Developing High Quality Student Assessments

Edward Roeber

November 2018

Version 2.0

Michigan Assessment Consortium

Introduction

The purpose of this document is to present strategies that educators can use to create common types of educational assessments. Although this guide will present a number of ideas about assessment – those that are effective and those that are not – the best assessments are the result of collaboration among assessment developers and usually consists of an iterative process of development, refinement, tryouts, and further refinement. Hence, the ideas presented here are simply the beginning of a much larger process for creating high quality assessments.

In preparing this paper, several sources were consulted and drawn on liberally. These included item-writing handbooks prepared by the Michigan Department of Education, Sharif Shakrani and others, textbooks, as well as the author.

Table of Contents

Introduction 2
Table of Contents 2
Characteristics of Good Assessments 3
General Guidelines for Writing Assessment Items 5
Tips for Developing Multiple-Choice Items 7
Tips for Developing Constructed-Response Items 11
Tips for Developing Scoring Rubrics 13
Tips for Developing Performance Events 15
Tips for Developing Performance Tasks 17
Considerations for Universally Designed Assessments 19
Appendices 22
Appendices 23
Item Review Criteria 23
Scoring Guide/Rubric Review Criteria 24
Glossary of Terms You Should Know 25
Multiple-Choice Item Development Form 28
Constructed Response Development Form 29
Performance Event Development Form 30
Performance Event Template Directions 32
Performance Task Development Form 34
Performance Task Template Directions 36
Scoring Rubric Template 38

Characteristics of Good Assessments

Good assessments have several distinguishing characteristics:

1. A clear and concise test blueprint was developed and used as the basis for planning the assessment. The assessment specifications in the test blueprint shows how many items of each type are to be used to measure each standard or expectation, as well as the total number of items of each type to be used overall. The blueprint should also describe the purposes of the assessment and how the results will be reported.
2. A clear set of standards, objectives, learning targets, or expectations were used as the basis for developing the assessment items.
3. The assessment items were developed by individuals with sufficient content knowledge to do a good job, and who were trained and mentored in good item construction techniques.
4. The items were edited by persons knowledgeable of the content area and good item development. These individuals will help to assure that the items are clearly worded, are not confusing, have one correct answer, and are stated in a straight-forward manner.
5. The items were also reviewed by persons with content knowledge to assure good alignment with the standards being assessed, and by others who reviewed the items for potential bias and sensitivity.
    a) The test items are unbiased. There is evidence that the test items have been reviewed for bias, using both statistical techniques and expert judgment, and any poor items have been revised or dropped.
6. The assessment items were tried out at least once and found to work:
    a) The items are clearly stated and do not contain any extraneous material not needed to answer the questions, as reviewed by content experts
    b) Students who do well on the overall assessment do well on each item in it.
    c) Each item is correlated with other items measuring the same content more highly than with items measuring different content.
7. The items that comprise the test have been tried out together. There is data to show that the set of items work together effectively.
8. The test is reliable – that is, it consistently measures what it is supposed to measure:
    a) The set of items are internally consistent
    b) If we were to give the assessment multiple times, it would produce the same results
    c) Items within a subtest are more highly correlated with one another than with items in other subtests.
    d) Different forms of the assessment yield comparable results.

9. There is evidence the test results can be used for the purposes intended (validity evidence is available). Types of validity:

    a) Content – the test measures the content standards it is intended to measure.

        i. Both depth and breadth of the standards are assessed.

        ii. Content experts have judged the alignment of the items to the standards

        iii. Two-way alignment between the standards and the test has been shown

    b) Construct – the test can be shown to measure the constructs that it is said to measure by virtue of producing similar results to other comparable assessments.

    c) Predictive – if the test is intended to predict some future behavior, there is evidence that it does so.

    d) Consequential – the test produces information that results in positive consequences for students, educators, and/or schools.

General Guidelines for Writing Assessment Items

1. Each item should measure a single standard, expectation or competency. Make sure the item is really addressing the standard, expectation or competency.
2. The problem or question should be a significant one, not a trivial one.
3. Item content should be of interest to the students being assessed.
4. The items should be written at a grade- or age-appropriate level so that the assessments do not rely on high levels of reading or writing skills in order to answer them.
5. The reading level of the item should be one or more grade levels below that of the students being assessed (unless the items are measuring reading achievement). In school, this should be reading levels one or more grade levels below that of the individuals being assessed. Out of school, try to keep the readability to about the ninth or tenth grades.
6. Carefully consider any technical terms that are introduced in the item. Is this terminology that the individual should be expected to know? If not, do not use it or provide a brief definition of it in non-technical terms.
7. The question should be presented as clearly and straight-forward as possible. Do not use unnecessary words or embed any additional wording in it. Ask a straightforward question that elicits a straightforward response that is consistent with the concept being assessed.
8. Use formal language rather than informal or slang language.
9. Avoid the use of contractions, idiomatic expressions, euphemisms, and clichés.
10. Avoid situations that may be interpreted as stereotypical. For example, if you are using baking as a context for a question, include both boys and girls. Have both boys and girls mowing the lawns, or working with tools.
11. Items should not use any material that would be viewed as offensive by any individuals. This includes political and religious topics.
12. Avoid items that advantage or disadvantage any group of individuals. Be careful about national, state, regional, community-type bias in the assessment items. Watch the balance of items that use rural versus urban settings.
13. Use common names found in the culture(s) being assessed. Using complicated names will only distract individuals being assessed.
14. Avoid references to expensive purchases, vacations, or possessions, since this will introduce socioeconomic bias in the items. For example, ask about a swimming pool in a community center rather in someone’s back yard.
15. Avoid potentially disturbing themes for items, such as death or violence. For example, a reading passage or graph about the number of people attacked by sharks should be avoided.

16. When a stimulus is used with an assessment item, individuals should not be able to answer the item without using the stimulus. If they can, then the stimulus is not needed and should be dropped, or the item should be re-written.
17. When several items are used with one stimulus (e.g., a reading passage, text, or chart or graph), there are several problems to avoid:
    a) One item should not provide clues to the correct answer of another item.
    b) The items must be independent of one another. That is, answering one item correctly should not depend on having answered a previous item correctly.
18. Assessment items should be written across a range of difficulty levels. Most should be average difficulty (in the middle of the range of difficulty). When writing harder items, make sure the difficulty comes from the level of thinking required, not from the use of obscure material or language.
19. Unless called for by the standard being assessed, items should not test the ability to define terms. Assessing understanding of the concept is usually preferred. For example, understanding the concept of federalism is very different from a knowing the definition of the term.
20. Be careful in the use of trademarked terms. These may require special permission to use them in assessment items, so be sure that the use of them (rather than their generic terms) is actually required in an assessment item. For example, is Clorox bleach really needed, or could chlorine bleach be used instead?
21. Use grammatical forms appropriate to standard written English. Watch out for problems such as:
    a) Incorrect subject-verb agreement: “Each of the boys were….” One of Mary’s friends were….” None of the sides are….”
    b) Incorrect pronoun-antecedent agreement: “The teacher asked each student to take out their book.”
    c) Comma splices
    d) Dangling modifiers
    e) Inconsistent verb terms

Tips for Developing Multiple-Choice Items

1. Provide an overview of the standard (content standard or performance standard) that will be assessed. It may be helpful to write this out in its entirety so as to keep this at the forefront as the item is being written.

2. The alignment of the item to the standard is essential. When the development of the item is complete, check to make sure that the item still measures the standard. Make sure that the item measures a very important piece of knowledge or skill, that it measures the "heart" of the standard.

3. A multiple-choice item should be written either as a question or an incomplete sentence. Do not use "fill in the blank" items.

Example with a question stem:

The development of the National Park System in the United States was closely related to which nineteenth-century reform movement?

A Labor

B Conservation*

C Business regulation

D Rural electrification

Example with a partial-sentence stem:

The development of the National Park System in the United States was closely related to the nineteenth-century reform movement that focused on

A labor.

B conservation.*

C business regulation.

D rural electrification.

4. The essence of the problem should be in the stem - not in the answer responses.

Example:

Washington was:

A. Admitted to statehood in 1989.

B. Part of the Territory of Oregon until 1912.

C. Made a state in 1811.

D. Part of the Northwest Territory until 1848.

Rewritten:

In what year did the State of Washington enter the Union?

A. 1889

B. 1811

C. 1848

D. 1912

5. Do not repeat a word of phrase in the options that could easily be incorporated into the stem.

Example

The development of the National Park System in the United States was closely related to the nineteenth-century reform movement that was

A focused on labor.

B focused on conservation.*

C focused on business regulation.

D focused on rural electrification.

Simply moving "focused on" to the stem corrects this issue:

The development of the National Park System in the United States was closely related to the nineteenth-century reform movement that was focused on

A labor.

B conservation.*

C business regulation.

D rural electrification.

6. State the stem and answer choices clearly and concisely. Once you draft the item, see if you can edit it down to fewer words and not lose or change the meaning of the stem.

Example:

In carrying out scientific research, the type of hypothesis that indicates the direction in which the researcher expects the results to occur once the data has been analyze is known as a(n)

Rewritten:

A hypothesis that indicates the expected direction of the result of an study is a(n)

7. 7. Be careful not to put information in the stem that will clue the correct response.

8. Avoid writing a stem in which the answer is the only option that makes sense given the context of the stem.

    a) Do not repeat the same word, phrase, or concept in both the stem and the correct answer option to avoid cluing the answer.

    Example: The stem asks about birds and only the correct answer refers to bird nests.

    b) Make sure all response options, not just the correct answer, fit grammatically, logically, and semantically with the stem. Individuals can easily eliminate a distracter that does not fit the stem. Conversely, individuals can be clued to the correct answer if only the correct answer fits the stem.

    Example: The stem asks for a plural response but one or more distracters are singular.

    Example: A test that can be scored by counting the correct responses is an ___________ test.

    A consistent

    B objective

    C stable

    D standardized

9. Do not write stems with phrases that ask for opinions, since any option could be considered correct for such questions.

    Example: "What do you think is the best…." "Why do you think…."

10. Word the stem in the positive whenever possible. Avoid negatively phrased stems.

11. In the stem, use which, not what, when there is more than one correct answer other than the correct choice you have listed. Use what, not which, when there is only one possible correct answer.

    Example: "Which of the following is the best reason for…."

12. Distracters should be plausible as well as attractive to students

    a) Distracters should represent classic mistakes that individuals might make or common misunderstandings they may have about a subject or concept.

    b) Do not use "throwaway" options – ones that are illogical, impossible, or humorous.

    c) Do not use distracters such as: "all of the above." "none of the above," "A and B only," All but C."

13. Make sure that all answer options are mutually exclusive and do not overlap.

Example

Zack wants to buy a pair of shoes that cost $69 including tax. He has already saved $53. When he receives his allowance this week, Zack will have enough money to buy the sneakers. Which amount could be Zack's weekly allowance?

A $15

B $16

C $17

D $18

In this example, the last three answer options are correct

14. Be careful not to make the correct answer conspicuous
    a) One of the major pitfalls item writers encounter is ensuring that the correct answer is specific enough to be correct while leaving the distracters more general.
    b) Stating the correct answer in greater detail or greater length than the other options gives the answer away.
    c) Avoid stating the correct answer in technical or textbook language, while the distracters are written more informally.
    d) Avoid making any one option, including the correct answer, unique in any way (e.g., the correct answer is stated in the positive and the distracters are stated in the negative; the correct answer is concrete, the distracters are abstract).
15. Response options should be parallel in grammatical structure (e.g., all noun phrases, or all verbs followed by objects)
16. Be careful not to put information in the distracters that will clue the correct answer.
    a) Avoid specific determiners (absolute terms) when possible. Examples include such terms as always, never, totally, all, none, absolutely, and completely. When a specific determiner must be used, maintain parallelism by using a specific determiner in two or all of the options.
    b) In general, do not use the opposite of the correct answer as one of the distracters.
    c) In a group of related items, avoid using the same distracter several times.
17. Whenever possible, put response options in a logical order. For example, arrange numerical options in ascending or descending order, except when doing so would clue the correct response. Put options in chronological order if one exists.
18. The response options affect the difficulty of the item. The finer the distinctions to be made among the options, the more difficult the item. Use fine distinctions among options sparingly.

Tips for Developing Constructed-Response Items

1. Provide an overview of the standard (content standard or performance standard) that will be assessed. It may be helpful to write this out in its entirety so as to keep this at the forefront as the item is being written.
2. The alignment of the item to the standard is essential. When the development of the item is complete, check to make sure that the item still measures the standard. Make sure that the item measures a very important piece of knowledge or skill, that it measures the "heart" of the standard.
3. Keep the stem of the constructed-response item as brief and succinct as possible.
4. Avoid teaching students about the concept(s) to be assessed. That said, if there is terminology used in the item not required to measure the standard, then briefly define this for students (or avoid using such terms in the first place).
5. Make sure that any stimulus materials used are really needed to respond to the stem. Try answering the question without the stimulus. If the item can be answered correctly without the stimulus materials, either drop the stimulus materials or re-write the stem.
6. A single constructed-response item may have several parts, but ask only one question in each part. If more than one part is used, make sure that each is independent of the other parts (e.g., it is not necessary to answer the first part in order to correctly answer the second part).
7. Keep the reading load (and level) of the stem as low as possible so that all individuals will be able to understand the task and carry it out. Adapt it to the level of the student.
8. Indicate the expected length of response in one of two ways – directly saying so or by the number of lines (and pages) provided.
9. Indicate to the student taking the item how individuals' responses will be scored in general terms. For example, are grammar, spelling and punctuation going to count? If so, say so. Is creativity called for? Say that it is.
10. Responses need not be just written responses
    - Visual art: Drawings, sketches
    - Music: Compose music
    - Science: Draw a graph that describes data collected in an experiment
    - Mathematics: Respond to a word problem with the steps needed to solve a problem
    - Social studies: Construct a map showing the states in the Midwest
    - Reading: Construct a concept map for a story that the student is asked to read
11. There are several formats for using constructed-response items:
    - Comparison of two or more things:
        - "Compare norm- versus criterion-referenced measurement."
    - Causes or effects of a situation
        - "What caused the entry of the U.S. into World War I?"

- Analysis of a situation:
  - “Why is it better to use a criterion-referenced test to measure student achievement of content standards?”
- Discussion of an idea
  - “Discuss the value of the United Nations to the U.S. in the following areas:
    - World peace
    - Children’s care
    - World health”
- Reorganization of facts
  - “Trace the development of industrial (versus laboratory) preparation of nitric acid.”
- Formulation of a new question (problems and questions raised)
  - “If the crunch in the U.S. credit markets continues, and if this situation spreads to both Asia and Europe, what is the likely scenario for a small business owner in the U.S?”
- Critique of the accuracy of a printed statement
  - “Accountability by test distorts the education of students and corrupts the system being held accountable.”

12. The scoring rubric should focus on one or more important dimensions of the task asked of individuals. Make sure that the scoring dimension are aligned to the standard being measured as well as the assessment(s) being used to measure the standard.

13. Consider the dimensions of the scoring rubric when writing the item, to assure that the rubric and the item are aligned.

14. Do not use the constructed-response item format for questions that only have one correct response – use the multiple-choice item format instead.

Tips for Developing Scoring Rubrics

1. Determine the types of responses that the item will elicit. Will they be short or lengthy? If short, consider a two-point rubric; if lengthy, a four- or six-point rubric may be feasible and desirable.
2. Make sure that the rubric is aligned with the item, which should be aligned with the standard being measured. In other words, the rubric should be aligned to the standard.
3. Consider the different types of scoring rubrics and select the most suitable one(s) to use:
    - Primary-trait – The most important trait being measured is used to set up the scoring rubric
    - Holistic – The scorer is asked to form an overall judgment on the student work and grade accordingly
    - Analytic – Various characteristics of a piece of student work are judged more or less independently and each is graded separately.
    - Norm-referenced versus criterion-referenced scoring – Are responses “forced” into a normal curve or allowed to “float” independently?
4. The rubric should measure one or more important dimensions of the responses from students.
5. The rationale for the correct answer and how this is developed across the levels of the rubric should be very clear.
6. Would experts in the field agree that you have defined the important dimensions of the essay question to be scored?
7. Use descriptive language to characterize the answers at each level of the rubric.
8. Avoid using terms such as “some,” “many,” “much,” “a lot” and so forth, since these are open to differences in interpretation.
9. Consider the different ways in which a student response might be scored. Should you use separate rubrics for these or combine them into one rubric? If the dimensions are independent of one another, then use separate rubrics.
10. Determine how many rubrics are needed to adequately judge the responses of students. Keep the number of rubrics per item reasonable. Remember, someone has to actually use them.
11. Have you defined a useful scale – can persons without extensive training actually use it, or is it so complex that it cannot be given to non-experts to use.
12. How teachers should score essay questions:
    - Check the model response(s) against those that the students provided - is it still relevant? Did students respond as expected?
    - Be consistent in grading - occasionally return to the first few papers to see if you have “drifted” in scoring.
    - Shuffle the papers randomly before you score.
    - Grade only one question at a time for all students.

- Shuffle the papers again before scoring the next question.
- Try to grade all papers for one question without interruption - to reduce variability in scoring.
- Grade the responses without knowing the students' identities - to avoid benefiting or hurting any students.
- Grade the mechanics of expression separately from the content - two different passes through the papers.
- If possible, have two independent readings of each paper, and use the total or the average of the two scores as the final rating.
- Provide comments and correct errors, in order to help students improve their performance.
- Set realistic standards, for the topic and for the maturity of the students.

Tips for Developing Performance Events

1. Provide an overview of the standard (content standard or performance standard) that will be assessed. It may be helpful to write this out in its entirety so as to keep this at the forefront as the item is being written.
2. Summarize the flow of the performance event for the assessment administrator. The summary should describe the resources needed to administer the assessment, the manner in which the assessment will flow, how students' responses will be recorded (and perhaps, scored) during the assessment.
3. Make sure to write out the entire performance assessment, step by step. Leave nothing to the imagination or determination of the assessment administrator.
4. Make sure that any materials needed to administer the assessment are carefully described (if they are to be provided by classroom teachers), described sufficiently to be acquired by the school, or provided in a testing kit.
5. Be sure to specify where the assessment will occur. Will it take place in the regular classroom, in a special room (if so, what makes it "special?"), or elsewhere? This needs to be spelled out in detail so that the assessment is administered in a standard manner across administration sites.
6. Are there any steps that the assessment administrator needs to take to set up the assessment situation in advance? How far in advance? Specify this in detail.
7. Define for the assessment administrator (e.g., the teacher) that students are to be assessed. Is it all students in a class or are students to be sampled in some fashion? If the latter, describe the manner for drawing the sample of students in the classroom or school.
8. Write the exercise in such a manner that the directions to the test administrator and those to the student(s) can be distinguished. One way to do this is to put directions to the test administrator in regular type and those to the student in bold print.
9. In some cases, where the item requires a series of activities to be carried out, it may be helpful to provide students with a checklist that they can use to make sure that they complete all parts of the item.
10. Be sure to give the assessment administrator complete directions on what they should say and what they cannot say, in case students do not respond, or their responses are off target or incomplete. This is important especially because students might 1) not respond, 2) respond off topic, or 3) respond with a very brief response. Is the assessment administrator allowed to "probe" these responses? If so, are they to use a standard probe that is given to them, or "ad lib?"
11. Provide detailed directions on how students' responses will be recorded. Will the assessment administrator use a video camera? Audio recorder? Write out their observations? Write out what students say (or describe what they do)? Collect work from students?

12. Students' responses will probably be scored according to a rubric. Determine what students should be told about how their responses will be scored. This may be a mention of the dimensions on which they will be scored, a summary of the rubric, an outline of the rubric, or a more complete description or even a handout showing a copy of the rubric. Remember, while it is fair to provide students with at least a summary of how their responses will be scored, a detailed rubric may simply make the item harder to respond to.

13. Students might be asked to self-assess their response. This may require information on the dimensions on which they are to judge their responses, as well as a form on which to record their responses.

14. Students might be asked to self-reflect on their learning from having participated in the performance assessment. Again, a form on which to record their responses may be needed.

Tips for Developing Performance Tasks

1. Provide an overview of the standard (content standard or performance standard) that will be assessed. It may be helpful to write this out in its entirety so as to keep this at the forefront as the item is being written.
2. Performance tasks usually include several parts, some of which may be carried individually, with other parts worked on in small groups. Some of the task may be done in school and other parts outside of school. Thus a summary of the steps in the task, each explained with a brief summary, is essential. Summarize the flow of the performance task for the teacher and the student. The summary should describe the resources needed to complete the assessment, the manner in which the assessment will flow, what is asked of students, the nature of students' response, how students' responses will be provided (and perhaps, scored) during the assessment. There may be different types of responses for different parts of the task.
3. Make sure to write out the entire performance task, step by step. Leave nothing to the imagination or determination of the student or the teacher. Consider that one summary is probably needed for the teacher as well as for the student. The teacher summary can describe what to say to students, what to say to students, and help to be offered.
4. Make sure that any materials needed to administer the assessment are carefully described (if they are to be provided by classroom teachers), described sufficiently to be acquired by the school, or provided in a testing kit.
5. Be sure to specify where the assessment will occur. Will it take place in the regular classroom, in a special room (if so, what makes it "special?"), or elsewhere? This needs to be spelled out in detail so that the assessment is administered in a standard manner across administration sites. What portions of the task will be done in school and what parts outside of school. How should the teachers monitor the out-of-school work?
6. Are there any steps that the assessment administrator needs to take to set up the assessment situation in advance? How far in advance? Specify this in detail.
7. Define for the assessment administrator (e.g., the teacher) the students that are to be assessed. Is it all students in a class or are students to be sampled in some fashion? If the latter, describe the manner for drawing the sample of students in the classroom or school.
8. Write the exercise in such a manner that the directions to the test administrator and those to the student(s) can be distinguished. One way to do this is to put directions to the test administrator in regular type and those to the student in bold print. Remember for the task, there probably will be a teacher guide (with directions to them and directions about what is to be read to students) and a student booklet, containing the directions for the task, and in some cases, a place for them to record their responses.

9. In some cases, where the item requires a series of activities to be carried out, it may be helpful to provide students with a checklist that they can use to make sure that they complete all parts of the item.

10. Be sure to give the assessment administrator complete directions on what they should say and what they cannot say, in case students do not respond, or their responses are off target or incomplete. This is important especially because students might 1) not respond, 2) respond off topic, or 3) respond with a very brief response. Is the assessment administrator allowed to "probe" these responses? If so, are they to use a standard probe that is given to them, or "ad lib?"

11. Provide detailed directions on how students' responses will be recorded. Will the assessment administrator use a video camera? Audio recorder? Write out their observations? Write out what students say (or describe what they do)? Collect work from students?

12. Students' responses will probably be scored according to one or more rubrics, and different rubrics might be used for different parts of the task. Determine what students should be told about how their responses will be scored. This may be a mention of the dimensions on which they will be scored, a summary of the rubric, an outline of the rubric, or a more complete description or even a handout showing a copy of the rubric. Remember, while it is fair to provide students with at least a summary of how their responses will be scored, a detailed rubric may simply make the item harder to respond to.

13. Students might be asked to self-assess their response. This may require information on the dimensions on which they are to judge their responses, as well as a form on which to record their responses.

14. Students might be asked to self-reflect on their learning from having participated in the performance assessment. If so, a form on which students record their responses may be needed.

Considerations for Universally Designed Assessments

Do the items:

Measure what it intends to measure?

- Reflect the intended content standards
- Minimize skills required beyond those being measured

Respect the diversity of the assessment population?

- Accessible to test takers (consider gender, age, ethnicity, and socio-economic level)
- Avoid content that might unfairly advantage or disadvantage any student subgroup

Have clear format for text?

- Standard typeface
- Twelve (12) point minimum for all print, including captions, footnotes, and graphs (type size appropriate for age group)
- Wide spacing between letters, words, and lines
- High contrast between color of text and background
- Sufficient blank space (leading) between lines of text
- Staggered right margins (no right justification)

Have clear pictures and graphics (when essential to items)?

- Pictures are needed to respond to the items
- Pictures with clearly defined features
- Dark lines (minimum use of gray scale and shading)
- Sufficient contrast between colors
- Color is not relied on to convey important information or distinctions
- Pictures and graphs are labeled

Have concise and readable text?

- Commonly used words
- Vocabulary appropriate for grade level
- Minimum use of unnecessary words
- Idioms avoided unless idiomatic speech is being measured
- Technical terms and abbreviations avoided (or defined) if not related to the content being measured
- Sentence complexity is appropriate for grade level
- Question to be answered is clearly identifiable

Allow changes to its format without changing its meaning or difficulty (including visual or memory load)

- Allows for the use of Braille or other tactile format
- Allows for signing to a student
- Allows for the use of oral presentation to the individual
- Allows for the use of assistive technology
- Allows for translation into another language

Does the test:

Have an overall appearance that is clean and organized?

- All images, pictures, and text provide information necessary to respond to the item
- Information is organized in a consistent manner (left-right, top-bottom flow)

In addition to the other considerations, a computer-based test should have these considerations

Layout and Design

- Sufficient contrast between background and text and graphics
- Color is not relied on to convey important information or distinctions
- Font size and color scheme can be easily modified (through browser settings, style sheets, or on-screen options)
- Stimulus and response options are viewable on one screen when possible
- Page layout is consistent throughout the test
- Computer interfaces follow Section 508 guidelines

Navigation

- Navigation is clear and intuitive; it makes sense and is easy to figure out
- Navigation and response selection is possible by mouse click or keyboard

Screen reader considerations

- Item is intelligible when read by a text/screen reader
- Links make sense when read out of visual context (“go to the next question” rather than “click here”)
- Non-text elements have a text equivalent or description
- Tables are only used to contain data, and make sense when read by screen reader

Test-specific options

- Access to other functions is restricted (e.g., e-mail, Internet, instant messaging)
- Pop up translations and definitions of key words/phrases are available if appropriate to the test
- Individuals are able to record their responses and read them back (and have them read back using text-to-speech) as an alternative to a human scribe, but only if the individual has experience with this mode of expression and chooses it for the test

Computer capabilities

- Adjustable volume
- Speech recognition available (to convert user’s speech to text)
- Test is compatible with current screen reader software
- Computer-based option to mask items or text (e.g., split screen)
- Computer software for test delivery is designed to be amenable to assistive technology

Source Thompson, S.J., Johnstone, C.J., Anderson, M.E., & Miller, N.A. (2005) Considerations for the development and review of universally designed assessments (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Appendices

Appendices

Item Review Criteria

1. Is the task aligned with content standard and performance indicator? Did it measure important aspects of these?
    - Does the item measure the intended content standard and performance indicator?
    - Does it measure other standards or indicators as well?
    - Does it focus on important or trivial aspects of these?
    - Is it worth the time and expense to administer and score?
2. Was the task understandable?
    - Did students appear to understand what was asked of them?
    - Were they able to do what was asked of them?
    - Are there any “trick” aspects to the item?
3. Was the task engaging?
    - Was the item interesting to students?
    - Did it engage them sufficiently that they appeared to give it enough attention?
4. Is the task appropriate?
    - Is the task age-appropriate for students?
    - Can they be reasonably expected to perform well on it?
    - Should students who received adequate amounts and types of instruction do well on the item?
    - Is the concept better assessed with another item format?
5. Does the task clearly communicate to students what is asked for and how they will be evaluated?
    - Did students understand the type(s) of responses they were expected to provide?
    - Did students understand the dimension(s) of the responses they were to provide?
    - Did the item indicate how student responses would be evaluated?
6. Is the task fair to all students?
    - Is the task something that all students would be equally knowledgeable about?
    - Have all biases (gender, social class, racial-ethnic, regional, community-type, and so forth) been avoided?
7. Is the task repetitive?
    - Is the item overly repetitive of an item type that is not needed?
    - Does the item have too many parts?
8. Is the task worth further development?
    - Did this item produce information worthy of the time and effort needed to administer and score it?
    - Is the time needed to continue to revise and refine the item worthwhile?

Scoring Guide/Rubric Review Criteria

1. Did the item elicit the types of student responses anticipated?
   - Did the item lead students to produce the quality and quantity of responses anticipated when the item was written?
2. Does the scoring guide measure what the item asked for? What the content standard and performance indicator suggest? Is there alignment?
   - Is the scoring guide aligned to the content standard and the performance indicator?
   - Is the scoring guide aligned to the item (that is, does it measure the important dimensions of the item)?
3. Is the scoring guide based on appropriate dimensions of the student responses to be elicited by the item?
   - Does the scoring guide measure a combination of dimensions (or separate dimensions) of the item appropriately?
   - Are the dimensions appropriately weighted in the scoring rubric
4. Is the scoring guide, particularly the rationale for correct response and the score level descriptors, clear?
   - Clearly written rationale statements and descriptions of each score level are needed in order to train scorers of the item.
   - Are these clear and understandable without sample student responses?
5. Does the task produce responses with the range of quality needed? Number of levels anticipated?
   - If the item produces only one level of performance, or no responses that are correct or incorrect, there may be problems with the wording of the item.
   - Does the item produce responses that can be categorized as excellent, good, fair, and poor?
   - Is this the number of levels intended when the item was written?
6. Can student responses be scored in more than one way?
   - On which dimensions can student responses be scored?
   - Should these be combined within one rubric?
   - Should multiple rubrics be used? If so, are these truly independent of one another?
   - Is there one rubric that uses more important criteria to judge student responses?
7. Is the scoring guide useful?
   - Does the scoring guide contain a rating scale that can easily be used?
   - Can training readily be provided to scorers so that they can quickly learn to score student responses?
8. Which student examples can serve as exemplars of each scoring category?
   - Unless the item has been extensively revised based on the tryouts, several of the student responses should be selected to exemplify each score level of the scoring rubric.

Glossary of Terms You Should Know

There are terms that refer to assessment items or their parts will often be used. In order to standardize how we talk about assessment items, the following terminology will be used. However, not all of these terms will be pertinent to each development project.

Item – An assessment question, problem, or exercise is often referred to as an item.

Item Format – This refers not only to the type of item, but also to its appearance on the assessment instrument. Item types include multiple choice, short constructed-response, extended constructed-response, performance, and performance task items. The format refers to the layout of the item on a page or computer screen, including for example, text and visual stimuli and how they are presented and labeled, as well as the items themselves.

Selected-Response Item – In this type of item, students select a correct answer from among several answer choices. This item type includes multiple-choice, true-false items, and matching items. The multiple-choice item format is the selected-response format most suitable for large-scale assessment programs.

> **Multiple-Choice Item –** This item type includes, as a minimum, a stem that can be an incomplete sentence or a question, and four viable answer choices, only one of which should be correct. This item type may or may not be accompanied by a stimulus.
>
> **Direct Question –** Students are asked a question in the stem and it concludes with a question mark.
>
> **Sentence Completion –** The four foils are used to complete the stem. One will do so correctly.

Constructed-Response Item – This item type requires the individual to create their own answer(s) rather than select from prewritten options. These items are open-ended, that is, there are usually several ways in which they can be answered correctly. Sometimes, they have only one or two specific answers. These items allow the awarding of partial credit for partial knowledge or partial completion of the task. These items are scored using a standardized scoring rubric that is objective and clearly defined.

> **Short Constructed-Response Item –** This is a constructed-response item in which individuals' responses are short – typically, a word, a phrase, and one or more sentences, up to a paragraph. This item might also call for students to work a brief problem, fill in the blank, make a quick sketch, and so forth. Each item of this type may take 1 to 5 minutes to complete.
>
> **Extended Constructed-Response Item –** This is a type of constructed-response item in which individuals' responses are lengthy. This could include an essay of several paragraphs (or more), an extensive problem to be solved, the development of a procedure to be carried out, and so forth. Each item of this type may take 15, 20 or more minutes to complete.

Performance Assessments – These are types of assessments that require the student to perform some activity. There are two types, distinguished by their complexity and the length of time students have to respond to them.

> **Performance Task –** On this assessment, students have days, weeks, or months to compose a response. Thus, these assessments may involve multiple responses of different types to multiple prompts. The resultant work may be lengthy and comprised of multiple parts. Embedded in the Task may be written response items, presentations, papers, student self-reflections, and so forth.

Performance Event – This is an on-demand performance assessment on which students are given little or no time to rehearse their performance nor limited opportunities to improve their initial performance. Such assessments may take a class period or less to administer.

Stimulus (Lead)– This refers to the text or artwork that conveys information to be analyzed before reading and answering the actual assessment item. Stimuli may accompany any type of item. If extensive (and time consuming to read), it is best if the stimuli is tied to a cluster of assessment items).

Text Stimulus – This refers to the text or passage that precedes an assessment item may provide context or additional information for the individual to use in determining their answer. Typically, items that require more complex thinking skills will require this context or information for individuals to analyze. This may be source material pertinent to the assessment situation.

Visual Stimulus – A diagram, chart, table, graph, photograph, illustration, or cartoon that must be analyzed in order to answer an assessment item. Visual stimuli can be of particular value if they are real source material used in the field being assessed. They are also effective with those individuals who have lower English language proficiency. However, care must be taken in using visual stimuli to make sure that they are accessible to visually-impaired individuals.

Prompt – Another word for the stimulus that students are asked to respond to.

Stem – This is the part of the assessment item that elicits a response by asking the individual to select, identify, describe, explain, solve, locate, choose, calculate, or evaluate. It is a part of both multiple-choice and constructed-response items. It may be accompanied with visual or text stimuli.

Direct Question – This is one or more sentences that end with punctuation, either a question mark or a period. In multiple-choice items, closed stems always end with a question mark.

Sentence Completion – This is one or more sentences that the response options complete. Leave the end of the open sentence blank (do not use a dash or an underline).

Response Options – In multiple-choice items, these are the prewritten answer choices and may consist of text, numbers, or graphics (e.g., artwork).

Correct Answer – The one best or clearly correct answer for a multiple-choice item.

Distracters or Foils – The incorrect response options in a multiple-choice item. These should be plausible or partially correct responses and are best when they represent classic mistakes that individuals might make or common misunderstandings they might make about the subject or concept.

Scoring Rubric – A set of guidelines for scoring constructed-response items. The rubric establishes the expectations for performance and delineates what a response must include in order to earn each of the possible points. Rubrics may offer 2, 3, 4, or more points. The requirements of a response to obtain each score point is described in the rubric, as well as sample papers drawn from actual responses used to illustrate responses at each level of the rubric.

Sample Responses – Actual individual responses to a constructed-response item used to illustrate each point level in the scoring rubric. These are very useful in training scorers, as well as in communicating the types of responses at each score level in the rubric.

Reading Level – Refers to grade level of reading established through one or more readability formulae. This serves as a rough guide to how accessible the item (including any stimuli) are to the individuals being assessed. These are rough approximations due to shortness of most assessment text being analyzed.

Cognitive Complexity – The type(s) of mental processing (i.e., thinking skills) required by an item or set of items. This may refer to the Depth of Knowledge (Webb), Bloom's Taxonomy, or other definition of thinking skills.

Standard/Expectation/Competency/Learning Target – The specific skill to be assessed. These can be referred to by any of these terms, or other terms as well (e.g., goal, skill, objective, outcome, and so forth).

Alignment – This refers to whether an assessment item measures any (ideally, the most important part) of a content standard, expectation, or competency. It also refers to how much of a set of content standards, expectations, competencies, or other set of skills that an assessment instrument measures.

> **Two-Way Alignment –** Refers to how much of the set of content standards, expectations or competencies is measured by the assessment instrument as well as whether the assessment instrument covers most if not all of the set of content standards, expectations, or competencies.

Item Difficulty – This refers to the percentage of individuals assessed who answer correctly or at a level deemed to be "passing" on a constructed-response item. Typically, this is express either as a percentage that ranges from 0% to 100%, or as a decimal fraction, which is also called a p-value.

P-Value – This is the decimal fraction that refers to the percent of individuals assessed who answer correctly or who received a score on a constructed-response item deemed to be correct or passing.

Cluing – Items in a test where one item will provide a clue or the answer for a subsequent item. This can occur through the item itself or accompanying stimulus materials. Cluing compromises the reliability and validity of the assessment.

Cuing – This is the use of a term in the stem and only the correct answer (in none of the distracters) which enables a test-wise individual to answer correctly without knowing the concept being assessed. This compromises the reliability and validity of the assessment.

Achievement Level – The standard of performance set through a standard-setting procedure. Also called a "performance standard." Defines how well students need to do on an assessment to meet or exceed predefined targets for achievement such as "proficient."

Multiple-Choice Item Development Form

Content Area ________________**Grade(s)** _____________**Developer** ______________________________

1. Stand-alone item or item cluster? Stand Alone Item Cluster ______________

2. Performance Standard(s) Assessed

3. Materials Required (List)

4. Stimulus/Stimuli Used in the Item

5. Stem

6. Responses (intended correct answer should be option A

   A.

   B.

   C.

   D.

   E.

7. Source(s) for Prompt Materials

Constructed Response Development Form

Content Area: D M T VA Grade: 5 8 HS Developer ________________________________

1. Performance Standard(s) Assessed

2. Overview of the Constructed-Response Item

3. List of Materials Required

4. Stimulus/Stimuli Used in the Task

5. Stem

6. Response Space (Number of lines; blank space, etc.)

7. Intermediate and Final Student Products to be Developed

8. Student Checklist and/or Reflection/Reaction Sheet (Optional)

9. Scoring Rubrics and Scoring Form(s)

10. Source(s) for Prompt Materials

Performance Event Development Form

Content Area ______________Grades ________Item Number ______ Developer ___________________

Editor ______________________

1. Performance Standard(s) Assessed

2. Overview and Outline of the Performance Event

3. List of Materials Required

4. Assessment Set-Up

5. Stimulus/Stimuli Used in the Task

6. Detailed Script with Teacher and Student Directions

Part 1

Directions to Teacher

Directions Read to Student

Part 2

Directions to Teacher

Directions Read to Student

Part 3

Directions to Teacher

Directions Read to Student

7. Probes that Teacher Can Use

Directions to Teacher

Probes Used with the Student

8. Intermediate and Final Student Products to be Developed

9. Student Checklist and/or Reflection/Reaction Sheet (Optional)

10. Teacher Scoring Rubrics and Scoring Form(s)

11. Sources for Copyrighted Materials

After drafting this item, respond to these questions:

12. What declarative knowledge and/or content is necessary to successfully complete this PE?

13. What Michigan GLCE(s) does this PE measure (please indicate no more than 5)?

Performance Event Template Directions

1. Enter the complete Performance Standard Assessed in this area, and make sure that your Event is related to this standard.
2. Provide a summary of the steps involved in the Event in the **Overview and Outline of the Performance Event** field. This field will be useful to persons not familiar with the Event to see the "flow" of it and be able to plan for the time needed to administer the Event.
3. Provide a complete list of the materials needed to administer the Event in the **List of Materials Required** field. These may be materials that teachers or students are expected to provide and/or they may be materials that will be provided for the teachers. Be sure to describe them completely so that the materials needed will be available for use in the assessment.
4. Describe the **Assessment Set-Up** completely, since the assessment administrators will not be familiar with the Event when they first administer it. The **Set-Up** includes space for students to perform and/or respond to the Event.
5. **Stimulus/Stimuli Used in the Event** refers to any additional materials used in the assessment. This might be a video, an audio clip, handouts, and so forth – anything beyond a student assessment booklet.
6. The **Detailed Script with Teacher and Student Directions** is where the steps in administering the assessment will be given. This section contains two types of directions – those for teachers administering the assessment (but not read to students) and those given to students – either verbally during the assessment or printed and given to students in a student assessment booklet, or both. This section may be lengthy, since some Events contain multiple parts, each requiring directions for both teachers and students. Teacher directions should be printed in plain text, while text to the read to students should be in **bold print**.
7. Probes that teachers can use for students who need assistance or are "stuck" should be described, especially if they are unique to the event.
8. The **Intermediate and Final Student Products to be Developed** is where you describe what students are asked to produce, either as final products or as interim ones on the way to such final products. Describe these clearly.
9. The **Student Checklist and/or Reflection/Reaction Sheet (Optional)** is just that – optional. For some Events that have multiple parts and/or for which students must carry out multiple steps to respond to them, it may be helpful to provide a **Checklist** for students' use to make sure that carry out other activities. For other Events, it may be important to elicit students' **Reflections/Reactions** on the processes they used in completing the Event and/or what they learned in doing so. Both (or either) may be used.
10. The **Teacher Scoring Rubrics and Scoring Form(s)** section is where you indicate the rubric(s) that will be used to score students' responses. Determine first which parts of the Event are to be scored, whether each part has one or more than one rubric for scoring the responses, and then, whether there is any overall score (across all

parts of the Event) to be given as well. Be sure to specify not only the name of the rubric but also how many levels each rubric will have (e.g., label these as “1” for the lowest level and “4” for the highest level). Then provide a description of each of the performance levels.

11. The **Sources for Copyrighted Materials** should be filled in. It is essential that we know of any copyrighted materials you used and where you found them so we can secure permissions to use the materials.

12. **What declarative knowledge and/or content is necessary to successfully complete the PE?** Note: these may form the basis of multiple-choice items to be written later. Note the important concepts that we want to make sure that students know.

13. **What Michigan content standards does this PE measure?** Be sure to note up to five Michigan standards measured by this PE.

Performance Task Development Form

Content Area ________________Grades ______Item Number ______ Developer ___________________

Editor _______________________

1. Performance Standard(s) Assessed

2. Overview and Outline of the Performance Task

3. List of Materials Required

4. Assessment Set-Up

5. Stimulus/Stimuli Used in the Task

6. Detailed Script with Teacher and Student Directions

Part 1

Directions to Teacher

Directions Read to Student

Part 2

Directions to Teacher

Directions Read to Student

Part 3

Directions to Teacher

Directions Read to Student

7. Probes that Teacher Can Use

Directions to Teacher

Probes Used with the Student

8. Intermediate and Final Student Products to be Developed

9. Student Checklist and/or Reflection/Reaction Sheet (Optional)

10. Teacher Scoring Rubrics and Scoring Form(s)

11. Sources for Copyrighted Materials

After drafting this item, respond to these questions:

12. What declarative knowledge and/or content is necessary to successfully complete this Task?

13. What Michigan content standards does this Task measure (please indicate no more than 5)?

Performance Task Template Directions

1. Enter the complete **Performance Standard(s) Assessed** in this area, and make sure that your Task is related to this standard.
2. Provide a summary of the steps involved in the Task in the **Overview and Outline of the Performance Task** field. This field will be useful to persons not familiar with the Task to see the "flow" of it and be able to plan for the time needed to administer the Task.
3. Provide a complete list of the materials needed to administer the Task in the **List of Materials Required** field. These may be materials that teachers or students are expected to provide and/or they may be materials that will be provided for the teachers. Be sure to describe them completely so that all of the materials needed will be available for use in the assessment.
4. Describe the **Assessment Set-Up** completely, since the assessment administrators will not be familiar with the Task when they first administer it. The **Set-Up** includes space for students to perform and/or respond to the Task.
5. **Stimulus/Stimuli Used in the Task** refers to any additional materials used in the assessment. This might be a video, an audio clip, handouts, and so forth – anything beyond a student assessment booklet.
6. The **Detailed Script with Teacher and Student Directions** is where the step-by-step directions to administer the assessment task will be given. This section contains two types of directions – those for teachers administering the assessment (not read to students) and those given to students – either verbally during the assessment or printed and given to students in a student assessment booklet, or both. This section may be lengthy, since many Tasks contain multiple parts, each requiring directions for both teachers and students. Teacher directions should be printed in plain text, while text to the read to students should be in **bold** print.
7. Probes that teachers can use for students who need assistance or are "stuck" should be described, especially if they are unique to the task.
8. The **Intermediate and Final Student Products to be Developed** is where you describe what students are asked to produce, either as final products or interim products on the way to such final products. Describe these clearly.
9. The **Student Checklist and/or Reflection/Reaction Sheet (Optional)** is just that – optional. For some Tasks that have multiple parts and/or for which students must carry out multiple steps to respond to them, it may be helpful to provide a Checklist for students' use to make sure that carry out other activities. For other Tasks, it may be important to elicit students' Reflections/Reactions on the processes they used in completing the Task and/or what they learned in doing so. Both (or neither) may be used.
10. The **Teacher Scoring Rubrics and Scoring Form(s)** section is where you indicate the rubric(s) that will be used to score students' responses. Determine first which parts of the Task are to be scored, whether each part has one or more than one rubric for scoring the responses, and then, whether there is any overall score (across all parts

of the Task) to be given as well. Be sure to specify not only the name of the rubric but also how many levels each rubric will have (e.g., label these as “1” for the lowest level and “4” for the highest level). Then provide a description of each of the performance levels.

11. The **Sources for Copyrighted Materials** should be filled in. It is essential that we know of any copyrighted materials you used and where you located them so we can secure permissions to use the materials.
12. **What declarative knowledge and/or content is necessary to successfully complete the PT?** Note: these may form the basis of multiple-choice items to be written later. Note the important concepts that we want to make sure that students know.
13. **What Michigan content standards does this PT measure?** Be sure to note up to five Michigan standards measured by this PT.

Scoring Rubric Template

Add or delete the “Dimension” lines to match the number of dimensions on which you propose to score the responses of students to a constructed-response item, or a performance event or task. Level 1 is the lowest score point and Level 4 is the highest. Note: your rubrics might have two, three, or four levels.

[Rubric Title]				
Dimension	**1**	**2**	**3**	**4**

[Rubric Title]				
Dimension	**1**	**2**	**3**	**4**

[Rubric Title]				
Dimension	**1**	**2**	**3**	**4**