
Segment exams should not suddenly be much more difficult or present an overall higher complexity level or a very different experience than what the students were expected to achieve on module exams. The exam characteristics should be similar enough that students who do well on the module exams should also do well on a segment exam in most cases. Project-based courses that give a traditional segment exam must also provide opportunities throughout the course for students to practice on traditional type test items and master the objectives that will be tested traditionally so they are prepared to succeed on the traditional exam.

When students take the segment exam, they respond to approximately 1–3 items per benchmark. For academic integrity purposes, a test bank is required.

If a project team wants to use a non-traditional segment assessment, the project manager will include the description and rationale in the project charter. In most cases, FLVS Global will require a traditional version of the segment assessment.

Segment exams count 20% of a student’s overall grade in the course. Students are no longer required to pass the segment exam in order to pass the course.

**Test Time Limits and Multiple Parts to a Test**

Since students testing in a virtual learning lab must complete an exam during a single class period, some module and segment exams must be divided into multiple parts. Students must be able to complete a designated part in 45 minutes or less (do not set a 45 minute time limit but do design the test so that most all students can finish in 45 minutes or less). A single exam should not be divided into more than three parts. AP and industry certification exams are exempt from the 45 minute time limit. FLVS tests are not speed tests; grades should depend on mastery of standards and benchmarks, not on speed in completing the test. Tests length must therefore be carefully designed. Several courses have worked on a design for this shortened test length with multiple parts; consult with the lead curriculum specialist or psychometrician for guidance on which courses provide good examples of how this can best be accomplished. Due to the number of benchmarks in some subject areas, the project team will make some collaborative decisions during blueprinting of tests regarding number of parts to a test, number of items on a test and complexity of items.

**Points on a Test and Weighting Parts of Tests**

The number of points awarded for correct answers of items on a test is specified by the CS. The CS decides that a multiple choice item gets 2 points and an essay gets 10, for example. The weighting of each part of a multi-part test is also specified by the CS.

Within Educator, a segment exam is structured so that it is worth 20% of the total points in a segment. Once the number of points the segment exam is worth is determined (say it is 300 points), then if each part of a 3 part test is weighted equally, each part will worth 100 points. Alternatively, if part 1 is worth 50% of the points and each other part worth 25%, then Part 1 will get 150 points and parts 2 and 3 will get 75 points each. Based on this example, the teacher can figure the passing score by adding the total points the student made on all 3 parts and dividing by 300.
RECOMMENDATIONS FOR NUMBER OF ITEMS ON A TEST

Plan so that the slowest students can complete the test in the time allotted (i.e., at least 90% of students should complete a test, preferably all students). About 20-25 MC items would likely work in a 45 minute class period, fewer if you also plan for a few short answer items or a limited essay. Here are a few guidelines and rules of thumb that must be moderated by specific item characteristic such as difficulty and complexity levels of an item:

- Time yourself or others (students would be great) to see how long it takes to read the test items (factor that time into the total testing time)
- Allow 5 or 10 minutes for a student to check over the work before submitting
- Once the above is estimated plan roughly on
  - about 1 minute per MC or FIB
  - 2 minutes for short-answer requiring a few sentences
  - 10-15 minutes for a limited essay question
  - 30 minutes for a question requiring one or two full pages to respond

NAMING CONVENTIONS FOR EXAMS DIVIDED INTO MULTIPLE PARTS

Module # (2 digits).Lesson # (2 digits) Segment (One or Two) Exam (Optional: Part One, Two, Three)

Examples:

- 05.10 Segment One Exam
- 04.11 Segment One Exam Part One
- 04.11 Segment One Exam Part Two
- 09.12 Segment Two Exam
- 10.13 Segment Two Exam Part One
- 10.13 Segment Two Exam Part Two

TURNITIN

If an exam part is purely essay (it has no other item types), it can be submitted to Turnitin. However, if the exam uses multiple item types, Turnitin does not have the capability to evaluate it.

DISCUSSION-BASED ASSESSMENTS

The objectives of the discussion-based assessment (DBA) are to ensure student mastery of concepts and to maintain the integrity of FLVS as an organization.

Here are some general guidelines for implementing DBAs in course development:

- A DBA is not connected to a specific assessment. If the information for the DBA is included on a lesson html page, it still includes a broad range of topics from a unit of learning.
In the development of a course, DBAs appear in every module or approximately every 4 weeks. If a project requires a deviation from this guideline, consult the functional manager for approval and include the approval and rationale in the course charter.

- Teachers may complete DBAs at various times during the student’s learning experience.
- Teachers must score each DBA using the DBA rubric [Appendix D]. The rubric requires teachers to assess a student’s knowledge of the content, ability to apply that knowledge, and 21st century skills of professionalism and commitment.
- Students must also have access to the DBA rubric as they prepare for DBA’s

**PRE- AND POST-TESTS FOR EFFICACY STUDIES**

The purpose of efficacy studies is to measure the effectiveness of FLVS courses by measuring learning gains from pre-test scores to post-test scores. The curriculum and community relations team work together to complete these studies with an external expert. FLVS leadership selects which courses will participate in an efficacy study.

**Advanced Placement**

**AP REDESIGN**

With the introduction of the new biology frameworks in 2012, the College Board launched the beginnings of an AP redesign process. As courses and tests are redesigned, AP is for the first time providing a comprehensive curriculum framework to help guide student learning. Instruction will focus less on memorization and too broad content coverage and more on developing deeper conceptual understanding and critical thinking and reasoning skills.

**AP ITEM TYPES SUMMARY**

Some AP exams require writing in response to specific prompts or writing AP position or argument papers, often selecting and using sources provided in a prompt. Common item types are as follows:

**Multiple Choice (MC)**

- Complexity levels based on AP course descriptions
- Four or five answer choices, depending on the subject area
- Analysis of stimulus – often involves analysis of scenarios, data, or illustrations presented in a stimulus

**Free Response Questions (FRQ)**

Students should have frequent formative assessments to practice their writing skills through a variety of assessment methods including

- analysis of FRQ item
• planning
• thesis writing, when required by the subject area
• peer and self-evaluation.

**Reusing Items from Pre to Post**

At the module level, it is permissible to reuse up to 30% of module exam items on a module level pretest. For AP only, it is permissible to reuse up to 30% of the segment exam items on a segment level pretest.

**Document-Based Questions**

In AP courses with a document-based question (DBQ) requirement, students should participate in frequent formative assessments including

• analyzing primary and secondary sources
• interpreting informational text
• determining validity of sources
• developing arguments using sources

**AP Published Scoring and Practice Resources**

Released exam questions, scoring guidelines, sample student responses, and scoring commentary from previous exams are available on the AP Website at [http://apcentral.collegeboard.com/apc/public/exam/exam_information/index.html](http://apcentral.collegeboard.com/apc/public/exam/exam_information/index.html). Courses should emphasize and explain how student’s can access these AP scoring resources and provide a description of what they are and why they are useful. Many FLVS AP courses have a review module with answer keys where these materials are especially helpful. In some cases on the AP Website, the expert commentaries are included in the same file with the student responses instead of being separate documents. AP practice lessons may make reference to these materials and may include strategies for success, but due to copyright concerns, the specific documents, rubrics, problems, etc. are not included in FLVS courses. Whenever possible, FLVS-developed free response and document based questions that emulate the released exam questions should be included in the course, complete with rubrics and other scoring guidelines similar to those used on the AP exam.

**Other FLVS AP Guidelines**

• AP exams are exempt from the 45 minute time limit per sitting required by other FLVS exams.
• DBA Rubric - FLVS Global requests that a DBA rubric be included in AP courses in the answer keys and student modules wherever DBAs occur as is provided for other FLVS courses.
• Choice - Assessments in AP courses are often written to mirror the College Board guidelines though AP writing assessments typically do not give students choice in essay questions, FLVS offers choice in AP courses, too. For example, we may ask students to write a question, create a rubric for a question, grade each other’s work, etc.
• Blueprints - AP test blueprints are designed based on the curriculum developed from the College Board course description or curriculum frameworks.

• Exemption Guides - Exemption guides are not used for AP courses. Diagnostic exams in AP courses are not used for exemption but to serve as a tool to extend learning. Teachers and students use diagnostic results, discussion-based assessments, and other student information to create a learning plan that will enrich the student experience. A diagnostic assessment allows the instructor and student to check attainment of learning and gives the student confidence that the high-level concept has been mastered. If not mastered, the student and teacher create an action plan that can lead to mastery of the concept or skill. The goal is to help students reach a higher level of mastery on each standard, not to just reach or show competency. Often on the AP exam students will have to analyze, synthesize, and evaluate content never covered to create the required product. The skills and knowledge of attacking the problem to reach a solution on a concept must be continually repeated with new scenarios; therefore, students are not exempted from course work.

• No Cumulative Segment Exams - The intended result in an AP course is for students to demonstrate mastery of AP concepts through application on the AP exam. Since many students take only one semester with FLVS, segment exams should cover material taught during the current segment and should not be cumulative across the entire course, even in cases where a student must prepare for a cumulative exam outside of FLVS such as an EOC or AP exam. It is recommended that AP courses blueprint and provide a cumulative practice exam that does not serve as the segment exam in the course but will provide the student with a complete AP practice experience.
# Appendix A

## Cognitive Complexity – Depth of Knowledge (DOK)

<table>
<thead>
<tr>
<th>Depth of Knowledge</th>
<th>Cognitive Complexity</th>
<th>FLVS Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>Low Complexity</td>
<td>40 - 50% for core courses* 50-70% for electives</td>
</tr>
<tr>
<td>Recall of a fact, information, or procedure (e.g., What are the Red Cross Emergency Action steps [check, call, care]?)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Level 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Basic Application of Skill/Concept</strong></td>
<td>Moderate Complexity</td>
<td>50-60% for core courses* 30-50% for electives</td>
</tr>
<tr>
<td>Use of information, conceptual knowledge, procedures, two or more steps, etc. (e.g., given an emergency scenario, students determine the care needed for a victim, and explain the reason for their actions)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Level 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Strategic Thinking</strong></td>
<td>High Complexity [minimum 5%]</td>
<td></td>
</tr>
<tr>
<td>Requires reasoning, developing a plan or sequence of steps; has some complexity; more than one possible answer; generally takes less than 10 minutes to do (e.g., stressed due to parents’ divorce; crunched for time; signs of stress – ways to relieve stress – why managing stress is important to health)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Level 4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Extended Thinking</strong></td>
<td></td>
<td>Typically not measured by MC items but instead by performance tasks and authentic type assessments, projects, etc.</td>
</tr>
<tr>
<td>Requires an investigation; time to think and process multiple conditions of the problem or task; and more than 10 minutes to do non-routine manipulations (e.g., Welcome to Health High – create fact sheet/brochure from research activity)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix B
Test Blueprint Template

Screen Capture of a Test Blueprint Template. The actual template that can be used to blueprint a test and a copy of the SOP (Standard Operating Procedure) for blueprinting a test are located in the Development Handbook on Sharepoint.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Algebra Foundations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Module 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Module Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of Low Complexity Items on Test</strong></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Moderate Complexity Items on Test</strong></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of High Complexity Items on Test</strong></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Items on Test</strong></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Meta-Exams are given the weight of 0.2 because they are less at most twice as the MCFE questions. However, there should be only one essay for each standard on the Module exam part</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Test Blueprint Template

| **Fluency and Expression** | 0 | 2 | 3 | 6 |
| **Fluency and Orality** | 0 | 2 | 3 | 6 |
| **Fluency and Grammar** | 2 | 1 | 3 | 6 |
| **Fluency and Vocabulary** | 2 | 1 | 3 | 6 |
| **Fluency and Composition** | 2 | essay | 2 | 3 |

## Rationale Comments
- **Mod 1 Test**
- **Mod 2 Test**
- **Mod 3 Test**
- **Mod 4 Test**
- **Mod 5 Test**
- **Mod 6 Test**
- **Mod 7 Test**
- **Mod 8 Test**
Exemption Guidelines

Students may be awarded lesson and assessment exemptions at the discretion of the instructor based on student performance on module pretests and discussion based assessments. It is recommended that students only exempt assignments where the student correctly answers all questions aligned to a particular benchmark.

Students may NOT exempt:

- Discussion Based Assessments
- Module or Segment Exams
- [Insert items specific to the course]

If the teacher determines that the student will be exempt and the exemption is not in the above categories, the teacher will collaborate with the Instructional Leader and Curriculum Specialist.

Students are accountable for the material in the entire module, and can expect to be assessed on the benchmarks of the course in subsequent assessments. Students may elect to complete exempted lessons.

Documentation:

The instructor will record exemptions in the student gradebook as follows:
Place “EX” in the numerical score box, and in teacher comments a detailed description for the evidence of student mastery of the standard.

The instructor will record exemptions in VSA documenting the lesson number, reason for exemption, and description of the evidence of student mastery of the standard.

Sample chart from an Exemption Guide

<table>
<thead>
<tr>
<th>Module Three Pretest Questions</th>
<th>Lesson exempted</th>
<th>Connected Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.00 Pretest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6, 13, 16</td>
<td>3.01</td>
<td>MA.7.G.4.3</td>
</tr>
<tr>
<td>1, 11, 14</td>
<td>3.02</td>
<td>MA.8.A.1.1</td>
</tr>
<tr>
<td>3, 10, 18</td>
<td>3.03</td>
<td>MA.8.A.1.5</td>
</tr>
<tr>
<td>2, 15, 20</td>
<td>3.04</td>
<td>MA.8.A.1.2</td>
</tr>
<tr>
<td>5, 9, 17</td>
<td>3.05 Not to Exempt</td>
<td>MA.8.A.1.2</td>
</tr>
<tr>
<td>NA</td>
<td>3.06 MVE Lab</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>3.07 Discussion</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>3.08 Practice Test</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>3.09 Discussion Based Assessment</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>3.10 Module Three Test</td>
<td></td>
</tr>
</tbody>
</table>
## Appendix D
### DBA Rubric

## Discussion-Based Assessment

<table>
<thead>
<tr>
<th></th>
<th>30–24</th>
<th>23–21</th>
<th>21–0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td>Clearly shows understanding of course content. Provides <strong>sufficient</strong> supporting evidence when needed.</td>
<td>Somewhat shows understanding of course content. Provides <strong>some</strong> supporting evidence when needed.</td>
<td>Minimally shows understanding of course content. Provides <strong>little</strong> supporting evidence though needed.</td>
</tr>
<tr>
<td><strong>Application</strong></td>
<td>14–12</td>
<td>11–10</td>
<td>8–0</td>
</tr>
<tr>
<td></td>
<td>Clearly makes connections to other relevant ideas, concepts, texts, and/or real world examples as appropriate.</td>
<td>Somewhat makes connections to other relevant ideas, concepts, texts, and/or real world examples as appropriate.</td>
<td>Minimally makes connections to other relevant ideas, concepts, texts, and/or real world examples though needed.</td>
</tr>
<tr>
<td><strong>Preparation</strong></td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Clearly demonstrates preparation (notes, questions, etc.)</td>
<td>Somewhat demonstrates preparation (notes, questions, etc.)</td>
<td>Minimally demonstrates preparation (notes, questions, etc.)</td>
</tr>
<tr>
<td><strong>Participation</strong></td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Fully participates in the conversation.</td>
<td>Somewhat participates in the conversation.</td>
<td>Minimally participates in the conversation.</td>
</tr>
<tr>
<td><strong>Commitment</strong></td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Initiates and completes the contact in a timely manner.</td>
<td>Completes the contact in a timely manner.</td>
<td>Does not complete the contact in a timely manner.</td>
</tr>
<tr>
<td><strong>Total Points</strong></td>
<td>/50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix E
List of “DO NOTs”

1. Do not create cumulative Segment 2 exams. Since many students take only one semester with FLVS, segment exams should cover material taught during the current segment and should not be cumulative across the entire course, even in cases where a student must prepare for a cumulative exam outside of FLVS such as an EOC or AP exam.

2. Do not write multiple choice or other objective type exam questions until course content is available to the item writers.

3. Do not group items on a test so that more than one item is selected from a single group for administration to students. For instance, form groups of three items from which one item will be selected; do not put six items together in a single group and choose two of them or nine items in a group and choose three. It is much easier to control difficulty levels in groups of three items.

4. Do not allow for a test completion length longer than 45 minutes (AP and industry certification exams are exempt from this time limit). Divide the test into multiple parts (no more than 3 parts) if necessary to achieve this goal. Students in VLL’s must be able to complete a designated part in 45 minutes or less (do not set a 45 minute time limit but do design the test so that most all students can finish in 45 minutes or less). FLVS tests are not speed tests; grades should depend on mastery of standards and benchmarks, not on speed in completing the test. Tests length must therefore be carefully designed.

5. Do not weight multiple choice items on a single test by cognitive complexity level. For example, do not award five points for a high-complexity multiple choice item but only two points for a low-complexity multiple choice test item.

6. Do not violate the guidelines established by FLVS Global as explained in their document entitled Guidelines for Content Development – FLVS Global. It includes an entire section on Quizzes, Exams and Final Exams and another section on Answer Keys. The link to this document in Box is https://flvs.box.com/s/mbyk2uv8t8c3eeyg540h
Appendix F
Generalized Rubrics

The descriptors in each of these sample rubrics are observable and measurable. They use consistent language across the scale, and indicate the degree to which standards are met. The rows from these 4 point rubrics that can be used alone (if you just want to measure understanding, for example) or in combination with other rows to create an analytic rubric. For instance the understanding and performance quality can be put down one after another as illustrated to create a performance product analytic rubric that measures performance on both characteristics for a single performance product. When designing a new course, writing the objectives in a manner that makes the assessments and rubric objectives align is important practice.

Example of a generalized rubric adapted by FLVS curriculum specialists from the work of Wiggins and McTighe in Understanding by Design: In this rubric, mini checklists are added to the score point descriptors to clarify the expectations for a specific assignment.

<table>
<thead>
<tr>
<th>Performance Criteria/Attribute</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
</table>
| Quality                       | __The performance or finished product is very effective and polished.  
___The ideas are presented and in an engaging and thorough manner.  
___The work shows a strong awareness of the audience and purpose. | __The performance or finished product is effective.  
___The ideas are clear and presented in a thorough manner.  
___The work shows an awareness of the audience and purpose. | __The performance or finished product is effective to some extent.  
___The ideas show some problems with clarity and thoroughness.  
___The work shows little awareness of the audience and purpose. | __The performance or finished product is not effective.  
___The ideas are unclear and confusing.  
___The work shows no awareness of the audience or purpose. |
| Understanding                 | __Shows an advanced understanding of the main ideas or processes.  
___The concepts, evidence, arguments, and methods used to support thinking and show comprehension are advanced.  
___The response goes beyond the grasp of key ideas typically | __Shows strong understanding of the main ideas or processes.  
___The concepts, evidence, arguments, and methods used to support thinking and show comprehension are appropriate.  
___There are few or no misunderstandings of key ideas. | __Shows partial understanding of the main ideas or processes.  
___The concepts, evidence, arguments, and methods used to support thinking and show comprehension are simple or not complete.  
___The response shows some | __Shows little understanding of the main ideas or processes.  
___The concepts, evidence, arguments, and methods used to support thinking and show comprehension are not correct.  
___The response shows major misunderstandings of key ideas. |

If needed, the attribute column can include information to customize the rubric to an assignment.
found at this age level.  misunderstandings of key ideas.

Example of a very simple rubric that measures two criteria – understanding and supporting details

<table>
<thead>
<tr>
<th>Scoring Rubrics</th>
<th>Excellent (4 Point)</th>
<th>Proficient (3 Point)</th>
<th>Average (2 Point)</th>
<th>Poor (1 Point)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Understanding:</td>
<td>Understanding is strong and clear.</td>
<td>Understanding is generally correct.</td>
<td>Limited or weak Understanding.</td>
<td>Question not attempted or is completely incorrect.</td>
</tr>
<tr>
<td>Supporting Details:</td>
<td>Many details to support concept.</td>
<td>Some details show understanding of concept.</td>
<td>Some extra information.</td>
<td>Few or no details, or only slightly related to topic.</td>
</tr>
</tbody>
</table>

Example of a rubric with generalized categories of “Breadth, Depth and Clarity” with a limited amount of customization to the assignment.
Rubric for Cell Tour Writing Assignment:

<table>
<thead>
<tr>
<th>Performance Criteria</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breadth:</strong></td>
<td>Includes descriptions of all of the expected parts of the cell.</td>
<td>Includes descriptions of most of the expected parts of the cell.</td>
<td>Includes descriptions of some of the expected parts of the cell.</td>
<td>Includes descriptions of few of the expected parts of the cell.</td>
</tr>
<tr>
<td><strong>Depth:</strong></td>
<td>Includes accurate and complete descriptions of the organelle structures.</td>
<td>Includes detailed descriptions of the organelle structures.</td>
<td>Includes some detailed descriptions of the organelle structures.</td>
<td>Includes little to no description of the organelle structures.</td>
</tr>
<tr>
<td><strong>Clarity:</strong></td>
<td>The finished product is extremely clear and organized.</td>
<td>The finished product is mainly clear and organized.</td>
<td>The finished product is somewhat clear and organized.</td>
<td>The finished product shows little to no clarity or organization.</td>
</tr>
</tbody>
</table>
## Appendix G
### FLVS Item Writing Guidelines

<table>
<thead>
<tr>
<th><strong>MAKE SURE THE ITEM</strong></th>
<th><strong>MAKE SURE THE RESPONSE OPTIONS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures a predefined benchmark</td>
<td>Include four options with one correct answer and three distractors. Including rationales and references for response options is excellent practice.</td>
</tr>
<tr>
<td>Has content that is accurate, up-to-date and aligned to course content. Users should be able to identify related content in the course.</td>
<td>Exhibit only one correct or “best” answer that can be agreed upon by content experts</td>
</tr>
<tr>
<td>Addresses the desired cognitive level</td>
<td>Include plausible and attractive distracters representing common errors</td>
</tr>
<tr>
<td>Is written clearly, using straightforward vocabulary and sentence structure</td>
<td>Are arranged in logical order; i.e., numerical, alphabetical, chronological, etc.</td>
</tr>
<tr>
<td>Avoids gender, cultural, and age bias</td>
<td>Avoid “all of the above” or “none of the above.”</td>
</tr>
<tr>
<td>Has content that is critically important or frequently used; avoids nice-to-know, trivial, or rote memory items; avoids lifting items straight from the text</td>
<td>Avoid repetitive words. If words are repeated in each option, move them to the stem if possible.</td>
</tr>
<tr>
<td>Has a difficulty (content, math, or reading) level appropriate for the students and purpose of the test; avoids “tricky” or uncommon technical language.</td>
<td><strong>AVOID THESE CLUES</strong></td>
</tr>
<tr>
<td>Is stated positively, avoiding “not,” “except,” etc.</td>
<td>Vary the position of the correct answer in the sequence of options throughout the test. (Alphabetizing randomizes the correct answer)</td>
</tr>
<tr>
<td>Avoids using absolute words such as “all,” “always,” ”only,” and “never.”</td>
<td>Avoid KLANGS - words in the stem that are repeated in one of more of the options thus making those response options potentially more attractive</td>
</tr>
<tr>
<td>Is independent – an item should not have to be completed before answering other items and should not help answer other items</td>
<td>Make options agree grammatically and conceptually with the stem to avoid easy elimination of distracters.</td>
</tr>
<tr>
<td>Exhibits correct grammar, spelling, punctuation, etc.</td>
<td>Make options grammatically and conceptually similar or parallel.</td>
</tr>
<tr>
<td>Follows the FLVS Proofreading Style Guide</td>
<td>Make option lengths similar and as brief as possible. Longer options are often the correct answer.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>MAKE SURE THE STEM</strong></th>
<th><strong>IF USING A DISPLAY, MAKE SURE IT:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Is a complete statement of the question or problem. A content expert should be able to answer the item without seeing the options</td>
<td>Is limited to use with no more than five corresponding items</td>
</tr>
<tr>
<td>Only contains information pertinent to the question or problem unless it is used for a specific purpose</td>
<td>Is necessary for answering the display-linked items</td>
</tr>
<tr>
<td>Is as concise and brief as possible</td>
<td></td>
</tr>
<tr>
<td>Is free of grammatical constructions that aren’t used in regular speech (“and/or,” “he/she,” etc)</td>
<td></td>
</tr>
</tbody>
</table>