

**Developing the Secondary Technical Skill  
Assessment System  
for Michigan**

Edward Roeber

Michigan State University

July 2008

# Table of Contents

<b>Section</b>	<b>Page</b>
Introduction	3
Section 1: Qualities of Sound Educational Measures	5
• Reliability	5
• Validity	5
• Bias-Free	6
• Feasibility	6
• Affordable	6
• Utility	7
Section 2: Preliminary Activities	8
• Identify the Purposes for Assessment	8
• Identify the Programs for Which Measures Are Needed	8
• Identify the Standards for Each Program	9
• Locating Suitable Assessment Instruments	11
• Collect Specimen Sets of Existing Measures	11
Section 3: Review of Existing Assessment Instruments	14
• Thoroughly Review the Assessment Materials	14
• Complete an Assessment Review Form for the Assessment	15
• Determine Whether the Assessment Meets Minimum Program Standards	15
• Determine Whether the Assessment is Well Aligned with Program Content Standards	15
• Determine Whether the Assessment Meets Minimum Technical Standards	19
• Determine the Feasibility of the Assessment	23
• Determine the Cost of the Assessment	24
• Determine Whether to Use the Assessment	25
Section 4: Creating New Assessment Instruments	
• Using an Assessment Contractor	27
• Leading the Assessment Project	27
• Others to be Involved in the Project	28
• Collaborating with Other Agencies	28
• Paying for the Project	29
• Preliminary Development Activities	29
• Tips for the Development of State and Local Assessments	37
• Assemble the Final Instruments	46
• Administration of the Assessments	48
• Reporting of Assessment Results	49
Summary	52
Attachments	
A: Michigan Technical Skills Assessment Review Elements	53

B: Michigan Technical Skills Assessment Review Form 56

C: Technical Skill Assessment Content Review 2008-09 Instructions 58



## Introduction

One of the requirements of the Perkins Career and Technical Education Act of 2006 is that states are to implement assessments in technical skill areas at the high school level. These requirements are at least for students who are program concentrators, that is, who have completed 50% of a program of study, enrolled in the next course in the sequence (even if they don't complete the program), and who will leave high school either in pursuit of employment in the technical skill area, or pursue post-secondary education in order to be qualified for employment. The requirement for secondary technical skills assessments is a substantial one for states, for several reasons.

First, states have not been funded specifically to develop and implement these assessments. Unlike other Federal programs, such as the No Child Left Behind act, where states were provided substantial federal resources for assessment development and implementation, no Perkins funds provided to states are set aside for these purposes. Second, the funding that has been allocated has been provided to local school districts. They seek to use the funds to improve the career-technical programs, not to develop and/or administer assessments. Third, each state has developed its skills for the various career-technical areas, so that the assessments created by one state may not be useable elsewhere.

Thus, the challenge that the Michigan Department of Education faces is to meet the federal requirement for assessments of technical skills in the most cost-effective manner possible. This means, first, screening available instruments for their suitability and applicability to Michigan's schools. Second, if this does not produce a viable instrument, then the state will need to construct such an instrument. It is hoped that the second instance will not occur often, but should it do so, the state will need to have a plan for how such assessments could be created in the shortest possible time and at the lowest cost feasible.

The purpose of this paper is to describe ways in which both of these activities could be accomplished. For each activity, options for how the work could be carried out will be presented and discussed. The goal is to help the Michigan Department of Education and its advisors to determine the most cost-effective approaches to meeting the federal requirements.

The paper is organized into several sections. The first section describes some of the characteristics of sound educational assessments. This section is followed by a section that describes some of the activities that should be carried out prior to decisions about whether to select or develop the needed technical skills assessments. The final two sections describe the steps in determining whether existing measures are suitable for use in Michigan, and if not, how the necessary assessments could be developed.

## Section One

### Qualities of Sound Educational Assessments

The field of educational measurement has established criteria for determining the appropriateness and soundness of educational assessments. In general, these criteria are as follows:

Reliability – This characteristic of assessments is a necessary but not sufficient condition for assessments. Reliability might be measured as stability of the assessment information over time, the internal consistency of an assessment measure, or the precision of the measure in decision-making, but regardless of the metric used, without reliability, an educational assessment is useless. Therefore, whether the assessment measure is purchased, used with permission from its owner, or developed by the state, it is essential that each of the assessment measures produces data that are reliable.

For the assessments being considered for use in Michigan, there are several types of reliability data that may be available. These include test-retest reliability correlational data to determine how stable the student scores are from one time to another when further learning has not occurred, internal consistency coefficients (KR-20 or coefficient alpha), and accuracy (standard error of measurement). In cases where multiple versions of the same assessment are available, the equivalence of the forms is another type of reliability that will be important to examine. Where constructed-response (written response) items are used, the accuracy of the scoring of these types of items should also be reported.

Validity – This characteristic is not so much about the assessment itself but about the use of the assessment results. That is, a test does not in itself possess “validity,” but if the evidence is sufficient and compelling, there may be evidence that the use of an assessment is valid. This suggests that users of assessment must collect evidence to support each major use of an assessment, and that evidence that supports one use (e.g., to tell students how they have learned from a set of standards) will not automatically support other uses (e.g., the likelihood that the student will be successful in a community college training program in that same area). Evidence to support this latter use must also be collected and analyzed to support the use of the assessment in this manner.

In the context of the Michigan technical skills assessments, two types of validity may be of the greatest concern: 1) content validity – does the assessment adequately cover the content standards that it is designed to measure, and 2) predictive validity – are students who do well on the assessment and who score above the cut score actually able to perform the tasks, jobs or other skills that the assessment is designed to measure.

Content validity is established using a judgmental approach, and is measured both by how many of the assessment items measure one or more the content standards, as well as how many of the content standards are measured by the assessment. More in-depth analyses include looking at whether the assessment covers the content standards at the same depth of knowledge or cognitive complexity level, whether the range of the content standards is represented in the balance among the assessment items, and whether the assessment produces results that correspond to how the content standards are organized (the structure of the content standards is used to report the assessment results).

Predictive validity for Michigan technical skill assessments is demonstrated through the use of the assessments in high school and relating performance on the assessments to subsequent success elsewhere – on later-occurring assessments (e.g., community college placement tests), programs (e.g., community college technical training programs in the same program area as the assessment) or work itself. One difficulty in measuring the predictive validity of an assessment is developing a criterion measure that is itself reliable and relevant to the predictor measure (the assessment used in high school). For example, “success on the job,” can be measured differently in several ways. Thus, lower predictive validity than desired may be due to the use of poorly defined criterion measures, lower relationship between the predictor and the criterion, or both.

Bias-Free – Bias-free measurement is important so that we can be assured that the assessment is measuring the intended skills in such a manner that it permits all test takers to show what they know and can do to the fullest extent possible. Bias is an extraneous variable that results in some students’ scores not representing the full extent of their achievement.

Bias is determined both through statistical methods and human judgment. Typically, during pilot testing or during initial actual use, the scores on the test items of various sub-groups are examined. Where one sub-group (or more) performs significantly differently, these items are flagged. The flagged items are then examined by a panel of experts who carefully study the item to determine whether differential item performance is due to instructional differences between groups of students, or whether the item is exhibiting bias that prevents one or more sub-groups from doing their best on the item. While this combined use of statistical and judgmental techniques is not an exact science, good assessments have used both procedures to determine that there are not extraneous factors that have a negative effect on student scores.

Feasibility – Whether an assessment is feasible or not is a matter of some interpretation. Different program administrators and teachers have different expectations about how much time should be devoted to the assessment, how much effort is necessary to learn how to administer the assessment in a reliable manner, and how easy it is for students to respond to it. Good assessments should not be too time-consuming, should be easy to administer, and should be easy for students to respond to.

If Michigan educators have pilot tested the assessment, the feasibility of the assessment can be determined by their responses to questions that are asked of them at the time the pilot testing occurs. If the assessment has not been pilot tested in Michigan, the feasibility of the assessment can be determined by determining whether such factors were considered when it was pilot tested elsewhere, as well as from an examination of the assessment administration manual and other materials. A panel of Michigan educators could be used for this purpose.

Affordable – The affordability of an assessment is also a matter of relative judgment, in part tempered by the nature of the assessment, how intricate the program area is, the nature of the student standards, and the usefulness of the assessment results.

For example, an assessment that is all multiple choice items, relatively short, and measuring only one set of skills should be less expensive to use than an assessment in which students must respond in writing or by carrying out a performance. The latter may have more intricate scoring guides and scorer training procedures and materials.

An online assessment may cost less (because it is a multiple-choice test) or may cost more (if the assessment requires on-site installation of software or involves automated essay scoring where the scoring engine needs to be calibrated through the input of hundreds of students' responses).

Therefore, the context of the assessment will determine the costs, but cost alone is not the only variable in determining the value of the assessment. Some assessments that are more costly may be more affordable, given these factors.

Utility – The overall usefulness of the assessment is dependent on several factors. These include how the assessment is administered (online assessments may be more useful than those requiring test booklets and answer sheets), how quickly scoring takes place, how quickly the results are returned both to students and to program administrators, and how extensive the results are for the student, for teachers, school administrators, as well as the Michigan Department of Education. The utility of the assessment also considers the costs of the program – in terms both of dollars and effort needed to collect the information. Statistical characteristics of the assessment, such as reliability and validity information, can also be important in terms of determining the usefulness of the assessment. This, however, is the ultimate judgment about the assessment.

## Section Two

### Preliminary Activities

There are a number of activities that need to be carried out before the Michigan assessments of technical skills can be put in place, whether through selecting existing measures or developing new ones. Each of these is described below:

Identify the Purposes for Assessment – The first task in selecting or developing assessments in order to develop an assessment program is to define the purposes for the assessment program. Purposes for assessment can be found in legislation, program regulations as well as policy created by the sponsoring agency or board. While sometime this appears to be obvious, such as to measure and report on student achievement often ascribed to most testing programs, the issue is that rarely is this the only purpose for most assessment programs.

Different assessment purposes suggest different types of assessment instrumentation as well as different means and methods of reporting assessment results. One way to distinguish assessment purposes is to consider the purposes for individual student assessment versus those for program assessment. For example, for students, assessments may be used to determine the extent of their achievement, which could lead to remedial instruction, to program placement, to grades, and selection for subsequent educational opportunities (such as community college placement).

For program evaluation purposes, assessments may be used to determine whether a program of instruction adequately taught students (holding instructors accountable for the level of achievement that students demonstrated), whether instruction provided has prepared students for subsequent education or work, or whether students have reached a level of proficiency to be certified as “proficient” (or some other important criterion).

Each of these purposes may suggest assessment instruments that differ in the number of skills, the number of items per skill, and ultimately, what types of results are reported. Therefore, it is essential that the sponsoring agency consider all of the purposes for the technical skills assessments before they are selected or developed. This will help assure that the selected or developed assessments most fully meet the various purposes for assessment.

Identify the Programs for Which Measures Are Needed – This is a somewhat complicated issue. The Department and its advisors will need to determine the level of each program for which measures will be sought. For example, the sixteen career clusters could be used to select the programs for which assessments will be implemented. Or, the more than 80 career pathways could be used instead. Each has advantages and disadvantages.

*Career Clusters* – If broader career clusters are used as the basis for student assessment, fewer areas for assessment will be selected. The broader, more general the program area selected, the larger the set of content standards that apply to the program. If there a larger number of content standards to assess, this may mean that the assessment will need to be lengthier or that the assessments will need to cover the needed skills at a more general level.

The more general the program level for assessment, the more challenging it will be to demonstrate the validity of the assessment – whether content validity or predictive

validity. This is due to potentially lower content coverage and therefore potentially lower correlations with criterion measures such as community college or work success.

*Federal Cluster Pathways* – If the more specific Federal cluster pathways are selected there will be more programs for which distinct measures will be required. This could have substantial cost implications for local districts and/or the state. The narrower the scope of the program being measured, the greater the fit between the program's content standards and the assessment. This will likely mean that the validity information for the assessments will be higher due to a greater match between the standards and the assessments. Potentially, fewer assessments may be available to measure each of the more narrow career pathways.

The more general the program level, the more likely that an assessment will cover the more important content standards; the more specific the program, the fewer assessments that are likely to be available, thus potentially increasing the number of assessments that Michigan needs to construct.

*CIP-Level Assessments* – If the CIP levels are selected as the basis of assessment, then there will be a substantial increase in the number of program areas to be assessed, which of course means a much larger set of assessment instruments are needed. The plus of assessing at the CIP level is that the assessments can focus on a smaller set of skills, so given equal testing time, CIP-level assessments will more accurately assess students' knowledge and skills in the smaller range of skills contained in any CIP area. However, it will mean that either more existing instruments will need to be located or potentially, more assessments will need to be created.

One way to determine for which programs assessments should be selected or developed is to determine the number of concentrators in each program. This would permit the state to start work on those programs that affect the largest number of students. Easiest of all will be those programs for which high-quality measures are already available at a reasonable cost. By selecting these programs, the state will be able to begin measuring the success of the maximum number of students in the shortest period of time.

Identify the Standards for Each Program – Once the list of program areas has been determined, the next step is to identify the content standards that each program should help students achieve. The selected programs may be separated into three different categories, based on whether content standards are available, and if so, how they were developed. These are:

1. Common Standards: The set of standards for a program area were written by an industry group or others and have been agreed to by the Michigan Department of Education.
2. Custom Standards: Michigan has developed a unique set of student standards that differ significantly from those developed elsewhere.
3. Standards Do Not Exist: Michigan has neither adopted others' content standards nor developed ones of its own. No existing standards may exist in these program areas, either.

*Common Standards* Those programs for which common standards exist are those where finding existing instruments is most likely to occur. Because other organizations outside of the state have developed student standards that Michigan has agreed to use, the search for existing instruments will likely be most fruitful here. One source of these

will be the organization that developed the standards; they may have also developed a corresponding assessment as well. Of course, it will be important to make sure that the standards (and any corresponding assessments) are suitable for use with secondary students.

*Custom Standards* For programs where Michigan has developed its own (customized) standards, the first question is whether this is an area in which no one has developed standards and therefore Michigan had to develop its own. If not, then the second question is if other organizations have developed content standards, why has Michigan chosen to develop its own? Is there a compelling content reason why the state chose to develop its own standards – perhaps, important content that was missing in others’ standards, an emphasis that was thought to be misplaced, or other reason?

The reason(s) why Michigan developed its own standards is important to understand because rarely will the standards developed in Michigan be completely different from those developed elsewhere. More likely, there will be some degree of overlap between the content standards developed by other organizations and those developed in Michigan. If this is the case, then the measures developed for those other programs may still at least partially suit Michigan’s needs.

Depending on the extent of overlap, the assessments developed to measure the other program’s content standards might be customized for use in Michigan. This could happen in one of two ways. First, the other organization’s assessment items measuring content standards not found in Michigan might be deleted from the assessment. This, of course, is subject to the approval of the organization that created the assessment (the owner of the copyright on it). Second, Michigan could add additional assessment items to measure unique Michigan skills not covered in the other assessment while not reporting the items from the other assessment that do not measure Michigan content standards. Again, this may require the approval of the organization that developed the other assessment, especially if the Michigan-created items are to be packaged with the other assessment. A lesser-expensive approach is to offer a Michigan assessment booklet along with the other assessment booklet, with students taking both assessments and combining the scores from the two before reporting.

Customizing an existing assessment is generally less costly than the creation of a new assessment from scratch. If the extent of customization is not too great, the reliability and validity data attributable to the original instrument may be used for the customized version. This could be important if this data includes industry-based verification of the importance of the content standards assessed, or information on the predictive validity of the assessment for subsequent training or work. Even if the extent of customization is not too large, however, it will be important to study how the addition of new Michigan items (and, perhaps, the deletion of some of the items from the other assessment) affects the overall reporting of the combined assessment in order to assure that the complete assessment is both reliable and valid for its intended purposes.

*Standards Need to be Developed* In cases where programs are selected for which no external standards exist, Michigan will need to create the content standards for these programs. It is essential when these standards are created that a representative group of educators, practitioners, and others be included in the development process, and that industry representatives be used to validate the content standards to the greatest extent possible.

Before Michigan embarks on this development by itself, it should seek to establish which partners might be available to help carry out the work. This includes other state education agencies, industry-based groups, higher education, and other interested parties. Since the goal is not just to create a set of content standards that serve the purpose of guiding assessment in the state but also to create standards that are linked to post-secondary education and work possibilities for students in the program, it is essential that representatives of these organizations participate in the development of the content standards in some fashion.

Interested organizations and individuals can serve in several ways. First, they can serve as members of the working committee that actually drafts the content standards statements. Second, they can serve as reviewers of the draft statements as they are produced. Third, they can serve as conduits to keep their organization and/or constituency informed of the progress in developing the content standards and recruit individuals to serve as reviewers of them. Finally, they may serve to collect evidence of the suitability of the content standards for post-secondary users of them.

Once the standards are developed in Michigan, the Department should consider how to make sure that the content standards will help students who accomplish them succeed in post-secondary educational programs, as well as on jobs in the program area. This might be accomplished in a couple of ways, including the use of surveys of post-secondary educators or employers who would employ program graduates, focus groups comprised of educators or employers used to review the content standards, and eventually, statistical studies of success on the assessment and success in post-secondary programs and employment.

Once public input has been gathered and synthesized, the content standards document needs to be finalized and approved by the Department (and by any other organizations for which approval might be sought).

Locating Suitable Assessment Instruments – Once the programs for which assessments will be needed have been identified and the content standards that apply to each program have been selected or developed, attention can turn to locating suitable assessment instruments. Sources for such information include journals that cover career technical programs, organizations that have developed career content standards, industry-based organizations that have helped to create content standards and skill certification efforts, and higher education faculty. Any of these organizations may have created assessments that could be used in this state. All of them should be searched to locate references to assessments, as well as listings of available assessments. Another aid to locating suitable assessments is the Internet. Using one of the popular search engines may turn up additional assessments that may be suitable, especially those that have not been publicized much.

Collect Specimen Sets of Existing Measures – Once the program areas have been identified and as content standards are being identified, customized, and/or developed, Department staff should identify the availability of assessments in each program area. This collection should utilize all available sources of information about potential assessments to track down as many available instruments as possible. Several sources should be consulted, including instruments that staff are familiar with, those that local educators are aware of, those suggested by colleagues at the state and local levels, and those provided by organizations involved in education, training and related programs. Internet searches should also be used to ferret out instruments that might have been developed by local organizations across the United States and elsewhere. The goal is

this search is to cast a wide net for available instruments that might be used within the career tech assessment system in Michigan.

The intention of this collection of information is to identify those programs for which existing measures are available, regardless of whether the state will be using the standards on which the assessments were developed for its programs, whether it might be necessary to create customized assessments, or is an area in which Michigan will need to build its own standards. This materials collection might assist in program areas where Michigan has customized others' content standards or in areas where Michigan will develop its own content standards, since the information collected may include both the assessment instruments themselves as well as the content standards on which they are based.

Once potential assessment instruments are identified, additional information needed to review the instruments should also be collected. In some cases, this will be relatively easy, since the assessment publisher is known and they have prepared sets of materials for potential reviewers. In other cases, the information needed may not have been prepared, or the needed materials may not be readily available. This could happen, for example, when an instrument is located but the owner/publisher is not readily known. While the assessment instrument may appear to meet the state's needs, it would be better if information about its technical qualities would be available, as well as whether the assessment is copyrighted and if so, which organization owns it.

To the extent possible, the following materials and information should be collected about each available assessment:

#### **Information to be Collected for Available Assessments**

##### *Materials*

- Specimen set of assessments (it may be necessary to sign a security/confidentiality agreement to secure permission to review actual assessments)
- Administration Manual
- Technical Report

##### *Information*

- Official title of the assessment
- Test publisher
- Publishing date
- Grade/age span of the assessment
- Suitability for secondary students completing a program
- Type of assessment (norm-referenced versus criterion-referenced)
- Testing mode (paper and pencil or online, or both)
- Number of test items of each type per grade/age
- Testing time per grade/age
- When assessment can be administered during the school year
- Cost (testing materials, scoring, and reporting)
- Restrictions on use (if any)

It will be helpful to the Department to use a form (see the example in Attachment A) on which to request information from the assessment publisher/owner. The form should be accompanied by a cover letter that clarifies the reasons why the Department is

requesting the information and assessments, what will be done with these, and how the materials will be disposed after the review is completed.

In some cases, the form will be used to provide the needed policy, curricular, and technical information. In other cases, the assessment owner will note that the needed information is found in the administration manual, technical report, or other publications related to the assessment. In either case, the publisher is more likely to provide the information needed for the review of the assessment if sent the form since this will prompt the publisher by indicating the types of information and materials that are desired.

Some of this information may be readily available, while other data may not be available until the sample or specimen set has been ordered and the technical report or the administration manual has been reviewed. Some publishers, especially of commercially available assessments, may readily provide this and other information, while others may be reluctant to provide much of this information. In addition, some publishers may not have the information desired in written form, so in this case, a telephone call or e-mail may be necessary to obtain the needed information.

Once the official information has been received from the organization providing it, potential users will also want to search the Internet to determine if there are other individuals, groups, or organizations that have used the assessment in local or state programs. This may help to fill in information about both the feasibility and technical qualities of the assessment, useful information for assessment reviewers to have.

Once the materials and information are received, they should be catalogued and placed in a secure location within the Department since some of the testing materials may have required someone from the Department to sign off on a security/confidentiality agreement in order for the materials to be provided to the Department. Publishers expect the Department to keep the materials in a secure location.

As each set of assessment materials and related information is received, it should be quickly reviewed to make sure that all needed information and materials were provided. If they were not, is the information needed located in either the assessment's administration manual or technical report? If so, the information can be noted and no further action would be necessary. If this information was not provided and is not contained in the publications, the publisher should be contacted in order to provide the missing information or materials.

## Section Three

### Review of Existing Assessment Instruments

Once one or more assessment instruments have been identified for review and information about them has been collected, there are a series of steps that should be used to review the materials in order to determine their suitability for use in the state's technical skill assessment program. Note: although these steps are presented in a linear manner, it is not necessary to follow the same order when reviewing the assessments. However, each of these steps should be carried out before a decision is made on whether to use the assessment.

Thoroughly Review the Assessment Materials – The first step in the review of existing assessment instruments is to review the available materials and information completely. Assuming that the review set is complete (see the list of assessment materials and information provided in the previous section), the review can begin. Start by examining the review form that was sent to the publisher. The assessment publisher may have provided information on it that is not contained in the administration manual or technical report.

Then, review the assessment information sent along by the publisher. Often, questions about costs, administration time, and effort needed to learn how to administer and to take the assessment will be pieced together from multiple sources. By looking at all of the information carefully, many questions about the use of the assessment can be answered.

Next, review the assessment administration manual along with a copy of the student assessment booklet. The purpose of this part of the review is to help the reviewer better understand the assessment administration process as well as the nature of the assessments. This can help to understand whether the assessment is comprised entirely of multiple-choice items, or whether constructed response items and/or performance assessments are also used. The assessment administration manual will also help to describe what, if any, special administrator training is needed to administer and score the assessments.

Finally, look at the technical report and other technical data provided by the publisher. The goal of this portion of the review is to determine how complete is the technical information provided by the publisher. Some technical reports can be quite thorough in their coverage of the technical data, and no further searching is necessary. Other times, however, the technical report may only provide information on the processes used to develop the assessment and does not provide much if any technical data about the assessments. The latter can be true when new assessments are first being made available, although if the new assessment has been pilot tested, this information should have been provided in the technical report (and updated when the first operational data is available after the first substantial use of the assessment).

If the technical data is sparse or non-existent, the publisher should be contacted to ascertain whether any data regarding the reliability or validity of the assessment has been computed (on pilot or operational testing). If so, the publisher may be willing to provide this data informally to the reviewer; at least, the publisher should be willing to indicate when the technical data will become available.

Complete an Assessment Review Form for the Assessment – Once the thorough overview of the materials and information is completed, the actual review can begin. The first step is to complete the “Michigan Technical Skill Assessment Review Form,” shown in Attachment B. This form, and the review elements shown in Attachment A, will provide a good basis for describing the assessment, the assessment administration process, the technical data that supports the assessment, and the results that the assessment produces.

The review form should be completed as fully as possible. Not all information will necessarily be pertinent to the assessment (for example, some types of reliability estimates may not be appropriate for the assessment being reviewed); other information may not be available. If questions arise about the information requested, the reviewer may wish to seek clarification from the assessment publisher or local users of the assessment (if any are known).

The sections of the review form pertaining to coverage of program content standards, meeting technical standards, feasibility and cost-benefits are the parts of the review form that are most critical to complete. If this information is not available, it may be necessary to plan to collect the information in other ways, such as using the assessment in a pilot test to determine its technical and practical adequacy.

Determine Whether the Assessment Meets Minimum Program Standards – The assessment under review should be appropriate for the assessment of students (“completers”) in career tech programs at the high school level. One question that reviewers will need to determine is whether the assessment can be used in this context. For some assessments, this will be a simple matter since the assessment was designed to be used in this manner. For others, this may be more of a judgment call, since the assessment may have been developed for use in a different context – for example, in a community college class. This won’t automatically eliminate the assessment from being considered, but reviewers will want to attend carefully to its appropriateness for high school juniors and seniors. This should include a focus on content appropriateness as well as vocabulary and language.

Determine Whether the Assessment is Well Aligned with Program Content Standards – The match of the assessment to the program content standards is a key determiner of the appropriateness of an assessment for inclusion in the state’s technical skills assessment program. There are both informal and formal means for determining the alignment of an assessment and the standards that it supposedly measures. Each method of alignment may be appropriate, given both when the review is taking place and the stakes attached to it.

Informal alignment methods rely on human judgment. Ideally, these human judges will be persons who are quite familiar with the occupations that the program is preparing students for as well as the educational programs that help them attain proficiency. The judges will be looking for: 1) are each of the key skills measured by the assessment, 2) is the number of assessment items per skill appropriate for the importance of the skill and/or the complexity of it, 3) do the assessment items measure the important portions of each of the content standards, and 4) does every assessment item measure one or more the content standards?

Even though this sort of review is labeled “informal” does not mean that the review is not structured in such a manner as to produce quality results. If at all possible, more than one judge should be used to review the assessment. Each judge should work

independently and use a structured review form to judge each assessment item, and then to review each content standard. Once each of the judges has completed their review independently, their reviews should be compiled. Then, the panel of judges should be shown the compilation of their reviews in order for them to come up with overall conclusions about the assessment. It is doubtful that this sort of review will result in a numerical rating of the assessment, but by asking the panel of judges to provide an overall review of the assessment, the Department can obtain a judgment about the assessment's alignment to the content standards.

In some cases, a test publisher may limit access to the instrument and related information so that more than an informal is not possible. In these cases, those who are empowered to make decisions about instruments should try to gain access to at least one copy of the test and related information and conduct at least the informal review outlined here. If even this access is not permitted, then the assessment might have to be eliminated from consideration by the sponsoring agency.

Several documents have been prepared by the Department for conducting this informal review process. These materials are shown in Attachment C and include a form that persons who are asked to judge the alignment of a technical assessment with the corresponding technical skills can use to record their judgments.

More formal methods of studying the alignment of assessments have been created in recent years, due in part to the requirements of the No Child Left Behind (NCLB) act that states demonstrate the alignment of the assessments used for Title I purposes to their academic content standards. Norm Webb of the University of Wisconsin created the most prevalent method used. His method examines four characteristics of assessments and standards:

#### **Webb Alignment Criteria**

The alignment between standards and assessments is based on four criteria. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had met the standards.

*Categorical Concurrence* – An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.

This criterion is judged by determining whether the assessment included items measuring content from each standard. The analysis assumed that the assessment had to have at least six items measuring content from a standard in order for an acceptable level of categorical concurrence to exist between the standard and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale.

*Depth-of-Knowledge Consistency* – Standards and assessments can be aligned not only by the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge (DOK) consistency between standards and assessment* indicates alignment if what is elicited from students on the assessment

is as demanding cognitively as what students are expected to know and do as stated in the standards. For consistency to exist between the assessment and the standard, as judged in this analysis, at least 50% of the items corresponding to an objective had to be at or above the level of knowledge of the objective: 50%, a conservative cutoff point, is based on the assumption that a minimal passing score for any one standard of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding objectives.

There are four levels of DOK for standards and assessments. These are

- Level 1 (Recall) – includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula.
- Level 2 (Skill/Concept) – includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps.
- Level 3 (Strategic Thinking) – requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3 activity.
- Level 4 (Extended Thinking) – requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking.

As mentioned above, the goal is for the assessment to match or exceed the DOK designation of the content standards.

*Range-of-Knowledge Correspondence* – For standards and assessments to be aligned, the breadth of knowledge required on both should be comparable. The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.

The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of objectives within the standard with one related assessment item/activity. Fifty percent of the objectives for a standard had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a standard.

*Balance of Representation* – In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally in both. The range-of-knowledge criterion only considers the number of objectives within a standard hit (a standard with a corresponding item); it does not take into consideration how the hits (or assessment items/activities) are distributed among these objectives. *The balance-of-representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another.*

There is one additional criterion that is used when the assessment gives some evidence of having an issue:

*Source-of-Challenge* – The source-of-challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted content skill, concept, or application. Cultural bias or specialized knowledge could be reasons for an item to have a source-of-challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed.

The Webb methodology requires a panel of impartial judges who work on computers (the alignment methodology is completely computerized). Typical alignment studies employ between 9 and 15 judges, each working for two to three days (depending on the number, length and complexity of the assessments that they will be reviewing). Before the session begins, much information has already been loaded into the computer system, including the content standards, their codes and the coding schema for the standards, as well as information about the assessment items – the content standard that each measures, the correct answer, and their weight (most multiple choice items have a weight of 1, but some constructed response or performance items may have carry a larger weight). By arranging all of this in advance, the results of the review are available immediately following the completion of the review by the judges.

The review begins with an orientation to the project and an overview of the criteria listed above. Then, the facilitators review the academic content standards themselves, describing the structure of the standards as well as their contents. Finally, the review leaders provide an overview of the assessments – the grades assessed, the types of results that are returned, and the nature of the assessments used.

At this point, the judges are ready to start their review. They begin by rating the Depth of Knowledge and Range of Knowledge of the academic content standards. This provides the frame against which to compare the Depth of Knowledge and Range of Knowledge ratings of the assessments. The assumption in doing this is that the content standards provide the base against which to compare the assessments. While this is “alignment” in the strictest sense, this methodology does not account for instances where either the depth or range of knowledge called for in the content standards are less than adequate.

Once the content standards have been rated, the judges turn their attention to the assessment itself. They systematically rate each item and the underlying computer program tallies the results. When the review is completed, there is inter-rater reliability data on the agreement of the judges in their ratings, and more importantly, summary information on the alignment of the assessment to the standards. A sample summary report is shown below

*Comparison of Three Grade 7 Assessments on Four Alignment Criteria Acceptable Levels in Relationship to Michigan Mathematics Strands for Grade 7*

ALIGNMENT SUMMARY GRADE 7 STANDARDS												
Standard	Alignment Criteria											
	Categorical Concurrence			Depth of Knowledge Consistency			Range of Knowledge Correspondence			Balance of Representation		
	TerraNova	ITBS	Stanford 10	TerraNova	ITBS	Stanford 10	TerraNova	ITBS	Stanford 10	TerraNova	ITBS	Stanford 10
Numeration	Y	Y	Y	Y	Y	Y	Y	Y	Y	W	W	N
Algebra	Y	Y	Y	W	N	Y	N	N	N	Y	Y	Y
Geometry	Y	N	Y	Y	N	W	N	N	N	Y	Y	W
Data/Statistics	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y

Y = Adequate alignment; W = Weak alignment; N = Inadequate alignment

The benefit of using such a formal approach to alignment is that the data that results from the study is quite robust, especially if the judges are independent of the sponsor of the review. The downside of using such a method is that it is expensive; judges often require an honorarium, overnight lodging, meals, and travel expenses (which may include air or surface transportation).

The value and cost of each method needs to be judged against the purpose and need for the review. If it is just to make a quick judgment about the suitability of an assessment for measuring a set of content standards, informal methods are more than adequate. However, if there is a need to thoroughly document the alignment of an assessment to a set of content standards, more formal methods such as the Webb methodology will be useful, since the findings from such a study are more incontrovertible.

Before proceeding with a more thorough review of the assessment instrument, it is important to at least informally judge that the standards measured by the instrument are related to the career tech program, and the assessment appears to be a good measure of that program's content standards. More formal alignment studies can occur later, but it is not cost-effective to do them in the absence of at least some informed judgment that the assessment is likely to be aligned to the content standards and cover them well. If that appears to be the case, then more detailed studies of the assessment (its alignment as well as its technical and other qualities) can occur.

Determine Whether the Assessment Meets Minimum Technical Standards – Once there is a determination that the assessment is likely to align with the program content standards, it is also important to take into account whether the assessment is technically sound. Ideally, this determination will be based on data that results from the use of the assessment in either pilot testing, field testing, or operational use. However, in the case of newly-developed assessments, this is not always the case.

Assuming that data is available, the reviewer will want to search for information on the following technical characteristics as shown on the review form given in Attachment A:

1. Reliability – Please describe how the reliability of the assessment has been established. These include any of the following:

- A. Stability (Test-Retest)
- B. Internal Consistency (KR-20 or Cronbach Alpha)
- C. Split-Half
- D. Alternate Form
- E. Inter-Rater Reliability (for constructed-response or performance items requiring handscoring)

What analyses have been performed and what data are available to demonstrate the reliability of the assessment. Is the data available suitable for the assessment?

2. Validity – Please describe how the validity of the assessment has been established for each type of inference for which the assessment is recommended.

- A. Content Validity – An essential element in the selection of appropriate instruments to assess Michigan technical skill standards is the alignment of the assessments and the Michigan technical skill standards. Alignment will be judged four ways:

1. How many of the Michigan technical skill standards are measured by the assessment,
2. How many assessment items measure one or more Michigan technical skill standards.
3. Does the assessment emphasize the most important skills in the Michigan technical skill standards?
4. Does the assessment assess students at a level appropriate for high school?

What analyses have been carried out to demonstrate alignment to the standards? Please provide information on any alignment study that has been carried out for this assessment to the Michigan technical skill standards. Please describe the results of evaluation of the content by business, industry and postsecondary institutions.

- B. Predictive Validity – Please describe the results of analyses of the predictive validity of the assessment. Is there information to show that successful performance on the assessment predicts success on the job? Do students who do better on the assessment do well in jobs or in postsecondary education in the same field?

Please describe the results of analyses examining the extent to which the assessment accurately discriminates between students with greater mastery compared to students with lower mastery (discrimination index)

- C. Concurrent Validity – Does performance on this assessment correlate with success on other related/comparable assessments?
- D. Construct Validity – Is the assessment designed to measure a more theoretical construct, is there evidence to support the proposed interpretation of the results?
- E. Sub-Scale Validity – If applicable, please describe the results of analyses of the validity of any subscales.

3. Bias and Sensitivity Reviews – Please describe the results of bias and sensitivity reviews conducted on the items.

The major characteristics for the reviewer to examine are the reliability, validity, and lack of bias in the assessment. The data that is made available may be found in the assessment's technical report, or in reports on the use of the assessment in various contexts. The latter types of studies might be found in a search of the Internet, although not all published studies will provide information on the technical characteristics of the assessment(s) used.

For reliability, data that might be shown could include the stability of the instrument (an estimate of its test-retest reliability), the comparability of alternate forms, the internal consistency of the assessment, and/or the reliability of administering performance items and scoring constructed-response items. Each of these are important, but for different reasons. While there is no accepted minimum level of reliability in the technical literature, users of test always seek instruments that have high reliability, no matter how measured. For test-retest reliability, a figure of .90 is sought after. Internal consistency reliability may be lower, especially where the instrument measures more than one skill. However, even in these cases, users will want the reliability to be .5,.6 or higher.

Test-retest reliability indicates how stable are the results from the assessment when it is administered twice (usually within a two-week period with no intervening instruction or learning). If the scores are stable, we can be assured that the estimate we have obtained of students' achievement will remain constant unless learning has occurred. This gives us more confidence about our estimates of student performance.

Alternate form reliability is similar in nature. If two or more forms of the assessment produce comparable results, we can be confident that it won't matter which form of the assessment we use when we assess students. Having an alternate form of the assessment is important in the long run since we want to make sure that students are learning the concepts covered by the assessment and not just practicing on the items on the assessment. If the same assessment form is used year after year, teachers and students may concentrate on the items on the test and therefore scores might increase artificially. Having comparable alternate forms will help assure that this does not occur.

The internal consistency of the measure may or may not be important. Some assessments are designed to cover several content standards within one measure (that is, the score on the assessment is a combination of measures of several skills). For these assessments, one would not expect to see high internal consistency, since several skills are being measured and reported on together within one test. Most often, however, assessments individually measure and report on each content standard, and so the internal consistency of these individual measures is important. Several factors can affect internal consistency of the assessment, somewhat artificially. These include whether the assessment has mixed assessment types (e.g., multiple choice and constructed response items), has items at significantly different difficulty levels (e.g., very difficult and very easy items), or other diverse ways of assessing the standard.

The final type of reliability that may be important concerns the administration of any performance items and the scoring of constructed-response items. Each of these also has an aspect of reliability to it. When performance assessments are administered, one question is whether the administration conducted by one trained test administrator

would be comparable to that provided by another trained individual. This is especially true when the assessment administrator is responsible for recording the responses of the student. Are the students' responses recorded completely and accurately?

When constructed-response items are used, they are typically scored by individuals who have received training using a scoring rubric and a scoring guide that includes many sample student responses. In spite of the efforts to provide comparable training for all scorers and to assess their ability to score pre-scored student work in a comparable manner, some differences in human scorers will be observed. The reliability of the scorers is important, however, in judging the accuracy of the resultant student scores.

If no reliability information is available for an assessment, the Department may wish to propose a pilot test of the assessment in order to gather student results that can be used to estimate the reliability of the assessment. Because reliability is a necessary but not sufficient consideration for determining which assessments to use, it is essential that reliability data either be located or secured.

For validity, again several types of information may be provided. These include content validity, predictive validity, concurrent validity, and construct validity. Another type of "validity," called consequential validity, may also be mentioned, although this is technically not an "official" form of validity in the technical standards that guide the work of assessment developers and users.

Content validity is the most prevalent form of validity evidence that achievement measures must muster. This can be done in several ways. One way is through the manner in which the assessment was created, especially the manner in which the accurate measurement of the standards was emphasized through item development, item editing, item reviews, pilot testing, and final item revisions. The process of item development is one key way to establish the validity of the assessments. Second, the types of alignment studies described in the previous section can also be used to verify the content validity of the assessments. Finally, expert panels can review the assessments to assure that they faithfully measure the content standards.

Predictive validity is another form of validity that may pertain to these technical skill assessments. This type of validity is important for assessments that claim to predict future events, such as readiness for post-secondary training or readiness for work (success at work). These claims must be substantiated by data that shows that the criterion (school grades or work success ratings) are correlated with the predictor (the assessment scores in question). This sort of validity is only pertinent for assessments that make claims about how performance on the assessment is related to some other performance measure in the future.

Concurrent validity is important only for those assessments that claim to produce comparable results to another assessment. Usually, such claims are made for assessments that are revisions of a previous one ('this test produces the same results as the previous version of this test') or that the new assessment is similar to an assessment that is more expensive, time consuming or challenging to administer.

Construct validity is often associated with assessments that purport to measure psychological constructs that are difficult or impossible to measure directly. Such measures are often difficult to directly validate, so a process of "triangulation" may be used. For example, a measure of test anxiety might be validated relative to other

measures of anxiety, observations of students' physiological responses in stressful situations, and observations of students taking tests. Together, this evidence might suggest that students who score high on the test anxiety scale also exhibit signs of anxiety. This sort of validation will not occur often in career tech programs.

The final technical issue that needs to be investigated is the steps that the assessment developer has taken to assure that the assessment is free of bias and does not contain sensitive nor stereotypical materials. These steps could include mentioning in a technical report how item writers were trained to avoid such biases, whether and how the items were specifically reviewed for bias and sensitivity, and what statistical means were used to flag potentially biased items. The usual statistical procedure used is called differential item functioning or dif. Dif statistics are run for important sub-groups likely to take the assessment, including sex, racial-ethnic group, community type, socioeconomic status, geographical region, and so forth. The essence of the dif technique is to determine if any sub-group scored statistically significantly lower on any test item. If the answer is affirmative, the flagged items are then reviewed by a bias committee comprised of experts in the detection of bias in test items.

Not all flagged items are eliminated, since differences in student performance could be due to instructional differences or opportunity to learn differences. In these cases, it is important not to delete the items because these data will be used by educators to provide greater and more effective learning opportunities for these sub-groups of students.

Determine the Feasibility of the Assessment – As mentioned earlier in this paper, the feasibility of the assessment is more of a subjective matter. Several factors go into helping to judge this attribute of the assessment. First, how complex is the assessment itself? How many parts does it have, how long is each of these, and how many types of assessment exercises are there in the entire assessment? Are some parts group administered and other parts individually administered? Anything that adds complexity to the assessment can reduce the feasibility of it in the minds of users. Of course, some of this may be offset by the perceived benefits of using the assessment. If the assessment uses assessment techniques that simulate on-the-job performance, users may be more willing to overlook the complexities of using the assessment in order to gain what they see as more useful information about student achievement.

Second, how difficult is it to learn how to administer the assessment? If the assessment is a paper-and-pencil one then likely it is not too complex. Even adding a constructed-response section to it will not add much to its complexity. However, an individually-administered section, with directions to the test administrator that must be read to students verbatim, adds greatly to the complexity of learning how to administer the assessment. Some online assessments may be challenging for assessment administrators, especially if the online system is not familiar to them and they need to set up the system so students can use it. Also, the more sections to the assessment, the more complex it is, especially if the assessment sections have to be administered in a particular order.

Third, how challenging is it to score the assessment? Some assessments require nothing more than returning a set of answer documents for machine scoring elsewhere. In other cases, however, teachers must first learn to use a scoring rubric and then score the constructed responses of the students, placing their ratings on the students' answer documents. This is time consuming and may add to the unreliability of the assessment. However, by teachers having a chance to score student responses, they

are able to better understand the instructional consequences of using these sorts of assessment items. In other cases, teachers may need to learn the scoring procedures and scoring rubric prior to administering the individually-administered portions of an assessment. Again, this is time consuming and may produce somewhat unreliable results.

If local scoring is desired, one way to do this is to set up scoring sessions at intermediate school districts or regional sites. In this manner, teachers from the region can be trained to score student responses and could work for a day, two, or more in scoring responses from students in their region. If scoring is done locally, however, there is the issue of teacher familiarity with the students and how this may affect the scores that students receive.

In other cases, the students' responses are returned to a test contractor who has hired staff and trained them to provide professional scoring of the assessments. While this certainly reduces teacher effort and expense, as well as increases reliability of the data, it is more expensive. Because teachers may not see actual student responses to these assessment items, the instructional value of using them may be somewhat diminished as well.

A fourth factor to consider in judging the feasibility of an assessment is what provisions have been made for the participation of students with disabilities and English language learners? Some assessments have made explicit provision for the assessment of these students including recommended (and prohibited) assessment accommodations. For those assessments that have not made such provisions, local educators may wish to use the same types of accommodations as permitted on the NCLB-required general assessments of mathematics, reading/English language arts, and science. These accommodations are generally available from the state education agency and can serve as guides to how these students participate in the assessments.

Once all of these factors are "added up," the user can make a judgment about the overall feasibility of using the assessment. As mentioned above, the feasibility is judged somewhat on the basis of the perceived value of the assessment in the learning process.

Determine the Cost of the Assessment – The financial cost of using an assessment is made up of several components. These include the costs for the administration manual and other assessment administration materials, the student booklets and answer documents (if any), the costs for scoring the assessments, and the costs of reporting the assessment results. There are other costs that might be incurred in some instances, such as the costs of any special assessment administration materials (such as those that might be used in an individually-administered assessment) or special costs for online assessment (the costs of setting up each testing site, for example).

The variable nature of these costs makes determining the actual costs of using an assessment a bit challenging to determine. This, in turn, makes cost comparisons between assessments challenging as well. Making the cost determinations even more difficult can be the discounts that users may receive if large quantities of materials are ordered or large numbers of students are assessed.

In order to determine the actual costs associated with the use of an assessment, the user should carefully delineate the numbers of students to be assessed, the number of unique situations (classrooms) where the assessment will occur, the number of

individuals who will receive results (i.e., teachers who will receive classroom results; administrators who will receive school results; administrators who will receive district or ISD results), which types of results are desired, and whether any ancillary features of the assessment are to be ordered. The user will also need to determine if substitute teachers need to be hired or if other teacher expenses need to be covered. By systematically laying out the assessment order in this fashion, it should be possible to accurately estimate the costs of using the assessment.

Should the Assessment be Pilot Tested? – There are a couple of reasons why the state might want to pilot test an existing assessment prior to adopting it for use within the state. These include the following:

- The assessment is new and no technical data exists for it. Although it is possible to use the assessment and obtain the needed technical information for it from the operational use of the assessment, there is the risk that the assessment will prove to be unreliable or not valid (e.g., predict future post-secondary educational success). Using the assessment and then finding this out means that students have been misled to some extent and perhaps programs poorly evaluated. While these aren't necessarily dire consequences, they could be prevented by pilot testing these new assessments first.
- While the assessment is not new, the existing technical data is not applicable to the state's intended use of the assessment. There will be occasions where an available assessment looks good, but it was not developed for use with high school completers. It may have been developed for an industry-based program or a post-secondary program. This does not make it inappropriate, but certainly questions could be raised about the suitability of it for the target group of students. One way to address these potential concerns is to pilot test the assessment with a sample of students
- The existing data for the assessment is dated and there is a need to obtain updated information for the assessment. If the assessment is a norm-referenced one and the normative information is over 5-8 years old, it may be desirable to obtain more recent normative information. In this case, obtaining newer data may be useful in determining whether the assessment is suitable for students at the current time. For example, pilot testing the assessment in a sample of schools can help answer whether the content of the assessment is suitable. More importantly, it will provide more recent normative data applicable to students currently. However, if the real desire is to create more recent norms, an informal pilot test, using volunteer schools and students, will not provide an accurate normative data for future use.
- The state's schools aren't sure whether to use the assessment, so pilot testing will allow them to "see the assessment in action." By trying out the assessment, teachers and administrators will have a chance the assessment as students take it. This will help them to decide whether the assessment is relatively easy to administer and score, as well as whether it provides useful information. While users can sometimes come to this conclusion just by examining a specimen set of materials, the actual use of the assessment will provide more definitive information about the assessment. This can help users make a more informed judgment about suitability of the assessment.

Determine Whether to Use the Assessment – When all of these factors have been considered and their relative merits have been weighed, it is then time to decide whether to select the assessment for use in the technical skills assessment program. Sometimes, this choice is an easy one. The assessment is well aligned to the content standards of the program, the technical data show that the assessment is sound, it appears to be feasible and not too costly. In these cases, the decision is easy.

Often, however, the data is not so clear cut. In these cases, the relative merits or strengths of the assessment need to be weighed against the areas of weakness. While one option is not to adopt an existing assessment but to develop one from scratch, this is not an easy nor inexpensive solution, as the next section of the paper will point out. This makes the decision even more difficult – do we accept a less-than-perfect measure or set out to create one of our own, which probably will be better, but won't be available for a year or more, will be somewhat costly to develop, and may prove not to be so superior to the existing measures.

It may be helpful to use an advisory committee – either a standing one or one assembled for this purpose – to help make the decisions about which existing tests to use. Such an advisory committee should be comprised of educators who work in the technical skills area, along with business and industry representatives, post-secondary educators, and others who would be able to help the agency determine whether the assessment is sound and whether the scores resulting from it would be useful – to employers and to educators.

Unfortunately, there is no existing formula to help users decide what to do in these instances. Decisions may be dictated by the immediacy of the need for data, by budgets, and the availability (or lack thereof) of resources to create assessments locally. All of these factors need to be weighed in making a decision which will neither be easy nor perfect.

## Section Four

### Creating New Assessment Instruments

If the user has arrived at this point, the assumption is that either no existing assessment meets the needs of the user, or that no assessment exists at all. Thus, the user is contemplating creating an assessment from scratch. This is not a step that should be taken lightly because of the time, effort and expense involved in creating (and maintaining) a new assessment. However, if a technical skills assessment is needed and none is currently available, there may be no choice but to proceed. Before doing so, there are several things that should be considered. These include:

1. Will the assessment development work be done with state and local staff or will an assessment contractor be used?
2. Who will actually lead the assessment development project?
3. What entities will be involved in some capacity in the development project?
4. Is it possible to collaborate with other agencies (states or ISDs) in creating and administering the assessment? If so, how?
5. How will the project be paid for?

**Using an Assessment Contractor** – The decision about whether to use an assessment contractor or to do the work with local educators and others is often one made on the basis for availability of funding and appropriate staff. Assessment contractors will typically be able to handle much of the development work and the logistics involved in the assessment project, thus relieving local and state educators from much of the work. The downside is that they are expensive – costing up to \$1 million per assessment. However, contractors can bring expertise that might not otherwise be available, can carry out much of the mundane (and not-so-mundane) aspects of the developmental work, and can bring a level of technical expertise to the project. Plus, they are more convenient.

Assessment development work can be done less expensively, but to do so requires locating suitable staff to direct and manage the project. Finding suitable staff who can devote the time necessary to successfully carry out a development project can be a challenge. The advantages of a “home-grown” project can include providing local educators with more development experience, creating assessments more suited to the state, and creating the assessments at lower cost. Balancing this is the much larger amount of work to manage the development project locally.

The staffing described below will be needed regardless of who develops the assessment. The difference will be how much time is needed from a project director, a project manager, and clerical staff. This will partially offset the cost savings from doing the work locally.

**Leading the Assessment Project** – A key decision that will need to be made at the outset of the development project is to select an agency and/or individual(s) to lead the developmental project. This may vary from assessment to assessment, and need not be the same for each of the technical assessments to be created (although there is something to be said for keeping the assessment work consistent by using the same entity).

What is important is the leadership of the assessment development activities should come from someone who is familiar with how assessments can be created and with the

assessment training and management experience to successfully plan the project and manage it to successful conclusion. Because managing a project as complex as an assessment development one is not always straightforward without issues, the experience will come in handy to develop workable solutions to the issues that will inevitably arise during the course of the development work.

Assessment project managers can come from several types of organizations. These include local and intermediate school districts, universities, non-profit and for-profit organizations, and retirees from such organizations. Note: it is not necessary for the project director to be someone who is familiar with the technical skill area for which the assessment will be created. The individual's assessment experience is more important than their content experience.

The assessment director position will undoubtedly be a part time one, unless the organization and the project director are involved in managing multiple development projects.

**Others to be Involved in the Project** – Although it is not necessary for the assessment project director to be a content specialist in the technical skill area to be assessed, it is important that individuals who have this experience be actively engaged in the project while the development work is occurring. Depending on the breadth of the technical skills area to be assessed, one person may serve in this role, or additional persons may be needed to cover the breadth of the area. The content specialist(s) will work part time on each assessment development project.

The individual(s) hired for this work should be ones that local educators and college specialists would trust to make content-related decisions about the project. Just as the assessment project director is a key position, so is the selection of the content specialist(s). Because this individual will be making decisions about the content of the assessments, it is important that this individual not only be a trusted individual, but that they understand the breadth of the content area and can make sure that it is suitably represented in the development of the assessment. This means that the specialist should be someone who will not impose their own values in determining the balance and inclusion of various assessment topics with the new assessment.

In addition to the assessment project director and content specialist(s), there is a need to hire an individual to serve as a project manager to assist the director in carrying out the developmental work. For example, when it is time to locate suitable facilities in which to develop the assessment items, the project manager will contact various facilities, determine their suitability for the work to be done, and determine which facility to contract with to carry out the needed work.

Finally, there will be need for project clerical staff throughout the project. This may include staff that is experienced in desktop publishing of the assessments. The number of staff needed, and their skills, will depend somewhat on what work project staff does and what work might be contracted out.

**Collaborating with Other Agencies** – One of the ways of reducing the costs of developing an assessment is to create collaboratives that share in the work to be done and the costs for doing so. There are a couple of ways in which other agencies could be engaged in the work. The first of these is to see if there are other states that are interested in the development of the assessment as well. If there is such interest, a group of states might work together to develop the necessary content standards that

would be shared by the states, and then to create the assessment to measure those content standards. One way in which this collaborative work might be done would be for the participating states to each nominate item writers who work under the direction of a contractor or central staff to create the assessments that eventually all of the states could use. Once the assessments are created, the availability of the assessments could be advertised to other potential users (e.g., other state education agencies) who could purchase the assessments at a cost that would allow the developing states to recoup at least part of their development costs and/or put money aside for further development of new forms of the assessment.

A second approach might be to work within the state to network intermediate and local school districts to share in the costs of developing the assessment. This approach might be realistic since local districts receive the funding for assessment-related activities, but the funding provided to any one entity would not be sufficient to permit them to develop the assessments. If a group of entities such as intermediate school districts banded together and shared the development costs, they could create an assessment that they could use, and which they could sell to other potential users within the state and elsewhere. A mechanism for such collaborative work – the Michigan Assessment Consortium (MAC) – already exists, and although the MAC was formed to foster the development of high school course assessments, it could easily turn to the development of technical skill assessments by interested intermediate and local school districts.

**Paying for the Project** – As mentioned above, there are several ways in which the costs for developing needed assessments could be covered. These could include one or more agencies paying the costs, collaboratives of states or collaboratives of intermediate and local districts sharing the costs, or seeking external funding – from governmental agencies (e.g., the U.S Department of Education), foundations, and industry groups. Each of these possibilities has advantages and disadvantages, and these should be explored in the early stages of project planning.

The advantage of a group of entities (such as states or intermediate and local districts) working together is that costs for such work can be shared among several participants. This means that either the assessment development work will be far less expensive for any one agency, or that the amount of work that can be done for the same price as working alone will be far greater.

There are several examples of collaboratives of states. These include the Council of Chief State School Officers' State Collaboratives on Assessment and Student Standards (SCASS). At the state level, the newly-formed Michigan Assessment Consortium could be a very good vehicle for collaborative work among intermediate and local districts.

**Preliminary Development Activities** – There are a number of steps that need to be taken when determining how to develop assessments of a technical skill area. The purpose of this section of the paper is to suggest processes and guidelines that should be used to determine if the assessments that are planned will be created in a sound manner. That is the purpose of this section of the paper – to indicate the steps that should be taken before actually creating these assessments.

Assessment Purposes – One of the first decisions that an assessment developer must make is what are the purposes of the assessments that will be created. Assessment often serves more than one purpose, so it is important to both list all of the intended purposes, and also to prioritize these. Some common purposes for assessment include:

- Guide the learning of individual students
- Determine what instructional needs remain for a group of students following a unit of instruction
- Assure that students are taught the breadth and depth of content in the course/credit area
- Certify student accomplishment of the key skills in a technical skills area
- Hold educators accountable for student learning
- Select students for enriched instruction or special programs
- Predict how well students will do on a subsequent examination
- Predict how well students will do in subsequent instruction, such as in college

These are just some of the possible purposes that could be ascribed to an assessment in a technical skills area. Each of the purposes should be determined first, and then the purposes should be rank ordered from most important to least important. The following is an example for a technical skills assessment:

1. Determine whether students have learned enough to earn credit for the course/credit area.
2. Predicting how well students will do in subsequent instruction, such as in community college
3. Hold educators accountable for student learning

Next, the assessment developers will need to determine whether the assessment purposes can be accomplished with a single examination or whether more than one assessment will be needed. One common conflict in purposes that makes use of a single examination problematic is when developers want to produce information to determine individual student learning needs versus producing data to hold educators accountable. The former purpose probably calls for assessments that are customized for each student and that are extensive in nature, while the latter purpose calls for a common assessment that is given to all students and may be much shorter. The accountability assessment will not be able to do a good job of showing individual student learning needs, just as the individualized assessments may make it difficult if not impossible to hold educators accountable. In this instance, different assessments may be required.

If more than one measure is used, the sponsoring agency will need to determine how to combine these two measures, both physically and psychometrically. Physically, the two tests might be printed as one instrument, or administered separately. Once the two measures have been used, the scores from each will need to be combined and reported on a common scale, a task for a psychometrician, as well as reported separately (if this makes sense).

It is typical to engage a representative advisory group to help the agency make these sorts of decisions. It can also help to determine how extensive the assessment program should be.

Assessment Reporting – It may appear odd to discuss the reporting of assessment results prior to a discussion of the assessment instruments, but assessment developers find it useful to do so. The types of assessment reports to be produced and the depth of detail on these reports should reflect the purposes for assessment, and will guide determinations of the length of the instruments to be used.

Here are some of the reporting levels that should be considered:

- Individual student reports for teachers – developers will probably want to report assessment results at the student level. Such reports could include how students did on each test item, each content standard, and/or each strand or cluster of expectations. Reporting at the student level could indicate that students have passed or failed each of items or expectations, or may be an indication of the total number of items passed.

Besides reporting student performance on the individual test items and perhaps content standards and strands, overall reporting may also occur. For example, overall student performance may be reported in multiple performance levels, and these performance levels may also be labeled (e.g., advanced, proficient, partially proficient, and not proficient). Or, student performance may be expressed solely in terms of pass-fail.

Overall scores may be reported in raw score terms, or the raw scores may be converted to scaled scores. The latter is useful if different forms of the assessment will be used each year, or across several years. It is virtually impossible to have the same raw score passing score on multiple forms. It is equally difficult to explain to lay users why one form has a passing score of 27 out of 35, while another form has a passing score of 28 out of 35. One will seem more challenging and the other easier, even though statistically, they produce comparable results when given to the same students.

One caution: each level (e.g., content standard or strand) chosen for reporting at the student level should have multiple test items measuring that level. The rule of thumb often used in the measurement community is to have about 10 score points per skill to be reported. Because there is measurement error inherent in any assessment, it is important to use multiple test items to measure the lowest level to be used in reporting at the student level. If only one item per expectation can be used, it will be important not to report at the expectation level, but perhaps at a strand level so that reports at this level are based on student performance on multiple test items measuring that strand. This provides more stability in the student scores.

The requirement to report scores that are stable (reliable) is why it is important to understand reporting requirements in the design of the assessment. This helps to prevent disappointment later when the assessment is actually used and reported for the first time.

- Student and/or parent reports – Another important report to provide is a report of the results to students and their parents or guardians. While students want to know whether they passed or not, this report can provide substantially more information both to them and their parents. Depending on the depth of coverage of the assessment instrument, this report might show how students did on each test item, on each content standard, and/or on each strand of standards. This data may be helpful to students who did not pass the assessment for them to determine what material that they have not as yet mastered that they should study for a subsequent re-test. In some cases, it may provide a longitudinal record of student performance, if the student has attempted the assessment more than once and the district has maintained a running record of students' performances.

In addition to item, standard and/or strand data, the individual student or parent report will also probably report the students' overall levels of performance on the assessment. This may be a simple pass-fail, or might include the students' performance relative to several performance standards. These performance levels might serve to differentiate among the students. For example, students who score at the "advanced" level might be recommended for a subsequent class, while students who score at the "not proficient" level might be expected to participate in remedial instruction either after school or during the summer in order to improve their performance. These uses of the overall student performance are optional at the district level, however.

- Classroom roster reports – These reports serve as a handy summary of student performance at the classroom level. If provided at the classroom level, teachers can quickly determine which students in each section need additional help, and they can equally quickly determine on which skills students need assistance. This can permit teachers to determine a plan of action – how to address common needs among all students, which sub-groups of students need assistance on a subset of skills, and which students need more intensive assistance.

These "classroom" reports can also be used in other ways. For example, the school can have students indicate their guidance counselor rather than teacher. This would permit each counselor to receive a report of results. This could serve the useful purpose of helping the student and counselor determine what classes the student should enroll in for the following semester or year. Other schools may wish to use the roster report by having students code their program rather than their teacher. This might permit schools to examine how students in career-tech programs do in comparison to students in general education programs. Still other schools will have students use just one code for all students, thus receiving an alphabetical list of all students taking the test in the high school.

- Item analysis reports – These reports show the total performance of all students in a group on each test item used in the assessment. This report has two useful purposes – to help educators determine the learning needs of groups of students, as well as to determine the quality of each of the items on the test. The first use occurs when teachers examine the performance of students on each test item. If the items have been carefully crafted, the incorrect responses built into multiple-choice items, or used as the basis for the scoring rubrics used with constructed-response items, can provide valuable insights into the thinking of students. These insights can provide guidance to the educators providing remedial instruction to students, as well as to teachers as they revamp instruction for the following school year.

Test item analyses can also be used to refine the assessment instruments. Items that more high performing students answer incorrectly than students who did poorly on the overall assessment may indicate a flawed item that should be discarded or revised. If one of the incorrect answer choices for a multiple-choice item does not attract any students, then that answer choice should not be replicated when subsequent versions of the assessment are created. This is especially invaluable if incorrect choices are based on presumed incorrect student reasoning.

In both cases, educators should examine the quality of the assessments after they have been used, and before the assessment results are released to students – and important decisions are made based on the assessments. These reports can aid in the post-administration examination of the quality of the assessments.

- School summary reports – These reports provide useful overall information on the performance of all students in the school (or district). Because parents and other members of the public are quite interested in the achievement of students in our schools, but most reports of results contain individual-student identifiable data on them, this report provides a way of depicting overall performance without disclosing confidential information about individual students.

The summary report might include just basic information about the percentage of students passing or failing the assessment or scoring at each performance level. In addition, this report could also show the percent of students passing each test item, and students' performance on each content expectation and strand.

This report might be an overall summary for all students, or disaggregated versions of it might be prepared for important subgroups of students. Examples of disaggregation categories include poverty, race-ethnic group, gender, special education status and English language status.

- Special Reports – The list of reports provided above covers many of the needs of students, teachers, other educators, parents, and other citizens. However, schools may have other reports that they wish to prepare. These reports should also be determined at the outset so that the assessment design used can provide the data needed to assure sufficient technical quality.

As mentioned above, the critical decisions at this point in assessment development deal not only with the types of reports of results to be provided, but also the *depth* of reports – at what levels will student and group performance be reported. These are essential questions to answer at the outset so that an assessment design (e.g., the number of items to be used at the lowest level of reporting of student results) supports the nature of the reporting that is desired. If an advisory committee is being used to advise on the technical skills assessment, the type and design of the reports of assessment results would be a good topic for them to consider and assist the agency with resolving.

Assessment Design – There are a number of important considerations when designing an assessment. Some of these are discussed below:

- Type of Assessment – One key feature of assessments is whether they attempt to cover a broad spectrum of skills or whether they hone in on a much narrower set of skills that are assessed in greater detail. This can be characterized as “breadth versus depth.” Essentially, a “broad” type of assessment is one that may cover a large number of skills, but very few items per skill. This survey-type of instrument might be developed to assess a two-semester sequence in a technical skill area. Given restrictions on testing time, such an instrument might assess fifty, sixty or more skills, each with a single test item. The result is that we can assess whether the student has learned the broad scope of the two-semester course.

On the other hand, the “depth” type of assessment is designed to assess a far smaller range of skills, but to do so with a much larger number of items. This type of assessment would allow a teacher to determine more fully whether students have acquired the depth of knowledge and skills that would commonly be associated with a unit of instruction. This type of assessment would be most appropriate where we want to make fine distinctions between students’ learning of a limited set of skills.

Both types of assessments are useful, but in different places. The survey assessment could provide a useful summative, end-of-course assessment for use at the end of instruction. The in-depth assessments would be useful in ascertaining student learning during the course of instruction. The survey assessment would most likely be reported at the strand level, while the in-depth assessment can usefully (and soundly) be reported at the individual test item level and summarized at the content standard level.

Of course, often times we want to accomplish both purposes for assessment – and that is when compromises in assessment design are needed. As mentioned earlier, the assessment design and assessment reporting are intricately linked.

- Speed or Power Tests – There are advantages to each type of assessment. Speeded (timed) tests are easier to administer (their length is fixed by the time students are allotted to finish it), but can place additional pressures on students to perform quickly as well as adequately. Timed tests are also used to help differentiate among students on norm-referenced tests.

Untimed (power) tests are more challenging to administer - what does the school do with a student who isn’t finished when the testing period is completed? However, the advantage of these assessments is that they permit each student to be completely assessed on each skill on the assessment.

- Length of the Assessment – The length of the assessment – whether expressed in terms of the number of test items, testing time, or both – is one of the most important pragmatic decisions made on assessment design. Some times the length of the assessment is dictated by convention, such as the length of a class period, or traditions that say final exams are two hours in length. The length of timed tests can be fixed, while the length of an untimed assessment is much harder to predict.

Before a test is used initially, it can be challenging to predict the amount of time students will take to complete it. Thus, in some settings, the field test of the instrument is administered in an untimed manner, and the time that each student takes to complete it is tracked. This can help the assessment developers determine how many items can be administered in a fixed period of time, or what time to allocate for an untimed test so that most students can finish it within that period of time.

- Types of Items Used in the Assessment – One of the most critical (and contentious) decisions in assessment design is the decision about what types of items to use on the assessment. Many tests use a combination of multiple-choice and constructed-response item types. However, there is a much broader array of assessment types that could be used. Each assessment type has its strengths and challenges, so a good design is one that does not rely exclusively

on one assessment type.

*Multiple-choice items* are useful in assessing content knowledge, which could be an important component of many assessments. They are quick and easy to administer, do not require any special scoring after administration, and can be readily understood by students. These items are susceptible to guessing. High quality multiple-choice items can also be very difficult to create, especially if the assessment developer wishes to avoid factoid-type questions.

*Constructed-response items* are good at assessing student understanding of concepts and the ability to use the knowledge learned. Because students have to construct a response, guessing plays little role in students' responses (although less knowledgeable students may try to "bluff" their way to a high score). Developers who build these types of items need to also create a scoring rubric and a scoring guide so that the student responses are scored in a consistent and reliable manner. Ultimately, after field testing, the rubric should contain a rationale statement for the preferred answers, a descriptive statement about each score level in the rubric, and samples of student work at each score level. While these items are easy to develop initially, there is much more work inherent in these items by the time the assessment is ready to administer.

*Performance items* can be useful when we want students to demonstrate a skill that they should have acquired. For example, we may want students to give a presentation, program a computer, repair a car, prepare a patient for a medical procedure, conduct an experiment, or discuss an important ethical issue in a small group. These are examples of the performance assessments that could be included in an assessment program. For each of these assessments, a detailed assessment administration protocol would need to be created, along with a detailed scoring guide which might include audio or video excerpts in order to adequately train the scorers of students' performances. These assessments also need detailed scoring guides with samples of student responses that might include photographs, video clips or audio clips.

*Portfolios* can serve the useful purpose of documenting the progress that students have made in learning. These may be working folios or may serve as a mechanism to display the best of each student's work. Portfolios can also be scored. If this is desired, a detailed scoring rubric would need to be created. It should describe how the work contained in the portfolio will be scored (as a whole or individually), what characteristics of the work will be examined (quantity, quality, measure of the intended content standards, or change over time are just a few of the dimensions on which a portfolio might be scored), and the logistics of gathering student work, selecting it for inclusion in the portfolio, and how it will be displayed. In addition, students might be asked to write an overview of the work in their portfolios. This overview might include a reflective piece on what the student had learned. This reflection can be presented in writing or might be given orally (or both).

- Test Administration Mode – There are several choices in how assessments might be administered. The most common mode of administration of tests containing multiple-choice and constructed response items is using a paper-and-pencil format. Typically, a separate test booklet and a scannable answer sheet or folder is used. More recently, computerized online test administration is becoming more common. The advantages of this administration mode include less cost for

administration and scoring, quicker return of results, and perhaps increased student motivation to take the assessment. However, online assessments are more expensive to set up initially.

Performance assessments might be presented to students on paper, using audiotape or videotape, or individually by the teacher or other professional. The mode of presentation will vary depending on the type of performance that the student is expected to produce.

- Student Response Mode – Students respond to most assessments by recording their answers on an answer sheet or folder. For performance assessments, students might respond individually or in small groups, and their responses might be spoken, written, or performed. In many cases, an observer is the recorder (and perhaps scoring the student response at the same time). The observer may simply record student responses (using a tape recorder, video recorder, or a checklist), or the observer may have a detailed scoring guide that they use to rate the student performance while it is occurring.

Developing the Assessment Blueprint – There are a number of steps that are involved in developing high quality student assessments, and a number of ways that these steps can be accomplished. However, good assessments begin with assessment blueprints, which serve as the road map to the assessment. The contents and level of specificity of the assessment blueprint will vary, but there are a few essentials that a good blueprint should contain. These include:

- An indication of which content standards will be measured by the assessment.
- A precise determination of how each of these skills will be measured. This could include simply that certain skills will be measured with multiple-choice items, while other skills will be measured with constructed-response items or performance measures. Or, this could involve detailed specifications about how each skill will be measured, such as which mathematical errors will be built into the answer choices of multiple-choice items measuring a particular skill.
- An indication of how many items of each type will be developed for each skill.
- An indication of the total length of the test, the number of test sections, and the composition of each test section.
- The manner in which the assessment results will be reported, and an indication that the test design will permit the types of reporting planned.
- An overview of the processes to be used to develop the assessments, including the manner in which draft items will be edited, pilot tested, reviewed, and how the final instruments will be assembled.

The assessment blueprint may be as brief as a few pages, or as lengthy as 100 or more pages, depending on the depth that each of the above-listed topics are dealt with.

Development Project Staffing – The types and levels of staffing to be used in the assessment development project will depend highly on whether the work is contracted out to assessment development organization or whether the work will be done by the sponsoring agency. If the former is the approach chosen, the sponsoring organization will need to have a project manager, to oversee the work of the contractor, at least one person familiar with the technical skills area, to serve as the specialist to respond to content questions that inevitably arise, and perhaps a clerical person to assist with the logistics of the project (e.g., locating suitable facilities for item development meetings, arranging sleeping facilities for item developers, and so on).

If the sponsoring organization chooses not to use an external contractor to support the development work, additional staff will be needed. These include a project director, who should be someone with a substantial assessment training and experience, particularly in item development, a project manager who can keep the myriad of details of running a complex, multi-faceted project straight, several content specialists who are individuals quite familiar with the technical skills area, and one or more clerical staff to handle logistical details of the project.

In addition to these project staff, it will be important to include two or three technical advisors (or to use an existing technical advisory committee) to assure the technical soundness of the assessments. Also, to assure the correctness of the technical skills assessments, a panel of about five content experts should be used to judge that the technical skills assessments created cover the essential areas of the technical skills area and that the assessments are worded effectively and scored correctly. Finally, one or two individuals who are aware of bias and sensitivity issues should be employed to assure that the assessments are free from bias and content that would be sensitive to any students.

**Tips for the Development of State and Local Assessments** – Once the assessment blueprint has been drafted, reviewed, and accepted, it is time to begin the development of the actual assessment exercises. As mentioned above, there are a number of ways in which assessment exercises can be created. What is presented here is one model, based on the manner in which the Michigan Department of Education creates the assessment items that are used in statewide assessments. There are other viable methods of item development, so readers are encouraged to try other strategies that may suit their particular situations. This process takes about two years, starting with development and ending with the use of field-tested items on a statewide assessment.

The typical test used at the state level costs somewhere between \$250,000 and \$1 million to create; however, it is possible for assessments to be developed for far less money. At the outset, however, it is important to point out that poorly developed assessments are more than a waste of time and money. They may in fact give educators, students and parents misleading information on student achievement that may in some cases underestimate what students know and can do, or in other cases, overestimate what students have accomplished. Either of these may have negative consequences for students.

An initial decision that test developers need to make is whether to create new items from scratch or whether to obtain items that have already been created and used elsewhere. A number of banks of items are available, some from commercial sources, some from items in the public domain. The advantage of using existing items is that the task of creating assessments might be substantially easier (and perhaps less expensive). However, commercial sources typically charge to use the items in their item banks each time they are used. Item banks may not cover the entire spectrum of skills that need to be measured. In addition, it may be difficult to assure that students have not already been exposed to items that are publicly available.

Even if item banks are used to start with, developers may still need to engage in item development – to fill in the gaps of coverage of the content standards to be assessed. In addition, it is critical to note that an assemblage of test items does not automatically make a reliable and valid test instrument, so the items should be field tested together in order to ascertain that the items “hang together” and make a coherent measure.

Thus, even those who pick items from item banks need to engage in at least some of the steps that are outlined below.

1. Determine the technical skill areas to be assessed – assessment developers will need to determine the areas in which assessments are to be created. Developers will also need to determine which assessments are to be created first. Priority might be given to assessments for technical skills areas to be most used or those for students in earlier grade levels. A schedule for development will help to assure that the resources needed for development are available.
2. Determine the type(s) of assessment to be created – There are several types of assessments that could be created. Each has a different purpose yet fit together to make a coherent and balanced assessment system. These include:
  - Summative assessments – These assessments are given towards to end of instruction and serve to provide an overall view of student achievement at the conclusion of a course of study. These might be traditional end-of-course assessments given in the spring. These assessments typically take about two hours in length, provide a survey of student achievement using a mixture of multiple-choice and constructed response items.
  - Interim assessments – These assessments are based on instructional units or occur periodically (e.g., on a quarterly basis). The purpose of these assessments is to provide information on student accomplishment of units of instruction. This will provide teachers with information on student instructional needs. These assessments typically also combine multiple-choice and constructed-response items, but might incorporate performance assessments as well. A two-semester technical skills area might include between eight to twelve such interim or unit assessments, and student accomplishment of them might replace the summative assessments in determining student technical skills area achievement.
  - Formative assessments – These assessments are used daily by classroom teachers during the course of instruction. The purpose of them is to help teachers (and students) to determine whether they are learning concepts, ideas, and skills as they are being taught. They provide immediate feedback to teachers so that they can adjust instruction and provide students with learning opportunities they need. Typically, these data are not formally recorded and used for accountability purposes, although they may well be used for student grading purposes.
3. Identify the content standards that will be assessed in each technical skills area – Once the technical skills area(s) are identified, the types of assessments to be created are chosen (summative, interim, and/or formative), and the item types selected (multiple-choice, constructed-response, or other), the assessment developers will need to determine which content standards will be measured. The type of assessment will undoubtedly dictate the length of the assessment, and this will largely determine the number of expectations that can be measured.

Typical summative, end-of-course assessments will run about two hours in length, measure about thirty to seventy skills, with about fifty to seventy multiple-choice and two to five constructed-response items total. Areas that require longer constructed-response items will probably use fewer (maybe as few as two or three) but with much longer prompts.

Interim assessments should take less than one class period each, which means that probably no more than ten to fifteen skills can be measured, each with multiple items.

A fifteen-skill test comprised of three multiple-choice items per skill and no constructed-response items should take no more than 45 minutes. If constructed-response items are to be used, fewer skills can be covered and/or fewer multiple-choice items per skill should be used.

Formative assessments typically are embedded within instruction and no set number or time allocation can be provided.

4. Draft an assessment blueprint for the potential assessments. See the section given above for more ideas. Determine:
  - a. The purpose(s) of the assessments.
  - b. How the assessment results will be reported – to whom and how
  - c. Determine what types of items will be used – multiple-choice, constructed-response, and performance
  - d. The assessment design that matches the purpose(s) and intended reporting – determine number of items and number of items per expectation

Assessment developers should develop a written assessment design, since this will serve to guide their work in developing the needed assessments. The assessment design should be the core of the assessment blueprint, which should contain the responses to the list of questions above as well as other important aspects of the assessment.

5. Determine the highest priority content standards for the identified grade levels – This step is necessary because often there are more standards than can be covered by an assessment in a reasonable period of testing time. Hence, it is necessary to select the highest priority expectations so that the assessment focuses on the most important areas of the content expectations. Other skills can either be assessed on a matrix sampling basis or not assessed at all.
6. Finalize the assessment design – the expectations to be assessed, the number of items per expectation, test length, reporting structure, and so forth. Once decisions have been made about which expectations to assess with what types of assessment types and for what purposes, this assessment design should be written up so that the plans can be shared with others who either will be affected by the assessment or who will help develop and implement it. This will also serve to assure that everyone understands the scope of the assessment development effort.
7. Determine if assessment items exist to cover the outline of the assessment design. If yes, go to step 8. If not, go to step 10. Often, there are assessment resources available that might be used to meet some or all of the needs for assessment exercises. This step will consist of a careful review of the content standards for which assessments are needed, along with the pool of available assessment exercises. Select the exercises that measure the expectations designated for assessment. Note that in carrying out this process, it is likely that there will not be available assessment exercises to measure each content standard selected for assessment. If this is the case, go to step 10 for the exercises that will need to be created.

8. If yes to step 7, select those to fill out the assessment design. In carrying out this step, make sure to be careful that the selected exercises adequately measure the standards selected for assessment.
9. Determine whether scoring rubrics and exemplars are available for constructed-response and performance exercises that were selected. If they are available, go to step 24. If not available, go to step 17.
10. If no to step 7, the needed assessments (and the scoring rubrics for constructed-response items) will need to be created. The following are steps that should be followed to create the needed assessments.
11. Select a panel of item writers who are familiar with the technical skills area. This might be a combination of K-12 educators, business and industry experts, and university technical skills specialists. The purpose of using classroom teachers is to use persons who are familiar with the students – the manner in which instruction is currently being provided, students' current levels of achievement and the types of activities that would engage students. Subject-matter specialists are used because of their awareness of current instructional trends, familiarity with state and national content standards, and their ability to assure that the assessments are technically sound and instructionally pertinent.

The number of panelists will depend on the number of exercises to be created, and the number of grades or course/credit areas to be covered. In the model laid out below, where panelists will meet twice for about six days total plus be expected to work for up to six days on their own between meetings, the typical item writer could be expected to write about 50 items (assuming that about 80% of these are selected-response items). To be safe, it is best to assume that only about 50% of each item writer's work will be useable, so project planners will want to produce twice the number of items needed in the end and then divide this number by 50 to determine the number of panelists needed.

Project planners will also need to determine whether one panel can produce items in more than one area. Having panelists work on assessments related to separate CIP programs within a pathway area is not a problem. Generally, having panelists working on item writing for adjacent grades (i.e., grades 11 and 12) is also not a problem. At the high school level, depending on the qualifications of the panelists, it may be possible for the same panel to write assessments for more than one technical skills area, or it may be preferable to have a separate panel working on assessments in each area.

12. Conduct an item writer meeting – typically two-to-three days. Train the item writers in the principles of good item development, then monitor their performance during the item writing session. The key to effective item development is to treat the item developers as if they are capable adult learners by providing them with quality instruction and then have mentor/coaches work with them to learn to apply these skills to their own work.

The first step is to prepare a good item development handbook. It should contain a description of the item development process(es) to be followed, a series of do's and don'ts about high quality multiple-choice, constructed-response, and performance assessments (including assessment administration directions and scoring rubrics, if

applicable), samples of good and not-so-good assessments of each of these types, and a checklist for item writers to use to check their work before submitting it.

The second step is to determine how item writers will develop and submit their items. Will paper-and-pencil be used? Computer-based development using word-processing software or item templates, or online item submission? Whatever method is used, the managers of item development will need to develop the necessary materials, as well as prepare for any computer-generated development. This may include locating a sufficient number of computers and assuring adequate access to the Internet.

Once the item developers are ready to begin, they should be oriented to the principles of good item writing. Following this orientation, the item writers are often divided into smaller work teams, where they begin to generate items. Once they have drafted a few items, a mentor should review the items and work with the item developer individually or in small groups to examine these fledgling efforts, determine areas where improvement is needed, and suggest methods for correcting these issues. Item writers should edit their work and generate new items. Then the process should be continued. This step is the key to producing high quality test items from new item developers.

Note: If the items are generated using word-processing software or specifically-created item generation software, the mentoring that teachers receive can come from either persons who are physically present at the item writing site, or could be off-site, at the contractor's work site, submitting edits and suggestions via the Internet, with telephone conversations if needed.

13. Edit the items that were created at the meeting and provide feedback to the item writers about how to improve the items – Once the initial meeting has concluded, the item developers should complete their initial development and submit it to the project contractor or director for review and light editing.

At this stage, the development staff should not invest too much time and effort to clean up the items; that is still the job of the item developers. However, the "light editing" mentioned above refers to pointing out to the item developers issues that remain with their items so that they can correct these issues before they come to the second item development meeting (and don't repeat these issues with new items that they will create before and during the second meeting).

14. Conduct a second item-writing meeting – At the second, three-day meeting of the item development team, each item writer will work on completing his or her assigned work. Depending on the progress that each developer has made, this might include completing their item development assignment, editing the work that they have already created, developing assessment administration guides for any individually-administered performance assessments, and/or developing the scoring guides and scoring rubrics for any constructed response or performance assessments.

At the second meeting, instructions on developing scoring guides and scoring rubrics will be provided to all item developers.

As at the first meeting, the development staff will need to be present to work with the item developers in small groups or individually. Each content specialist should

spend the time with individual developers reviewing their work, suggesting ways to improve it, and making sure that any assessment administration directions are worked out in their entirety, and that the scoring guide and scoring rubric are complete. One way that they can do this is to have the developer “administer” the performance item to them, or to suggest possible student responses to a constructed response item to make sure that all plausible student responses are covered by the guide and rubric.

Before leaving the second meeting, item developers should have completed their assigned items, or have scheduled with development staff when shortly after this meeting they will have completed their work.

15. Edit the items for pilot testing – Following the second meeting, it is now time for the development staff to take stock of the items that they have. Are all assignments completed? Were the needed items developed? Were all of the content standards covered with the numbers of items needed? How good do the items look – will they require a substantial amount of editing?

Once these questions have been answered, the project staff can begin their work of editing the items. They need to initially determine what if any art work or graphics will be needed, since these need to be ordered at the outset so that the work can be completed as the editing and desktop publishing is taking place. If the items use any copyrighted materials (i.e., text, photographs, graphs, charts, and so forth), this needs to be identified at the outset as well, since the owners of these copyright will need to be contacted in order to obtain permission to use the work in the test items. If permission is not forthcoming, or the cost of permissions is high, the item editors may need to re-write the item(s) so that it isn't necessary to use copyrighted information.

Each assessment item that was submitted by an item developer needs to be read carefully and revised where necessary. Editors should seek to make sure that the item does indeed measure the intended content standard, the question or stem of the test item is stated as succinctly as possible, that there is clearly one and only one correct answer to the question (and that this option is not the longest), that the answer choices are of similar length and parallel in construction, and that overall, the test question measures an important concept.

If the item editor spots any issue with the item, the editor should make the changes needed to improve the item. If flaws are found but corrections are not obvious, it may be necessary to discard the item, or at least set it aside to see if other viable items are on hand that make further work on the troubled item unnecessary.

If the item developers were able to submit their work electronically, it will be much easier and faster for the editors to use this electronic version to edit the item. If the items were submitted on paper, it will be necessary at this stage to prepare an electronic version of the assessment exercises. Clerical staff may do this work after the editors have marked up the item during the editorial process.

Ideally, before the editing process is completed, the artwork and graphics as well as permissions will be obtained. Adding the artwork to the item will allow the editors to layout the item in a suitable manner, assuring that the artwork fits the item in the space allocated, and if not, the item is adjusted so that it does. Knowing the status

of the permissions will assure that items that require such permissions going forward will not need to be changed substantially or dropped in the future.

16. Conduct bias/sensitivity and content area committee reviews – Once the assessment items have been edited and are completed electronically, they need to be reviewed for content accuracy and for bias and any material that might be sensitive. Typically, two different committees are used – one for content and another for bias and sensitivity.

The content advisory committee reviews the assessment items for content accuracy and completeness. The intent of this committee is to assure that each assessment item is of the highest quality, and that the overall set of assessment items measure the content standards effectively. The committee, typically comprised of 5 to 10 content experts in the technical skills area being assessed, will be drawn from K-12 educators and higher education content specialists. This review will take about two days for a set of 300 items.

The bias and sensitivity review will typically take less time. This group focuses on whether any of the material in the assessment might be found to be offensive by any subgroup of students taking it, as well as whether any of the items might appear to disadvantage any group of students. These reviewers, usually around five in number, need to be individuals who are sensitive to the types of materials student subgroups might find offensive or that might disadvantage any subgroup of students. This committee may be able to review 300 items in a day.

The reviews held at this point are without data, since the assessment items have not been pilot tested. However, these reviews are important to try to assure that any flawed assessment items (that are biased or that may cause negative student reactions) are not exposed to students even in pilot testing.

17. Recruit schools to participate in the pilot testing – As the materials for pilot testing are being prepared, the assessment developer should recruit schools to participate in the pilot testing. Typically, each participating classroom will give one pilot test form to all students, and the pilot test will be packaged so that the piloting will take only one class period. For pilot testing, where the purpose is to see if the item seems to work, the number of students who should participate can be fairly small – 100 to 200 students per item. Because a larger field test will occur later, the pilot testing does not need to involve large groups of students nor do they need to be randomly selected. For 100 students, this means recruiting four classrooms, ideally from four different districts (assuming that 25 students per classroom will actually complete the assessment).

18. Prepare pilot test booklets – Once all of the reviews have taken place, it is time to clean up the items and prepare for pilot testing. This involves several steps. First, each of the test items needs to be cleaned up – making the changes suggested during editing, content reviews, and bias/sensitivity reviews. The production staff should make the changes to the electronic version of the items. If the items have been entered into an item bank, the changes can be readily be made there. If an item bank is available, but the items have not been entered into it, this would be an ideal time to do so.

Second, once the items have been cleaned up, the items to be pilot tested need to be “pulled.” This involves reviewing the test blueprint for the content standards to

be measured in the final assessment instrument, and then over-selecting items for each content standard to assure that should any item not work, there is still sufficient coverage of each of the content standards after pilot testing. The number of extra items to be selected will vary, but it is generally safe to pilot one extra item if two or three will be used in the final instrument and two extra items if the final assessment will have five or six items. Of course, if there is not a large number of items to be pilot tested and there are available pilot test slots, trying out extra items will not hurt, since these can be retained until they are needed for future test forms.

Third, the selected items should be packaged for pilot testing. This generally involves placing the various items for any content standard together in the same test booklet. Trying all of the items for a content standard is preferable, since this will permit the assessment developer to select the best set of items for the operational assessment. Each pilot test booklet may contain the items for several content standards, depending on the testing time being used for each booklet. Public schools often prefer to have the pilot testing not take more than one class period, so the number of assessment items per booklet might be 30 or so multiple-choice items or a fewer number of these items if they are packaged together with one or two constructed-response items. Performance assessment items will need to be packaged separately, since one item will usually have to be administered to each student.

Fourth, the assessment administration materials will need to be developed. At a minimum, these consist of a school coordinator's manual, an assessment administration manual (for the teacher who administers the assessment), and answer document(s). The coordinator manual will provide directions to the school (or district) coordinator, who in turn will select classrooms for pilot testing, distribute and collect testing materials, and assure that the pilot testing occurs as scheduled. The assessment administration manual will provide directions to the teachers who will administer the tests. The answer document(s) will be used by students to record their answers – a scannable sheet for multiple-choice items and a booklet for them to record their written responses to constructed-response items.

Fifth, the necessary test booklets will need to be printed. In addition, the manuals and answer document(s) will also need to be printed. Extra copies of each of these should also be printed, so that an overage can be provided to each pilot school, and extra copies of the test booklets are available for review, too.

Sixth, package the testing materials in classroom sets. This involved packaging one assessment administration manual, 30-35 test booklets and 30-35 answer documents, and any other testing materials together in a single box. A school chosen for pilot testing may be selected for more than one classroom, but each classroom should be given a different form. Each classroom should be boxed separately to make it easy for the school coordinator to distribute the materials to each classroom (and to assure that the materials to pilot one form are not mixed up with those for another form). Each of these classroom boxes should be placed inside one or more larger boxes. In the first box, include one coordinator manual for the school. Mark the entire shipment "Box 1 of X", "Box 2 of X," and so forth.

Seventh, ship the pilot testing materials to the district or school (at the preference of the district or school coordinator). The materials should be shipped via a shipment procedure that requires a signature upon receipt so that errant materials

can be traced. This will reduce the probability of lost materials. The materials should be shipped to districts so as to arrive at least one week prior to the start of pilot testing.

19. Pilot test the items – The group of districts that have volunteered to administer the assessments should be given at least two weeks to administer the assessments to students. This should be sufficient time to administer an assessment comprised of multiple-choice items. If the assessment contains assessments that have to be individually-administered to students, more time may have to be allowed.

Educators who are administering the assessments should also be asked to complete an assessment administration observation sheet to share with developers any observations about the assessment items and assessment administration. This might include assessment items that students don't understand, ones that students complain about or mention, items that teachers feel were not fair to ask of students, and any other observations.

At the end of the allotted time for assessment administration, teachers should return the testing materials to the school coordinator, who should box up all of the testing materials and return them to the district coordinator or directly to the scoring contractor. Schools should have about a week after the testing period has concluded to return the materials for scoring.

20. Analyze the pilot test data – Once the assessment materials are returned for scoring, the multiple-choice items should be scanned, while the constructed-response and performance items should be separated for hand-scoring. The scoring contractor should be able to scan the different assessment forms' answer sheets rather quickly, producing a file of item responses.

During item development and item editing, scoring rubrics and scoring guides should have been developed. One purpose of the pilot testing is to not only try out the constructed response and performance items, but also to determine if the scoring rubrics are adequate as well. Hence, it is important if possible to collect student responses from as diverse a group as possible (given the limits of the small sample size). Then, the developer and the scoring staff should carefully apply the rubrics to the corresponding items to determine if the items worked (produced the full range of responses anticipated), whether the items need to be edited, whether the rubrics were adequate or whether the rubrics need to be modified or clarified.

Another aspect of applying the scoring rubric to each constructed response and performance item is to select exemplars for each score point in the rubrics. This will complete the scoring guide and assure that high quality samples of each scoring point will be available for use when the assessment becomes operational. Samples of student responses are needed for three purposes: 1) to provide examples of the responses to each score point for initial scorer training (calibration), 2) to verify the adequacy of scorer training before scorers begin scoring actual student responses, and 3) to use to periodically check to make sure that scorers don't "drift" during the scoring process by scoring pre-scored samples and verifying their agreement with these samples.

21. Conduct statistical analyses – The purpose of analyzing the pilot test data is to determine which test items worked and which did not. The goal of this phase is to

produce item analysis data to look at the performance of each item statistically. There are several analyses that the scoring contractor may carry out. These include:

- A. Traditional item analysis in which the performance of high- and low-scoring students on the entire test is displayed on each test item. This will help show whether high-performing students do well on each test item. If they don't, there may be a flaw in the item.
- B. IRT scaling of the items to determine the shape of the item characteristic curves of the set of items.
- C. Reliability analyses of each sub-test on the assessment, to determine the internal consistency of the set of items.
- D. DIF analyses to determine whether the items show differences in performance for important subgroups of test takers.
- E. Additional analyses needed to determine the viability of the items or to place them on the statistical scale in use for the assessment instrument.

Another, non-statistical analysis that should be carried out is to compile the responses of the teachers who administered the pilot tests so as to determine if there is any consensus on their comments about individual items or item types. Therefore, the comments should be compiled by test form and item.

- 22. Edit the items that survived the pilot test process – Once the statistical analyses are conducted, the developer and other project staff should review the statistical and non-statistical data to determine whether each test item is okay as is, whether it should be revised in any way, or whether the item appears so flawed that it should be dropped. Items that need to be revised should be edited and those that are fatally flawed should be deleted.
- 23. Conduct bias/sensitivity and content area committee reviews – Once the items have been fully edited, they should be reviewed for bias and sensitivity for the second and final time. Since DIF analyses were run by the scoring contractor, items that evidenced significant differences in sub-group performance should be flagged for review by the bias/sensitivity committee. In most cases, differential performance is not a sign of bias in the items but differences in opportunity to learn. Items that show the former should be revised or dropped, while those that show the latter should be retained and used to help assure future students are given greater opportunities to learn the important content assessed.

The content advisory committee should review the final versions of the items to assure that editing and other work on them did not materially alter their content accuracy nor their importance. At this point, usually only minor changes should be made to items since they have been reviewed before. However, the pilot test data may reveal flaws in the items that the content advisory committee may need to fix, so some editing of the items is to be expected.

- 24. Prepare the final version of the items – Once all of the reviews are completed, the assessment developer should make all final changes to the items in the item bank. Then, the final version of each assessment item should be prepared. Conduct final desktop publishing of the items so that at the conclusion of this phase, the full item set is prepared, ready to use them in future assessments.

### **Assembling the Final Instruments**

At this point, some assessment programs may use the assessment items as if they are final items ready to use in operational assessments. Ideally, however, the assessment items should first be formally field-tested, using statistically representative samples of students. This is essential in cases where new forms of the assessment are being developed, since it will be necessary to place the new items on the same existing statistical scale. An ideal way of doing this is to embed the new items in the operational forms of the existing assessment.

Embedded field-testing has the advantage of being able to administer the field test items in such a way that students don't know operational versus field test items, and therefore will put forth equal effort in responding to both types of questions. This will help assure the comparability of the results and assist in the equating of the new items with the existing ones.

Regardless of how the items are used, the following steps apply.

25. Select the items for the final assessment – The assessment items that will be field-tested need to be selected from the item bank of new items. The items should be selected to cover the range of content standards to be assessed, as well as the types of items specified for use in the assessment blueprint. If the items will be used to prepare a new form of an existing assessment, attention may need to be paid to matching the types of assessment items in the existing assessment, including the content standard, assessment type (Multiple-choice, constructed-response, etc.), difficulty level, types of response mode, and any other pertinent variables.
26. Package the items by test booklet – There are several ways in which the selected assessment items can be packaged. First, if the new items are to be field tested as a stand-alone test, they can be packaged either by content standard or by difficulty level (starting with easier items). If the stand-alone test is designed to match an existing measure, the assessment items should be ordered by standard to match the order of assessment items and standards of the existing assessment.

Second, the items to be field tested may be embedded into existing instruments so that a few of them are included in each form of the assessment. If this is the case, the number of items to be embedded in each field test form should be decided. If the operational tests are spiraled in the same classes, the number of items should be comparable across each form, so that each test form will take students about the same amount of time to complete. If constructed-response items are to be field tested as well in some forms while multiple-choice items are to be included in other forms, and all of these forms will be spiraled together and used in the same classrooms, the concern should be to equalize the testing time for each field test form, not the number of assessment items. One constructed-response items may equal 10 to 15 multiple-choice items in terms of testing time.

Typically, embedded field-test items are placed in several locations within an operational test, and the same positions are used for items to be field tested across multiple forms. This makes the preparation of the test forms easier, since for proofing purposes, the operational items are always in the same positions in the test forms, so only the items in the field test positions need to be carefully proofed.

Once the items are selected, the items should be assembled (packaged) to be able to provide to the test publishing staff an example of each test form with the items shown in the correct order.

27. Desktop publish the test booklets – Once the items to be field tested are selected and the manner in which they are to be assembled is determined, the publishing staff should desktop publish each of the test forms. This step may mean retyping the items, pulling the items from an item bank, or other means, but the goal is to end up with final test forms for testing.

Once the publishing staff has produced each test form, the forms should be carefully proofed. The items should be read by someone unfamiliar with the items checking for page layout (making sure that any passages or written material needed to answer any question are on the same page or adjacent page as the test items), there are no misspellings, grammatical errors, and that the sentence structure is straight-forward. Two or more individuals, also unfamiliar with the items, should take the test to determine if there is one correct answer to each test question. Their responses should be compared with each other as well as to the answer key to determine if there are any inconsistencies.

After corrections are made to the various proofed tests, they should be carefully examined to make sure that all of the changes noted are actually made, and that no new changes were inadvertently introduced. At this point, it is essential to have some mechanism for “versioning” these assessment forms to make sure that any changes that are needed are made to the most recent version of each of the assessment forms.

### **Administration of the Assessments**

28. Develop the materials needed to administer the assessments – There are several types of material that will need to be prepared for administering the assessments in schools. These include the creation of a coordinator manual (district and school), an assessment administration manual, one or more answer sheets or answer folders and other assessment materials such as teacher or school identification sheets, and so forth.

At a minimum, the assessment administration materials consist of a school and district coordinator’s manual, an assessment administration manual, and answer document(s). The coordinator manual will provide directions to the school (or district) coordinator, distribute and collect testing materials, and assure that the testing occurs as scheduled. The assessment administration manual will provide directions to the teachers who will administer the tests. The answer document(s) will be used by students to record their answers – a scannable sheet for multiple-choice items and a scannable or non-scannable booklet for them to record their written responses to constructed-response items.

29. Print the needed assessment materials – The necessary test booklets will need to be printed. In addition, the manuals and answer document(s) will also need to be printed. Extra copies of each of these should also be printed, so that an overage can be provided to each district and school, and extra copies of review can be available, too. At a minimum, an overage of 10% of the test booklets and answer documents should be printed.

30. Distribute the assessment materials to schools – The testing materials should be packaged in school sets. This involved packaging one coordinator manual, one or more assessment administration manuals, and a sufficient number of test booklets and answer documents, and any other testing materials together in one or more boxes per school. Each school should also receive an overage of test booklets and answer documents

A school participating in the assessment may be administering more than one assessment, but it is helpful to package each assessment form separately. Each school should be boxed separately to make it easy for the district coordinator to distribute the materials to each school. Each of these school boxes might be placed inside one or more larger boxes, or bundled together to ship on a pallet to the district. If more than one box is needed for any school, include one coordinator manual for the school in the first box and mark the entire shipment “Box 1 of X”, “Box 2 of X,” and so forth.

31. Ship the testing materials to the district coordinator or directly to each school (at the preference of the district or school coordinator). The address for the shipment should be ascertained in advance to assure that they are sent where it is most convenient for the district to distribute to schools. The materials should be shipped via a shipment procedure that requires a signature upon receipt so that errant materials can be traced. This will reduce the probability of lost materials. The materials should be shipped to districts so as to arrive at least one week prior to the start of pilot testing.

32. Monitor the assessment and respond to questions from the field – As local educators are administering the assessment to their students, the assessment development staff should monitor how the assessment is being implemented. This may include observation of the assessment administration in selected schools (especially valuable for atypical assessments such as individually-administered performance assessments), as well as responding to issues that arise during testing. It may be valuable as well to contact some of the assessment sites during the assessment administration period to see if they have faced any issues, and if so, how they have handled them. Being proactive about assessment administration may help to anticipate issues and seek assistance from local assessment administrators about how to best handle those issues.

33. Collect scorable and non-scorable assessment materials – At the conclusion of the assessment period, the scorable assessment materials should be collected. In addition, the non-scorable assessment materials may be collected or districts might be directed to either recycle or destroy these materials. Some of these non-scorable materials may be secure testing materials, so it is advisable to collect these materials to verify that none of them remain in the hands of local educators.

The scorable materials should be separated into those that will be scanned (such as scannable answer sheets) and those that will be hand-scored (including constructed response and performance assessments). This will facilitate the scoring of these materials.

### **Reporting of the Assessment Results**

34. Scan scorable answer documents – The scannable answer documents will be sent to a facility that can scan them and produce electronic files of student responses.

These responses will need to be combined with the hand-scored student responses to prepare a complete file for analysis purposes.

35. Hand-score student constructed responses and responses to performance assessments – Any open-ended response, whether a written response to a constructed response question or students' responses to performance assessments will need to be hand-scored by persons with sufficient background and training to score them in a reliable manner.

Hand scoring starts with a high quality scoring rubric and guide. The guide needs to contain one or more scoring rubrics, each of which describes a characteristic on which students' responses will be scored. The language of the rubric should be clear and concise, describing in measurable terms the different levels of the dimension(s) on which students will be scored. The scoring guide will add samples of student responses at each level of the rubric. These samples will be useful in training the scorers, as well as in verifying that their training has been sufficient for them to score student responses in a reliable manner. The samples of student work should be selected from the pilot test samples.

The training of scorers generally consists of a thorough review of the scoring rubric, point by point, and a review of one or more samples for each score point. Then, scorers should be asked to score one or more pre-scored samples to see if they have internalized the rubric. This continues until the scorers demonstrate a high enough level of reliability. At this point, each scorer is given another set of papers to score and if they score at least 90% of them accurately, they are then ready to begin scoring for real.

Once scoring does occur, it is important to monitor scorers so that they do not drift – from the original calibration nor from one another. One way to do this is to have some or all of the papers double-scored in a blind fashion (where the second scorer does not know what score the first scorer gave each paper). Then, someone needs to compare the two scores and ascertain whether the level of agreement (exact or adjacent) is sufficiently high. If not, the paper should be scored by a third scorer who is more highly trained (e.g., a scoring table leader). It is not necessary to double score all papers in order to calculate inter-rater reliability; a 20% or more sample is sufficient for this purpose.

To make sure that all scorers are not drifting away from the original standards, sample pre-scored papers (similar to those used in training) should be given to all scorers at set times throughout the scoring process. By doing this periodically, it will be possible to monitor the quality of the scoring for each scorer, and this will help to identify any scorers who need to be retrained or dismissed.

Upon completion of hand scoring, the students' responses should be combined with those from the scanning of their multiple-choice items to form a complete student file. This file will then be used to analyze students' performances and to prepare reports of assessment results.

36. Conduct statistical analyses of the assessments – Once all of the students' responses have been assembled in one electronic file, a variety of analyses can be performed on them and then various reports of results can be prepared. The first step is to assure that the assessments were technically sound. This means ascertaining that the assessment items performed as expected by running item

analyses. Then, various test analyses should be performed to assure that the items on each section of the test “hung together” sufficiently to be used to report results separately. Usually, this step is perfunctory, that is, the items and sub-parts of the assessment performed as expected (and as they have in past). The analyses will be useful in preparing a technical report on the assessment administration, which is the expected way to document the development and administration processes used and to demonstrate the soundness of the assessment as it is used.

37. Produce reports of results – Upon completion of the analyses of student performance on the assessment, the reports of student results should be prepared. This will include the student, parent, classroom, school, and district reports of results. The accuracy of these reports of results should have been verified by “rollin up” mock data from the student to the classroom to the school to the district level in a data verification process some time before the production of actual reports begins. This will assure that the reporting system is capturing and summarizing the data accurately and reporting of results can proceed without a hitch.
38. Produce a technical report on the assessment development and administration – A technical report that documents how the assessment was developed (the steps used to develop it and data collected during the pilot and field testing) and administered (including data from at least the first actual use of the assessment) should be produced. This document is essential in demonstrating the soundness of the assessment for its intended uses and should be written in language accessible to typical users of the assessment. Although it may contain material that is technical in nature, these data and procedures should be explained in non-technical language.

## Summary

This paper has described the process for the development of a technical skills assessment program for use statewide. This process includes the determination of which technical skill areas to include in the assessment program, the selection or creation of the technical skills for each selected area, the review of existing measures and/or test items to determine their suitability for use, and the creation of new measures where existing measures are either not available or are found to be inadequate.

There are a number of steps involved in this work, but carried out carefully, this work will result in a sound and viable assessment program that aides in the instruction of students and encourages their learning. While a program that covers a number of technical skills areas will require the investment of resources to complete, suggestions are given for how this work can be carried out, including strategies for having districts and/or states collaborate on developing the work so as to reduce effort and costs.

Given the number of areas where assessments are needed, not all areas will be available at the same time. However, with diligence (and some sharing among collaborating agencies), it should be possible to fill out the assessment design in two to three years, and therefore contribute to the improvement of career technical education of students in the state.

## ATTACHMENT A

### Michigan Technical Skill Assessment Review Elements

In order to select technical skill assessments for adoption by the State of Michigan, the Office of Career and Technical Education will review potential assessments that are under consideration for adoption according to the following elements and criteria.

#### A. General Assessment Information

1. Assessment Name
2. DOT
3. O\*NET
4. Career Cluster Designation
5. CIP Code
6. Assessment Publisher/Owner
7. Publication Date – Date the Assessment was published? How often is the assessment is revised, updated and re-aligned with standards?
8. Type of Assessment – Criterion-referenced or norm-referenced. If the latter, when was the assessment most recently normed?
9. Nature of the assessment – paper and pencil; online; both?
10. Assessment Cost – What are the costs associated with the assessment? If the assessment was used statewide, is a lower price available?
11. Administration Time – What is the total testing time for students? Is the test timed or untimed?
12. Recommended Grades/Ages – For which grades or ages is the assessment designed and/or recommended for use? Do you consider the assessment to be appropriate for evaluating the technical skills of secondary students?
13. Available Information – What information is available for the assessment – administration manual/directions; technical report/information; other?

#### B. Logistics

1. Equivalent Forms – Please describe the number of equivalent alternate forms available,
  - A. How many forms of the assessment are currently available?
  - B. How often are new alternate forms developed?
  - C. Has the equivalence of the alternative forms been established (see B1D above)
  - D. Has equivalence at the subscale level been established?
2. Security of the Assessment – Is this assessment secure? If so, please describe the measures used to maintain the security of the test forms and items.

3. Assessment Administration – Please describe the administration schedule and processes for the assessment.
  - A. Can the assessment be administered according to district or state timelines?
  - B. If not, when can the assessment be administered?
  - C. Please describe the processes, rules and procedures for administration. Address how easy or difficult it will be for school districts to administer the assessment.
4. Assessment Accommodations – What accommodations or modifications in the assessments or assessment process have been provided for students with disabilities or English language learners? How do these affect the reports of results for these students?

**C. Technical Information**

3. Reliability – Please describe how the reliability of the assessment has been established. These include any of the following:
  - F. Stability (Test-Retest)
  - G. Internal Consistency (KR-20 or Cronbach Alpha)
  - H. Split-Half
  - I. Alternate Form
  - J. Inter-Rater Reliability (for constructed-response or performance items requiring handscoring)

What analyses have been performed and what data are available to demonstrate the reliability of the assessment. Is the data available suitable for the assessment?

4. Validity – Please describe how the validity of the assessment has been established for each type of inference for which the assessment is recommended.
  - A. Content Validity – An essential element in the selection of appropriate instruments to assess Michigan technical skill standards is the alignment of the assessments and the Michigan technical skill standards. Alignment will be judged four ways:
    5. How many of the Michigan technical skill standards are measured by the assessment,
    6. How many assessment items measure one or more Michigan technical skill standards.
    7. Does the assessment emphasize the most important skills in the Michigan technical skill standards?
    8. Does the assessment assess students at a level appropriate for high school?

What analyses have been carried out to demonstrate alignment to the standards? Please provide information on any alignment study that has been carried out for this assessment to the Michigan technical skill standards. Please describe the results of evaluation of the content by business, industry and postsecondary institutions.

- B. Predictive Validity – Please describe the results of analyses of the predictive validity of the assessment. Is there information to show that successful performance on the assessment predicts success on the job? Do students who do better on the assessment do well in jobs or in postsecondary education in the same field?

Please describe the results of analyses examining the extent to which the assessment accurately discriminates between students with greater mastery compared to students with lower mastery (discrimination index)

- C. Concurrent Validity – Does performance on this assessment correlate with success on other related/comparable assessments?
  - D. Construct Validity – Is the assessment designed to measure a more theoretical construct, is there evidence to support the proposed interpretation of the results?
  - E. Sub-Scale Validity – If applicable, please describe the results of analyses of the validity of any subscales.
3. Bias and Sensitivity Reviews – Please describe the results of bias and sensitivity reviews conducted on the items.
- D. Reports of Assessment Results** – What types of assessment results does the assessment produce?
- 1. Availability of Student Results – Please describe how individual student assessment results may be obtained by school districts and/or the State of Michigan.
    - A. Please describe how the assessment results are provided to students and parents in an understandable manner.
    - B. What interpretive materials are available to help explain the results?
  - 2. Availability of Group Data – By what date(s) will the results be available to school districts and the state.
  - 3. Clarity of Scales and Norms
    - A. Are the scales used for reporting clear and carefully described?
    - B. Is the population to which the norms, if any, apply clearly defined and described?
  - 4. Utility of Assessment Results – Please describe how the results of the assessment will be useful:
    - A. To students.
    - B. For program improvement.
    - C. Are utilized by business and industry for hiring, promoting or evaluating employees.
  - 5. Value of Assessment Results – Address the value of the assessment results to business and industry.
  - 6. Postsecondary Institutional Use of Results – Do any postsecondary institutions accept a passing score on the assessment for credit in their institution?
- E. Other Statistical Information**
- 1. Please provide a copy of the inter-item correlation matrix showing all items, including those within subscales, and the item-total correlations for each item within the assessment.
  - 2. Please provide any additional test statistics that may help in evaluation of this assessment.
- F. Summary**

**ATTACHMENT B**

**Assessment:** \_\_\_\_\_

**Michigan Technical Skill Assessment Review Form**

**A. General Assessment Information**

14. Assessment Name –
15. DOT –
16. O\*NET –
17. Career Cluster Designation –
18. CIP –
19. Assessment Publisher/Owner –
20. Publication Date –
21. Type of Assessment –
22. Nature of the assessment –
23. Assessment Cost –
24. Administration Time –
25. Recommended Grades/Ages –
26. Available Information Reviewed –

**B. Logistics**

1. Equivalent Forms –
2. Security of the Assessment –
3. Assessment Administration –

**C. Technical Information**

5. Reliability
  - K. Stability (Test-Retest) –
  - L. Internal Consistency (KR-20 or Cronbach Alpha) –
  - M. Split-Half –
  - N. Alternate Form –
  - O. Inter-Rater Reliability –
6. Validity
  - A. Content Validity –

- B. Predictive Validity –
- C. Concurrent Validity –
- D. Construct Validity –
- E. Sub-Scale Validity –
- 3. Bias and Sensitivity Reviews –
- D. Reports of Assessment Results –**
  - 1. Availability of Student Results –
  - 2. Availability of Group Data –
  - 3. Clarity of Scales and Norms –
  - 4. Utility of Assessment Results –
  - 5. Value of Assessment Results –
- E. Other Statistical Information**
  - 3. Inter-item correlation matrix –
  - 4. Any additional test statistics –
- F. Summary**

**ATTACHMENT C**  
Technical Skill Assessment Content Review  
2008-09  
Instructions

Thank you for agreeing to serve as an expert content reviewer for this technical skill assessment. The purpose of this review is to obtain input from secondary and postsecondary instructors regarding the appropriateness of the assessment instrument and items in terms of:

- Alignment to the state secondary technical skill standards
- Ease of use
- Appropriateness of the assessment for use with Michigan secondary CTE students

As a content expert, you will be provided with a copy of the assessment (either on paper or online), a copy of the assessment instructions or manual(s), where such materials exist, and a copy of the technical skill standards that the assessment instrument is intended to measure. You will have an opportunity to take the assessment yourself and receive a score report. You will be asked to review and critique the assessment utilizing the review form provided by the Office of Career and Technical Education (OCTE). Any additional comments or insights regarding the assessment are welcome. Your content expertise will be given considerable weight in determining whether an assessment is adopted, rejected or adopted in a modified form as the state technical skill assessment. Therefore, please review the assessment and any instructions carefully and provide considered and detailed feedback.

It is important that you maintain the integrity of the assessment process by maintaining the security of the assessment. Do not discuss with, or disclose to anyone outside of the reviewer meeting with OCTE, any aspect of the testing materials or questions you will review as part of this process. In order to keep the assessments secure, access to the test items is highly restricted and there are a limited number of content reviewers. Do not retain notes, photocopies or any other records of any of the materials you will review. All materials must be returned to designated OCTE staff members.

1. Please complete and fax the enclosed/attached confidentiality & security agreement to Jill Kroll at 517-373-8776 before accessing or viewing the assessment.
2. Review the technical skill standards for this program (provided by OCTE).
3. Read all instructions and directions provided with the assessment before beginning the test.
4. Read through the assessment reviewer questionnaire to familiarize yourself with the questions to be answered.
5. If you will be accessing the assessment online, prepare your review materials including a pen or pencil, note paper, the reviewer questionnaire, and instructions prior to accessing the assessment online. Find out whether you must complete the assessment in one sitting or if you can save your assessment and return at a later time. Leave plenty of time to carry out your review.
6. Open your assessment booklet or follow the instructions access the assessment online. Take the assessment and review each item carefully. Determine whether the item aligns to one or more of the program technical skill standards. Complete the questionnaire with comments about the appropriateness of the assessment in terms of alignment to the program standards, appropriate to secondary CTE student completers, ease of use, other comments.
7. Meet with OCTE and other reviewers (in person or by phone) and review questionnaires and reviewer impressions.

If you have questions regarding the content review, please contact:

Jill Kroll, Ph.D.  
Education Research Consultant  
Office of Career and Technical Education  
Michigan Department of Education  
PO Box 30712, Lansing, MI 48909  
[KrollJ1@Michigan.gov](mailto:KrollJ1@Michigan.gov)  
517-241-4354

# **CONFIDENTIALITY & SECURITY AGREEMENT**

## **Michigan Department of Education**

### **Office of Career and Technical Education**

To maintain security of the assessments being considered for the state technical skill assessment, the Office of Career and Technical Education (OCTE) requires groups or individuals who wish to review either current tests, or items that may be used on future test forms, to sign this Confidentiality & Security Agreement before they review test items.

You are personally responsible for maintaining strict confidentiality of any information related to the test materials or questions you will review here.

Please follow these security standards:

1. Do not discuss with, or disclose to anyone outside of the reviewer meeting with OCTE, any aspect of the testing materials or questions you will review here.
2. Do not retain notes, photocopies or any other records of any of the materials you will review during this meeting. All such materials must be returned to designated OCTE staff members.

The Michigan Department of Education and OCTE appreciate your cooperation in this important activity. Please review and sign this form.

---

I have read and understand these confidentiality and security standards and agree to abide by them. I acknowledge and agree that all tests and related materials developed or adopted by the Michigan Department of Education and OCTE are highly confidential and their contents are not to be divulged to anyone. I further understand violations of this Confidentiality and Security Agreement may lead to disciplinary or legal action by the Michigan Department of Education.

Name (Please Print): \_\_\_\_\_ Date: \_\_\_\_\_

Signature: \_\_\_\_\_

CONTENT REVIEW INSTRUMENT

Assessment: \_\_\_\_\_

CIP Code \_\_\_\_\_

Yes No 1. Does each item measure one of the secondary CTE technical skill standards listed on the Technical Skill Standards document for this CIP Code? List any assessment items that do not appear to measure one of the standards listed for this CIP Code.

_____	_____
_____	_____
_____	_____
_____	_____

Yes No 2. Are all of the Michigan secondary CTE technical skill standards for this program assessed by at least one assessment item? List any standards NOT assessed by at least one item on the assessment.

_____	_____
_____	_____
_____	_____
_____	_____

Yes No 3. Does the assessment place its emphasis where it should be for measuring the technical skill attainment of Michigan secondary CTE students who have nearly completed their program? If no, please explain any problems you see with the emphasis of the assessment:

_____
_____

4. On a scale of 1 to 5, with 5 being highest and 1 being lowest, to what extent would you say that the content categories on the assessment are the same as the content categories in the standards?

Circle one: (lowest) 1      2      3      4      5 (highest)

Comments: \_\_\_\_\_

Yes No 5. Based on your review of the assessment items, would you say that what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards? If no, please explain any problems you see with the assessment:

---

---

Yes No 6. Based on your review of the assessment items, do you think that secondary students who are well-prepared for postsecondary education or employment in this program area will do well on this assessment? Circle one: Yes No If no, please explain any problems you see with the assessment:

---

---

Yes No 7. Were all of the test items clear, well-worded and understandable? Please list any items that were problematic and describe the problem:

---

---

Yes No 8. Were any of the test items ambiguous or confusing? Please list any items that were problematic and describe the problem:

---

---

Yes No 9. Do you foresee any practical or logistical problems with administering this assessment to secondary Career and Technical Education students? Please describe any problems you foresee and recommend solutions, if any:

---

---

Yes No 10. Do you think this assessment should be adopted as the state assessment of technical skills for this program area? Please explain why or why not:

---

---

11. Please provide additional comments or observations regarding this assessment:

---

---

---

---

THANK YOU

Please refer to the document entitled “Michigan Secondary Career and Technical Education Technical Skill Standards for CIP Code” listed above for the review of this assessment.