Implementing a District-wide Standards-Referenced Assessment System

A Report from the Assessment Literacy Task Group
At the Request of the Iowa Department of Education, September 3, 1998

Table of Contents

Iowa Assessment Literacy Task Group

Dennis Brown
Iowa Department of Education

Paul Cahill
Iowa Department of Education

Nina Carran
Iowa Department of Education

Thomas E. Deeter, Ph.D.
Des Moines Public Schools

Bill Dutton, Ph.D.
Grant Wood AEA

Gail Galbraith
Arrowhead AEA

Kathy Hinders, Ph.D.
Iowa Department of Education

Connor Hood
Arrowhead AEA

Ann Johnson, Ph. D.
Ankeny Community School District

Joanmarie McGuire, Ph.D.
Iowa Department of Education

Laurie Phelan
Iowa Department of Education

Gary D. Phye, Ph.D.
Iowa State University

W. David Tilly, Ph.D.
Iowa Department of Education

Jon S. Twing, Ph.D.
National Computer Systems

Mike Vanderwood, Ph. D.
University of Wisconsin-Milwaukee

A Letter to Iowa Educators from the Assessment Literacy Task Group

The business of education is to provide each child opportunities to develop the capacity to leave this world a better place than when he or she entered it. A fundamental premise of educational systems is to improve the teaching-for-learning process for the benefit of the students who attend our schools.

The Iowa Department of Education continues to champion the philosophy that the best decisions are those that are made closest to the students, i.e., at the local (Local Education Agency) level. Through negotiations with federal officials, Iowa Department of Education staff has been able to protect the local freedoms that continue to benefit students.

Iowa's school improvement cycle has evolved into a systemic and systematic process that continues to focus on locally developed standards, benchmarks and accountability processes. One goal is to avoid the complacency that can accompany the long-standing tradition of success that Iowa has enjoyed relative to other states in the nation. In addition, the integration of federal guidelines surrounding the Individuals with Disabilities Education Act, the Title I reauthorization and school-to-work initiatives (among others) has not been a small task.

Local autonomy is accompanied by the responsibility to clearly articulate the learning expectations that we have for our children, to establish the processes to help them achieve those expectations, and to implement sound assessment and evaluation systems that will let us know if we are achieving our goals. Iowa educators have an opportunity to help create an educational system that will allow local autonomy to be maintained, and to serve as a model for other states dealing with similar issues. Failure to act may result in losing the privileges associated with local control.

The Iowa Department of Education continues to bring together representatives from Local Education Agencies, Area Education Agencies, post-secondary education institutions, educational associations, private corporations, as well as parents and other members of the public, the combined efforts of which will build the capacity of educators to reach every student.

Education in Iowa has reached a defining moment. This is likely the most challenging process in which Iowa educators will engage for a long time. It is so formidable that no one can do it alone. We must capitalize on what each of us can bring to the table to contribute to shaping Iowa education for the future. To do it right, it will take your knowledge, your experience, and your willingness to make tough decisions, keeping the focus on the welfare of the children. It will take your heart and your passion. It will require the courage of your conviction. It is time to step forward. Are you up to it?

I. Purpose

The purpose of this document is to provide information, outline the necessary steps and present supporting references so that a local school district can successfully implement a district-wide standards-referenced assessment system (DSRAS.) Such a system is different from a district-wide testing program, program evaluation or other monitoring program in several regards.

- First, the DSRAS is all encompassing: it will inform instruction, which will lead to enhanced teaching and improved student learning. It includes aspects of current testing, current evaluation, and current instructional practices.
- Second, it is appropriate for monitoring school and individual student improvement over time through the evaluation of multiple facets of student learning.
- Third, it is used to fulfill new state and federal guidelines regarding Title I.
- Fourth, it provides for a "state of the art" evaluation system, applicable to all students such that instruction and evaluation are linked.

The goal of such an assessment system is improved learning through informed instruction. Be assured, this is a very challenging goal, and one that will require the commitment and dedication of all those involved: state agency personnel, local administrators, local teachers, and students. With the help of this document and the additional resources identified within, this task should be achievable.

This document is not the final word regarding what represents a "good" or a "bad" assessment program. This document is neither all-inclusive nor exhaustive. The committee responsible for this document sees it evolving, as examples of district-wide assessment systems become available.

A. Best Practices and Procedural Guidelines

This document is not intended to prescribe how a district should assess students, monitor progress over time, or implement a local assessment system. It is intended, rather, to provide a description of "best practice," providing information regarding what to consider in designing and implementing a district-wide standards-referenced assessment system. For example, when a teacher constructs a test to provide feedback regarding a recent instructional topic, he or she is typically not concerned with the psychometric principle of

reliability. Clearly, the teacher desires the best and most accurate assessment possible, but typically does not conduct research on the technical properties of the test before using it in the classroom. However, a "state-of-the-art" assessment used at the district level will require the collection of evidence to defend the implemented assessment for the types of judgments resulting from the scores. Educators are likely to find themselves in need of additional information, support and guidance regarding such "technical" aspects of an assessment system. Hopefully, by presenting them what is considered best practice in this regard, educators will be better equipped to collect such information as reliability evidence that is required of an assessment system.

This document is not a "cookbook" for determining what "ingredients" should be added to the assessment system at what time and in what quantity. Users of this document will have to carefully evaluate the procedures outlined in light of their own annual goals, local content standards and performance standards. In this regard, this document can be seen as providing some procedural guidelines with topics to consider when implementing a standards-referenced assessment system, and not a list of things "to do." For example, if your annual improvement goal requires students to be able to generate text at some level of proficiency, then your locally constructed assessment will probably require a writing sample. Such a sample will undoubtedly be scored by having judges read student responses. If so, you will need to collect evidence of scoring consistency (inter-rater reliability or scorer reliability) in addition to demonstrating that the writing outcome is meaningful, useful or otherwise a worthwhile endeavor. This document will not tell you how to do this, but it will tell you the steps taken to do it in ideal, best practice or state-of-the-art systems. You may then choose to follow these steps explicitly, adapt them to your local circumstance or modify them as you see fit to meet your needs. Again, it is not so much a matter of how you collect such evidence of scorer reliability as it is to realize the importance of such information and that documenting and/or demonstrating such information might distinguish an exemplary system from others.

B. Local Control, the Iowa Model and Legislation

Nationally, there is evidence of a trend toward increased accountability as evidenced by the Improving America's Schools Act of 1994. The 1994 law linked Title I accountability requirements with state reform efforts. The 1994 law dictates a "standards

referenced" model of assessment and requires states to define "adequate yearly progress" in at least the subject areas of reading and mathematics for local school districts.

Local Control

However, Iowa believes that a great deal of accountability should rest at the local level rather than at the state level. Iowa schools have a long record of local control and a strong sense of community ownership. Communities in Iowa have historically set high expectations for their students. Iowa students have consistently scored well on numerous national tests, including the ACT and SAT. The State of Iowa's role is to provide support for local school districts and accredited nonpublic schools.

The Iowa Model

Iowa's strong heritage of local control and existing state legislation allowed the Iowa Department of Education to reach an agreement called, the Iowa Model, with the federal government that meets the intent of the requirements inherent in the Improving America's Schools Act of 1994.

A thorough explanation of the Iowa Model and the requirements of the model were provided to schools in a letter dated April 16, 1998, from Judy Jeffery of the Division of Early Childhood, Elementary and Secondary Education. The expectations of the schools can be summarized as the following:

- By September 15, 1998, each accredited nonpublic school or school district must report for reading and mathematics their content standards, achievement data for all students, subgroup achievement data for race and gender, and their annual improvement goals;
- The achievement data must be reported for at least three grade levels (3-5, 6-9, 10-12.) Recent passage of House File 2272 requires reporting data for grades 4, 8 and 11 in mathematics and reading, as well as grades 4 and 8 for science;
- The achievement data must also be reported for at least three desired levels of student performance;
- The achievement data must be reported for at least reading and mathematics, and now science;
- Data should be collected for both long and short-term goals and from various (multiple) assessments.

The expectations from the Iowa Model for the Iowa Department of Education can be summarized as the following:

- Establishment of criteria for determining the clarity, rigor and quality of the content standards;
- Establishment of criteria for determining the clarity, rigor and quality of the performance standards;
- Review of student achievement data in mathematics and reading (and science per Senate File 2272) submitted by school districts regarding the district's expected growth for children;
- Working collaboratively with and providing technical assistance to local schools in need of improvement to complete a self-study.

In summary, The Iowa Model provides for local control regarding the definition, implementation and execution of local content, performance standards, and assessment measures while maintaining qualifications required for federal funding, particularly with regards to Title I. The Iowa Model does this without requiring the implementation of a state mandated testing program, as other states have been required to do.

State Legislation

In 1988, the Iowa legislature passed Iowa Code sections 280.12 and 280.18, which requires all local school districts and approved nonpublic schools to assess local needs and establish local student achievement goals with evaluation of progress, which is to be reported to their public and the state. This legislation provided one base to meet Title I reporting requirements.

In 1998, the Iowa legislature passed House File 2272 requiring the State Board of Education to develop and adopt rules, by July 1, 1999, incorporating accountability for student achievement into the standards and accreditation process described in section 256.11. The rules provide for the following:

a. Requirements that all school districts and accredited nonpublic schools develop, implement, and file with the department a comprehensive school improvement plan that includes, but is not limited to, demonstrated school, parental, and community involvement in assessing educational needs, establishing local education standards and student achievement levels, and, as applicable, the consolidation of federal and state planning, goal-setting, and reporting requirements.

b. A set of core academic indicators in mathematics and reading in grades four, eight, and eleven, a set of core academic indicators in science in grades eight and eleven, and another set of core indicators that includes, but is not limited to, graduate rate, post-secondary education, and successful employment in Iowa. Annually, the department shall report state data for each indicator in the condition of education report.

c. A requirement that all school districts and accredited nonpublic schools annually report to the department and the local community the district-wide progress made in attaining student achievement goals on the academic and other core indicators and the district-wide progress made in attaining locally established student learning goals. The school districts and accredited nonpublic schools shall demonstrate the use of multiple assessment measures in determining student achievement levels. The school districts and accredited nonpublic schools may report on other locally determined factors influencing student achievement. The school districts and accredited nonpublic schools shall also report to the local community their results by individual attendance center.

II. Background

Before more details are provided regarding the implementation of a standards-referenced assessment system, a general background of the movement toward standards-referenced assessments is needed. As alluded to earlier, the national trend toward standards-referenced accountability models is quite apparent. In fact, Iowa is the only state left without such mandated assessments.

The following sections of this chapter help to clarify how the requirements of Title I, the Iowa Code and The Iowa Model fit together with a locally controlled and implemented standards-referenced assessment program. Included in these sections is information regarding the role the Iowa Tests of Basic Skills and the Iowa Tests of Educational Development could play (in addition to other multiple-measures) in partially fulfilling the requirements of the legislation.

A. Title I Regulations

The reauthorization of Title I (PL 103-382) defines "adequate yearly progress" (AYP) as:

> "The measure set by each state to assess performance of Title I schools and districts. The definition of adequate yearly progress will vary from state to state, and is expected to result in continuous and substantial yearly improvement of each school and local district sufficient to achieve the goal of all children served under Title I, particularly economically disadvantaged and limited-English-proficient children meeting the state's proficient and advanced levels of performance."

The demonstration of adequate yearly progress should be the goal of all education, but is no small undertaking and one which is fraught with potential pitfalls. These problems are not new. Educators under Chapter 1 legislation struggled with the need to define and track individual student growth from year to year. Most Chapter 1 coordinators at the local school level relied on differences in normal curve equivalents (NCEs) as supplied by test publishers in the form of pretest/posttest comparisons. This required the tracking of students from year to year (regardless of student mobility), using the same assessment instrument, and was independent of student population and school background characteristics or changes in those characteristics across the years. In addition, the psychometric properties of such difference scores were questionable. In fact, sufficient evidence of reliable score interpretations using differences was hard to come by due to

the fact that the error component of a difference score is as large or larger than the error components of either part (Allen & Yen, 1979.) Finally, a large amount of supporting logistics were required in compiling, tracking and reporting the progress of students in Chapter 1. Taking all of this in total, it is easy to see why measuring "adequate yearly progress" has been a difficult task.

Title I of the Improving America's Schools Act of 1994 (IASA) was intended to provide the flexibility states have desired. State departments of education are no longer required to follow a prescribed procedure, but rather, are required to establish their own procedures such that *all students* can reach challenging academic standards. As Roeber (1996, pg. 2) points out, states are given "considerable flexibility" in determining "what the standards will look like, how they are developed, and how they are implemented." Under the guidelines of Title I, this means that the same rigorous standards used to monitor all students will be used for those receiving services under Title I. Hence, the states will only have to deal with "a single system of standards and assessments" (Winter, 1996), answering one of the ongoing criticisms of the previous legislation. Also, the new Title I provisions dictate that challenging content and performance standards be established, that a system of monitoring the progress of students in attaining these standards be developed, and that students make adequate progress (yearly) toward those standards. The Title I provisions do not describe how the states should go about these three tasks. In fact, discussions regarding the possible designs of such systems generate more questions than answers. These are some of the questions you might ask, though some have already been answered via the reporting requirements placed in legislation (see the section on the Iowa Model in the previous chapter):

- What information should be included in measuring adequate yearly progress (Winter, 1996)?
- Are three levels of performance standards sufficient (Carlson, 1996)?
- What role will nonacademic variables play (Winter, 1996)?
- How will adequate yearly progress be reported?
- How can adequate yearly progress data be used to improve instruction?

- How much progress is adequate and within what time frame (Carlson, 1996)?
- What will be done with grade levels not assessed (Winter, 1996)?
- Are cross-sectional or longitudinal data collection designs required (Carlson, 1996)?
- Should adequate yearly progress be based on all students or only those receiving Title I services (Winter, 1996)?
- Is a composite index to be developed to monitor adequate yearly progress? If so, how will the weighting of the various pieces be made in the combination (Carlson, 1996)?
- Will a sampling plan be used or will census testing be required (Winter, 1996)?
- How will school size, mobility and other "uncontrolled" factors be used in measuring adequate yearly progress?
- How will multiple measures be combined to determine adequate yearly progress (Winter, 1996)?
- How will fair disaggregations be made (Winter, 1996)?

As has been discussed, while the State of Iowa does not mandate how school districts should measure adequate yearly progress, it is still the responsibility of the local school districts to put into place such a system.

B. Iowa Code 280.12 and 280.18

The basis of most of the reporting requirements of the Iowa Department of Education to date has been established by Iowa Administrative Code 280.12 and 280.18. This code establishes the responsibility of both public and accredited nonpublic schools to: perform needs assessments, develop long-range goals and plans, establish short-term and intermediate goals, evaluate progress in meeting the goals, report progress, and communicate the plan and progress back to the state. The text of the code is provided as the following (this information was taken from the Iowa Department of Education CD-ROM: Standards Development for School Improvement in Iowa, 1997):

280.12 Goals and plans—evaluation—advisory committee.

1. The board of directors of each public school district and the authorities in charge of each nonpublic school shall:

a. Determine major educational needs and rank them in priority order.

b. Develop long-range goals and plans to meet the needs.

c. Establish and implement short-range and intermediate-range plans to meet the goals and to attain the desired levels of pupil performance.

d. Evaluate progress toward meeting the goals and maintain a record of progress under the plan that includes reports of pupil performance and results of school improvement projects.

e. Report progress made under the plan at least annually to the advisory committee appointed under subsection 2, the community and the department of education. Make other reports of progress, as the director of the department of education requires.

2. In meeting the requirements of subsection 1, a board of directors or the authorities in charge of a nonpublic school shall appoint an advisory committee to make recommendations to the board or authorities. The advisory committee shall consist of members representing students, parents, teachers, administrators, and representatives from the community.

[C75, 77, 79, 81, § 280.12]
85 Acts, ch. 212, §8
1997 Iowa Code

280.18 Student achievement goals.

The board of directors of each school district shall adopt goals to improve student achievement and performance. Student achievement and performance can be measured by measuring the improvement of students' skills in reading, writing, speaking, listening, mathematics, reasoning, studying, and technological literacy. In order to achieve the goal of improving student achievement and performance on a statewide basis, the board of directors of each school district shall adopt goals that will improve student achievement at each grade level in the skills listed in this section and other skills deemed important by the board. Not later than July 1, 1989, the board of each district shall transmit to the department of education its plans for achieving the goals it has adopted and the periodic assessment that will be used to determine whether its goals have been achieved. The committee appointed by the board under section 280.12 shall advise the board concerning the development of goals, the assessment process to be used, and the measurements to be used. The periodic assessment used by a school district to determine whether its student achievement goals have been met shall use various measures for determination, of which standardized tests may be one. The board shall ensure that the achievement of goals for a grade level has been assessed at least once during every four-year period. The board shall file assessment reports with the

department of education and shall make copies of these reports available to the residents of the school district.

87 Acts, ch 224, §56

C. House File 2272

Legislation reflecting The Iowa Model can be found in House File 2272 (as outlined previously and reproduced here) and works with the requirements of the Iowa Code previously outlined. The text of this bill, as provided from the legislative internet web site: (http://www2.legis.state.ia.us/GA/77GA/Legislation/HF/02200/hf02272/980330.html):

Section 1. Section 256.7, code 1997, is amended by adding the following new subsection:

NEW SUBSECTION. 21. Develop and adopt rules by July 1, 1999, incorporating accountability for student achievement into the standards and accreditation process described in section 256.11. The rules shall provide for all of the following:

a. Requirements that all school districts and accredited nonpublic schools develop, implement, and file with the department a comprehensive school improvement plan that includes, but is not limited to, demonstrated school, parental, and community involvement in assessing educational needs, establishing local education standards and student achievement levels, and, as applicable, the consolidation of federal and state planning, goal-setting, and reporting requirements.

b. A set of core academic indicators in mathematics and reading in grades four, eight, and eleven, a set of core academic indicators in science in grades eight and eleven, and another set of core indicators that includes, but is not limited to, graduate rate, post-secondary education, and successful employment in Iowa. Annually, the department shall report state data for each indicator in the condition of education report.

c. A requirement that all school districts and accredited nonpublic schools annually report to the department and the local community the district-wide progress made in attaining student achievement goals on the academic and other core indicators and the district-wide progress made in attaining locally established student learning goals. The school districts and accredited nonpublic schools shall demonstrate the use of multiple assessment measures in determining student achievement levels. The school districts and accredited nonpublic schools may report on other locally determined factors influencing student achievement. The school districts and accredited nonpublic schools shall also report to the local community their results by individual attendance center.

D. The Iowa Testing Programs Connection

One-advantage students attending school here in Iowa have is a state of the art standardized achievement-testing program offered by the Iowa Testing Programs at the University of Iowa. The fact that so many schools in Iowa use either the Iowa Tests of Basic Skills (ITBS), the Iowa Tests of Educational Development (ITED) or both, was also in no small part the reason for the successful agreement with the U.S. Department of Education referenced here as The Iowa Model. Though not required, it is anticipated that many school districts will choose to continue to use the ITBS and/or ITED as part of their assessment system. In fact, the Iowa Testing Programs have already provided both an Achievement Levels Report and Interpretive Supplement for the Achievement Levels Report (Iowa Testing Programs, 1997-1998 Revision.) This document outlines what schools must do if they choose to use ITBS and/or ITED and the Achievement Levels Report to help meet their requirements regarding fulfilling the previously mentioned legislation. In addition, the authors of this document provide guidelines regarding the ways in which the report can best be used for monitoring the achievement of grade groups and the reporting of achievement results of school buildings (Iowa Testing Programs, 1997-1998 Revision.)

E. Setting Performance Standards

Either implicitly stated, as in the Title I legislation, or explicitly stated, as in the Iowa Code, establishing levels of required student achievement or performance standards is a major aspect of the assessment system. The Iowa Department of Education realized this and provided support via the CD-ROM "Standards Development for School Improvement in Iowa." This CD-ROM outlines the steps necessary in considering performance standards along with additional supporting references and documentation. However, the current document would be incomplete if the "best practices" in standard setting were not addressed. Hence, the remaining text in this chapter describes the standard setting process that can be applied to most assessments.

Traditionally, standard setting methods have fallen into two camps: test-centered methods and examinee-centered methods (Jaeger, 1989.) Test-centered methods simply reflect the fact that the standard setting judgment is primarily made about the test itself, usually based on an inspection of the actual test items. Whereas, examinee-centered methods call

for judgments to be made about the performance of examinees. As Kane (1995) points out, "all standard setting is based on judgments." Performance standard setting should use a process from which the best judgments are obtained from the people in the best position to make those judgments. Typically these are content-experts, people familiar with the skills and knowledge to be learned.

Berk (1996) provides a list of guidelines or recommended procedures for the judgmental standard setting process. While this list was intended for use with item-level procedures, it applies nonetheless to the examinee-centered approaches.

Recommended Steps in Developing Performance Standards

1. Select a broad-based sample of the most qualified and credible judges you can find. *(For all intents and purposes, this would presumably be teachers.)*
2. Train these judges to perform the standard-setting tasks to minimize instrumentation effect and maximize intra-judge consistency. *(This means that we must make teachers comfortable and familiar with the process and make sure they can consistently apply judgments themselves before comparing to other judges.)*
3. Use a multistage-iterative process whereby judges are given one or two opportunities to review and refine their original decisions based on new information to maximize inter-judge consistency. *(This means that teachers should apply their judgments, see the effect of their judgments via empirical feedback and impact, discuss their judgments with other teachers and be allowed to change their judgments.)*
4. Require judges to provide explicit behavioral descriptions for each achievement level with corresponding anchor items. *(This means that the teachers will have to operationally define their meaning of the different levels of the standards.)*
5. Determine the judges' decision policy based on the objectives or dimensions measured. *(This means that the rules the teachers use in making decisions should be documented and should be based on relevant facts.)*
6. Provide judges with feedback on their individual and the panel's decisions. *(This means that initial individual decisions need to be compiled and supplied to the entire standard setting panel.)*
7. Supply judges with meaningful performance data on a representative sample or appropriate sub-sample of examinees to "reality-base" the ratings. *(This means that the teachers should be shown the effects of their judgments on students in relation to the levels of the standards.)*
8. Allow the judges the opportunity to discuss their decisions and pertinent data without the pressure to reach consensus. *(Many teachers will have different backgrounds and different skill levels of students and different educational goals, as such; it will not benefit the standard setting process to try to force all these diverse considerations into consensus.)*
9. Solicit judges' content-related decisions about achievement levels via consensus but all item and test score decisions via independent ratings to avoid pressuring judges into alignment or the influence of dominant judges. *(This means that intense discussions are allowed as long as the final vote is private.)*
10. Compute the cut-score(s) from the mean or median item or test scores based on the judges' ratings. *(Use the average of the groups' decisions to determine the cutting score.)*

The advantages of following such standard setting guidelines, as presented by Berk (1996), are many:

- Teachers are asked to operationally define the performance standards. This makes the standards more meaningful for the teachers during subsequent use of the standards statements;
- Teachers will use their own students' behavior as a basis for the classification of students into the various performance levels. Clearly, this allows the teachers to maximize what they know best and the most about: the interaction between content requirements and student performance;
- The process can start before actual student data is collected, expediting the standard setting process, though empirical student performance is needed for the actual standard setting;
- Teachers get an opportunity to see and review their judgments in relation to actual student performance on the assessment as required by most standard setting processes;
- The multiple-choice items and open-ended items can be placed on the same common scale along with the students' ability estimates (from a statistical process), so that total student performance is used in determining the empirical performance of each group;
- The resulting standards will be points on the ability scale (i.e., a scaled score) and will be, through the equating process, applied to future test forms, making future standard setting unnecessary;
- Because the teachers will review actual student performance and because the distributions will be generated from student ability, the standard setting process will accommodate responses from modified assessments, such as Braille forms and large print booklets.

Because the implementation of any standard setting procedure requires the use of actual student performance, the determination of the cutting scores cannot be calculated before scoring has been completed. Therefore, careful consideration regarding the coordination of the testing date, standard setting and score reporting must be made.

The following table provides the steps typically involved in establishing empirical performance standards:

Required Steps in the Standard Setting Process

1. Identify curriculum to be assessed.
   (*Determine what is to be measured)*
2. Construct content standards referenced on curriculum.
   (*Determine the importance of that being measured*)
3. Construct measuring instrument in alignment with curriculum and content standards.
   (*Develop items, performance tasks and create the assessment*)
4. Select teachers for the standard setting process.
   (*Sample representatively, selecting all types of background characteristics*)
5. Provide teachers with general statements of the standards and instructions.
   (Provide content standards and procedures before meetings)
6. Administer measure to acquire actual student data (i.e., field test assessment.)
   (Conduct a representative sample, field test)
7. Convene standard setting panel.
   (Teachers construct performance descriptors of the proficiency levels)
   (Teachers classify students' performance by item using the performance descriptors)
   (Teachers discuss their ratings and receive empirical student performance feedback)
   (Teachers revise their classifications if desired)
8. Empirical performance standards are derived from teacher revisions.
   (Cutting scores are calculated as the sum of average teacher ratings)
9. Performance standards are reviewed by stakeholders, in light of actual student performance from the field-testing, and adjusted if necessary.
   (Opportunity for policy-oriented changes for alignment purposes)
10. Standards are applied to field test results and disseminated to field test participants.
    (Standards and their results are provided for public consumption)

III. A Conceptual Assessment System

This chapter provides the reader with a conceptual district-wide standards-referenced assessment system. The goal of this chapter is to make the reader understand that many aspects of such a system have already been thought about at the local level and that the implementation of such a system can be achieved.

It is also a goal of this chapter to provide some rationale for why such a system will be of benefit to students, teachers, administration, parents and the public alike because it exploits the same desires and goals of all stakeholders of education.

A district-wide standards-referenced assessment system is a system of assessment tied directly to established content-standards that provides enhanced student learning through informed instruction. As such, this system will incorporate as many different measures, or assessment components, as necessary to address the content-standards. The results will be used not only by teachers, but by all stakeholders.

A. Why Measure with Assessments?

One criticism often charged at those building and implementing assessments is that they are not needed and get in the way of real evaluation. Often these critics cite anecdotal records of poorly constructed or poorly used measures that misrepresented the skill level of the individual involved. Other critics will recall when you could "get a real education" through hard work and strict discipline. Unfortunately, we can probably all recall bad experiences from our educational past in general and past assessment activities in particular. Despite these previous errors, assessments offer many advantages, particularly at the district level, over less rigorous and less formal measures. For example, many teachers still construct formative assessments for use in their classrooms on an almost daily basis. This is due to the need to guide instruction, monitor student progress and to report feedback to the student and the student's parent regarding the student's performance. These teacher-made assessments come in many shapes and varieties. One teacher might administer pop-quizzes once a week. Another teacher might construct elaborate performance assessments requiring the students to engage in data collection or the application of mathematical problem-solving tasks outside of the classroom. Other teachers will resort to work sheet exercises while others will assign and

collect homework. Each of these activities are take place in an uncoordinated way, with each teacher designing and constructing his or her own assessments. If the district has a current testing program, it is not unlikely to hear comments about how much "real" instruction these assessments take away from class time. The implementation of a district-wide assessment system puts the power of the assessments into the hands of the teachers. Teachers will be able to model their assessment activities to be in tune with the district-wide assessment. If the district selects multiple measures and multiple modes of assessment, the teachers will be able to offer input into what types of assessments are best for the type of content being instructed. For example, perhaps a performance-assessment in mathematics could be constructed (or purchased) as one of the multiple measures comprising the district-wide assessment. Perhaps student portfolios, essays regarding mathematical experiments, as well as "on demand" assessments could comprise the district-wide assessment system. Provided the district follows the guidelines offered in this and other documents and that these assessment components meet the requirements of a large-scale assessment regarding reliability, validity and fairness (as will be explained in upcoming chapters) there is no reason not to include these assessments.

What better way to ensure that your content-standards are being addressed than by selecting and implementing multiple-measures of those standards. In addition, the standardized directions and data collection forms associated with a district-wide assessment will ensure that all students have the same chance to show their best work. By selecting a variety of assessments, provided proper reviews are conducted, only the most fair and equitable assessments will be used for the purposes of evaluation.

A district-wide assessment system offers an opportunity to coordinate instruction and assessment much more closely, thereby opening the door to improved student learning through informed instruction. The multiple measures will give more opportunity for students to show their best work while the impact of a "bad day" during any one assessment will be reduced. Clearly, the district-wide assessment system offers more local control regarding the types of measures implemented. It also, however, requires additional work and responsibility.

B. Link to the Content Standards

By now, most districts should have in place their content standards. Content standards are statements about what is most important instructionally and will guide both educators and assessment specialists. Content standards describe the knowledge and skills to be instructed and to be learned.

The following definition of content-standards was taken from a paper presented at the Midwest Regional ESU Conference: Building Leadership for High Performing Schools (Whisler, 1998):

> "A standard is a general statement of information or skill. It identifies or articulates what students are expected to learn. Specifically, a standard articulates what students should know or understand and the skills they should have."

The CD-ROM distributed by the Iowa Department of Education entitled "Standards Development for School Improvement in Iowa" also provides information regarding the establishment of content standards.

Regardless of the actual content of the standards, the processes followed in their establishment and the way they are communicated, they are little use if the selected measures of student learning are not directly linked to the content-standards. A district-wide assessment system will allow for the best possible link to these standards in several ways.

- First, the use of multiple-measures of student achievement will provide the flexibility to measure all aspects of the content-standards without sacrificing important content that may not otherwise be convenient to assess. For example, some districts have relied on the editing skills assessed by standardized multiple-choice assessments when in reality their curriculum or content-standards called for direct measures of student writing. In a district-wide assessment, both measures could be required, offering a broader coverage of the desired content.

- Second, presumably the people involved in determining the content-standards will also be the people providing input to the components of assessment in the district-wide system. This provides the motivation and the ownership to ensure that the district-wide assessments are the best that can be constructed.

- Third, because the content-standards explicitly state what is important for instruction and learning, they should also guide the reporting. Teachers will want reports that answer specifically where students are having trouble and hence, where additional instructional emphasis is warranted. The public will probably want to know what students are prepared to do and how well they do it. The Iowa Department of Education, as well as Title I, will want to know how well the district is meeting its annual improvement goals. Because of the link to the content-standards, developers of the district-wide assessment system can pick and choose reporting mediums that meet the needs of the constituency. For example, the Achievement Levels Report from the Iowa Testing Programs (ITP, 1998), provided this matches with the content-standards could meet the needs of some stakeholders. Another report might include annotated scoring-rubrics associated with a writing measure. These might be useful to the teacher trying to improve student learning through informed instruction. The list of possible reporting options is unlimited.

C. Progress Indicators

Another advantage that a district-wide assessment system offers users is the ability to provide a host of progress indicators from the various components, which are relevant for the different needs of the district. The goal of assessment in general is to provide feedback to those involved such that enhanced student learning can take place through informed instruction. This goal will have many mileposts depending upon the user of the results. For example, the teacher's need with regard to progress indicators is quite different than the need of the superintendent or the Iowa Department of Education. The teacher will probably want to know, through feedback from many different sources (including assessments), what instructional strategies are working, where student deficits lie, which student exemplary skill should be exploited, etc. The state, on the other hand, wants to be sure that the school is making progress toward reaching its annual improvement goal. The teacher will need one set of progress indicators whereas a clearly different set will be needed by the state. The teacher will need direct diagnostic information in a timely manner that is relevant to that particular section of the content-standards being taught. The state will need information regarding the percentages of students falling into the performance standard categories, how this compares to the baseline performance, and what plans are in place to continue improvement.

Another use of the various progress indicators will be to assist in decisions regarding the placement of individuals into different programs. New students into the district will require decisions regarding reading level, gifted-and-talented programs or special needs

programs. One consideration in this regard is to include various "on demand" assessment components as part of the district-wide assessment system. In such a way, all such decisions about placement within the district can use the same assessments, thereby reducing the amount of individual effort needed for each decision.

Non-cognitive measures can also be part of the assessment system. The system can use such measures as dropout rates, absenteeism, and tardiness, for example, to make statements about school improvement. Provided a documented procedure is in place, such evidence could be provided supporting the general improvement plan. Other non-cognitive indices such as participation in "science fairs"; regional, state and national competitions (such as spelling bees), as well as other extracurricular activities could be cited as indicators that education plans are having a real impact. Again, the success of these indicators depends almost entirely on a systematic, documented and reproducible procedure for showing their effects.

D. Program Evaluation

Logic dictates that if you desire an improvement in the outcome of some process, you implement changes to the process in hopes of affecting an improvement. Hopefully, these improvements are based on a study of what was perceive as "wrong" with the program in the first place. Then, hypotheses were generated, "tried out" and deemed successful long before they were actually implemented as changes to the program. Surely, it would be a simple task to wait and see if the changes did indeed affect improvement. As Frechtling (1989, pg. 479) points out, nothing could be farther from the truth! What is one evaluator's success is in many cases another's failure. One trivial example might be when one evaluator uses participation rate as the key index to the successfulness of a program, while another uses a score from an assessment given at the end of the "program" as the index of success. If participants are "sold" on attendance only to find the "program" boring and are not allowed to drop out, they will stay to the end and do poorly on the assessment. Hence, there is a lot of participation (good by one measure), but poor scores on the assessment (poor by another measure.) What this example points out is that there is debate even regarding what is considered a success and what is considered a failure in a program evaluation.

Program evaluation means different things to different people, but as Mehrens and Lehmann (1987) point out, program evaluation differs from student evaluation in terms of the decisions which are ultimately made from the results. For example, evaluations yielding results used for individual student decisions are clearly student evaluations and not program evaluations. Decisions regarding how well students perform is part of student evaluation, whereas decisions regarding at which grade to teach a subject is part of program evaluation (Mehrens and Lehmann, 1987, pg. 423.)

Mehrens and Lehmann (1987, pg. 423) provide the following list of things they see as appropriate under program evaluation:

- Why were student goals achieved (or not achieved)?
- What are the goals of the evaluation...what are the objectives of the evaluation?
- Are there unintended outcomes of the program?
- What impact does the program have on persons other than students?
- Is the program cost-effective?
- What aspects of formative evaluation are evident in addition to summative evaluation?

Program evaluation is the collection of evidence that desired outcomes have transpired and undesired outcomes have been minimized. Program evaluation is not as rigorous as scientific research, though it might use the scientific process and require standardized, documented procedures. Those implementing a district-wide assessment system will need to demonstrate what outcomes they desire, which ones they wish to avoid and to propose evidence collection systems that will provide evidence to support judgments regarding these outcomes. Often, the data collected for program evaluation will be similar to if not the same as student evaluation. As Mehrens and Lehmann (1987, pg. 423) state, student progress is only one dimension of program evaluation. Frechtling (1989, pg. 479) states that despite the "institutionalization" of test scores as one index of program evaluation, problems still exist. The key to a successful program evaluation of a district-wide assessment system will be to clearly delineate the goals of the system

(program) and to collect evidence supporting the results of the system in meeting those goals, while not introducing unintended and perhaps negative consequences.

IV. Different Assessments for Different Purposes

The reason people choose different assessments can at times be quite interesting. For example, some people select a particular assessment because: "it's what we have always done." Other people might choose an assessment because: "it was within our budget." Still others maintain that all assessments are the same anyway, so: "...just pick one." While these might all be justifiable reasons in at least the selector's eye, they are not good decisions if the content-standards and the purpose for which assessment is required has any bearing. The fundamental concern in selecting an assessment instrument is the purpose for measuring. This purpose is often predicated, especially in achievement testing, on a need to measure attainment of the content-standards.

Given that the purpose for measuring must drive the selection or construction of an assessment, and that this is mainly based on a district's content-standards, the remaining sections of this document review different assessments. The perceive purposes, as well as the advantages and disadvantages of each, will be outlined. In addition, the impact each may have on planning and implementing a district-wide standards referenced assessment will also be discussed.

A. Classroom Based Assessments

Teachers have been using classroom-based assessments since the beginning of formalized education. These may take the form of "pop quizzes," oral readings, graded homework, book reports, essays, projects, experiments, as well as both formal and informal tests. Teachers typically construct and score such instruments, though some publishers produce "curriculum packages" with embedded assessments. These assessments are typically the most relevant instructionally because teachers know what they teach and are best able to determine what is appropriate to ask their students. In addition to being more instructionally relevant than non-teacher made tests, they provide for a continuous "stimulus / feedback" loop wherein the teacher may modify instruction based on student learning as seen from the measurement. Another measure is taken and the instruction may be modified or directed again, and so on. This allows for very timely and powerful information for the teacher regarding what aspects of instruction (strategies, aids, sequence, pacing, etc.) are working best for this subject and this group of students.

Teacher constructed classroom-based assessments are perhaps the best possible example of true formative evaluation.

Implications for a District-wide Standards-Referenced Assessment System

It is not the case that a simple way to construct a district-wide assessment is to use the "best" of classroom assessments. Remember that the purpose of the assessment must dictate why it is selected or constructed. What is the best classroom assessment for one teacher (in one school, in one subject area, with a particular group of students) may be a poor assessment for use in a different school or class. The teacher in a different classroom will have a different needs, will interpret the content-standards a different way, will have different students with various behavioral and / or cognitive needs. This does not mean, however, that teachers cannot develop a component of the district-wide assessment system. In order to do this, a panel of teachers must decide which pieces of the content-standards the particular component of the assessment should address. They then must construct the items / tasks to be measured (these could be multiple-choice items, writing essays, performance assessments, open-ended or student response items, or a combination of all.) These items must be evaluated by, ideally, a different panel of teachers to ensure that they do indeed measure the identified content standards. Once the items / tasks are constructed, then directions, answer documents, and other logistics of testing must be developed or outlined. Then, the items (as well as the directions, scoring rules and administration procedures) must be "field tested" by administering these items / tasks to groups of students, preferably in different environments. The items and the results of the pilot testing must be reviewed, particularly with respect to possible differential functioning (i.e., possible bias.) The items remaining can be placed in a pool for ultimate selection into an assessment component. Understand that the psychometric properties of reliability, validity, fairness, etc., must be fulfilled and documented. In fact, the test construction process outlined in another chapter of this document should be followed if these assessments are to be constructed locally. Finally, after all of these criteria are established the assessment may be used as a component of the district-wide assessment. However, the requirements of using the assessment continue. After the assessment is constructed, consideration regarding the administration district-wide must be taken into account. Such issues as the dissemination of test forms, the collection of

student responses, the scoring of student responses, the distribution of results, as well as the documentation and presentation of reliability, validity and "fair use" information must be made. Implicit into all of this are considerations regarding printing or copy facilities, scoring options (especially in judgmental scoring, as is required for a writing essay), transportation for distribution, etc.

Clearly, the development of a district-wide assessment is far more difficult than that of a classroom-based assessment and may be an overwhelming task for a many school districts. It can be done, however, provided the correct procedures are followed, sound judgments are made along the way (ones which are documented), and the people in the district: teachers, administration and pupils alike, are committed. Sections of this document describing the test construction process, the collection of reliability and validity evidence, and the reporting of scores will help those who desire to construct their own assessment components.

B. Standardized Norm-referenced Assessments (NRTs)

Nationally standardized norm-referenced assessments provide perhaps the soundest measurements from a psychometric perspective as can be found in testing. Authors of such tests spend years constructing the items, trying the items out to ensure their functionality, and conducting research to support their use (i.e., equating multiple test forms, building growth scales, deriving reported score metrics like grade-equivalents, etc.) Provided these tests match the district content-standards, they may be an easy and relatively cost effective way to add assessment components to the district-wide assessment system.

NRTs are typically written to a "consensus national curriculum" such that they will likely match many of the key aspects of the content-standards put in place by the district. However, it is doubtful that they will cover all of the content standards, nor are they likely to cover the content-standards in enough depth to be the only component. Additionally, often the results of the assessment are in terms of "status" scores, showing student performance relative to other students (i.e., are norm referenced.)

A potential disadvantage to using nationally standardized norm-referenced assessments is they may provide a relatively narrow picture of the content-standards. They provide a

single point of time "snapshot" of student performance. They require absolute standardization in administration and student answer documents are typically returned to the publisher or other contractor before they can be scored.

Some advantages include the state of the art evidence of reliability and validity (for the purposes outlined in the test manual.) They have an outstanding array of score reports, some of which can usually be tailored to meet the needs of the district. They are usually "on-demand" assessments, meaning that they can be given at the time desired by the district. Finally, they provide for very easy comparisons at the district, state and national level. Many offer performance levels that are already established. In the case of ITBS / ITED, these reports have been tailored to schools in Iowa (See the Interpretive Supplement for the Achievement Levels Report, 1998, ITP.)

Implications for a District-Wide Standards-Referenced Assessment System

The biggest concern for districts wishing to add a national norm-referenced assessment as one component to the district-wide assessment is the need to select such an assessment aligned with the content standards. Many publishers will provide such a match provided the local curriculum (i.e., the content standards) are well documented and easy to understand. Most districts will want to perform this matching themselves, probably at the item level, in order to ensure an accurate match, but also to understand which content standards are not covered by the assessment or are covered only sparsely.

Another concern will be the work selecting or purchasing an assessment that will cover the remaining content standards. Even if the match between the selected NRT and the content standards is good, there may be so few items associated with any particular content standard that the district would desire additional measures of student progress across all content standards. This would also take care of the problem of the assessment providing only one "point in time" measure. By adding additional components to the district-wide assessment system beyond the selected NRT, the district should be able to provide a wide-ranging (broad content coverage) but yet still fairly in-depth (many items / tasks per content standard) measurement.

C. Standardized Criterion–Referenced Assessments

Criterion-referenced assessments are well suited to a standards-based assessment system because they usually measure more specific content-related criterion (i.e., content standards.) As such, they are usually constructed such that many more items / tasks are available for any particular aspect of the content or specific content-standard.

Most CRTs that are in use today are constructed explicitly for large-scale assessment, typically as part of a state-mandated assessment (Minnesota Comprehensive Assessment System, Michigan Educational Assessment Program, Alabama Writing Program, etc.) However, some CRTs are available commercially, though often these are in specific subject areas for use as "end-of-course" examinations.

As with the NRTs, the CRTs also provide only a single point-in-time "snapshot" of student performance. However, unlike the NRTs, they typically provide more items or tasks per content-standard they purport to measure. CRTs provide for more items per content-standard because they typically measure fewer pieces of content, thereby reducing their coverage of all content-standards more than the NRT.

Perhaps the biggest advantage CRTs offer districts over other assessments is that results are often expressed in terms of what students can or cannot do and are not based on a norm-referenced comparison. For example, one of the most trivial, yet very often used, reports is the number of items per specific content cluster a student answered correctly versus how many items were asked in total. Some educators interpret this as a statement regarding the domain of all such items the student would be likely to answer.

CRTs, where available, also allow for comparisons between local, district and statewide performance through the establishment of performance standards. This will typically require that a performance-standard setting be conducted as described in a different section of this paper.

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

Districts developing an assessment system will find the appeal of a criterion-referenced assessment very tempting. However, it is unlikely that a commercially available CRT can be found that will address even the most important district content-standards, let alone all of them. Therefore, districts should consider their use to "fill in the gaps"

regarding which pieces of the content-standards are not addressed by other aspects of the assessment system. For example, if the district content-standards call for assessing student progress in the performance arts, a commercially available CRT could possibly be found that would match the particulars of these standards. If so, this assessment could be added to the assessment system broadening the coverage of the standards at, presumably, a reasonable cost.

If districts desire to construct their own CRTs, they will undoubtedly improve the ability of the assessment system to cover the content standards but will take on a great deal of development costs. Each test construction step described in a previous section will have to be followed, as well as the additional need to collect all of the reliability, validity and test fairness data typically associated with large scale assessments. As stated previously in this section, such a task is not for the faint of heart.

D. Performance Assessments

The recent trend toward more "authentic" measures of student performance has yielded a new type of assessment termed the "performance assessment." While the concept of performance assessment means different things to different people, it is generally considered to be constructed of "real world" tasks, tasks that require students to generate more than answers but often to demonstrate judgment, problem solving and other organizational skills. One example of a performance assessment might be an integrated reading and writing task. This task may require a student to read a passage, answer some multiple-choice questions about the passage, discuss the passage with other students, generate a list of questions about particular aspects of the passage (as a pre-writing exercise), and write an essay about the meaning of the passage. Furthermore, the task could also require the student to edit the essay for grammar, content and organization and possibly discuss the essay with other students. Across this task, any or all student-produced work might be scored and submitted as individual student measures.

Performance assessments have particular appeal for teachers because they are so closely linked to good instructional practices (as evident from the previous example.) In addition, because the assessments typically require the generation of actual student work the inference made about actual student skill is presumed to be more valid (though this is

a hotly debated topic in many assessment circles.) Despite these points, it is true that, due to their close link with instruction, they will probably provide a much better match to the district's content standards.

The availability of commercially produced performance assessments is on the rise though their quality and their match to a particular set of content standards is unknown. Most of these performance assessments provide for standardized administration and scoring, thereby improving their reliability. Often performance assessments or parts of them are imbedded within NRTs and / or CRTs. Regardless of where the district obtains a performance assessment, the assessment still requires that the sound psychometric principles of reliability, validity be demonstrated. While it may seem very logical that an assessment that requires students to write would yield results that would be "intrinsically-rationally valid" (Ebel, 1983) (i.e., generate inferences about writing skill), the evidence must still be collected. In fact, performance assessments have been criticized in the past because of poor reliability evidence and poorly documented validity evidence.

Implications for a District-wide Standards-Referenced Assessment System

Scoring rubrics, rules and procedures are often very complicated with almost all performance assessments requiring judgmental scoring procedures to some extent. This means that the district will have to find funds or personnel available for the scoring. Training, materials, space and time are all resources the district will have to spend in order to score these relatively complicated assessments. In addition, the "turn-around time," time between when the assessment is administered and when results are available for dissemination, is also much longer due to the judgmental scoring.

In addition to the actual physical acts of getting ready to score, scoring and reporting the results of the performance assessments, districts will be burdened to demonstrate the reliability of the judgmental scoring process, often referred to as inter-rater reliability or rater/reader/judge agreement. This issue is discussed further in the chapter on reliability that appears later in this document.

In summary, performance assessments offer the advantages of instructionally relevant assessments which can be linked directly to the content-standards. These assessments produce student generated work that may increase the validity of resulting score

interpretations provided appropriate procedures are followed. Accurate scoring of these assessments typically requires a large commitment in terms of time, money and effort. Demonstration of reliable, valid and fair score use is still required as too is evidence of accurate scoring.

E. Program Evaluation

As stated in the previous chapter, program evaluation is the collection of evidence that desired outcomes have transpired and undesired outcomes have been minimized. Program evaluation is not as rigorous as scientific research, though it might use the scientific process and require standardized, documented procedures. Program evaluation is discernable from student evaluation because of the object of the evaluation.

Districts implementing an assessment system will need to demonstrate what outcomes they desire, which ones they wish to avoid and to propose evidence collection systems that will support judgments regarding these outcomes. Often, the data collected for program evaluation will be similar to, if not the same as, student evaluation. As Mehrens and Lehmann (1987, pg. 423) state, student progress is only one dimension of program evaluation.

Mehrens and Lehmann (1987, pg. 423) provide the following list of things they see as appropriate under program evaluation:

- Why were student goals achieved (or not achieved)?
- What are the goals of the evaluation...what are the objectives of the evaluation?
- Are there unintended outcomes of the program?
- What impact does the program have on persons other than students?
- Is the program cost-effective?
- What aspects of formative evaluation are evident in addition to summative evaluation?

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

Districts will want to put into place some aspects of program evaluation to facilitate, document and provide evidence that the implementation of a district-wide assessment

yields desirable outcomes. Districts should be careful not to confuse aspects of program evaluation with other aspects of institutional research and especially not with individual student evaluation or assessment. Readers should reference the section on program evaluation presented in the previous chapter.

F. Survey Instruments

Survey instruments as used in this section are typically referred to as self-report measures or self-report inventories (Linn and Gronlund, 1995, pg. 284.) These "surveys" should not be confused with survey achievement tests batteries. Most people are familiar with simple phone surveys or "Likert" type surveys. In the former, people simply ask you your opinion about some topic and record the responses via some standardized reporting metric. The Likert-type scale requires the respondent to provide reactions such as "Agree," "Neutral," or "Disagree" to a series of statements on a particular topic.

Surveys are often used with nonscientific samples (i.e., samples that are not intended to represent a real population of interest but are convenient because participants are available to respond. As such, these surveys typically have large error and do not generalize very well beyond the particular group sampled. This is not true of all surveys. For example, the "USA Today" poll that appears occasionally in the news is a scientifically generated survey with a documented probability sample and associated margins of error reported.

Surveys are often used in the assessment of attitudes. The goal of a survey is to ask the respondents what their reaction is to a particular topic or question. Because attitudes are not "curriculum" or content-standard specific, they are much more difficult to assess than student academic achievement. This is one of the reasons surveys are used.

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

Districts implementing a standards-referenced assessment system will probably find surveys of limited (if any) use as a component of the assessment system itself. However, surveys could be an invaluable aid in collecting evidence of consequential validity. For example, one of the unanticipated consequences (or maybe a planned consequence) of implementing a district-wide standards-referenced assessment system is an improvement in the attitudes of teachers, administrators and students toward the prospects of

assessment. Surveys could detect this. Additionally, districts conducting program evaluations may find surveys a good means of detecting perceived improvements in education by the community and cite these as evidence that the goals of the program are being fulfilled.

G. Needs Assessments

"Needs Assessment" is a phrase applied to many settings which generally means ascertaining what pre-requisites are required before a desired outcome can be achieved. However, the specific meaning of a "needs assessment" requires a clear understanding of the situation. For example, when a business conducts a needs assessment, such an assessment typically identifies an area of problems and collects information about what needs to happen in order to eliminate the problem: "A needs assessment was conducted to determine why the copy center was continuously late in delivering products. Results indicated that due to attrition, the copy center is two machines short of the basic equipment required to perform the task." In this example, the question was very specific (why is the copy center late), the results were very specific (because they don't have enough equipment) and the outcome is very clear (buy additional copy machines.) What is not clear and simple is the process to determine what the needs were, who collected the information (in what format) to determine the needs, and how to explain or document clearly what the need. Unfortunately, when the issue is determining the needs of a district related to education and specifically meeting the annual improvement goals, this task is much more complicated.

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

It is beyond the scope of this text to outline how districts should determine what improvements are necessary and what their annual improvement goals should be. However, districts should consider what their needs are not only in showing improvement, but doing so in light of the content-standards adopted. For additional information regarding the "On-going Needs Assessment" outlined in Iowa Code Section 280.12, see the Self-Assessment 280.12 and 280.18 Annual Report for the 1997-1998 School Year as provided by the Iowa Department of Education.

H. Procedural Guidelines in Assessment Development

The following table provides an example of the steps taken to develop a "large scale" criterion-referenced assessment. While this example was taken from a state-mandated assessment program, many of the steps are relevant to locally constructed standards-based assessments. Following these steps in planning for and implementing a standards-referenced assessment should ultimately maximize the quality of the assessments constructed.

Clearly, the steps presented in the table will be modified for a host of reasons: purpose of the assessment, number of students served in the district, skills and expertise of personnel involved, funds available, etc. However, the example represents a realistic list of the steps required in the construction of a quality assessment and provides a comprehensive overview.

Example Test Development Process

- **Develop Assessment Objectives (Content-Standards)**

  Committees review the curriculum to develop appropriate assessment objectives and targets of instruction. Committees provide advice on assessment models and methods to align assessment with instruction.

- **Develop Assessment Specifications**

  Committees develop measurement specifications. These specifications outline the requirements of the development of the assessment such as eligible test content, item types and formats, and may include sample items. These specifications are then distributed to teachers as a guide to the implementation of the assessment program.

- **Develop Assessment Blueprint**

  Committees develop a test blueprint. The assessment blueprint defines many practical aspects of the assessment, including the length of the assessment, the number of items or tasks per objective, etc.

- **Develop Task Specifications and Example Items**

  Committees construct procedures and rules for developing items and tasks. For example, the specifications might state that no more than three decimal places be used in mathematics, for multiple-choice item types. Another specification may state that writing essays requires a pre-writing activity. Once the rules are established, the teachers develop sample items and tasks for dissemination.

- **Item and Task Development**

  Using the specifications, committees develop items and tasks.

- **Item Content Review**

  All members of the assessment team review the developed items, discuss possible revisions and make changes.

- **Item Content Review Committee**

  Committees review the items (some of which were revised during content review) for appropriate difficulty, grade level specificity and to eliminate potential bias.

- **Field Testing**

  Items are taken from the item content review committees with or without changes and are field tested as part of the assessment program. Data are compiled regarding student performance, reliability, validity and possible bias.

- **Data Review Committee**

  Committees review the items in light of the field test data and make recommendations regarding the inclusion of the items into the available item pool.

- **New Form Construction**

  Items are selected for the assessment. This selection is based on content requirements (such as matches to the test blueprint), as well as statistical (predicted passing rates, predicted test form difficulty), and / or psychometric (reliability, validity) considerations.

V. Standards-Referenced Assessment System

There are several characteristics that distinguish a district-wide standards-referenced assessment system from other assessments. Some of these characteristics include:

- Multiple assessment components;
- Assessment components which are matched to the content-standards;
- Performance-standards which are attached to the results of the components or to a composite of all components;
- The standards (both content and performance) apply to all students.

In addition, different stakeholders will use the results from such a system in a variety of ways. As such, many different types of information will be presented in a variety of ways. Some of the uses may be:

- Teachers will use the results to inform instruction;
- Principals and superintendents will use the results to evaluate how well the district is performing relative to the annual improvement goals;
- The public will use the results for a variety of different purposes;
- The Iowa Department of Education will use the results as outlined by legislation;
- Students will use the results for self-evaluation.

As the needs of the stakeholders are expressed, each district-wide assessment will change to meet those needs. This in turn will change the characteristics of the assessment system, especially with regards to reporting. For example, one district may have content standards that strongly emphasis student generated work. As such, the assessment system will probably require more open-ended, task oriented assessment components than some other district. This will impact both the number and types of multiple measures, as well as when and how the results will be reported.

The remainder of this chapter provides information regarding select characteristics of a standards-referenced assessment system. These are not the only characteristics but some

which will help set the stage for thinking about how a district might go about designing such as system which will be elaborated on in the next chapter.

A. Standards-Referenced Assessments

In some ways, all assessments are standards-referenced. The real question is: which standards? For example, when a teacher constructs an exam for use in the classroom, the exam will cover particular and often explicitly stated pieces of the curriculum (i.e., content requirements.) Also, the grading scale will often be explicitly stated: 90 percent is an "A," 80 percent is a "B." etc. (i.e., performance requirements.) Commercially published tests like the ITBS or ITED also have explicitly stated content requirements (usually seen in the test blueprint, table of specifications, or in the objective / cluster level reporting.) Also, these tests can have performance requirements either explicitly or implicitly stated. For example, the ITBS and ITED have explicitly stated achievement levels (See the Achievement Levels Report, ITP, 1998.) They may also have some implicit performance standards in the grade-equivalent or percentile-rank scales (e.g., if the national average is the 50^{th} percentile then we may desire that all students in a particular district score above this value.) The particular content to be covered by the assessment and the level of performance required may or may not be explicitly stated before the assessment. If these assessments have, at least arguably, inherent content and performance standards, then how do they differ from those required by a district-wide standards-referenced assessment system? For one thing, both of these examples allow the content that is being assessed to be determined by a specific person for a specific application. The teacher decided what to assess in the first example and the test publishers decided what to assess in the latter example. Hence, if implemented district-wide there could be much disagreement regarding the content selected. This disagreement is essentially a lack of match between what is being assessed, what is being taught, or what is perceived as important and should be taught. If the content-standards were determined for the district as a whole and agreed upon prior to assessment then there would be a consistent understanding of what is to be taught as well as a "road map" of what the subsequent assessment must cover. Similarly, because people's expectations will differ regarding student performance, a systematic way to establish performance

standards would standardize what is required for all students and would remove any doubts about what performance is considered "good enough" for a particular purpose.

The two characteristics of a district-wide standards-referenced assessment system that have been discussed to this point (content and performance standards) are elaborated upon further in the next few paragraphs. The reader should pay attention to the differences between the characteristics of an assessment system as described in this section and the more general characteristics of assessments.

Everything starts with content standards. It is impossible to conceptualize what a district-wide standards-referenced assessment system will look like without a clear understanding about the content to be assessed. Content standards should already be in place in the districts. The CD-ROM provided by the Iowa Department of Education (Standards Development for School Improvement in Iowa, 1998) provides the following definitions:

> "Content standards specify what students should know and be able to do in identified disciplines or subject areas."
>
> "Performance standards describe how good is good enough and describe at least three levels of student performance. The federal Elementary and Secondary Education Act (ESEA) requires that at least three levels of performance be established to assist in determining which students have or have not achieved a satisfactory or proficient level of performance..."

The CD-ROM also states that districts have to option to develop more than three levels of the performance-standards, as explained in another chapter of the current document.

These content and performance standards are applicable to all students. Previously, schools may have been required to document student performance (and any gain or loss) for the purposes of Chapter 1 or Title I. Students receiving these special services were typically the ones for which performance gains were considered. Now with the implementation of a district-wide standards-referenced assessment system, all students will use the same content-standards and measures of progress (via the performance-standards) will be collected for all students.

Using the performance-standards, districts will document their progress in meeting their annual improvement goals. This documentation will be either via a profile showing changes in the percentage of students falling into each performance-standard level on

each component of the assessment system, or via a composite across components. Discussions regarding the differences between profile and composite indices are provided in the sixth chapter of this document.

It is doubtful that a district will find a single assessment which will measure all of the content-standards of the district. If a single assessment is found, it is doubtful that the coverage of the content-standards will have very much depth (i.e., there will be relatively few items / tasks per each content-standard.) Therefore, districts will need to use multiple measures of student performance via multiple assessment components. These multiple components should be selected to maximize the coverage of the content standards while still providing for rich measurement of any particular content standard. The use of multiple assessment components, which match and exploit the content-standards of the district, from which performance-standards can be established, essentially defines the shell of a district-wide standards-referenced assessment system. This is elaborated upon further in the next section.

B. Standards-Referenced Assessment System

The three characteristics of a standards-referenced assessment system (as opposed to an assessment), are:

- Clearly defined, documented and understood **content-standards**;
- Clearly defined, documented and understood **performance-standards**;
- Multiple measures of the content standards (**multiple assessment components**.)

These are not the only characteristics of an assessment system but they are the ones that make it most distinguishable from assessments typically used.

The need to use multiple measures of student progress comes from several factors. First, it is always better to make individual student decisions on as much information as possible. Hence, if results from the assessment system are to inform instruction they should provide as much information as possible about the strengths and weakness of student skills. Second, as alluded to in the previous section, the need to cover all district-

wide content standards with as much depth (i.e., with many items / tasks per content standard) as possible will simply necessitate multiple measures of student achievement.

Other than the required match to the content-standards, the district may select the multiple measures seen as most appropriate. Many districts will select a norm-referenced multiple-choice assessment from a test publisher. This will provide a broad-range measure of student skills collected in a standardized fashion, with high quality reporting. The districts, depending upon how they plan to monitor student progress, will probably have to establish performance-standards to apply to these measures. A match between the table of specification, test blueprint, objective / cluster reports or via an inspection of the items themselves will reveal where (i.e., in which areas) the content standards are not being addressed or are being sparsely covered. These are the areas the district may wish to start with in considering which of the multitude of additional measures will be used.

The district may consider adding a commercially available criterion-referenced assessment or "end-of-course" test as one of the components of the assessment system. These assessments typically have explicitly stated content objectives, are usually standardized and often have supported scoring and reporting services. Typically, the publishers or providers of these assessments will work with the districts to ensure that the tests / tasks / items selected match the content-standards. Additionally, many of these service providers are willing to customize the assessments specifically for a district when costs allow.

Some districts may have district-wide assessments already in place. Provided these assessments match the stated content-standards and this match is documented, they may become components of the assessment system. Districts should ensure that such "locally constructed" or locally used assessments meet all of the psychometric, scoring and reporting requirements outlined in this document. The advantage of adding such measures as components to the system is that they will probably provide a very strong match to the content-standards, will be economically advantageous since they are in place already, and should have good recognition within the district thereby eliminating any reservations associated with "starting up" a new assessment program.

Districts should be careful when considering adding an existing district-wide assessment as a component of the assessment system for several reasons.

- First, the match to the content-standards must be documented. Some districts are surprised to find that the assessments they constructed themselves to (measure their curriculum) sometimes do not match with the content-standards as well as desired.

- Second, these assessments are not teacher made, classroom-based measures. They must fulfill all psychometric requirements regarding evidence of reliability, validity of resulting score interpretations, fairness, etc. (see the sixth chapter of the current document for more detail.) Districts should be able to score and report to all stakeholders involved, and these reporting requirements will probably be different if the assessment is added to the system than when it was used district-wide as a stand alone.

- Security is another concern. If the assessment has been used district-wide, what assurance is there that the items or tasks have not been compromised. It would be a shame to develop a system made up of multiple assessment components only to discover that the interpretation of the results is questionable because one of the components security was compromised.

Some districts will find the need to add multiple-measures to the assessment system to cover content-standards which require student-generated work. Writing essays, science projects, performance assessments, portfolios, etc., would provide these "authentic" measures of student performance desired by some districts. The district should be careful regarding the amount of resources expended in developing, scoring and reporting such assessments. For example, if a writing essay is collected for all students in the district, will this be scored by a professional scoring agency or will the district desire to score the essay themselves? Either consideration will cost money, time and personnel hours that may not have been planned for. Additionally, all of the psychometric requirements alluded to in the previous section will still have to be fulfilled. Evidence of reliability, validity, and fairness will have to be collected. In addition, because the measure collects examples of student work, this does not mean it implicitly matches the content-standards. Documentation of the match between the content-standards and these assessments, if they are to be components of the assessment system, will have to be made.

Districts will need to consider how other information may, or may not, be applicable to the district-wide assessment system. For example, such non-cognitive measures as

attendance, dropout rates, student award programs, science fairs, etc., may have a place in the district-wide assessment system. These indices may not serve as components to demonstrate student performance on the content standards, but may be cited as evidence of consequential validity, monitored for impact as potential confounding error components, or may simply be aspects of a systemic program-evaluation.

After a district has selected the components of the assessment system, thoughts should be directed toward the steps needed to document student progress across these various assessment components in terms of the performance standards. Information regarding establishing performance-standard levels and monitoring student progress across those levels is presented in the next chapter. Consider for now, however, that once individual student performance on these components is tied to performance-standards, this data can be used to make judgments about the degree to which the district is meeting the annual improvement goals. For example, if the district's annual improvement goal is to move five-percent of the student population from one performance standard category to another, this can be documented by aggregating the student level data. In addition, because each component of the assessment system is matched to the content-standards and because performance-standards will be established on these components, the districts will be able to make statements about the type of student skills required for each performance standard.

VI. Critical Issues in Designing Standards-Referenced Assessments

This chapter provides much of the information relevant to those wishing to implement a district-wide standards-referenced assessment. Such important issues as how to establish performance levels, how many levels are best, assessment logistics including standardization, data collection and reporting, reliability and validity as well as other topics are considered.

Some of the most often discussed topics regarding educational measurement are those surrounding the terms reliability and validity. These are terms that measurement practitioners purport to know yet have a very difficult time defining, especially in easy and clearly understood ways. Because of their importance, they are presented early in this chapter and in much detail. Because the concepts of reliability and validity go hand-in-hand, the remainder of this introduction serves as a preface before each topic is discussed individually.

Stated simply, reliability is the consistency of the results from some measurement. However simple this may sound we have to be careful to distinguish this definition of reliability from that of validity. Linn and Gronlund (1995, pg. 48) make the following distinction:

> "In all instances in which reliability is being determined, however, we are concerned with the consistency of the results, rather than with the appropriateness of the interpretations made from the results (validity.)"

Notice that the statements about reliability have in both instances referred to the reliability of the results from a measurement and not a measuring device or instrument. Often we hear people speak about how reliable some test is, and this is simply not correct. The real things in question are the results from the measurement, are they consistent? Additionally, we must first be able to measure consistently before we can make appropriate interpretations (i.e., show evidence of valid use of the results.) Linn and Gronlund (1995, pg. 48) make this distinction in the following way:

> "Reliability (consistency) of measurement is needed to obtain valid results, but we can have reliability without validity. That is, we can have consistent measures that provide the wrong information or are interpreted inappropriately."
>
> "...reliability is a necessary but not sufficient condition for validity."

Validity is a concept applied to the interpretation of the results or the use of the scores from an assessment. The appropriateness of these interpretations must take into account the purpose for which measures were collected. Linn and Gronlund (1995, pg. 47) provide the following:

> "...validity is always concerned with the specific use of assessment results and the soundness of our proposed interpretations of those results."

The following chapters will carry these distinctions further, but the reader should understand that a sound assessment system will require evidence of both reliability and validity regarding the use and interpretations made of the results of an assessment.

A. Reliability

Reliability is a term used by many people in a variety of ways. Unfortunately, each of these different uses conveys different meanings. Hence, it is not always clear when a person states that a particular observation "was reliable," as to just which aspects of the observation they are referring. While this jargon may be confusing, the concept itself is actually quite simple. The concept of reliability in its simplest form refers to the consistency of the observations over repeated measures.

Introduction

Many textbooks, research papers and journal articles have been devoted to the concept of reliability. It is beyond the scope of this document to provide an in-depth explanation of reliability theory. Rather, this section provides some conceptual examples, outlines some of the procedures used to collect reliability evidence and provides some guidelines about what to look for when considering how to collect reliability evidence for a district-wide assessment system. Along the way, the reader is pointed to references for further detail, including the upcoming technical companion to this document.

A Conceptual Example

Consider the following conceptual example. If one were to gather measures of a person's bowling ability without changing (i.e., improving) the person's true ability, we would expect those scores to be fairly consistent. A simple way to do this is to observe the individual bowling on several occasions and to record the score. A simple average of all the scores across all the games would be an estimate of that person's bowling ability. We know that it is unlikely that the individual would score the same in every game, after all,

human interactions are quite complex and are subject to influences of the situation. Still, if a person were to average a score of 100 across all previous games, we would be surprised if they bowled a 200 on any particular game and very skeptical (to the point of disbelief) if they bowled a perfect game of 300. Our disbelief is fueled by what we know to be a fairly stable or consistent index of the person's bowling skill, namely their average across previous games. In reality, we are making a statement about the reliability of the results from the measures. We expect this person to earn a score on any one game similar to the average.

Some Important Properties of Reliability

Most people strive for continuous improvement in nearly every endeavor they pursue. Such an improvement plan is also closely related to the concept of reliability as can be seen from the work of W. Edwards Deming regarding quality control in manufacturing (See for example, Deming, 1986.) Deming's idea of continuous improvement can be summarized as the following: If manufacturing specifications call for tolerances, (i.e., errors) of only one-quarter inch during the current round of production, cut this in half during the next production round. Deming's idea was for constant improvements in accuracy through error reduction. In other words, Deming wanted to increase reliability (consistency) by reducing error and he wanted to do this continuously. This example highlights two important properties associated with reliability: 1) the collection of evidence of reliability is an ongoing and continuous process; and 2) many different sources of error will reduce reliability and all such sources need to be investigated and controlled when possible.

Another way to conceptualize reliability is to consider the degree to which we can generalize a score collected on one occasion, under a particular set of circumstances, to another occasion where the circumstances my be slightly different (Mehrens and Lehmann, pg. 54, 1987.) In the bowling example we used the average of all past scores from a bowler to judge the consistency with the score resulting from a specific game. Another way to do this would be to compare the scores from two games. For example, suppose the bowler scored 106 on the first game and 112 on the second game. Why do these scores differ? First, we should recognize that they are very similar in reference to how the points in bowling are earned. Second, how did the circumstances change

between the first and second games bowled? Did the bowler need the first game as a "warm-up" indicating that the first score was too low? Did the bowler learn something about the particular lane being used? If the games occurred during different weeks at different bowling alleys, then such a difference of six points might not be perceived as very large. Additionally, if the bowler used different balls, different shoes and bowled at different times of the day across these two occasions, we might feel quite confident that the bowler's true ability is some where in the range of about 106 to about 112. The particular circumstances that could potentially lead to error (error components) mediate our confidence in generalizing from one observation of a behavior to another.

Another Conceptual Example

Using another example inspired by Mehrens and Lehmann (1987, pg. 54), we expect that fluctuations in our personal weight between, say, Tuesday and Wednesday morning to reflect true individual differences and not be attributed to random changes in the measurement device (scale) used. The measurement of physical weight is a good example to use when discussing reliability because it too is impacted by inconsistencies or errors that can be attributed to different components that are either implicitly (without thought) or explicitly (purposefully) held constant. For example, most people weigh-in at about the same time of day, usually in the morning. This "holds constant" any fluctuations associated with time. Most people also weigh-in wearing the same outfit (i.e., in about the same state of dress or undress.) In fact, many people purposely weigh when wearing the same clothes in order to "hold constant" any differences in weight due to clothing. Additionally, people tend to stand on the scale in the same way and, use the same scales from day to day, with the scales in the same physical location each time. All of this control adds to reliable measures of your weight. We would probably not weigh ourselves on the morning of one day and the evening of the next because we suspect that such variation will have unknown impact or consequence on the accuracy of our measure of weight. This reluctance to *generalize* from what we perceive as unstable measures is due to our desire to be consistent and ultimately accurate. In summary, it is important to identify and control variables that would otherwise reduce the consistently of our measures (i.e., interject error), thereby lowering reliability.

Reliability Evidence for Academic Achievement

Our examples of reliability so far have dealt with the relatively easy concepts of athletic and physical measurement. This does not mean that such measurement is easy, only that measuring such intangible things like student academic progress or achievement is much more difficult. Using our bowling analogy again, perhaps in a single evening we can observe and record the scores from as many as five games before the person's behavior changes (i.e., the person's performance decreases due to fatigue or other changes to the actual condition of the bowler.) The average would be the best estimate of the bowler's "true" bowling ability. The range of scores (presumably the bowler would score differently in each game) represent the inconsistency in measures of this ability. In fact, the mathematical interpretation of this range is via a statistic called the standard error of measurement (which will be presented later in this chapter.) Each game was not exactly the same, nor was it the same as the estimated true bowling ability, the average of all games. This is the case for any number of reasons: individual differences in performance; different distractions; different lanes; different contexts (i.e., a different sequence of bowlers), different pin configurations and so on (e.g., different error components.) Obviously, when it comes to the results of measures of achievement, we will have trouble getting repeated observations of student behavior (scores) through repeated testings. If we gave the same test a second time to a group of students, for example, fatigue would likely contaminate the student's performance. Additionally, the student's behavior will also be contaminated by how they responded when they saw the test the first time (a latency effect.) This influence could produce either higher or lower scores, but would certainly not yield independent results. Because of problems with repeated observations, psychometricians, measurement specialists and mathematicians provide ways to estimate reliability without contaminating the student scores through repeated testing. This is possible through classical and modern reliability theory (See for example: Mehrens and Lehmann, 1987.) These mathematical estimates will be considered briefly in the remainder of this section. However, additional more technical information will be forthcoming from the Iowa Department of Education in a companion technical document. Please note that reliability evidence is inherently collected in a

mathematical way, requiring an understanding of some basic statistical and measurement concepts (Linn and Gronlund, 1995.)

Implications for a District-wide Standards-Referenced Assessment System

For the purposes of constructing a district-wide standards-referenced assessment system, the sources of potential error contributing to inaccurate judgments and interpretations of assessment results need to be identified and their effects eliminated or reduced when possible. One easy way to reduce error is to standardize the assessment opportunities where possible. For example, do not collect writing scores in one year of the program on a Monday and then collect them the second year on a Friday. You may not know if such a "day of the week" effect would really impact the resulting student performance, but why take the chance? Standardized administration procedures, as well as standardized directions for administration, are good ways to reduce the impact of potential errors on the consistency of observed measurements: hold all potentially impacting sources of error constant when possible.

An Operational Definition of Reliability

Before we visit the mathematically derived estimates of reliability, additional conceptual frameworks are necessary. These will be provided in the form of an operational or working definition of reliability. As alluded to in the previous paragraphs, reliability is the consistency of measures obtained primarily though the removal of influences by mitigating or intervening circumstances.

- *Reliability refers to the consistency of the results between two measures of the same thing.*

This consistency can be seen in the degree of agreement between two measures on two occasions. When this agreement is high, it is likely due to the lack of error modifying the individual measures. In the bowling example presented previously, the agreement would be between, say, the first game and a second. Operationally, such comparisons are the essence of the mathematically defined reliability indices. These indices provide the reliability coefficient so often cited in technical manuals from test publishers as well as general textbooks on test construction and measurement theory. Before we look at the variety of mathematical estimates of the reliability, we will explore the components of the reliability coefficient.

The Coefficient of Reliability

In the examples provided so far, emphasis on increasing reliability was based on increasing the consistency between two observations or measures by removing error components. Specifically, we are likely to get similar (consistent) measures of our weight if we hold constant (control) the time of day of the two different measures. Three things are implicit in these examples:

- Consistent measures in a controlled environment is enviable and will increase measures of reliability;
- Control of the circumstances reduces the potential for differential impact of error on the measures;
- It is impossible to identify, let alone control (eliminate), all possible influences (error) on the measure.

These characteristics of measurement, taken together, lead to a fundamental conclusion that all measures consist of an accurate or "true" part and some inaccurate or "error" component. In fact, this is the fundamental premise of classical reliability analysis as well as classical measurement theory. Stated explicitly, this relationship can be seen as the following:

$$\text{Observed Measure} = \text{True Score} + \text{Error}.$$

To facilitate a mathematical definition of reliability, these components can be rearranged to form the following ratio:

$$\textit{Reliability Index} = \frac{\textit{True Score}}{\textit{True Score} + \textit{Error}}.$$

Clearly, when there is no error the reliability index will be the true score divided by true score, which is unity. However, as more error influences our measure, the error component in the denominator of the ratio will increase thereby decreasing the reliability index below the perfect value of one. It is this type of ratio that is estimated when people discuss reliability indices associated with various measures as, say, ranging between 0.80 and 0.90.

Summary

A quick review of the points made regarding reliability in this section is warranted:

- Reliability is a general concept associated with the consistency between different measures of the same thing;
- The more we are able to reduce the error impacting the measures by controlling intervening variables or circumstances, the more consistency will be seen and an increase in reliability will result;
- Observed measures consist of a "true" part, typically referred to as the true score, and components of error.
- A reliability index can be operationally defined as the ratio of the true component of the observed measure divided by the observed part (true score plus error.)

The reader should not let the concept of a "true score" in this conceptual definition be too troublesome. As we shall see from the upcoming sections, the true score can never be known and in real practice is not needed!

Implications for a District-wide Standards-Referenced Assessment System

It is important to remember that each time we collect information from a student, the resulting "scores" are comprised of two general parts: the true part (true score) and an error component. The error component is itself made-up of many different sources. For example, students who did not sleep well the night before the assessment will probably have larger error components contributing to their scores than students who are well rested. Students who are easily distracted will probably have scores with larger error components than their less distracted peers, especially if the testing environment is full of distractions. It is important that the developer of an assessment system list these potential error components and evaluate their impact on the resulting measures. If, for example, the assessment will be collected in one large group setting, potential error could include: distractors in the environment (noise from within the room or coming from an adjacent room); distribution of the assessment (some students will get the stimulus before others); unexpected interruptions (disconnect the school bell so it will not ring during testing.) Obviously, the list will never be exhausted and is limited only by the lack of ability to anticipate each possible error. The list will serve as a starting point in controlling these confounding error components.

Classical Estimation of Reliability

Due primarily to the unknowable "true-score" component of a person's observed measure, various estimates of the reliability have been derived. Allen and Yen (1979, pg. 76) provide three general classifications of these estimates: test / retest, parallel forms, and internal consistency. So far, all examples have used the test / retest classification with the assumption that repeated measurement did not impact student's performance. As we shall see, this is not always the case and is often an assumption that may not hold in practice. Additionally, each different estimate of reliability accounts for different components of error and as such, each may lend themselves to different applications in the practical collection of reliability evidence.

Test / Retest Reliability Estimates

As Allen and Yen (1979) point out, test / retest estimates are based on the notion of an examinee taking the same measurement twice. A simple comparison of the results (usually in the form of a mathematical correlation) provides an index to the degree of agreement or consistency between the two measures. For example, a simple listing of the rank orderings from the first measure compared to a similar listing from the second measure provides an estimate of reliability. If the lists are the same (i.e., each student scored in exactly the same order on both measures) then there is perfect agreement between the measures and, conceptually, reliability would be unity (a mathematical correlation of 1.0, within the limitations of the correlation), as pointed out by Allen and Yen (1979, pg. 76.)

Unfortunately, especially with regards to academic and achievement oriented tasks, it is not likely that a student can respond to the same assessment twice without being influenced to some unknown extent by the assessment itself. The act of testing itself introduces inconsistency into a test / retest reliability estimate. Additionally, depending upon the time frame, the students might legitimately improve during the time period between testings. Hence, the second testing will be different from the first because of this improvement, but this will be depicted as unreliability. Again, as Allen and Yen (1979, pg., 77) point out, because of the influences of repeated testing and due to circumstances encountered during the time period between testings, the test / retest model

of reliability is best when used for traits which are stable across time (i.e., not related to direct instruction.)

Parallel-Forms and Alternate-Forms Reliability Estimates

The idea behind both of these reliability estimates is to correlate student performance on two different forms of the same measurement device. The same group of students could then be given both forms and the correlation between student performance on these forms would be an estimate of the reliability for either of the resulting measures. According to Allen and Yen (1979), it is not possible to verify when two versions or forms of an assessment are parallel. Strictly speaking, two forms of an assessment are parallel when they fulfill a variety of strong statistical requirements that are almost never possible in practice. A more detailed explanation of these requirements can be found in Lord and Novick (1968, pg. 37.) Alternate test forms are simply a less rigorous implementation of the requirements for parallel tests. Allen and Yen (1979, pg. 78) define alternate test forms as any test forms constructed to be parallel, but that do not achieve the equality in statistical indices required under the definition of parallel. Test publishers usually construct test forms that come the closest to fulfilling the parallel forms criteria, but they too are usually considered alternate forms.

Internal-Consistency Reliability Estimates

Internal consistency estimates are derived from scores resulting from a single test administration. According to Allen and Yen (1979, pg. 78), the most often used implementation of this method is to correlate student performance on the even items with that from the odd items. This procedure is referred to as the split-half procedure (Allen and Yen, 1979, pg. 78.) Clearly, the biggest advantage of such a procedure is that only one administration of the assessment is needed. However, as Allen and Yen (1979) point out, there are requirements regarding how the halves are assigned and not all splits are equal. Further discussion of these concepts will be provided in the companion piece providing technical information.

Implications for a District-wide Standards-Referenced Assessment System

While all of the reliability estimates presented in this section have been conceptually complex, their calculation is rather simple and straightforward. The technical companion piece to this document will examine in great detail the procedures and mathematical

manipulations involved with collecting evidence of reliability. However, this does not mean that persons implementing a district-wide standards references assessment need not pay attention to reliability until after the scores are collected. Indeed, the most useful way to increase reliability is to anticipate all sources of error which may impact the measure. With these sources identified, design your assessment such that their influences are reduced or at least held constant across all assessments. Test / retest estimates are particularly sensitive to changes in the scores between the first and second measurements. Hence, collect these reliability estimates on traits that are fairly stable such as visual or auditory acuity (Allen and Yen, 1979.) Parallel-forms or alternate-forms reliability estimates are particularly useful when more than one assessment of the same thing is administered. Both measures must be obtained from forms constructed for the same purpose and used in the same way. Attention should be paid to the interval between testings, content of the assessments, as well as testing conditions. Internal consistency estimates of reliability are perhaps the easiest to collect and are those typically reported by test publishers. Collect these estimates when only one test administration is possible.

Reliability in Generalizability Theory

Generalizability theory (G-Theory) is a conceptual extension of classical reliability theory as provided by Feldt and Brennan, 1989. Generalizability theory is an analytical procedure used to identify and quantify error components, which reduce the reliability of virtually all measures. G-Theory analyses will require trained individuals, and as such will not be elaborated upon here. However, interested readers should consider several different sources for information: Feldt and Brennan, 1989; Shavelson and Webb, 1991; and Brennan 1983. Further information regarding G-Theory will be provided in the technical companion.

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

Studies can be conducted to quantify where and which sources of error contribute the most to the unreliability of a measure via a generalizability study. For agencies with access to trained staff, these studies could include looking at error components associated with the types of item formats, numbers of essays, numbers of scoring sites (schools), numbers and types of passages, etc.

Standard Error of Measurement

One of the biggest problems with indices of reliability is that they have no inherent meaning. For example, is a reliability coefficient of 0.82 sufficient? One way to determine the meaning of 0.82 is to compare it to known quantities or "rules of thumb." For example, the Iowa Tests of Basic Skills typically provides reliability evidence (internal consistency estimates) in excess of 0.90 for all domain total scores regardless of test length (Hoover, et. al., 1996: ITBS, Grade 6, Level 12, pg. XIX.) So, compared to ITBS a coefficient of 0.82 is lower. However, such comparisons can often be misleading for several reasons, including differences in test length. Perhaps the biggest limitation to interpretation of such coefficients is their lack of application to individual student scores. If a measure has lower reliability than some other measure, it is influenced by error to a greater extent. Hence, the scores resulting from that measure are less accurate. The standard error of measurement uses the information from the test along with an estimate of reliability to make statements about the degree to which error is influencing individual scores. The standard error expresses unreliability in terms of the reported score metric. Using the standard error of measurement, an error band can be placed around an individual score indicating the degree to which error might be impacting that score.

Again, much more detail regarding how to calculate a standard error of measurement will be provided in the technical companion to this document. In the meantime, interested readers can refer to Allen and Yen (1979), Feldt and Brennan (1989), or Traub (1994).

Implications for a District-wide Standards-Referenced Assessment System

Just as considerations must be given to potential error impacting measures and the control of that error, the reporting of the unreliability of a measure is a fundamental requirement of a district-wide standards-referenced assessment system. While it may seem like the reporting of the reliability coefficient(s) from each component of the assessment system would suffice, the standard error of measurement should also be reported. In fact, this highly useful metric will allow for a better interpretation of the error inherent in any measure. Additionally, the Standards for Educational and Psychological Testing (APA, 1985, pg. 19) call for the reporting of both the reliability coefficients as well as the standard errors of measurement.

Decision Consistency Reliability Estimates

One of the main reasons a district-wide standards-based assessment system is being implemented is to make better decisions regarding the level of student competencies. From these, informed instruction will lead to improvements in student learning. What this implies is that the consistency or reliability of a measure may not be as important as the accuracy with which the measures are used to classify students into these competency categories. Simply stated, if an assessment classifies a student into a level of competency, based on some standard setting or cut-score policy, because the measure is fallible (unreliable to some degree), these classifications are going to be in error some extent of the time. For example, there will usually be a nonzero frequency of "false masters" (students classified above their actual competency) and "false non-masters" (students classified below their actual competency.) It is not a matter of indifference regarding the direction of these errors, districts should consider costs associated with both types of errors.

Like classical reliability theory and generalizability theory, there is an area of study dedicated to the understanding and quantification of errors in misclassification associated with such decisions. Again, the mathematical formulas used to estimate these errors are beyond the scope of this text and will appear in the technical companion. However, the interested reader is referred to the following sources for more detail: Traub (1994 pg. 70); Huynh (1976) and; Berk, R. A. (1984.)

Implications for a District-wide Standards-Referenced Assessment System

Because many of the measures collected as part of the district-wide assessment system will be used to generate classification decisions for individual students, it is important that some evidence regarding the accuracy of classification be collected. This could be one of the mathematical indices referred to in the previous paragraph, or simple estimates of the percentage of false-masters and false-nonmasters. Additionally, the steps taken to ensure that misclassification is minimized should be documented. Finally, the costs associated with misclassification should be described and considered when making the final classification decision.

Scorer Consistency and Inter-Rater Reliability

To this point, the discussion regarding reliability has been confined to objective multiple-choice type tasks. However, the need to collect reliability evidence for non-objectively scored tasks, such as writing essays, is also important. Clearly, the reliability of a writing essay will have the same components of error affecting the resulting scores as the multiple-choice items *plus* some degree of inconsistency added though the judgmental scoring process (i.e., assigning scores to the essays.) This potential for additional error is inherent in all scorings using a judgmental process and is not limited to writing essays.

One index of the degree of error or unreliability added to a measure from judgmental scoring could be obtained by having two judges read the same set of essays, with each assigning scores independently. Presumably, these judges would follow the same rules in determining the score (i.e., use the same scoring rubric.) The percent of agreement between these readers would be an indication of the consistency of the application of the scoring rules (rubrics) to determine the student scores. If the readers consistently assigned the same score, it is more likely that the judges are applying the scoring rules in a consistent manner thereby eliminating error and increasing reliability. Another index used to determine the degree of association between a first set of judgments and a second set of judgments is to simply calculate a mathematical correlation between these pairs of scores. Recall that this is similar to the concept of the test / retest reliability coefficient where the first set of judgements is analogous to the test and the second set of judgments is analogous to the retest. If the first set of judgments agrees, in the most part, with the second set, this estimate of reliability would be positive and large.

As was true with the classical concept of reliability on a multiple-choice assessment, a great deal of effort has been used to study the error associated with human judgments particularly in the scoring of essay responses. Generalizability theory (as referenced in a previous example) is only one of many ways to investigate the variability of judgments applied to scoring. The purpose of this section is to acquaint the reader with some of the simpler ways to investigate the degree that unstable judgments may impact the scores resulting from a measure.

Implications for a District-wide Standards-Referenced Assessment System

The first step to reducing error associated with the scoring of open-ended or essay type assessments is to define a clear understanding of the scores. For example, assigning a four to the best paper and a one to the worst paper might produce the results desired for the scoring of a classroom assessment, but it is unlikely to be useful in a district-wide assessment for several reasons.

- First, the definition of best and worst is unknown. Teachers, students, parents and the public will want to know why a particular paper got a score of four and another only got a score of three.
- Second, unless there is only one person to assign the scores to all of the papers, the definition of best and worst will probably change as the scoring continues or will be misapplied. For example, somebody else may consider what one person considers best as less worthy.
- Third, the definition of best should remain constant for future assessment. As student-writing skills improve, without some anchor to what is considered "best" this year, it is unlikely that best will mean the same thing next year.

In order to avoid these pitfalls, a scoring rubric is usually developed with clearly delineated student behavior required at each and every score point. For example, instead of stating that the best paper would earn a score of four, the rubric would say something about the writing required to earn a four: student writes in complete sentences; writing has a clear beginning, middle, and end; writing flows with no grammatical errors; writing engages the reader with several alliterations; etc. Once the rubric is generated, discussed by the appropriate content staff and agreed upon, specific examples of student writing which would earn the various score points on the rubric should be selected. These exemplars or anchor papers provide further clarification regarding just what writing is required to earn the various score points. In addition, anchor papers could be selected representing the transition points (i.e., papers "on the bubble" between, say, a score of three and a score of four.) These papers could be discussed and annotated to document why they ultimately received a particular score.

In addition to developing scoring rules or rubrics, consideration must be made regarding who will ultimately make the judgments. For example, teachers in the content area being assessed would be most familiar with the measures and may be the best judges.

However, these teachers probably have little or no experience in scoring a district-wide performance assessment. Therefore, training will have to be provided to instruct them on how to use the rubrics and assign the scores. In addition, due to the possible contamination of resulting scores from the subjectivity of judgment, an individual score might come as the sum (or average) of two readings. This would also provide an "built-in" way to obtain inter-rater reliability or to assess the agreement between readings.

Perhaps of even greater concern than the training of readers and checks for the consistency of scoring between readings is the need to outline a procedure and document the process used to obtain the measures, score the measures and return the scores and papers to those needing the information. Careful attention needs to be paid to such logistical issues as: packaging the student responses if they are moved off-site to be scored; and coding both the papers and the score forms such that the appropriate students get the credit they earned. Providing enough time to complete the scoring, taking into account when the scores are needed, will also be an administrative burden. Decisions regarding what to do when judges disagree while assigning scores will also have to be anticipated and planned for. Also, the amount of human effort required to score all student responses across the district is no small consideration. Such an endeavor will take time to plan and require resources of time and money.

B. Validity

As stated in the introduction to this chapter, assessment results must show evidence of reliability for the purpose for which they were intended before they can show evidence of validity. Hence, the concept of validity is presented after that of reliability. This does not mean that validity is less important. In fact, the Standards for Educational and Psychological Testing (here after referred to as "the standards") state that validity is .".the most important consideration in test evaluation" (APA, 1985, pg.9.) The main concern is that collecting evidence of appropriate interpretations of the results from a measurement is mainly judgmental. Reliability, for the most part, lent itself well to estimation through statistical means, though judgment was required in trying to reduce the error components influencing a measure. Validity evidence on the other hand is often judgmental. This is especially true in the area of educational achievement where questions regarding how

much content a student had learned is the paramount interpretation desired from assessment results.

In the past, distinction was made between different types of "validity." For example, the terms "content validity," "construct validity" and "criterion-referenced validity" were used. This generated more confusion than was necessary. Validity is, and always was, a "unitary concept" (APA, 1985; Linn and Gronlund, 1995, pg.49.) The way the evidence was collected to demonstrate valid use and interpretations of assessment results took on many different and specific classifications. While this distinction between the unitary concept and the different types of validity evidence offered might seem trivial, it is important to remember that all types of validity evidence should be collected when possible. This evidence should document and demonstrate that the interpretations being made from the results of the assessment are appropriate. The standards provide the following (APA, 1985, pg.9):

> "An ideal validation includes several types of evidence, which span all three traditional categories (content-related, criterion-related, construct-related.) Other things being equal, more sources of evidence are better than fewer. However, the quality of the evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable quality."

The standards go on to state that it is professional judgment that should determine which evidence should be collected and documented as evidence of valid score use and interpretation. The standards also state that resources should be investigated in obtaining the evidence that "optimally reflects the value of a test for an intended purpose" (APA, 1985, pg. 1.)

<u>Implications for a District-wide Standards-Referenced Assessment System</u>
Districts should make the distinction between reliability and validity and must collect evidence to document that the consistency of the results obtained from various measures and also the results are interpreted and used in the appropriate manner in light of the purpose of assessment.

The primary concern for many, if not all, districts will be the match between the test and the content-standards that should already be in place. Clearly it will do little good to have a very consistent (reliable) measure of well documented content when that content is not what is desired for the students to learn or what is being taught through the

curriculum. Collecting evidence of content validity will be of paramount concern for most districts. The procedures required to do this are presented in the paragraphs that follow.

Content-Related Evidence Regarding the Use of Assessment Results

As alluded to in the previous paragraph, most districts will be concerned with establishing evidence for valid use of assessment results based on judgments regarding the content being measured. For example, the content-standards define key aspects of the curriculum which are important. The assessment presumably measures the degree to which these content are learned. Inferences are made about how much content is learned based upon a student's score on the assessment. If the assessment does not measure the content-standards being taught then these inferences (i.e., the interpretations) made from the assessment results will be misleading. It is therefore very important to ensure that the assessment, be it a commercially available assessment or one developed by the district, matches the content being taught as documented in the content-standards. There are many different ways to ensure that the content of the assessment is aligned with the content-standards, but perhaps the most often used procedure relies on expert judgment (APA, 1985, pg. 10.) Presumably expert teachers in the content areas were involved in establishing the district-wide content-standards. These same teachers could serve as judges in assigning components of the assessment to the content standards. Such a match would provide evidence the assessment does indeed measure the content-standards and to what degree. A simple table showing the number of assessment components, tasks or items matching to each content-standard could serve as this documentation. Additional judgments will then be required regarding the sufficiency of coverage of the content-standards by the assessments. For example, perhaps there is little, if any, match between a content-standard regarding student writing skill and a selected measurement device. If this is true, then it is doubtful that appropriate interpretations of the assessment results can be made regarding this (or these) content standard(s.) In fact, this process might lead the district assessment team to consider additional assessment components to add to the district-wide assessment. An assessment is simply a single-point in time sample of all possible tasks that might be constructed for a given content-standard. The judges must

determine if the particular sample of items or tasks on the current assessment is a fair representation of all content possible under that content-standard.

Judgments must be made regarding the format and environment in which the student responses are collected. For example, if the content-standards state that the student is to construct multiple solutions to a problem in a particular sub-domain of mathematics but the assessment is a multiple-choice test, this should be documented in the match. It is doubtful that valid score interpretations are being made regarding student generated solutions in this example if on the assessment students do not generate their responses but rather pick them from a list.

Unlike the reliability estimates previously presented, there are no general mathematical indices to establish evidence of content validity. As Mehrens and Lehmann (1987, pg. 78) point out, careful consideration of the match between the content-standards and each test item is probably the best way; it is nonetheless subjective. While it may be possible to develop some sort of scoring rubric and generate some inter-judge agreement data, it is probably better to collect judgments from as many experts as possible and to allow for group discussion and / or consensus building. Additional steps can be taken to minimize the work involved if the assessment is a published piece. Test publishers usually provide very detailed classifications of their assessments down to the sub-objective or "content cluster" level. The district should obtain such classifications prior to matching the items to the content-standards. Additionally, publishers of assessments have been known to provide "on demand" matches, particularly for larger districts, in which they will match their test to the district content-standards.

There are procedures that can be used to collect evidence of content-related validity other than expert judgments (Mehrens and Lehmann, 1987, pg. 78.) The problem with these other procedures is that they are complicated and may require measurement expertise beyond that typically found in the district or even the area educational agency. Generalizability theory, as briefly outlined in the previous section on reliability, is one potential tool that could be used to investigate issues of content validity. For example, Mehrens and Lehmann present a context they call "content reliability" inspired by the research of Robert Ebel (1975.) They suggest that one could construct two tests from the

same pool of items (i.e., matching the same content-standards), give both tests to students and correlate the results. Correcting for the unreliability of error, the correlation between scores by these students on these two test forms would provide a "validity coefficient" per se. Again, these alternatives are possible but judgmental procedures will probably still be required.

Implications for a District-wide Standards-Referenced Assessment System

The district should first have in place content-standards before content-related validity evidence can be collected. With content-standards in place, judgments regarding the degree of match between the assessment components of the district-wide assessment system and the content-standards can be made. These judgments should be made by content experts familiar to both the district-content standards and the assessments being used. These judgments must be made in light of the purpose of the assessment as well as the format and environment of the assessment.

While content-related evidence of valid score use and interpretation will probably be the most important for each district, other types of evidence are also allowed and desired. All such validity evidence should be collected if it is relevant.

Criterion-Related Evidence Regarding the Use of Assessment Results

Criterion-related validity evidence, according the Standards (APA, 1985, pg. 11), attempts to answer the following question:

> "How accurately can criterion performance be predicted from scores on the test?"

The Standards go on to state that the key to criterion-related validity evidence is the degree of relationship between the assessment items or tasks and the outcome criterion (APA, 1985, pg. 11.) Furthermore, this relationship must be systematic and predictable.

Often, measurement experts trying to collect evidence of appropriate criterion-related score interpretations are confronted with several problems. First, the outcome criteria is determined by the district or, more to the point, by the purpose of the assessment. For example, often the degree of relationship between "end-of-course" exam results and final course grade (criterion) is disappointingly poor. However, is this the fault of the assessment results or the criterion? Secondly, if the criterion is performance on, say, the ACT Assessment (Ziomek and Svec, 1995), how does the district account for the fact that

not all examinees will necessarily participate in the criterion. Clearly, careful consideration regarding the purpose of the assessment and the definition of the criterion must be made. Mehrens and Lehmann (1987, pg. 80) state the following:

> "One of the most difficult tasks in a study of criterion-related validity is to obtain adequate criterion data. Gathering such data is often a more troublesome measurement problem than constructing the test..."
>
> "Criterion measures, like all other measures, must have certain characteristics if they are to be considered adequate. First of all, they should be relevant. A second desired characteristic of a criterion is that it be reliable."

Independent of a clearly defined criterion, we would still like to see that the results of, say, a science performance assessment would agree, for the most part, with the results of a standardized science assessment. Hence, a district could correlate scores on both assessments and provide this as criterion-related validity evidence. In other words, if the inferences about student performance based on the science performance assessment were valid we would expect them to be in general agreement with the results from other measures of science content.

People collecting criterion-related validity evidence often cite two types of evidence: concurrent and predictive. The only difference between these procedures for collecting validity evidence is when they are carried out. Typically, concurrent evidence is collected from both the assessment and the criterion at the same time. An example might be in relating the scores from a district-wide assessment to the ACT assessment. In this example results from the district-wide assessment and the ACT assessment would be collected in the same semester of the school year. Predictive evidence is usually collected at different times. For example, if the ACT assessment results were used to predict success in the first year of college, the ACT results would be obtained in the junior or senior year of high school whereas the criterion (say college grade-point average) would not be available until the following year.

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

When collecting criterion-related evidence of valid interpretations and use of the results of the assessment system, the district should be concerned not only with the reliability of the results, but must also consider the reliability of the criterion. The criterion must be selected carefully and in light of the purpose for which the assessment is designed. Many

different pieces of information regarding the relationship between the assessment and the criterion are possible, but only those that are relevant will yield evidence of valid score use. Traditionally two kinds of criterion-related validity evidence have been collected: predictive and concurrent. These types of evidence only differ regarding the timeframe from which the data was collected.

Construct-Related Evidence Regarding the Use of Assessment Results

Collecting construct-related evidence of valid score use is probably the most difficult and misunderstood types of validation evidence typically reported. This might be due to the misunderstood and often misused term "construct" itself. Simply stated (Linn and Gronlund, 1995, pg. 67):

> "A *construct* is an individual characteristic that we assume exits in order to explain some aspect of behavior."

Linn and Gronlund explain that when we infer a particular individual characteristic from the assessment results, we are generalizing or making an interpretation in terms of some construct. For example, problem solving is a construct. When we infer that students who master the mathematical reasoning portion of an assessment are "good problem-solvers" we are interpreting the results of the assessment in terms of a construct. As such, we will need to demonstrate that this is a reasonable and valid use of the results.

The Standards (APA, 1985, pg. 10) suggest that construct-related validity evidence can come from many sources:

- High inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain or construct;
- Substantial relationships between the assessment results and other measures of the same defined construct;
- Little or no relationship between the assessment results and other measures which are clearly not of the defined construct;
- Substantial relationships between different methods of measurement regarding the same defined construct;
- Relationships to non-assessment measures of the same defined construct.

Linn and Gronlund (1995, pp. 68-70) more explicitly define the process of collecting construct-related validity evidence. They say that there are three general steps in the process of construct validation:

- Identifying and describing the meaning of the construct;
- Deriving hypotheses about the performance on an assessment from the theory underlying the construct;
- Verifying the hypotheses by empirical and logical means.

Like most other validity evidence, the collection of construct-related evidence is a continuous and ongoing process. It is paramount that the assessments be constructed in light of research regarding the construct being assessed. In addition, the underlying theory regarding how the construct is defined and how it is typically measured must be well understood. Finally, Linn and Gronlund (1995, pp. 69-70) provide the following guidelines:

- Define the domain or tasks to be measured: well defined assessment specifications will aid in the understanding of the construct being measured;
- Analyze the mental process required by the assessment tasks: provide a "field test" or "pilot test" in which students describe how they answered the items. Builders of the assessment can then judge if the students are doing what they desired or if the items are evoking measures about the desired construct;
- Compare the scores from known groups of students: a simple comparison of the assessment results for a group of instructed and uninstructed students will reveal the degree that the construct is being measured.
- Compare scores before and after some learning activity: we would clearly like to see continued improvement in the construct being measured as more learning takes place;
- Correlate the scores with other measures: the results of the current assessment purportedly measuring a defined construct should correlate highly with the results from another measure of the same construct.

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

Before construct-related validity evidence can be obtained, steps must be taken to ensure the assessment measures the desired construct. This begins with research at the time the assessment is constructed. Perhaps the best place to start collecting construct-related

validity evidence is in the definition of the construct. The construct must be meaningful and clearly defined. Districts could then collect information from known groups of students (such as instructed and uninstructed groups), compare scores before and after instruction, and establish relationships to other measures of the defined construct.

Consequential Validity Evidence

The underlying goal of this document is to provide educators with tools which, when implemented, will ultimately lead to improved student learning. Improved student learning will ultimately lead to a windfall of associated benefits. It should therefore be quite common to think about how a district-wide assessment system impacts not only student learning but teaching, as well as any other unanticipated consequences. Messick (1989, pg. 20, Table 2.1) refers to such impacts as the general consequential basis of test score use and interpretation. If an assessment system has been put into place to ultimately generate good consequences, to what extent has it fulfilled its mission?

Linn and Gronlund (1995, pg. 72) point out that considerations regarding the consequences of assessment score use and interpretation are clearly evident in the move toward more authentic performance based assessments. This includes both the intended use of the assessment (i.e., a better look at actual student performance) as well as the unintended consequences (such as delayed reporting time due to the judgmental scoring process required.) Linn and Gronlund (1995, pg. 73) suggest that the consequences of an assessment be considered in light of the following:

- Do the assessment tasks address key learning objectives or content-standards? Emphasis on important and not secondary aspects of the content-standards is a desirable consequence.
- Is there reason to believe that students study harder in preparation for the assessment? Increased student motivation is a desirable consequence.
- Does the assessment artificially constrain the focus of the student's study? The narrowing of the content-standards through over-emphasis or "drill and kill" activities is an undesired consequence.
- Does the assessment encourage creative modes of expression? Students exploring new ideas is a desirable consequence.

Implications for a District-wide Standards-Referenced Assessment System

Systematically generating a list of questions like those presented by Linn and Gronlund (1995) could help in documenting the consequential aspects of an assessment. Districts should ensure that only the most relevant issues are addressed and that it is in the best interests of the students to partake and succeed on the assessment. Additional attitudinal surveys could capture other unanticipated consequences.

Summary

This chapter has defined validity as the evidence generated in support of appropriate use and interpretation of the results of an assessment. This use of the results cannot be independent from the intention of the assessment. Reliability was shown to be a necessary but not sufficient condition of validity. Reliability is to consistency as validity is to accuracy; we must measure consistently before we are able to measure accurately.

While validity is a unitary concept, four different types of validity evidence were discussed.

- First, content-related validity evidence questioned the degree to which the assessment results were interpreted appropriately regarding the content-standards and what aspects of the content-standards were to be assessed.
- Second, criterion-related validity evidence asked the question of how well results of the assessment agreed with or could be used to predict a criterion outcome measure.
- Third, the issue of construct-validity was addressed. Construct-validity evidence relates the results of the assessment to individual student characteristics and does so in a way that is clearly understood.
- Finally, consequential validity was discussed. Districts should document the consequences of the assessment, both intended and unintended.

Districts should further strive to develop assessment systems that provide for the best consequences.

While the concepts of both reliability and validity are related, this section has clearly documented the differences between the two. Reliability is a primarily mathematical concept that is empirical in nature and is a necessary condition for validation. The concept of validity, on the other hand, is primarily judgmental and a unitary concept, requiring different types of evidence.

C. Fairness

Commercial test publishers typically spend an immense about of time, effort and money to ensure that their assessments are fair for all students. The process begins even before items or tasks are developed. For example, most publishers develop a set of "item development specifications" which list the steps needed to ensure that each item and task developed is fair for all students. This list includes such things as:

- Avoid depicting members of minority groups in stereotypic settings such as women cooking or sewing or men working outside of the home;
- Avoid using pictures that may constitute stereotypic interpretations such as men playing horseshoes at a picnic while women prepare the meal;
- Avoid using reading passages which differentiate between groups based on background experiences such as taking an airplane trip.

In addition to item development specifications, publishers spend time reviewing the items internally, piloting or field testing the items, and establishing "bias" review committee meetings of minority educators to review and discuss the items. Finally, the publishers conduct extensive statistical analyses and review of these analyses to help identify any items that may be performing in an unfair way for a particular subgroup of students. Any item with the possibility of being unfair is removed from the pool of items that will ultimately comprise the test.

While the districts may not have the resources and the flexibility in item analyses that a test publisher has, there are many similar steps that can be taken to eliminate unfair items from the district-wide assessment system. For example, minority educators as well as other educators within the district probably have a better understanding about the challenges faced by minority students than anyone else, including publishers. As such, a review of the items and tasks by a committee of these educators and the documented results of their review will provide a great deal of evidence regarding the fairness of the items. All districts should implement such procedures and the burden in terms of time and effort should be very reasonable.

For districts with access to more sophisticated statistical expertise and software, a host of procedures exist for investigating differences in performance on items and tasks between

minority and majority groups. Such analyses are typically referred to as differential item functioning or DIF investigations. While the scope of the current document prohibits a detailed explanation of these statistical procedures here, the companion technical piece will provide such information in greater detail. The interested reader should review the following for more explanation in the meantime: Camilli and Shepard, 1994; Wainer and Braun, Section III: Testing Validity in Specific Sub-populations, 1988.

Districts should investigate differential impact by each population subgroup. While it may not be possible to explain the differences between groups regarding their performance, such an investigation will provide an additional check of the fairness of the assessment system. This check should be documented such that those interested could review the work and cite it as evidence that the resulting differential performance was not attributed to poorly constructed or biased test items or tasks.

Implications for a District-wide Standards-Referenced Assessment System

Districts should obtain evidence that the items and tasks comprising the district-wide assessment system are fair for all students. The majority of this evidence may already exist if the assessment component is a commercially available assessment. However, even if this data does exist, the district should collect additional data showing that the items or tasks are fair for their students. In order to do this, the district should convene panels of minority educators to review the items for cultural, gender, or race bias. Additionally the district should investigate any items or tasks that show a disproportionally high number of minority students performing poorly. Such reviews should be conducted at both the item / task level and for the entire assessment form (including directions and support materials.)

Districts with the statistical expertise may consider empirical investigations regarding differential item functioning. While these analyses are not simple, they can be implemented in light of the district-wide assessment system. Districts should be aware, however, that such statistical investigations will not substitute for panel reviews.

Districts should develop procedures to investigate differences in results by population subgroup and to document what reasons exist, if any, for these differences which cannot be attributed to unfair or biased items or tasks.

D. Establishing Performance Levels and Monitoring Progress

This chapter will provide information regarding the number of performance levels needed and the impact differing numbers of levels may have on a district's ability to show yearly student progress. In addition, the use of profile or composite score indices for combining the results from several measurement components into one statement about student performance is also discussed. The advantages and disadvantages of the procedures for monitoring student progress are provided. Implications for districts and, more specifically, for the development of a district-wide assessment system are also pointed out.

This chapter contains extensions to and sometimes overlap with work previously provided in the CD-ROM from the Iowa Department of Education entitled: Standards Development for School Improvement in Iowa. It is assumed that most if not all districts have put into place (at the very minimum) relevant content-standards. It is also hoped that some thought has been provided regarding performance standards and the standard setting process as presented in the CD-ROM. While there are sections of this document that elaborate on standard setting, particularly with respect to performance standards, a review of the CD-ROM is warranted.

Introduction / Optimal Levels of Performance

The expectations of schools regarding the number of required performance levels was clearly outlined by the Division of Early Childhood, Elementary and Secondary Education as the following (See the section of this document on The Iowa Model):

> The achievement data must also be reported for at least three desired levels of student performance.

While this statement outlines the minimum expectations, many users of a district-wide assessment system will require more than three performance levels. In order to show progress toward meeting the annual improvement goal, the number of levels selected must be sensitive to changes in the achievement level of the population. In districts with a wide range of student skill levels across the different school buildings, particularly in large districts, a fourth, fifth or even greater number of levels may be required. For example, as presented by Carlson (1996, pg. 13), many people are dividing the lowest performance level into two separate categories such that progress for the low performing

students will be better detected. In fact, Carlson goes on to point out that there is some debate about the meaning of "three levels." Three levels, if interpreted literally as "cut scores," would imply four levels of performance: the lowest level up to the first cut, the level between the first and second cuts, the level between the second and third cuts and the level above the fourth cut. However, for schools in Iowa, it is understood that three performance levels, which imply two cut scores and three different ranges of performance, will suffice. In a different setting, schools may choose to add a fourth performance level by dividing the top performance category into two sections, in order to be most sensitive to the detection of student progress between the top two performance categories. A school might consider such an option if they have the majority of their students near the upper-end of the achievement continuum. Once the number of levels is selected, then a procedure of determining how to measure progress using these levels needs to be established. This topic, among others, is addressed in the paragraphs that follow.

Implications for a District-wide Standards-Referenced Assessment System

Based on the requirements outlined in The Iowa Model, districts should select at least three performance levels to report progress meeting annual improvement goals using achievement data. However, based on the general achievement level of the population of students in the district, or for other reasons, the district may adopt additional levels of performance. Districts with a large proportion of lower achieving students may wish to add a fourth performance level which would be most sensitive to progress shown by these students. Districts with a large number of high achieving students may wish to add a fourth performance level to be sensitive to this particular group's progress.

Multiple Measures and Source of Data

Showing progress across the selected performance levels toward the annual improvement goal depends upon many things. The ability to show this progress not only depends upon the composition of the student population but on the kinds and types of data collected, as well as the judgments made about how much progress is needed. In this regard, Carlson (1996) provides an outline linking the judgments necessary about individual students and their relationship to judgments about school or district progress. For example, Carlson (1996, pg. 14) shows that individual student level judgments will be made about which

performance level a student attains. These judgments will be based on the student's performance to multiple-measures such as: standardized tests (Iowa Tests of Basic Skills / Iowa Tests of Educational Development); District-wide assessments which may include multiple-choice items, open-ended tasks, experiments, writing essays as well as other data collections; student portfolios, etc. The first task will be to make judgments about student's progress relative to the established performance levels across this wide range of different assessment components. Once these judgments have been made at the individual student level, they must be translated into statements about how well the school or district is showing progress. Typically this later requirement is determined by comparing percentages of students in each performance category and monitoring the change in these percentages across the years. As the reader will recognize, neither the first nor the second task is simple.

Implications for a District-wide Standards-Referenced Assessment System

It is clear that the district-wide assessment system will require the development of performance standards (as described in a previous chapter of this document) and the district should decide how many levels of these performance standards are required. Use as a guide the general achievement level of the population of students and select the number of performance levels, which will be most sensitive to changes in this population. The various assessments or assessment components will provide information for making these judgments provided that the manner in which progress is monitored is consistent with the data provided by these assessments.

Setting Performance Levels and Making Performance Judgments

Carlson (1996, pg.15) makes a very valid point when addressing the entire issue of measuring progress toward meeting the annual improvement goal:

> "The focus of the new Title I program is on the progress of schools, but the judgment of that progress is based on the performance of students—expressed in the form of increasing percent of students reaching the proficient and advanced performance standards."

Given that it is likely districts will have a variety of measures of individual student progress, Carlson's points are thought provoking: how can these multiple-measures (each providing a different type of information) be combined to make the judgment of progress at the school level (Carlson, 1996, pg. 16)? Unfortunately, no clear statement regarding

the best procedure to combine such information is available. However, Carlson points out some of the key characteristics of such a system (Carlson, 1996, pp. 18-19):

- The process of the combining of such data must be documented and understood by users and stakeholders. This is not necessarily a mathematical formula adding together scores from the separate assessment components, but a process for validating the combination judgments and training others to use these judgments;
- Use the same assessment components for all students. Follow the same "standardization" rules (implementation, directions, coding.) Understand the strengths and weaknesses of the data before combining;
- When combining assessment components a transformation to a common scale is most desirable for a bias free interpretation of the composite;
- Some differential weighting of the assessment components might be required in order to meet instructional or curriculum emphasis.

Two general combination procedures are available:

- forming a linear composite via a transformation and weighting;
- using patterns of the results between the different components across the monitoring period.

Carlson calls the procedures used to combine the scores "extremely straightforward and rely on simple computational methods' (Carlson, 1996, pg.19.) In fact, Carlson goes on to describe some of these procedures as explained in the paragraphs that follow.

Carlson begins by stating that such important student level decisions cannot be made on single, one point in time measures, but rather must be based on multiple measures. As such, the procedures require an understanding about the form of the data to be used in making the final judgment regarding which performance level the student will attain. Carlson (1996, pg. 86) describes four procedures that he cites as: "rules based or used to arrive at decision-rules uniformly; and are fair for all students." The specific procedures outlined by Carlson (1996, pp. 88-94) are described and elaborated upon:

Create a Weighted Composite Score
First, those desiring a weighted composite score must convert the score from each component to a "common" score scale. Perhaps the easiest way to do this is to construct a transformed standard-score (non-normalized t-score) for each assessment component (the procedure for finding the standard score will be

presented in the technical companion to this piece. Note however, as Carlson points out, almost any standard textbook on measurement theory will describe this process.) After transforming each component score to a standard metric, consider how much weight to give each part.

The following table provides an example. For this example, assume we are using an ITBS Battery score, a writing essay scored 1 to 4, and a student portfolio also scored 1 to 4 (note that these numbers have been constructed for this example and have no relationship to actual scores a student might have earned.) If the final composite score is 100 percent, perhaps it was decided to give 50 percent of the weight of the composite to the standardized testing component (ITBS), 25 percent to the writing assessment and 25 percent to the student portfolio. The resulting conversions will look like those in the following table:

Example of a Transformed Weighted Composite Score			
	Assessment Component		
	ITBS	Essay	Portfolio
Earned Raw Score	72	3	3
Transformed Standard Score	55	65	60
Weight	0.50	0.25	0.25
Weighted Component	27.5	16.25	15.0
Weighted Composite = 27.5 + 16.25 + 15.0 = 58.75 rounded to 59			

The final score of 59 will be this student's transformed weighted composite score and will be used to assess where the student falls with regards to the performance standard. Obviously, after the transformed and weighted composite is constructed, there is little, if any, meaning associated with the actual score. In order to determine the mapping of the composite score to the performance levels, careful consideration must be provided regarding what type of student behavior is required to earn various scores on the composite. In other words, the performance standards must be translated or placed on the transformed weighted composite score scale (i.e., stated in terms of the composite score.) In fact, Carlson (1996, pg. 89) suggest that "...broad participation and intensive analysis of the assessment exercises (components) and student's work" be carried out. In addition, because of the differential weighting, consideration regarding the compensatory aspects of the composite must be undertaken. For example, will a good writer be able to compensate on the composite score for poor content knowledge (i.e., low ITBS score) due to the weighting of the writing component

relative to the standardized component? Answers to these questions are most crucial in order to defend the weighted transformed composite as being fair for all students. Those choosing to construct such a composite are required to provide evidence of the validity of score interpretations resulting from it.

Creating a Weighted Composite of Separate Judgments

An alternative procedure to constructing a transformed weighted composite score (i.e., finding a single index score) is to make judgments regarding attainment of the performance standards on each assessment component first, and then combine these separate judgments with differential weighting if desired. Using an example similar to the previous one, make a judgment about the attainment of the performance standard on the standardized assessment component, followed by the writing essay and finally on the portfolio. Then, designate the lowest performance level as having a value of 1, the next a value of 2 and the highest a 3. If a student reached the highest performance level on the standardized assessment, the second level on the writing essay and the lowest level on the portfolio, this student's score would be as depicted in the table below:

Example of a Weighted Composite of Separate Standards Judgments

	Assessment Component		
	ITED	**Performance Assessment**	**Portfolio**
Earned Raw Score	301	3	1
Performance Level Attained	Level 3	Level 2	Level 1
Designated Value of Performance Level	3	2	1
Weight	0.50	0.25	0.25
Weighted Component	1.5	0.5	0.25
Weighted Composite = 1.5 + 0.5 + 0.25 = 2.25			

Presumably, the resulting composite score of 2.25 can be compared directly to the value system attached to the performance levels. Here, the value of three was attached to Level 3, while a value of two was attached to Level 2. Therefore, the weighted composite for this student of 2.25 is somewhere between a Level 2 and a Level 3.

This procedure may seem straightforward but there is a hidden cost associated with transforming or establishing the performance standards on each of the

assessment components. This is tantamount to conducting a different standard setting for each component of the assessment, which is no small task indeed as outlined in another chapter of this document.

Both of the weighted composite score procedures previously outlined have their share of potential problems and advantages. However, again as Carlson (1996, pg. 90) states, there is little experience regarding the practical implementation of either of these procedures to offer much guidance. Some of the advantages and disadvantages are listed below:

Advantages of a Weighted Composite Score:

- A standard score scale allows for investigations into which components of the assessment students performed their best as well as their worst;
- Only one standard setting is required which will map the performance levels onto the transformed but common score scale.

Disadvantages of a Weighted Composite Score:

- The construction of a standard score scale is an additional complex task required for each assessment component;
- Judgments mapping the performance standards onto the common score scale are complex and must take into account student behavior and student performance across different components;
- Attention must be paid to the type and degree of compensatory student behavior. In other words, how do students earn composite scores and how much contribution did each component add?
- Decisions must be made regarding the weighting of the various pieces;
- There is little if any inherit meaning in the values of the weighted composite scores. In fact, great efforts will need to be made to communicate this meaning to the various interested constituencies.

Advantages of a Weighted Composite of Separate Judgments:

- It may be easier to make judgments about the performance standards separately for each component of the assessment.
- Scores are essentially in the metric of the performance levels allowing a more meaningful interpretation of the composite;
- The additional computational burden of establishing a common score scale is not required.

Disadvantages of a Weighted Composite of Separate Judgments:

- Three separate judgments mapping each component to the performance in the example presented standards (i.e., three separate performance standard settings) are required;
- Attention must still be paid to the type and degree of compensatory student behavior. In other words, how do students earn composite scores and how much contribution did each component add?
- Decisions must be made regarding the weighting of the various pieces.

In addition to the two composite score procedures provided by Carlson (1996) described in the previous paragraphs, Carlson suggests two possible profile or pattern approaches as outlined in the following paragraphs.

Mapping Scores into Performance Levels

Interpreting student score profiles is not a new way to investigate student performance. In fact, many nationally norm-referenced tests, psychological assessments and personality inventories use profiles to aid score interpretation and use. Consider the following example as motivated by Carlson (1996):

Suppose there are two components associated with the district-wide assessment system. One is a writing essay scored on a 1 to 6 point scale and the other is the Iowa Tests of Educational Development. A simple way to conceptualize a profile and to map the district performance standards onto this profile is to construct a two-way matrix or box of possible scores. For example, consider the table on the following page.

By considering student performance simultaneously on both ITED and the writing essay, judgments can be made about the combination of scores to represent each of the three performance levels. In the table, a writing essay score of one, regardless of performance on ITED, was judged to represent only the lowest level of the performance standards. Similarly, a score of 4 on the writing essay in combination with being in the top third of the national ITED distribution (i.e., the top three stanines) was judged to reflect attainment of the highest performance level.

Example of Mapping Student Scores into Performance Levels (Performance Levels Appear in the Body of the Table)									
	ITED National Stanine*								
Writing Essay	1	2	3	4	5	6	7	8	9
1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
2	*1*	*1*	*1*	*1*	*1*	*1*	2	2	2
3	*1*	*1*	*1*	2	2	2	2	2	2
4	2	2	2	2	2	2	*3*	*3*	*3*
5	2	2	2	2	2	2	*3*	*3*	*3*
6	2	2	2	*3*	*3*	*3*	*3*	*3*	*3*

Note: The ranges, score points and categories have been created for illustrative purposes only and do not reflect actual student performance or expectations.

* National "standard-nine" percentile rank categories.

Next, a different type of profile or mapping is considered by Carlson (1996) in which scores on each component of the assessment are first mapped into performance standards which are then mapped into a single overall performance standard.

Mapping Individual Judgments into Performance Levels

This model, as well as the previous profile model, is analogous to the steps taken for the composite score calculation. Instead of considering the performance of the two assessments together and then making a judgment about what performance level is represented by that combination, this procedure requires that the performance standards be inferred separately for each assessment component. Once the performance standards have been established separately for each assessment component, these separate standards are then mapped into one overall or general statement regarding the attained performance standard.

Again, consider the next table inspired by Carlson (1996.) In this table, it can be seen that students have already been placed into the performance standard levels on both components of the assessment. This means that judgments regarding what students must achieve on each component (i.e., a standard setting) have already taken place. Judgments were then made, looking at the work associated with both components and necessary to attain a particular performance level on the component, regarding what combinations of separate standards on each component were required to be placed into specific overall performance standard categories. This can be seen from the entries in the body of the table. For example, it was judged that the combination of being in Level 3 on ITED while only in Level 1 on the Writing Essay would yield an overall performance standard

classification of Level 2. Similarly, being in Level 3 on the Writing Essay while earning only a Level 1 performance standard on ITED would yield an overall performance standard classification of only Level 2 and so on.

Example of Mapping Individual Judgments into Performance Levels

		Attained Performance Standard on ITED		
		Level 1	**Level 2**	**Level 3**
Attained Performance Standard on the Writing Essay	**Level 1**	1	1	2
	Level 2	1	2	3
	Level 3	2	3	3

Note: The ranges, score points and categories have been created for illustrative purposes only and do not reflect actual student performance or expectations.

Both of the procedures using a pattern or profile approach also have both advantages and disadvantages. Again, little research and / or experience regarding these procedures can be found (Carlson 1996, pg. 90.) Hence, careful consideration regarding the complexity of the tasks associated with these procedures is warranted.

Advantages of Mapping Scores into Performance Levels:

- Allows for the consideration of the different work required by each component in making the judgment regarding performance level.
- Allows for "implicit" weighting of the aspects of the assessment component deemed most important (i.e., judges will give "more credit" to success on what they see as the more important component.)
- Since the content and student performance is explicitly stated, the combined statement of performance standard can be easily communicated, and expectations regarding required performance are known.

Disadvantages of a Weighted Composite Score:

- Judges must simultaneously consider student performance on two different components and this may be a conceptually challenging task.

- Some judges may differentially weight the different components in unforeseen and inappropriate ways (i.e., a writing zealot may allow only the writing component to influence judgment regarding attainment of the performance standard.)

- Attention should be paid to the type and degree of compensatory student behavior. How will students' compensation for poor performance on one component affect their performance on the other component?

- Decisions must be made regarding the weighting of the various pieces, either implicitly or explicitly.

- The conceptual task of considering more than two assessment components quickly becomes overwhelming if not impossible to consider.

Advantages of Mapping Individual Judgments into Performance Levels:

- It may be easier to make judgments about the performance standards separately for each component of the assessment.

- Progress toward the annual improvement goal can be documented at both the component level as well as the overall level.

- Judgments about the final mapping depend upon the judges understanding of the student behavior required to attain a particular performance standard on each component of the assessment.

Disadvantages of a Weighted Composite of Separate Judgments:

- Requires an additional performance standard setting, one for each component, but also one for the overall performance.

- Attention must still be paid to the type and degree of compensatory student behavior.

- Decisions must be made regarding the weighting of the various pieces either explicitly or implicitly.

<u>Implications for a District-wide Standards-Referenced Assessment System</u>

Districts should decide how to map and track performance across the various components of the district-wide assessment system. Either a composite score can be constructed or a profile can be made. In either case, judgments regarding individual student performance on the components relative to the performance standards must be made (i.e., an empirical standard setting conducted.) Districts should then decide on different aspects of weighting of the components, again regardless of model. In the case of the composite indices, the weighting is more explicit, but implicit weighting will still take place with the

profile procedures. Attention should be paid to issues regarding student compensatory behavior. For example, will a student who scores extremely well on the standardized assessment but poorly on the writing piece score similarly to a student who performs well on the writing piece but scores poorly on the standardized assessment? It is likely the computational burden of constructing a standard or common scale will prohibit the usefulness of the composite procedures for some districts; these districts should probably rely on a profile approach. Finally, the procedures outlined in this section have little in the way of academic research or experience to support their implementation. Also, there are an infinite number of ways to arrive at a composite score or profile standard and certainly the procedures described in this section are only a trivial fraction of all the procedures that could ultimately be used.

E. Assessment System Logistics and Database Management

Many details must be considered in the actual implementation of a district-wide assessment system. While it is beyond the scope of this document to anticipate all such details, the following paragraphs reflect thoughts and current practice regarding some of the issues that need to be considered.

Full Participation for All Students

The Individuals with Disabilities Education Act (IDEA), Public Law 105-17, is very specific regarding the legal expectations of education for all students, including those with disabilities as outlined in the purpose of the law (Individuals with Disabilities Education Act Amendments of 1997, Part A-General Provisions, Section 601: Purpose, Paragraphs 1, 2, 4):

> "...to ensure that all children with disabilities have available to them a free appropriate public education that emphasizes special education and related services designed to meet their unique needs and prepare them for employment and independent living;"

the law further outlines its purpose:

> "...to ensure that educators and parents have the necessary tools to improve educational results for children with disabilities: by supporting systemic-change activities; coordinated research and personnel preparation, coordinated technical assistance, dissemination, and support, and technology development and media services"

and further:

> "...and to assess and ensure the effectiveness of efforts to educate children with disabilities."

An additional requirement of the IDEA, individualized education programs (IEPs) must be written for each student with a disability who is receiving special education or other related services. It is not within the scope of this document to describe and define completely the steps necessary to fulfill all of the requirements of the IDEA. However, an understanding of the requirements regarding student needs and the steps required to establish, monitor and maintain an IEP will help in anticipating trouble spots in implementing a district-wide assessment system suitable for use with all students. Note that this means we are suggesting the design of a single system for use by all and, as such, should any part of the system fail to lend itself to continued improvement (be it informed instruction or increased learning), the system itself will fail.

The Iowa Individualized Educational Program Guidebook (1998)

Much of the work regarding the establishment of the individualized education programs (IEP) can be found in a document supplied by the Iowa Department of Education called the Iowa IEP Guidebook (1998.) This guide, while specific to the IEP portion of the IDEA legislation, shares common ground with the goals of a district-wide standards-referenced assessment system. For example, the Educational System Goal for the State of Iowa is to improve learning, achievement and performance of all students (Iowa IEP Guidebook, pg. 3, 1998.) This goal is based upon guiding principles outlined in the guidebook. Namely, that all students will succeed, all students will reach their full potential, high expectations will be maintained for all students, education involves parents and families, and that all students have an equal opportunity to participate (Iowa IEP Guidebook, pg. 3, 1998.) It should be the goal of any district-wide assessment system to incorporate these principles.

The Iowa IEP Guidebook (1998) outlines the requirements of an IEP in the State of Iowa. Among these requirements are things also desirable for a district-wide assessment system. For example, the IEP must contain statements about the current performance level of individual students. Similarly, in order to show progress toward the annual improvement goal, gains in student performance will have to be noted by the district-wide assessment. This means that individual student level "baseline performance"

information will be collected, presumably, during the first year or two years (biennium) of the assessment program. Also required by the IEP is a statement of the measurable annual goals required of the student. Similarly, the district-wide assessment system will also provide annual progress goals, usually stated in terms of aggregated student performance. (The IEP requires students to participate in the district-wide assessment. If the team recommends that a student not participate in district-wide assessments, the team must record the reasons for non-participation.) Likewise, the goal of the district-wide assessment is to include all students, and documentation of those not participating in the various parts of the assessments should be kept. Finally, both the IEP and the district-wide assessment require measuring and reporting individual student progress.

Given that the general goals of both the IEP and a district-wide assessment are similar how might this commonality be shared in designing the district-wide assessment? This can be seen in the next few paragraphs.

Least Restrictive Environment

The goals of both the IEP and the district-wide assessment system can be summed up in this short if still misunderstood phrase: "least restrictive environment." This means that all students, those falling under the auspices of the IDEA legislation as well as other students, should be provided a learning environment in which they can do their best. (This means during their instructional training and during informal and formal assessment activities.) For example, it would be unfair to learn how to calculate mathematics problems with the classroom lights turned on, only to have them turned off during a formal assessment. While this example might sound a bit far fetched, the story of a mathematics class using calculators during instruction only to be denied them during a formal assessment is probably not. Both examples are those in which the least restrictive environment was either not present or was not allowed during some critical phase of the learning process. The Iowa IEP Guidebook goes on to list several areas that will need to be addressed in making a decision about an individual IEP. Several of these areas have relevance for the development of a district-wide assessment system and, as such, are discussed in the following sections.

Assistive Technology

An assistive technology device, as defined by the Iowa IEP guide (and as defined by the IDEA legislation) refers to any piece of equipment that is used to increase the capabilities of a child with a disability. Similarly, an assistive technology service is defined as any service which assists an individual with a disability in the selection, acquisition, or use of a device (Iowa IEP Guide, pg. 42, 1998.)

The key to designing a system for all students is to anticipate what assistive technologies might be needed and to "build those into" the assessment system where applicable. For example, if a writing collection is desired as part of the district-wide assessment consider several assistive technologies that could be included in the assessment. Perhaps dictated responses could be collected and scored. Perhaps some students would benefit from creating their responses using a computer. Others would probably need to generate their essays by supplying written responses. Whatever the method, it is incumbent upon the developer of the assessment system to find out what assistive technologies are being used during daily instruction and to provide for such use during subsequent assessment.

Limited English Proficient Students

Some school districts will have annual improvement goals that encompass many students who are limited English proficient or English language learners (LEP/ELL.) Therefore, a comprehensive district-wide assessment will need to consider these student's needs. For example, since English language acquisition does not progress at the same rate across the curriculum as other instructed skills, perhaps more informal measures will need to be taken more often. In addition, Spanish language versions (or other language versions) of formal assessments may be required. Another example may be the use of language-translated dictionaries. Again, the goal is to provide an assessment system that will offer the least restrictive environment to assess all students.

Deafness or Hard of Hearing Students

Consider the range of communications skills of all students. Providing directions to an assessment, for example, both orally and in written form may help many students understand the directions. However, some students who have difficulty reading may also have difficulty hearing. For these students, additional steps must be taken to provide for the least restrictive environment. Again, find out what procedures are used during

instruction. Do the students read lips? Is signing used? Anticipate the need to incorporate such activities into the assessment. The best assessment system is one that provides accurate feedback while still being instructionally relevant.

Blind or Visually Impaired Students

Because the range of visual problems is so great, it is very challenging to provide instruction for students with visual disabilities. In addition, the reading of Braille documents is no small feat for the blind student, and one that takes considerable time and effort, especially during assessment. In designing a district-wide assessment system, consider the range of visual impairments. Take into account such things as type size for all students, not just the visually impaired. Be ready to implement Braille and large-print versions of assessments where appropriate. This means that sufficient lead-time will need to be given to the development of such documents, as well as for their scoring. Consider alternatives, for example, collect oral or dictated responses from the students' reading or writing assessments. Consider the collection of oral portfolios in the relevant subject areas. Again, consider how these visually impaired students are taught, as well as how the other students are taught, and find the assessment strategy that is least restrictive to the entire group.

Accommodations

In light of maintaining the least restrictive environment both instructionally and for an assessment, remember that an accommodation is a way of making the assessment fair for all students, and one that gives all students an equal chance to do their best.

The Iowa IEP Guide (1998, pg. 98) states:

> "...The purpose of an accommodation is to help compensate for the student's disability and attempt to level the playing field. The intent of the accommodation is to address a specific need, and not simply to provide the student an opportunity to score better."

Some of the best assessments already have some of the accommodations built right in. For example, if additional student time is the most often used accommodation in the classroom and it is the accommodation used most often for assessment, why not use it for all students? Why not provide an untimed component of the assessment for all students, thereby eliminating the need to accommodate students?

The Iowa IEP Guide (1998, pg. 98) goes on to point out the following questions used during development of the IEP. While these questions are important for the IEP team, they are also important to consider in designing a district-wide assessment system:

- Is the accommodation typically used in the classroom?
- Does the accommodation address a specific need of the student?
- Does the use of the assessment provide a better picture of what the student knows and can do?

Not all accommodations will meet these needs as outlined by the Iowa IEP Guide (1998, pg. 98):

> "An example of an accommodation that would not pass these questions is oral reading of a reading comprehension test for a student with a reading disability. This accommodation is not typically used in the classroom and does not give a better picture of the student's skill in understanding and comprehending written words. Although the accommodation addressees the student's reading disability, the use of other accommodations such as extended time might more appropriately address the student's needs."

Reporting of Results

The goal of a district-wide standards-referenced assessment system is to enhance student learning. Instruction informed by performance information will lead to increased student learning. However, there are ultimately many users of the assessment data, each with a specific need. For example: Title I regulations require the reporting of progress toward the annual goal across each of the performance standard categories; teachers will want to see what students wrote on an essay regardless of the score received; principals may want to know how well their school ranked in comparison to other schools in the state; and finally, parents may want to know if their son or daughter is ready to take a college level course. These different needs of the users of assessment results will drive the selection, design, format and dissemination of results.

<u>Different Reporting for Different Needs</u>

The need to consider the variety of users of the data when considering reporting assessment results is not new. In fact, the reporting of results has a critical impact on the usefulness of the scores and as such, impacts the validity of how the scores will be used (Linn, 1988, pg. 5.) Test publishers have been cognizant that assessments will be used

for many purposes, however, as Frisbie (1991, pg. 306) points out, all achievement tests are tools for instruction and their results are to show attainment of the goals of instruction. Frisbie (1991, pg. 306) goes on to point out that scores resulting from an assessment are not a "be-all / end-all" in themselves, but rather are to be used to support teacher judgments regarding instruction. The audience, who will use these results, must be taken into account before the results are disseminated.

Perhaps the variety of needs for assessment results is why such a variety of reports exist for current assessments. For example, ordering score reports for a publisher's test may be one of the most complicated functions associated with testing. As Frisbie (1991, pg. 309) points out, the variety of score reports and services offered by many publishers is so great that schools have difficulty choosing the ones that best fit their needs. This wide variety includes not only individual student reports, but reports to the parent, item-analysis reports, as well as aggregated and disaggregated results. Typically aggregated results include a classroom, school, district and statewide summary. Typical disaggregations include summaries for Title I students only, summaries for LEP students only, summaries for Special Education Students, summaries by gender, and summaries by ethnicity. In addition to the burden of determining which pieces of information to report to whom and in what format, each potential user of the results may also have different aggregation / disaggregation requirements. The following paragraphs should help the reader determine which data should be reported to whom, thereby making judgments about the ultimate reporting format applicable for these groups.

Who are the Users of the Results?

Given your district's annual improvement goals, your locally constructed content standards and the types of assessments either constructed or selected to measure attainment of these goals, who will be interested in the results? Who will care about your ability to meet your annual improvement goals? One simple answer is the Iowa Department of Education as outlined in Legislative Code 280.12 and 280.18 as well as House File 2272 (See the section on the Iowa Model presented previously in this document):

- By September 15, 1998 each accredited nonpublic school or school district must report for reading and mathematics their content standards, achievement

data for all students, subgroup achievement data for race and gender, and their annual improvement goals;

- The achievement data must be reported for at least three grade levels (3-5, 6-9, 10-12), though House File 2272 will require data for grades 4, 8 and 11;
- The achievement data must also be reported for at least three desired levels of student performance (i.e., for each performance standard category).

So, the state will require reporting of achievement data in at least reading and mathematics, for all students and by gender and race, for three grades at each performance level. In fact, this might be satisfactory information for reporting to the public. The newspapers might find this information the most suitable to print. While this provides some direction, the description of achievement data is unknown. Some districts will decide to report information similar to that described by the Achievement Levels Report from Iowa Testing Programs (ITP, 1998.) Others might construct a specific document to meet the needs outlined in Iowa legislation. Still others might choose different options in order to fulfill these requirements. This is only one use of the assessment data, and it is unlikely that reporting this data will fulfill the needs of the other constituencies. Teachers and students, for example, will want direct feedback regarding performance on tasks specific to the district-wide assessment. Students and parents will probably want this information individually, while teachers will probably desire both a roster of individual performance and also some sort of summary aggregation (i.e., a classroom summary) where appropriate. From such a summary teachers can identify areas needed for further instructional emphasis, areas needing new instructional strategies, or areas requiring a shifting of current instructional practice; in other words, assessment data tied to content standards. On the other hand, school administrators will probably have no need for such specific information as student performance at the task level, but could very well be interested in aggregations of student performance at a broader level such as a school subject or domain (e.g., reading, writing, mathematics, etc.) This need would dictate another level of reporting of the results, including how to aggregate and disaggregate scores. Finally, the Title I reauthorization requires additional reporting breakouts including LEP, migrant status, students with disabilities and economically disadvantaged students. To the extent that each of these

groups also represent constituencies within the local district, their needs must also be taken into account during reporting.

Disaggregation

Once the users of the results of the assessment system have been identified, consideration regarding the design, formatting and mode of presentation can be considered. As Carlson (1996, pg. 67) points out:

> "The obligation on states and LEAs is to monitor the progress of all groups separately; that is, to ensure that the progress of some groups, and therefore the overall average, does not mask the lack of progress of other groups. This can be done in a various ways. One of the most straight-forward methods is to simply display the percentage of students who meet the goal for each subgroup."

For example, if the annual improvement goal is to move 2 percent of all the students from the lowest level of the performance standard (say Low Performance) to the middle performance standard (say Intermediate Performance) each year, then a simple solution is to report such movement for all students and for students belonging to each breakout group.

Another method suggested by Carlson (1996) is to look at growth of each breakout group separately in light of the overall goal and require an appropriate improvement for each breakout group (See for example, Carlson, 1996, Figure V-1, pg. 68.)

The problems inherent in reporting any assessment results are still an issue when reporting the results for a district-wide assessment. First, confidentiality of individual students as well as teachers should be maintained, especially when using disaggregation. Remember that students belong to multiple classifications or breakout groups and as the level of disaggregation becomes large, the number of students falling into any one category could become very small. One rule of thumb is not to report summary data (including percentages) on less than 10 students. Secondly, such small groups are simply not very stable. For example, two LEP students in grade 4 this year will not very likely represent the ten LEP students who were in grade 4 last year. Hence, impact on measures of annual progress will be large regardless of the reporting metric. Unless a more rigorous longitudinal design is used for tracking annual progress for these students, there is little that can be done regarding this instability. Finally, depending upon the number of

measures used in the district-wide assessment-system and if a composite or profile index will be used to measure progress (See the section on Establishing Performance Levels), issues regard the breadth of coverage of the various assessments will have to be addressed. For example, it could be that such group comparisons as those presented previously would have to be provided for each type of assessment used in the district-wide assessment system.

F. Publisher's Material

One of the advantages of selecting a commercially available assessment as one component of the district-wide assessment system, provided it matches the content-standards, is the host of supporting material typically associated with such assessments. The district will have a wide range of score reports to chose from. These reports typically provide different results from the assessment for different intended audiences. For example, a confidential student level report may provide the information a student needs to know about where his or her strengths and weaknesses are. A report to the parent may provide a general profile of how well the student is performing and not go into the diagnostic detail the student requires. The teacher may receive an item analysis report indicating which items a student selected, which items were correct, and in the case of some reports what likely errors caused the student to select the particular incorrect response option.

In addition to reports, test publisher's provide a wealth of documentation outlining the technical characteristics of the assessment. Tables and tables of reliability, validity, and fairness data are often provided. Granted that these data are only relevant for the purposes for which the authors state the test should be used, but they are nonetheless very powerful documents of the quality of the assessment.

Publishers typically offer many support services. For example, such things as interpretive guides which will help the teachers use the results of the assessment to inform instruction are available. Related performance assessments as well as, in some cases, actual curriculum related instructional topics or modules can be acquired, linking the assessment, results and instruction all together. At the very least, test publisher's provide tables of specifications outlining what the test purports to measure. This, along

with careful review by the district, could provide the documentation of the match to the content standards.

When a district is considering adding a commercially available assessment as one component of the district-wide assessment, it may want to ask the following questions:

- For what purpose was the test designed, and how well does that match with the district's purpose?
- What is the match between content as assessed on the test with that outlined in the district content-standards?
- What evidence of reliability, validity, and fairness is provided and is it relevant for the purposes the assessment will be used in the district?
- Is the publisher willing to help with the match to the content-standards and will it require additional costs?
- What are the available score reports and how will they meet the needs of the multiple stakeholders?
- How easily can the results from this assessment (one component of the system) be merged with results from the other components if required?
- What additional support materials are offered by the publisher and at what costs?
- Can the assessment be scored locally and if so, what support will the publisher supply and at what cost?
- How old is the assessment and when will new content be available?
- What are the required administration times if any, and what are the logistics in administration (testing time, number of sittings, personnel requirements, etc.)?
- Is "out-of-level" testing allowed or even appropriate given the purpose of the assessment?
- Is the test appropriate for all students? What accommodations, modifications, alternative forms are available?
- Is the test available in different languages and if so are the resulting scores equivalent to the English version?

VII. Use of Assessment Data

This chapter outlines how the data generated from a district-wide assessment system may be used to build a capacity at the LEA and AEA level to support continued improvement in education. Different uses of assessment data to inform instruction, which in turn should lead to enhanced student performance will be provided. It must be emphasized that this is not going to be an "overnight" phenomenon. Improvements in education will come as more and more experience is gained via the district-wide assessment system and as districts strive to fulfill their annual improvement goals.

A. Enhanced Student Learning is the Goal

So much of this document has addressed legislative mandates, federal rules under Title I, the psychometric requirements of assessment data, etc., that we must not loose site of the purposes of the assessment system. First and foremost must be our responsibility to the students impacted by the assessment. Perhaps the most important result from a successful implementation of a district-wide standards-referenced assessment system is the following:

> Results from the assessment system will be used by teachers to inform instruction. Informed instruction in turn will lead to enhanced student performance and an improved educational experience for everyone.

Arguably there are many different purposes behind implementing such a system and many different stakeholders each with a different need to be fulfilled from the assessment data. Be this as it may, failure to use the information from this system to improve student performance is a statement of failure about the system itself.

Teachers have requested different types of feedback from assessments for quite some time. This is evident in the wide range of types and styles of score reports seen from assessments across the nation. Teachers typically use these reports not to communicate to parents student strengths and weaknesses or to tell the students where they must focus their attention, but rather to show where the teacher should look to change instructional strategies. Based on these reports, teachers may try different motivational techniques, a different focus on basic or advanced skills, or generally modify their role in light of the interaction with a particular student. Despite the perceived complexity of a district-wide assessment, there are many aspects of the system which will allow teachers to use the

assessment data in precisely this way. For example, the multiple components of the assessment system will allow the teacher a "many faceted" view of student performance. Depending upon which components a district selects or constructs for the assessment system, this could provide the teacher with multiple measures of student behavior collected in similar ways, or it could provided a "multiple-trait, multiple-method" picture of student behavior. For example, data from two components of the assessment system, say a standardized norm-referenced assessment and a standardized criterion-referenced Algebra assessment at grade 11, provide two similar measures of some mathematics skill. These measures are both objectively collected, one is a general measure and another is more specific, but both are nonetheless measures of mathematics presumably in line with the district-content standards. Teachers can use the information provided by both of these components to help inform them about what works best with a particular student in the district. However, the results available from these two components provides additional information. The teacher may want to know, for example, why a student scored particularly well on the problem-solving portion of the algebra assessment but not so well on the problem-solving section of the norm-referenced component. It could be that the student was more engaged when answering the algebra assessment questions, it could also be that the student had a "bad day" when responding to the norm-referenced component, or it could be that the algebra assessment provided more structure to support the student's problem solving skills. Obviously, the multiple measures from this example have provided many more things the teacher needs to consider when interpreting and using the results from the assessment system to tailor instruction, than for either assessment alone.

An alternative to these previous examples can be seen in the following. Suppose that the district has selected a norm-referenced mathematics assessment as one component of the assessment system. Suppose further that, due to an understanding of the content-standards, a mathematics performance assessment was selected as another component. Presumably there is something inherent in the content-standards requiring the student-generated work necessary for most performance assessments in mathematics. Here the teacher will need to consider not only the profile in performance between the two measures of mathematics, but also the type of performance. Perhaps a student does

particularly well on the problem solving aspect of the norm-referenced assessment, but does quite poorly on the same portion of the performance assessment. This information may, or may not, be indicative of a problem in generating responses as opposed to selecting alternatives. Maybe the student does not have the organizational skills necessary to complete the performance assessment tasks. Regardless, the teacher has additional information from the assessment results that he or she can use to tailor instruction. This additional information is the comparison between the different components across the different "methods" of data collection. A multiple measures based system of assessment will provide such a wealth of data for the teacher to use to inform instruction.

B. Continuous Workshops

The multiple-measures used as components of a district-wide assessment system provide far more information than the single point-in-time "snap shot" of student performance which is typical of "on demand" assessments. In fact, a well-designed assessment system could intentionally embed assessments into the instructional process through out the year, offering teachers many different data collection and interpretation points. Such a plan would allow teachers the opportunity to tailor their instruction, not only toward the needs of individual students, but toward the needs of the entire class on an almost a continuous basis. For example, a district may choose to implement a series of nine-week "end-of-section" assessments. These assessments could be one component of the district-wide assessment system but one that could be used by the teachers to modify and adapt the instructional process based on very recent and direct feedback from the assessments. Additionally, these assessments could document student progress throughout the school year.

Teachers could be involved with the scoring of these assessments at the local level. This would reduce both the cost of the assessment and increase the input teachers would have on the assessment process. Furthermore, teachers would be trained in the scoring of the assessments providing another opportunity for professional development. While such a scenario has potential, there are still the requirements of the district-wide assessment regarding issues of reliability, validity and fairness to name a few. This means that the district will have to take the appropriate steps to ensure that teachers are trained and

ready to score the assessments locally, that evidence for reliability, validity and fairness are collected at the local level, and that test security, logistic and reporting requirements are fulfilled.

Data from the district-wide assessment system, regardless of when the components of the assessment are administered, can be used in workshops conducted to show teachers the many different ways the data could be used to inform instruction. In fact, such workshops provide the opportunity for teachers to share their experiences regarding what instructional interventions work or don't work based upon data from the assessment system.

C. Capacity Building

It should be clear from this chapter (if not the other sections of this document) that implementing a district-wide assessment system is going to be a lot of work. In addition, collecting evidence of reliability, validity and fairness will require expertise that may not be evident in the district currently. It is necessary, even for districts with trained personnel on staff, that the capacity of these people be expanded through professional development to better enable them to help implement a district-wide assessment system.

Collaboration between those involved implementing such assessment systems needs to begin immediately in order to avoid duplication of effort and reduce redundant mistakes via this capacity building. For example, LEAs will probably want to work with AEAs to discover what other districts are doing and to learn from the experiences of the collective whole. Many districts are probably going to struggle with the same issues regarding the use of assessment data. Ideally, a central "clearinghouse" that could document what various districts are doing and share this information with other districts would reduce a lot of the potential duplication of effort as well as errors. Local capacity would grow as more and more shared information was disseminated to the district. This shared information would reduce the learning curve, increase resource capacity and facilitate a quicker implementation of the assessment systems locally.

Capacity building, particularly with regards to the use of assessment data, should be continuous. As districts learn which types of data and which ways these data are used to inform instruction are best, improvements in student performance will continue. As

improvements are realized, additional and new uses of the assessment data will be discovered, which in turn will build capacity even further.

VIII. Future Directions

This document, Implementing a District-wide Standards-Referenced Assessment System, is really the second document or informational piece provided to the schools in Iowa as part of an *Assessment Initiative*. The first component of this series was the CD-ROM provided by the Iowa Department of Education entitled: Standards Development for School Improvement in Iowa. This media provided an overview of school improvement plans, standards and standard setting, as well as information regarding the implementation of the standards through modifications to the curriculum, assessments, staff development and reporting.

The current document is the second part of the *Assessment Initiative* as outlined in a letter to District Superintendents on July 29, 1998, from Nina Carran, Chief, Bureau of Instructional Services. The goal of this initiative (and consequently, this document) is to provide support in developing a district-wide standards-referenced assessment system and to design a process that focuses on Area Educational Agency (AEA) and Local Educational Agency (LEA) collaboration. The outcome of this initiative is to build statewide capacity at the area and local levels to meet the challenge of assessing and reporting through statewide training.

This initiative calls for the construction and dissemination of two additional support documents: the Technical Manual for Developing District-wide Assessment Systems, and the Consumer Assessment Brochure. The technical document will go into far more detail regarding such issues as reliability, validity, test fairness, monitoring student progress, as well as other issues. The technical document will focus primarily on the psychometric issues involved in developing a district-wide assessment system, and is intended as a resource for both LEA and AEA assessment specialists. The consumer's document will be designed to help stakeholders understand the strengths and weaknesses surrounding the use of assessment data.

In addition to these documents, workshops and a series of seminars will be provided to bring support of a more technical nature to the assessment specialists within the AEAs and selected LEAs. The dates and locations of these seminars have yet to be determined, but will be provided in the near future.

The following figure provides an overview of the implementation of Iowa's District-wide Standards-Referenced Assessment System. This figure explains the relationships between the three component pieces: Implementing a District-wide Standards-Referenced Assessment System; Technical Manual for Developing a District-wide Assessment System; and Consumer Assessment Literacy Document. During the 1998-1999 school year, assistance will be provided through the Connecting School Improvement Institutes and a series of special seminars of a more technical nature. Additional presentations, conferences and workshops will be provided. Finally, during the summer of 1999, seminars will be provided specific to data analysis and utilization which will augment the skills required to use data to improve instruction.

Implementing Iowa's District-wide Standards-Referenced Assessment System

Purpose

The DSRAS Initiative will help districts develop coordinated assessment systems that are aligned with content standards at the LEA level in support of the LEA Continuous Improvement Accreditation Process.

Activity	AEA/LEA Capacity Building for DSRAS	AEA/LEA Assessment Specialist Capacity Building for DSRAS	Consumer Assessment Capacity Building for DSRAS
Document	The **DSRAS document** serves as an anchor for professional development activities and provide a self-assessment to guide LEAs. This document covers topics such as the purpose of the assessment, designing coordinated assessment systems, and reporting results to multiple stakeholders	The **Technical Manual for Developing District-wide Assessment Systems** will focus on the psychometric components of assessment systems. The document will contain formulas and specific procedures to computer test characteristics such as reliability and convergent validity coefficients. This document will serve as a resource for LEA and AEA assessment specialists.	The **Consumer Assessment Literacy Document** is intended to help multiple stakeholders understand the strengths and limitations of large-scale assessment data. The document will also outline the reasons for having a reliable and valid assessment system.
Assistance 1998-1999	The **Connecting Schools Improvement Institutes** will focus on developing AEA and LEA assessment competency using the DSRAS document as a springboard for professional development activities.	**Seminars** designated to facilitate technical assessment skill development will be offered for AEA and LEA staff. These seminars will use the Technical Manual for Developing DSRAS as the primary source for training and professional development activities.	Presentations at conferences, meetings, workshops, etc., will be offered to various stakeholder groups. The Consumer Assessment Literacy document will be the primary source for these activities.
Technical 1999-2000	**Data Analysis and Utilization seminars** will be offered in the summer of 1999. The intent of these seminars is to help develop the skills necessary to use large-scale assessment data to improve instruction and system supports.		

IX. Glossary of Terms

The following glossary of terms as used in this document is provided to assist the reader regarding language that may not be familiar. Where possible, terms were taken from existing documentation from the Iowa Department of Education.

Accommodations

Supports or services provided to help a student access the general curriculum and validly demonstrate learning.

Achievement Levels Report

A report for monitoring the achievement of student grade groups, both at the building level and system wide, and for the reporting of the progress of those groups to others. This report is provided by the Iowa Testing Programs and is further described in the document: Interpretive Supplement for the Achievement Levels Report (1997-1998) Revision.

Adequate Yearly Progress (AYP)

The measure set by each district to assess performance of schools and the district. It is expected that each measure will be set sufficient to achieve the goal of all children regarding proficiency.

Alternate Forms

Two tests constructed to measure the same thing from the same table of specifications and to the same required psychometric and statistical properties. These forms are not strictly parallel due, primarily, to differences in the statistical properties of the two forms.

Anchor Papers

Examples of student generated work which demonstrate selected points on the scoring rubric. Typically used to assist in judgmental scoring in defining what student behavior is required to earn various scores.

Annual Improvement Goals

Goals which describe the district's desired rate of improvement for students.

Assistive Technology

Any item, piece of equipment or product system, whether acquired commercially or off the shelf, modified, or customized, that is used to increase, maintain, or improve the functional capabilities of a child with a disability.

Baseline Performance

To measure baseline performance, a point in time is selected from which one can monitor changes or improvement in student performance.

Benchmarks (Major Milestones)

Major milestones which specify skill or performance levels a student needs to accomplish.

Biennium

A two-year data collection period. Instead of defining adequate yearly progress from a baseline established on one year of data, a district could calculate a biennium (two years).

Compensatory

Typically describes the ability of a student to compensate for deficiencies in one area by relying on strengths in another. Compensatory student behavior should be studied before establishing a composite index for use in performance-standards.

Composite Index / Composite Score

Typically a linear summation of various assessment components to form a total. These components may or may not be differentially weighted before they are combined.

Content-Standards

Content-standards describe the goals for individual student achievement. Content standards specify what students should know and be able to do in identified disciplines or subject areas.

Consequential Validity

Evidence that the implemented assessment or assessment system results in the planned and desired consequences and that unanticipated consequences do not detract from the goal of the assessment.

Construct Validity

Evidence that performance on the assessment tasks and the individual student behavior that is inferred from the assessment shows strong agreement, and that this agreement is not attributable to other aspects of the individual or assessment.

Differential Item Functioning (DIF)

A term applied to investigations of test fairness. Explicitly defined as difference in performance on an item or task between a designated minority and majority group, usually after controlling for differences in-group achievement or ability level.

District-wide Assessment

A large-scale, academic achievement assessment.

English Language Learner

See Limited English Proficiency

Examinee-Centered Standard Setting Methods

Process used to establish performance-standards which is based upon actual student performance to the test items. A general classification of judgmental standard setting procedures.

Formative Evaluation

Evaluation used to guide instruction, to inform teachers, students and parents about the needs of an individual regarding specific skills. Evaluation which results in a specific action plan.

General Curriculum

A description of the content-standards and benchmarks adopted by an LEA or schools within an LEA that applies to all children. It is the basis of planning instruction for all students.

Generalizability Theory

A procedure for the study and classification of the components of error.

Individual Education Program (IEP)

Individual education plan which must be written for each student with a disability who is receiving special education or other related services.

Inter-Judge Agreement

Consistency statistics describing the relationship or degree of agreement between two judges scoring an open-ended assessment.

Inter-rater (Inter-reader) Reliability

Consistency statistics describing the relationship between scores on an open-ended assessment assigned by more that one judge. Typically these statistics are a simple correlation between judges, but other more sophisticate estimates are possible.

Inter-Judge Consistency

See inter-rater reliability and inter-judge agreement.

Improving America's Schools Act (IASA)

The 1994 federal law linking Title I accountability requirements with state education reform. This law dictates a "standards referenced" model of assessment requiring "adequate yearly progress" in at least reading and mathematics for at least three levels of proficiency.

Individual's with Disabilities Education Act (IDEA)

Public Law 105-17 sets the legal expectations of education for all students, including those with disabilities. The law states that all students with disabilities have available a free and appropriate public education which emphasizes special education and related services designed to meet their unique needs and prepare them for employment and independent living.

Internal Consistency Reliability Estimate

A statistic, which represents the correlation between scores, obtained from one measure when compared scores obtained from the same measure on another occasion. Typically this estimate comes as a correlation between different halves of the same test (split-half) method, thereby requiring only one test administration.

Iowa Model

An agreement reached between the Iowa Department of Education and the federal government that meets the intent of the requirements inherent in the IASA.

Least Restrictive Environment

To the maximum extent appropriate, children with disabilities are educated with children who are not disabled.

Limited English Proficiency (LEP)

An individual's primary language is something other than English.

Modifications

Changes made to the content and performance expectations for students.

On Demand Assessment

Typically a standardized assessment designed to begin at one specific point in time and to end shortly thereafter. Assessment that is not embedded into the instructional process.

Parallel Forms

Two tests constructed to measure the same thing from the same table of specifications with the same psychometric and statistical properties. True parallel test forms are not likely to ever be found. Most attempts to construct parallel forms result in alternate test forms.

Performance-Standards

Performance standards describe how good is good enough and describe at least three levels of student performance. The federal Elementary and Secondary Education Act (ESEA) requires that at least three levels of performance be established to assist in determining which students have or have not achieved a satisfactory or proficient level of performance. Districts may decide to provide more than three performance levels.

Portfolios

Collections of student work, usually student generated which may include many different modes and types of student performance and are usually considered some of the student's best work. Information in a portfolio could be collected over the course of a few weeks or the entire school year.

Reliability Coefficient

A mathematical index of consistency of results between two measures expressed as a ratio of true-score to observed-score. The as reliability increases, this coefficient approaches unity.

Standard

A clear statement that expresses what students are expected to know and be able to do. In Iowa, local school districts and communities are responsible for setting high quality standards.

Standard Error of Measurement

Statistic which expresses the unreliability of a particular measure in terms of the reporting metric. Often used to place score-bands or error-bands around individual student scores.

Summative Evaluation

Evaluation that results in a general statement or summary of the current skill level of the student.

Test / Retest Reliability Estimate

A statistic which represents the correlation between scores obtained from one measure when compared scores obtained from the same measure on another occasion.

Test-Centered Standard Setting Methods

Type of process used to establish performance-standards that focus on the content of the test itself. A more general classification of some judgmental standard setting procedures.

True Score

That piece of an observed student score that is not influence by error of measurement. The true-score is used for convenience in explaining the concept of reliability and is unknown in actual assessments.

Validity

A psychometric concept associated with the use of assessment results and the appropriateness or soundness of the interpretations regarding those results.

X. References

Allen, M. J., & Yen, W. M., (1979.) *Introduction to Measurement Theory.* Monterey, CA: Brooks/Cole Publishing.

American Psychological Association, (1985.) *Standards for educational and psychological testing.* Washington, D. C.: American Psychological Corporation.

Berk, R. A., (1984.) *A guide to criterion-referenced test construction* (Ed.) Baltimore: Johns Hopkins University Press.

Berk, R. A., (1996.) Standard setting: The next generation (where few psychometricians have gone before.) *Applied Measurement in Education, 9* (3), 215-235.

Brennan, R. L., (1983.) *Elements of generalizability theory.* Iowa City, IA: American College Testing Programs.

Camilli, G., & Shepard, L. A., (1994.) *Methods for identifying biased test items (Vol. 4.)* Thousand Oaks, CA: SAGE Publications.

Carlson, D., (1996.) *Adequate yearly progress in Title I of the improving america's schools act of 1994.* Council of Chief State School Officers: State Collaborative on Assessment and Student Standards. Third in a Series of Papers on the Promises and Challenges of IASA Title I. CCSSO, Washington, D. C.

Deming, W. E., (1986.) *Out of the crisis.* Cambridge, MA: Massachusetts Institute of Technology.

Ebel, R. L., (1983.) The practical validation of tests of ability. *Educational Measurement: Issues and Practices, 2,* 7-10.

Ebel, R. L., (1975.) *Prediction? Validation? Construct Validity*? Mimeograph.

Feldt, L. S., & Brennan, R. L., (1989.) Reliability. In R. L. Linn (Ed.), *Educational Measurement*, 3rd Edition. Washington D. C.: American Council on Education.

Frisbie, D. A., (1991.) *Essentials of educational measurement (5th ed.)* Prentice Hall: Englewood Cliffs, NJ.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Bray, G. B., Lewis, J. C., & Qualls, A. L., (1996.) *Norms and conversion tables with technical information, ITBS Form M Complete Battery, Levels 5-14.* Chicago: Riverside Publishing.

Huynh, H., (1976.) On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 254-264.

Improving America's Schools Act of 1994 (Public Law 103-382.) 103rd Congress. (1994.)

Individuals with Disabilities Education Act Amendments of 1997 (Public Law 105-17.) 105th Congress, (1997.)

Iowa Testing Programs (ITP 1997-1998 Revision.) *Interpretive supplement for the achievement levels report.* The University of Iowa, Iowa City.

Jaeger, R. M., (1989.) Certification of student competence. In R. L. Linn (Ed.), *Educational measurement*, (3rd Edition, pp.485-514.) New York: American Council on Education / Macmillan.

Kane, M., (1995.) Examinee-centered vs. Task-centered standard setting. In the *Proceedings of joint conference on standard setting for large scale assessments*, Washington DC: National Assessment Governing Board and the National Center for Education Statistics.

Linn, R. L., (1989.) Current Perspectives and Future Directions. In R. L. Linn (ed.), *Educational measurement* (3rd ed.) Washington, DC: American Council on Education.

Linn, R. L., & Gronlund, N. E., (1995.) *Measurement in assessment and teaching*, 7th edition. New Jersey: Prentice-Hill.

Mehrens, W. A., & Lehmann, I. J., (1987.) *Using standardized tests in education*, 4th Edition. New York: Longman.

Messick, S., (1989.) Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* Washington D.C.: American Council on Education.

Roeber, E. D., (1996.) *Designing coordinated assessment systems for Title I of the Improving america's schools act of 1994*. Council of Chief State School Officers: State Collaborative on Assessment and Student Standards. Fourth in a Series of Papers on the Promises and Challenges of IASA Title I. CCSSO, Washington, D. C.

Shavelson, R. J., & Webb, N. M., (1991.) *Generalizability theory: A primer*. Newbury Park, NJ: SAGE Publications.

Traub, R. E., (1994.) *Reliability for the social sciences: Theory and application.* Thousand Oaks, CA: SAGE Publications.

Wainer, H., & Braun, H. I., (1988.) *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

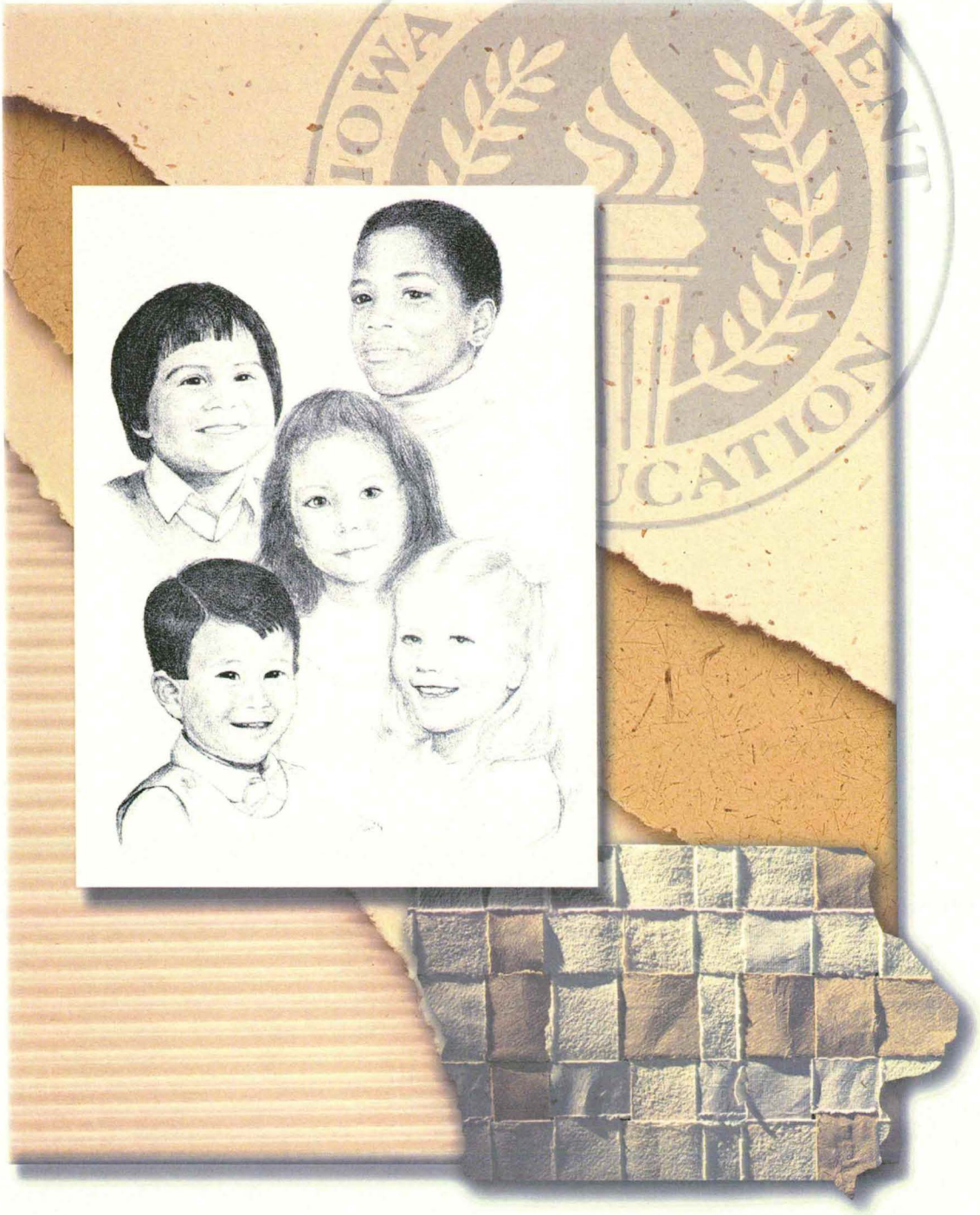
Whisler, J. S., (1998.) *Designing leadership of learning: Implementing standards and benchmarks to promote learning.* Paper presented at the Midwest Regional ESU Conference Building Leadership for High Performing Schools, Okoboji, Iowa. Aurora Colorado: McRel Institute

Winter, P., (1996.) *Implementing the adequate yearly progress provisions of Title I of the improving america's schools act of 1994*. Council of Chief State School Officers: State Collaborative on Assessment and Student Standards. Second in a Series of Papers on the Promises and Challenges of IASA Title I. CCSSO, Washington, D. C.

Ziomek, R. L., & Svec, J. C., (1995.) *High school grades and achievement: Evidence of grade inflation.* ACT Research Report Series, 95-3. Iowa City: The American College Testing Program.

XI. Index

Implementing a District-wide Standards-Referenced Assessment System (DSRAS): Technical Manual



Iowa Department of Education

STATE OF IOWA

THOMAS J. VILSACK
GOVERNOR

SALLY J. PEDERSON
LT. GOVERNOR

DEPARTMENT OF EDUCATION
TED STILWILL, DIRECTOR

DATE: May 26, 1999

TO: District Superintendents
AEA Technical Support Personnel

FROM: Nina Carran NC

RE: Enclosed Technical Assistant Document

The enclosed document has been designed to assist districts developing coordinated assessment systems aligned with content standards. This document was created to extend the contents presented in the Districtwide Standards Referenced Assessment System (DSRAS) document, specifically in the areas of reliability and validity.

We encourage you to share this document with the individual or individuals in your district or AEA who have responsibility for the districtwide assessment of your content standards.

We hope you will find the contents useful as you develop your assessment system aligned with your content standards. If you have any questions, please feel free to e-mail me at nina.carran@ed.state.ia.us or contact me at 515/281-4158. You may also contact Kathy Hinders at kathy.hinders@ed.state.ia.us or call 515/281-3517.

GRIMES STATE OFFICE BUILDING / DES MOINES, IOWA 50319-0146
FAX (515) 242-5988

Implementing a District-Wide Standards-Referenced Assessment System (DSRAS): Technical Manual

A Report from the Assessment Literacy Task Group
at the Request of the Iowa Department of Education, May 28, 1999

Table of Contents

I. Reliability

Introduction

Reliability is a term used by many people in a variety of ways. Unfortunately, each of these different uses conveys a different meaning. Hence, when a person states that a particular observation "was reliable," it is not always clear as to just which aspects of the observation he or she refers. For example, does the person mean that another measure is likely to reproduce similar results; under what circumstances? While the jargon may be confusing, the concept of reliability is quite simple. The concept of reliability refers to the consistency of observations over repeated measures.

Many textbooks, research papers and journal articles have been devoted to the concept of reliability. The document published by the Iowa Department of Education titled *Implementing a District-Wide Standards-Referenced Assessment System (DSRAS)* provides introductory information regarding reliability. Much of that information is repeated in this document. However, this document goes beyond the DSRAS and incorporates much more detail. Although this document is rather technical, it has been constructed so the reader (with a little diligence) will be able to comprehend many (if not all) of the nuances surrounding reliability. Additionally, many computational formulas, computer code, examples and explanations are provided. This multi-faceted approach is intended to provide ample support for the reader to actually embrace the concept of reliability from the theory, to the formulae, to the practical task of designing reliability studies and collecting reliability evidence.

A Conceptual Example

Consider the following conceptual example. If one were to gather measures of a person's bowling ability without a change (i.e., improvement) in the person's true ability, we would expect those scores to be fairly consistent. A simple way to do this would be to observe the individual bowling on several occasions and to record the scores. A simple average of all the scores across all the games would provide an estimate of that person's bowling ability. We know that it is unlikely that the individual would score the same in every game, after all, human interactions are

quite complex and are subject to influences of the situation. Still, if a person were to average a score of 100 across all previous games, we would be surprised if that person bowled a 200 on any particular game and very skeptical (to the point of disbelief) if he or she bowled a perfect game of 300. Our disbelief is fueled by what we know to be a fairly stable or consistent index of the person's bowling skill, namely, the average across previous games. In reality, we are making a statement about the reliability of the results from the measures. We expect this person to earn a score on any one game that is similar to the person's average score.

Some Important Properties of Reliability

People strive for continuous improvement in nearly every endeavor they pursue. Such a practice is also closely related to the concept of reliability, as can be seen from W. Edwards Deming's work in quality control manufacturing. (See for example, Deming, 1986.) Deming's idea of continuous improvement can be summarized as follows: If specifications call for tolerances (i.e., errors) of only one-quarter inch during the current round of production, cut this in half during the next production round by monitoring and controlling the manufacturing process. Deming's idea was for constant improvements in accuracy through reducing error from multiple sources. In other words, Deming wanted to increase reliability (consistency) by reducing error components, and he wanted to do this continuously. This example highlights two important properties associated with reliability: 1) the collection of evidence of reliability is an ongoing and continuous process; and 2) all sources of error need to be investigated and controlled to maximize reliability.

Another way to conceptualize reliability is to consider the degree to which we can generalize a score collected on one occasion, under a particular set of circumstances, to another occasion where the circumstances may be slightly different (Mehrens and Lehmann, p. 54, 1987). In the bowling example we used the average of all past bowling scores to judge the consistency of a score resulting from a specific game. Another approach would be to compare the scores from two games. For example, suppose the bowler scored 106 on the first game and 112 on the second game. Why do these scores differ? First, we should recognize that they are very similar in reference to how the points in bowling are earned. Second, how did the

circumstances change between the first and second games bowled? Did the bowler need the first game as a "warm-up" indicating that the first score was too low? Did the bowler learn something about the particular lane being used? If the games occurred during different weeks at different bowling alleys, then such a difference of six points might not be perceived as very large. Additionally, if the bowler used different balls, different shoes and bowled at different times of the day across these two occasions, we might feel quite confident that the bowler's true ability was somewhere in the range of about 106 to about 112. The particular circumstances that could potentially lead to error (error components) mediate our confidence in generalizing from one observation of a behavior to another.

Reliability Evidence for Academic Achievement

Our examples of reliability so far have dealt with the relatively easy application of measurement to sports. Measuring such intangibles as student academic progress or achievement is much more difficult. We have trouble getting repeated observations of behavior (scores) through repeated testing. If we give the same test twice to the same group of students on the same day, for example, fatigue would be likely to contaminate the students' performances the second time around. Results of the second administration of the test would also be influenced by the students' first experience with the test. This influence could be either positive or negative. In any event, the second administration would not be independent of the first administration (independent means *not influenced* by). To get around the repeated measures problem, psychometricians provide ways to estimate reliability based on a single test administration. In a single evening we might observe and record the scores from as many as five bowling games before a person's behavior changes (i.e., the person's performance decreases due to fatigue or other changes to the actual condition of the bowler). The average of the observations would be the best estimate of the bowler's "true" bowling ability. The bowler's range of scores (presumably the bowler would score differently in each game) would represent inconsistency in measures of this ability. The mathematical interpretation of this range is via a statistic called the **standard error of measurement** (which will be presented in great detail later in this document). Each game's score was not exactly the same as the other, nor was it the

same as the person's estimated true bowling ability, the average of all games' scores. This is the case for any number of reasons: individual differences in performance; different distractions; different lanes; different contexts (i.e., a different sequence of bowlers, different pin configurations and so on). Each of these reasons for inconsistent performance represents different error components.

An Operational Definition of Reliability

Before we visit the mathematically derived estimates of reliability, an additional conceptual framework is provided in the form of an operational definition of reliability. As alluded to previously, reliability is conceptualized as the consistency of measures.

- *Reliability refers to the consistency of the results between two measures of the same thing.*

Consistency can be seen in the degree of agreement between two measures on two occasions. When this agreement is high, it is likely due to the lack of error influencing the individual measures. In the bowling example, one would examine the agreement in scores between two games, say, the first game and the second. Operationally, such agreement is the essence of mathematically defined reliability indices. These indices provide the reliability coefficients so often cited in technical manuals and textbooks. Let us now explore the components of the reliability coefficient.

The Coefficient of Reliability

In the bowling example and the continuous improvement example provided previously, reliability was increased by increasing the consistency between measures of observations or by removing error components. Error components are the undesirable things influencing the measure (such as fatigue, distractions in the classroom, etc.). Three characteristics are implicit in these examples:

- *Consistent measures in a controlled environment will increase reliability;*
- *Control of circumstances reduces the potential for error;*
- *It is impossible to identify, let alone control (eliminate), all possible influences (error) on the measure.*

These characteristics of measurement, taken together, lead to the fundamental conclusion that all measures consist of an accurate or "true" component and some inaccurate or "error" component. In fact, this is the fundamental premise of classical reliability analysis as well as classical measurement theory. Stated explicitly, this relationship can be seen as the following:

Observed Measure = True Score + Error.

We can think of reliability as the ratio of the true score to the observed measure:

$$Reliability = \frac{True\ Score}{Observed\ Measure}$$

To facilitate a mathematical definition of reliability, these components can be expanded as follows:

$$Reliability\ Index = \frac{True\ Score}{True\ Score + Error}.$$

Clearly, when there is no error the reliability index will be the true score divided by the true score, which is unity, or, one. However, as more error influences our measure (the observed score), the error component in the denominator of the ratio will increase; this will decrease the size of the reliability index making it less than one. It is this type of ratio that is estimated when people say that the reliability indices associated with various measures range say, between 0.80 and 0.90.

Classical Estimation of Reliability

Due primarily to the unknown (and unknowable) "true-score" component of a person's observed measure, various estimates of reliability have been derived. As shall be seen, some of these estimates are applicable only to tests comprised of dichotomously scored items, while others can be applied to all assessments. Allen and Yen (1979, p. 76) provide three general classifications of these estimates: test / retest, parallel forms, and internal consistency. So far, all examples have used the test / retest classification with the assumption that repeated measurement did not impact students' performance. As we shall see, this assumption may not hold in practice. Additionally, each different estimate of reliability accounts for different

components of error and may lend itself to different applications in the practical collection of reliability evidence.

Test / Retest Reliability Estimates

As Allen and Yen (1979) point out, test / retest estimates are based on the notion of measuring the examinee's performance twice. A simple comparison of the results of the two measurements (usually in the form of a mathematical correlation) provides an index to the degree of agreement or consistency between the two measures. For example, a simple listing of the rank orderings from the first measure compared to a similar listing from the second measure provides an estimate of reliability. If the lists are the same (i.e., each student scored in exactly the same order on both measures), then there is perfect agreement between the measures and, conceptually, reliability would be unity, as pointed out by Allen and Yen (1979, p. 76.)

Unfortunately, especially with regard to academic and achievement oriented tasks, it is not likely that a student can respond to the same assessment twice without being influenced to some unknown extent by the assessment itself. The act of testing itself introduces inconsistency into a test / retest reliability estimate. Additionally, depending upon the time frame, the students might legitimately improve during the time period between testings. Hence, the results from the second testing will be different from those of the first because of this improvement, but this difference will be interpreted as unreliability. Again, as Allen and Yen (1979, p. 77) point out, because of the influences of repeated testing and due to circumstances encountered during the time period between testings, the test / retest model of reliability is best when used for traits that are stable across time (i.e., not related to direct instruction).

Assuming that the act of taking a test did not change the students (a strong assumption indeed), a test-retest reliability study would collect data from the same sample of students who responded to the same test on two different occasions. Suppose such a study was conducted with the following results for a 30-item test.

	Student Scores	
	Test	**Retest**
Student 1	28	30
Student 2	26	22
Student 3	23	19
Student 4	21	23
Student 5	19	18
Student 6	17	17
Student 7	15	17
Student 8	12	14
Student 9	11	10
Student 10	11	12
Mean	18.3	18.2
Standard Deviation	5.85	5.51
Test-Retest Reliability	0.92	

In this case, the Test-Retest Reliability estimate is simply the mathematical correlation between the scores on the test and retest. A correlation simply quantifies the degree of the relationship or agreement between the two sets of scores. Such a correlation is easy to calculate. You can do so using spreadsheet programs for personal computers as well as more sophisticated statistical software such as SAS© and SPSS for Windows©. For example, in SPSS for Windows© you can choose correlation from the statistics drop-down menu and point and click. It generates the following syntax:

```
CORRELATIONS
 /VARIABLES=TEST RETEST
 /STATISTICS DESCRIPTIVES
 /MISSING=PAIRWISE.
```

The same information can be generated in SAS© via the following syntax:

```
PROC CORR NOMISS;
     VAR TEST RETEST;
```

Most spreadsheets can generate a Pearson correlation from internal functions. For example, with Microsoft Excel 97© you would select the TOOLS menu, the DATA ANALYSIS menu (which is a Microsoft Excel 97© "Add-in"), and choose CORRELATION. You would then select the two columns of data from your spreadsheet representing student scores on the test and retest.

Finally, the computational formulas for the Pearson correlation are rather simple and some users could generate these correlations either explicitly in a spreadsheet or by hand. The computational formula is:

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{S_X S_Y}$$

$$= \frac{\frac{1}{n}\left[\sum_{i=1}^{n} X_i Y_i\right] - \overline{XY}}{S_X S_Y},$$

where:

r_{XY} = Pearson Correlation Coefficient,
$\overline{X}$ = Mean of the Test,
$\overline{Y}$ = Mean of the Retest,
S_X = Standard deviation of the Test,
S_Y = Standard deviation of the Retest,
n = number of students,
X_i = Each student's score on the Test,
Y_i = Each student's score on the Retest.

In fact, applying this formula to the student scores from the previous table reproduces the reliability coefficient as presented in the next table.

Student	(X_i - Mean of X)	(Y_i - Mean of Y)	Cross-Product
1	(28-18.3) = 9.7	(30-18.2) = 11.8	9.7 * 11.8 = 114.46
2	(26-18.3) = 7.7	(22-18.2) = 3.8	7.7 * 3.8 = 29.26
3	(23-18.3) = 4.7	(19-18.2) = 0.8	4.7 * 0.8 = 3.76
4	(21-18.3) = 2.7	(23-18.2) = 4.8	2.7 * 4.8 = 12.96
5	(19-18.3) = 0.7	(18-18.2) = -0.2	0.7 * (-0.2) = -0.14
6	(17-18.3) = -1.3	(17-18.2) = -1.2	(-1.3) * (-1.2) = 1.56
7	(15-18.3) = -3.3	(17-18.2) = -1.2	(-3.3) * (-1.2) = 3.96
8	(12-18.3) = -6.3	(14-18.2) = -4.2	(-6.3) * (-4.2) = 26.46
9	(11-18.3) = -7.3	(10-18.2) = -8.2	(-7.3) * (-8.2) = 59.86
10	(11-18.3) = -7.3	(12-18.2) = -6.2	(-7.3) * (-6.2) = 45.26
Sum of the Product =			297.4
Sum of Product Divided by the No. of Students =			297.4 / 10 = 29.74
Product of the Two Standard Deviations =			5.85 * 5.51 = 32.23
Correlation Coefficient (Reliability Estimate)			29.74 / 32.23 = 0.92

While Test-Retest reliability is not likely to be a concern for most districts, it was provided first in order to "set the stage" for discussions of other reliability indices. The concept of estimating the amount of agreement between two sets of scores, and

the factors that contribute to any disagreement, is the essence of reliability analysis. Furthermore, the correlation coefficient is often used as the basis of many of the different reliability formulations. This includes such formulations as inter-rater and intra-rater reliability (within judge and between judge agreements) typically encountered with performance assessments, writing essays and other measures requiring judgments in scoring.

Parallel-Forms and Alternate-Forms Reliability Estimates

The idea behind both parallel and alternate forms reliability is to correlate student performance on two different forms of the same measurement instrument. Forms K and L of ITBS are examples of parallel forms. It is important to know that students' scores will be highly related regardless of which form they respond to. Typically, such agreement statistics are collected by having the same group of students respond to both forms. The correlation between student performance on both forms would be an estimate of the reliability, or an estimate of parallel-forms reliability.

According to Allen and Yen (1979), it is not possible to verify when two versions or forms of an assessment are parallel. Strictly speaking, two forms of an assessment are parallel when they fulfill a variety of statistical requirements that are almost never possible in practice. A more detailed and technical explanation of these requirements can be found in Lord and Novick (1968, p. 37). Alternate test forms are simply a less rigorous implementation of the requirements for parallel tests. Allen and Yen (1979, p. 78). define alternate test forms as any test forms constructed to be parallel, but that do not achieve the equality in statistical indices required under the definition of parallel. Typically, only test publishers construct test forms that can be considered parallel, most other multiple forms are alternate. While these distinctions are not critical for users of a district-wide standards-referenced assessment, knowing the difference between strictly parallel forms and simply alternate forms may make the information clearer.

Publishers of standardized assessments, as well as constructed response and performance assessments, spend a lot of time and money establishing evidence that the forms are parallel. This means that it should be a matter of indifference to your students which form they respond to. A student's score is not expected to change

significantly from what was received on Form A should he or she take Form B instead. The Stanford Achievement Test Series, Ninth Edition (Stanford-9) for example provides "Alternate-Forms" reliability coefficients, standard errors of measurement and summary statistical information for each test and subtest by form. Typically, these reliability estimates range in the high 0.80s to low 0.90s, but change depending upon grade and subject (HBEM, 1997).

In the implementation of a district-wide standards-referenced assessment system, there may be the need locally to collect information about student performance on the same content standards at two different occasions. Hence, districts selecting assessments or constructing their own assessments for this purpose will have to provide evidence that the forms are indeed parallel and that they provide highly consistent (i.e., reliable) score interpretations. If a district selects an assessment with multiple forms from a commercial vendor, they must be sure to obtain technical information documenting how the forms are parallel (i.e., how they measure the same thing) and what evidence of reliability is provided.

If a district constructs its own assessment with multiple parallel forms, the district must document how they are parallel (i.e. how they consistently measure the same thing). Typically this is a two step process. First, the district must show that the two test forms were constructed in similar ways with the same number and type of items each of which clearly matches the same content standard and benchmark. Such things as item format, question complexity, reading passage length, paper quality, artwork, directions, administration, time length, among other considerations must be the same. Secondly, the results from administration of these forms should be almost identical. The raw score statistics (means, standard deviations, etc.) should be as similar as possible. The resulting frequency distribution of scores should also be of similar shape and provide similar distribution statistics. Finally there should be a high degree of relationship between the two forms as indicated by the correlation between the scores on the two forms. Clearly, the construction and use of parallel test forms locally is no small task.

Internal-Consistency Reliability Estimates

Districts are most familiar with reliability estimates that are derived from a single test administration. These are typically referred to as "internal-consistency" reliability estimates. According to Allen and Yen (1979, p. 78), the most often used implementation of this method is to correlate student performance on the even items with that from the odd items. This procedure is referred to as the split-half procedure (Allen and Yen, 1979, p. 78). Clearly, the biggest advantage of such a procedure is that only one administration of the assessment is needed. However, as Allen and Yen (1979) point out, there are requirements regarding how the halves are assigned and not all splits are equal.

Conceptually, the collection of internal-consistency reliability estimates is simple, all students take all items and you calculate a correlation coefficient between each student's performance on the first half with that of the second. While the odd/even split is often useful because it often leads to equitable splits, it is not always the case. For example, on speeded tests where the items near the end are not likely to be answered by all students, the split-half estimates will include this as an additional error component that will reduce the reliability estimate. Another problem is that splitting the test in half reduces the effective length of the test, and thereby also reduces the reliability estimate. There are correction formulas that can be applied, but the fundamental issue is that of the degree of agreement (or relationship) between the scores on the different halves of the test, or the "internal consistency" of the scores.

There are four generally accepted procedures for estimating internal consistency reliability which share properties to the concept of a "split-half" reliability estimate, each of which is different to some extent for the reasons to be discussed in the following paragraphs.

Split-Half Reliability Estimates

Consider the following set of student scores on one 12 item/task assessment:

Students	Items 1	2	3	4	5	6	7	8	9	10	11	12	Score
1	1	1	1	1	1	1	1	1	1	1	1	0	11
2	1	1	1	1	1	1	1	1	1	1	0	0	10
3	1	1	1	1	1	1	1	1	1	0	0	0	9
4	1	1	1	1	1	1	1	1	0	0	0	0	8
5	1	1	1	1	1	1	1	0	0	0	0	0	7
6	1	1	1	1	1	1	0	0	0	0	0	0	6
7	1	1	1	1	1	0	0	0	0	0	0	0	5
8	1	1	1	1	0	0	0	0	0	0	0	0	4
9	1	1	1	0	0	0	0	0	0	0	0	0	3
10	1	1	0	0	0	0	0	0	0	0	0	0	2

Note that the table looks unlike data we see in the real world only because it has been sorted with students and items renumbered such that the best scoring students appear first (in the first rows) and the easiest items appear first (in the first columns). This data is actual student performance on selected items from an educational psychology course. Consider the agreement between student scores when compiled from the even items and the odd items separately as depicted in the next table.

Students	Scores " Even Items"	" Odd Items"
1	5	6
2	5	5
3	4	5
4	4	4
5	3	4
6	3	3
7	2	3
8	2	2
9	1	2
10	1	1
Mean =	3.00	3.50
Standard Deviation =	1.41	1.50
Correlation =	0.94	

Even without the correlation coefficient, it is clear from looking at the scores that the students are rank-ordered in the same way regardless which score is used (total, even split or odd split). We can also see that the odd items were somewhat easier than the even items. The correlation between the scores resulting from the odd items with those resulting from the even items is an internal consistent reliability estimate for

this test, namely, a split-half estimate. Again, there are formulas for adjusting this estimate to reflect the fact that it was from a test which was "cut in half", but the use of these formulas require a determination about the degree to which the tests are parallel. One such formula, the Spearman-Brown "prophecy" formula, allows for the estimation of the reliability for the total length test. For the current purposes of exploring reliability as agreement between two sets of test scores, the Spearman-Brown formula will not be provided.

Kuder-Richardson Formula 20 (KR20)

The concept of a "split-half" internal consistency reliability estimate can apply to tests comprised of multiple-choice or other objectively scored items and performance assessment tasks with polytomously scored items. Attention must be paid to ensure that each test half equally represents the test as a whole and that the performance assessment components do not all fall into the same half.

When the test is comprised of truly dichotomous items such as in the multiple-choice

$$KR20 = \left[\frac{K}{K-1}\right] * \left[\frac{\sigma_X^2 - \sum_{i=1}^{K} p_i(1-p_i)}{\sigma_X^2}\right],$$

format, the formulas derived by Kuder and Richardson (1937) can be used.

where:

p_i is the proportion of students answering item i correctly,

K is the number of items, and

σ_X^2 is the total test variance (variance of the raw scores).

Supposing that the data from the previous example was collected from a test that was all dichotomously scored multiple-choice items. The resulting item and test form information can be seen in the table that follows:

Items	P_I	$(1-P_i)$	$P_i * (1-P_i)$
1	1.00	0.00	0.00
2	1.00	0.00	0.00
3	0.90	0.10	0.09
4	0.80	0.20	0.16
5	0.70	0.30	0.21
6	0.60	0.40	0.24
7	0.50	0.50	0.25
8	0.40	0.60	0.24
9	0.30	0.70	0.21
10	0.20	0.80	0.16
11	0.10	0.90	0.09
12	0.00	1.00	0.00
Sum of $P_i * (1-P_i) = 1.65$			
Total Test Mean = 6.50			
Total Test Variance = 8.24			

Application of the KR20 formula yields the following results:

$$KR20 = \left[\frac{12}{12-1}\right] * \left[\frac{8.24-1.65}{8.24}\right] = 0.87.$$

Several things can be seen from this example. First, classroom teachers, as well as district school personnel, have all of the information they need to calculate the KR20 once the assessment is scored. This is one of the reasons the KR20 coefficient is so popular, because it is simple to compute. Second, the KR20 estimate of 0.87 is less than the previously calculated split-half estimate of 0.92, even without adjusting for test length. This points to some of the problems in obtaining internal splits which result in true parallel half tests. As you can recall, the mean and standard deviation were quite a bit different for the odd-split. Also, KR20 is a "lower bound" estimate of the test's "true" reliability index in that it is always lower than the true but unknown index of reliability.

Kuder-Richardson Formula 21 (KR21)

For historical reasons, in the age before pocket calculators and desktop computers, a computationally simpler version of the KR20 was derived called the KR21. Again, when all of the items on a test are dichotomous the following formula provides an estimate of reliability known as KR21:

$$KR21 = \left[\frac{K}{K-1}\right] * \left[\frac{\sigma_X^2 - N\bar{p}(1-\bar{p})}{\sigma_X^2}\right],$$

where:

$\bar{p}$ is the average proportion of students answering each item correctly,

K is the number of items,

σ_X^2 is the total test variance (variance of the raw scores).

In the current example, the KR21 reliability estimate is:

$$KR21 = \left[\frac{12}{12-1}\right] * \left[\frac{8.24 - 12(0.54)(0.46)}{8.24}\right] = 0.70.$$

Because KR21 uses less information when it substitutes the mean proportion correct it provides a poorer estimate of the true reliability index. In addition, KR20 will never be less than KR21. The only time KR20 and KR21 will be equivalent is when all item difficulties (p-values) are the same, which is not likely to occur in practice.

Coefficient Alpha (α)

The most general estimate of the internal consistency reliability which is applicable to both multiple-choice and performance assessment type data is coefficient α (Cronbach, 1951). The following formula provides coefficient α:

$$\alpha = \left[\frac{K}{K-1}\right]\left[\frac{\sigma_X^2 - \sum_{i=1}^{K}\sigma_{y_i}^2}{\sigma_X^2}\right],$$

where,

X is equal to the observed score for a test with K items/tasks,

σ_X^2 is the total test variance (variance of the raw scores),

$\sigma_{Y_i}^2$ is the variance of each item or task i, and

K is the number of items or tasks.

If we revisit the data from the current example:

Items/Tasks	Student Responses	Item Variance
1	1 1 1 1 1 1 1 1 1 1	0.00
2	1 1 1 1 1 1 1 1 1 1	0.00
3	1 1 1 1 1 1 1 1 1 0	0.09
4	1 1 1 1 1 1 1 1 0 0	0.16
5	1 1 1 1 1 1 1 0 0 0	0.21
6	1 1 1 1 1 1 0 0 0 0	0.24
7	1 1 1 1 1 0 0 0 0 0	0.25
8	1 1 1 1 0 0 0 0 0 0	0.24
9	1 1 1 0 0 0 0 0 0 0	0.21
10	1 1 0 0 0 0 0 0 0 0	0.16
11	1 0 0 0 0 0 0 0 0 0	0.09
12	0 0 0 0 0 0 0 0 0 0	0.00
Sum of Item Variances = 1.65 **Total Test Variance = 8.24** **Coefficient α = 0.87**		

Like KR20 and KR21, Coefficient α is also a "lower-bound" reliability estimate.

Summary

Several internal consistency reliability estimates were examined in order to understand how each could be calculated and to provide an example of what influences impact their calculation. Using the current data set, the "split-half" estimate was 0.94 with an adjustment for test length, KR20 was 0.87, the "quick and dirty" KR21 was 0.70 and the generic Coefficient α (alpha) was 0.87. While these estimates were calculated by hand in these examples, the use of a spreadsheet would simplify almost all of these computations. If a spreadsheet is used be sure to index the items in columns and the students in rows. Fortunately, both SPSS for Windows© and SAS© will perform simple internal consistency reliability calculations. The following is the syntax necessary to produce the SPSS for Windows© reliability analyses:

```
RELIABILITY
 /VARIABLES= ITEM01 ITEM02 ITEM03 ITEM05 ITEM05
             ITEM06 ITEM07 ITEM08 ITEM09 ITEM10
 /SCALE(ALPHA) = ALL / MODEL=ALPHA
 /STATISTICS = DESCRIPTIVE SCALE CORR
 /SUMMARY = MEANS VARIANCE CORR.
```

Other methods of reliability estimation can also be conducted with SPSS for Windows© including split-half, parallel forms analyses and a variety of scale and sub-scale analyses.

The syntax for calculating simple correlation analyses and Coefficient α, is fairly simple in SAS©. The following syntax may be helpful:

```
PROC CORR NOMISS / ALPHA;
     VAR ITEM01-ITEM10;
```

Reliability in Generalizability Theory

Generalizability theory (G-Theory) is a conceptual extension of classical reliability theory as provided by Feldt and Brennan (1989). Generalizability theory is an analytical procedure used to identify and quantify the error components, which reduce the reliability of virtually all measures. G-Theory analyses require advanced training; however, interested readers should consider several sources for information: Feldt and Brennan (1989); Shavelson and Webb (1991); and Brennan (1983).

In G-theory, procedures typically found in "analysis of variance" are used to quantify which factors or facets of a measurement situation are associated with the true component being measured and which are components of error (i.e., contributing to unreliability). Conceptually, when we collect a measure all factors influencing the assessment, other than student achievement, are lumped together as a single error component. This is quantified in classical reliability analyses as the error variance. G-theory, on the other hand, attempts to break this error down into component parts. Hence, when estimating the components of error in a generalizability study, the error could be quantified as that attributed to different forms of the assessment, different testing occasions, different readers or raters, different items or tasks and so forth. The power of a generalizability study is that once these separate error components are estimated, variations on how to collect the measures can be made such as to reduce the contribution of any or all error components. For example, the number of raters could be increased or the number of occasions reduced. It is only through the quantification of this error can such control be exerted.

Standard Error of Measurement

One of the biggest problems with indices of reliability is that they have no inherent meaning. For example, is a reliability coefficient of 0.82 sufficient? One way to determine the meaning of 0.82 is to compare it to known quantities or "rules of thumb." For example, the Iowa Tests of Basic Skills typically provides reliability evidence (internal consistency estimates) in excess of 0.90 for all domain total scores regardless of test length (Hoover, et. al., 1996: ITBS, Grade 6, Level 12, p. XIX). So, compared to the ITBS reliability coefficient of 0.90, a coefficient of 0.82 is relatively low. However, such comparisons can often be misleading for several reasons, including differences in test length. Perhaps the biggest limitation to interpretation of such coefficients is their lack of application to individual student scores. If a measure has lower reliability than some other measure, it is influenced by error to a greater extent. Hence, the scores resulting from that measure are less accurate. The standard error of measurement uses the score information from the test along with an estimate of reliability to make statements about the degree to which error influences individual scores. The standard error expresses unreliability in terms of the reported score metric. Using the standard error of measurement, an error band can be placed around an individual score, indicating the degree to which error might be impacting that score. Interested readers can refer to Allen and Yen (1979), Feldt and Brennan (1989), or Traub (1994) for further elaborations regarding the standard error.

The basic definition of the standard error mathematically can be expressed as the following:

$$\hat{\sigma}_E = S_E = S_X\sqrt{1 - \text{Re}\,liability\ Estimate},$$

where:

S_E is the estimated standard error of measurement,

S_X is the total test standard deviation, and

Reliability Estimate is the estimate of total test reliability.

Once the standard error of measurement has been calculated, an error band placed around a student's score can be found via the following:

$$X \pm \hat{\sigma}_E$$

such that a particular student's performance is no longer expressed as a single point, but rather as a range of "most probable" score points. For example, if a student earns a score of 32 on a particular test with a standard error of measurement of 4 then it is more likely that the student's real or true score is somewhere in the range of 28 to 36. In fact, it is a common rule of thumb that students who have score bands that overlap are performing essentially the same on the test despite the actual scores they might have earned. For example, if the student from the previous example were compared to a student with a score of 35 from the same test, we can see that their "error bands" overlap. The first student had an error band ranging from 28 to 36, while the second student had an error band ranging from 31 to 39. Since these two students' error bands do indeed overlap, we would conclude (using our rule of thumb) that their difference in scores is such that it may be due to unreliability of the measure and not necessarily reflect "real" differences in student performance.

Decision Consistency Reliability Estimates

One of the main reasons a district-wide standards-referenced assessment system is implemented is to make better decisions regarding the level of student competencies. From these, informed instruction will lead to improvements in student learning. What this implies is that the consistency or reliability of a measure may not be as important as the accuracy with which the measures are used to classify students into these competency categories. Simply stated, if an assessment classifies students into levels of competency based on some performance standard setting or cut-score policy because the measure is fallible (unreliable to some degree). These classifications are going to be in error some of the time. For example, there will usually be "false masters" (students classified above their actual competency) and "false non-masters" (students classified below their actual competency). The direction of these errors is not a matter of indifference. Districts should consider costs associated with both types of errors in establishing policy regarding such things as remediation as well as the difficulty of the performance standards.

Like classical reliability theory and generalizability theory, there is an area of study dedicated to the understanding and quantification of errors in misclassification associated with decisions placement into "competency" levels. Interested readers are referred to the following sources for more detail: Traub (1994 p. 70); Huynh (1976); and Berk, R. A. (1984).

The purpose of a district-wide standards-referenced assessment is not to generate individual student classifications of proficiency. In fact, the usefulness of posting an individual student's score indicating that the student is at some level of proficiency district-wide ("Basic" for the lowest level, "Proficient" for the middle level, and "Advanced" for the highest level, for example) is dubious. Clearly, the intention of a district-wide assessment system is to make statements about how well the district is doing toward meeting its annual improvement goal and how quickly it is moving groups of students from lower levels of proficiency to higher levels. Nonetheless, because individual student classifications will need to be aggregated to the district level and because such classifications will ultimately contain error, it is necessary to consider the role this "error of classification" will ultimately have on the decisions made in the district. In this regard, decision consistency indices are more helpful than traditional reliability estimates. Decision consistency can be described as (Nitko, 1989 p. 458):

> "Decision consistency indexes describe the extent to which students are likely to be classified the same when either the same test form is re-administered or an alternate form of the test is administered."

It is the consistency of the classification into performance levels that is important and not the consistency of the test scores internally as with an internal consistency reliability estimate.

Coefficient Kappa (κ)

Perhaps the most often used index of decision consistency (one which attempts to quantify if students have been consistently placed into a proficiency level or performance standard category, which is too high or too low), is coefficient κ (kappa) (Cohen, 1960):

$$\kappa = \frac{\sum P_{ii} - \sum P_{i.} P_{.i}}{1 - \sum P_{i.} P_{.i}},$$

where,

the first subscript is the column and the second is the row with a '.' indicating to sum over that particular row or column;

$P_{ii} = F_{ii} / N$;

$P_{.i} = F_{.i} / N$;

$P_{i.} = F_{i.} / N$.

Classification on First Measure

Classification on Second Measure	**1**	**2**	**3**	**4**	
1	F_{ii}				$F_{.i}$
2		F_{ii}			$F_{.i}$
3			F_{ii}		$F_{.i}$
4				F_{ii}	$F_{.i}$
	$F_{i.}$	$F_{i.}$	$F_{i.}$	$F_{i.}$	

When there is perfect agreement, (depending upon the marginal values) the kappa coefficient will be unity as demonstrated in the following simplified example:

Classification on Writing Essay 2

Classification on Writing Essay 1	**1**	**2**	**3**	**4**	
1	25 (0.25)				25 (0.25)
2		25 (0.25)			25 (0.25)
3			25 (0.25)		25 (0.25)
4				25 (025)	25 (0.25)
	25 (0.25)	25 (0.25)	25 (0.25)	25 (0.25)	100 (1.00)

The kappa coefficient is the following:

$$\kappa = \frac{1.00 - 0.25}{1.00 - 0.25} = 1.00.$$

As the number of cases migrates away from the diagonal in the previous table, the value of kappa will decrease. If the values of the table are in complete disagreement

(i.e., there are an equal number of students in each cell of the table), then the coefficient will drop to zero.

Scorer Consistency and Inter-Rater Reliability

To this point, the discussions regarding reliability have been confined for the most part to objective multiple-choice items. However, the need to collect reliability evidence for tasks scored by raters, such as writing essays, is also important. Clearly, the reliability of a writing essay will have the same components of error affecting the resulting scores as the multiple-choice items, *plus* some degree of inconsistency added though the judgmental scoring process (i.e., assigning scores to the essays). This potential for additional error is inherent in all scorings using a judgmental process and is not limited to writing essays.

An index of the degree of error or unreliability added to a score from judgmental scoring could be obtained this way: Have two judges read the same set of essays, each assigning scores independently. Presumably, these judges would follow the same rules in determining the score (i.e., use the same scoring rubric). The percent of agreement between these readers would be an indication of the consistency of the application of the scoring rules (rubrics) to determine the student scores. If the readers consistently assigned the same score, it is more likely that the judges are applying the scoring rules in a consistent manner thereby eliminating error and increasing reliability. Another index used to determine the degree of association between a first set of judgments and a second set of judgments is to simply calculate a mathematical correlation between these pairs of scores. Recall that this is similar to the concept of the test / retest reliability coefficient where the first set of judgements is analogous to the test and the second set of judgments is analogous to the retest. If the first set of judgments agrees, in the most part, with the second set, this estimate of reliability would be positive and large.

As was true with the classical concept of reliability on a multiple-choice assessment, a great deal of effort has been used to study the error associated with human judgments particularly in the scoring of essay responses. Generalizability theory (as referenced in a previous example) is only one of many ways to investigate the variability of judgments applied to scoring. The purpose of this section is to acquaint

the reader with some of the simpler ways to investigate the degree to which unstable judgments may impact the scores resulting from a measure.

An Example

Typically, student scores from a judgmental scoring process are a combination of the points awarded by multiple judges or raters. This combination is usually a summation or an average of two judges' ratings. Suppose, for example, the task being evaluated is a writing essay. Also suppose that the final student score will be the sum of a single rating provided from two judges. When the two judges disagree, a third judge is asked to decide which judge's score is most correct, thereby replacing the outlying judge. In such a way, all students' final scores will be the sum of two ratings. Again suppose that the writing essay was a purchased product which included scoring guides: scoring rubrics (definitions of each of the four possible score points, 1 through 4), anchor papers (examples of actual student work receiving each score point) as well as some general guidelines about scoring the essay. Finally, suppose that three teachers have volunteered to score all of the essays for the entire 100 students in the district in a particular grade and content area. Also suppose that one additional teacher agreed to resolve the scores that differed by more than one score point. Remember, 100 students mean at least 200 separate evaluations of student work because each paper will be rated by at least two judges.

Once the writing is completed and the essays have been collected and transported to a central scoring site, the judges are ready to begin. The first task is to ensure that the teachers reading the essays do not have the opportunity to see the students' names associated with the responses. This is to prevent a host of things that may influence the judgments. A good way to do this is to assign each student an identification number and to remove the cover sheet (if one exists) from the response document. Judges would then go through the training manual to learn the scoring process. Scoring guides typically outline what is important to pay attention to and what distractions should be avoided. The judges then perform a "practice scoring" set provided with the scoring guide. This gives the judges an opportunity to practice scoring, checking their accuracy using the annotated student responses provided with the scoring guide. Additionally, the judges have the opportunity to discuss the

scoring endeavor with each other before the actual scoring of student responses begins. Once the judges are clear about the process, they select rater identification numbers. They will each record their number along with the student identification number each time they score a document. The judges should divide up the student papers for the first round of ratings such that one judge takes half and the other judge takes the other half. On separate sheets of paper the raters record their identification, the score they assigned and the student identification number. Once a paper has been scored for the first time, the judge will return it to a different location indicating that it is ready for the second reading. Once all of the scoring has been completed for the first time, the process starts over. This time the judges "trade" papers and each scores the other half, thereby assigning two scores for each student. When the second round of ratings has completed, the score sheets (papers with the student identification numbers, the judges identification numbers and the students scores) should be entered into a spreadsheet (or some other computer program) for manipulation. It is possible to do it all by hand but the record keeping can become rather labor intensive. The example spreadsheet provided later in this section depicts how such a compiled list, with statistics, might look.

Resolution Scores

The primary purpose of a resolution score is to increase the agreement between ratings provided on the assessment by different judges. Any time scores from two different judges on the same student response are non-adjacent (i.e., differ by more than one score point), a resolution scoring is required. This resolution scoring can simply be a third judge who scores the student response. The outlying score from the three ratings (first judge, second judge and resolution judge) is then eliminated. Of course, different decision rules can be applied such that the two most consistent scores are reported for the individual. As can be seen from the example spreadsheet, only one resolution was needed and this was for student number 999856642. For this student, Rater 01 provided a score of 1, while Rater 02 gave this student a score of 4. On a four point score scale, these ratings are as discrepant (inconsistent) as is possible. Certainly, both of these scores cannot be correct. In order to determine which ratings are most likely to be correct and ultimately determine the student's

score, a resolution rating was provided by the department chairperson (who also participated in the scoring training). This person gave the student response a score of 2 which is adjacent (within one) to the score provided by Rater 01. Hence, the score provided by Rater 02 (4) was seen as inconsistent and was replaced by the resolution score. This can be seen by the total score the student earned.

While the resolution scoring provides a mechanism to increase rater agreement, it is itself an index of the agreement or consistency in scoring. For example, in the present case only one resolution score was necessary, indicating that nine of the ten pairs of scores (one from each rater) were within one point of each other (i.e., 90 percent were adjacent). Clearly, if half of the ratings needed resolution this would be indicative of an inconsistent scoring process.

Correlations Between Readings

Another simple statistic that can be calculated and used to describe the degree of consistency in such a judgmental scoring process is a simple correlation between the first score given a student and a second score. This is analogous to the "split-half" method of estimating internal consistency reliability but instead of correlations between two half tests, the correlations are really between two "half-scores."

As can be seen from the example data, the correlations between the first and second reading (provided from the two judges) is 0.27 before resolution (i.e., using the unresolved scores), but jumps to 0.70 after resolution. This is a good example of the effects of extreme scores particularly when the number of cases is relatively small.

Because the correlation coefficient is a mathematical function constructed to describe the degree of relationship between two sets of data, its interpretation must be limited. The correlation is not only affected by the presence of extreme scores and the number of cases, but also the degree of similarity in the characteristics of the group as well as other factors. For example, groups of students who tend to score similar to each other typically show lower correlations between ratings than do more diverse sets of students. All of these limitations mean that the interpretation of the correlation coefficient should never be made in isolation.

	First Reading		Second Reading	
	Rater		Rater	
Student ID	01	02	01	02
149524986	4			3
148563214		3	2	
125569874	3			3
123584562		2	2	
523697849	2			2
897523146		2	1	
999856642	1			4
123568963		1	1	
856236951	1			2
282814963		3	3	

	Totals			Final
Student ID	Score 1	Score 2	Resolution	Score
149524986	4	3		7
148563214	3	2		5
125569874	3	3		6
123584562	2	2		4
523697849	2	2		4
897523146	2	1		3
999856642	1	4	2	3
123568963	1	1		2
856236951	1	2		2
282814963	3	3		6

	Differences	
Student ID	Before Resolution	After Resolution
149524986	1	1
148563214	1	1
125569874	0	0
123584562	0	0
523697849	0	0
897523146	1	1
999856642	3	1
123568963	0	0
856236951	1	1
282814963	0	0

Correlation Between Readings	
Before Resolution	After Resolution
0.27	0.70

Percent Agreement Before Resolution	
Perfect	0.50
Adjacent	0.40
Beyond	0.10

Percent Agreement After Resolution	
Perfect	0.50
Adjacent	0.50
Beyond	0.00

Percent Agreement

Another very simple to calculate, yet informative, index regarding the agreement between judges is the percent agreement index. This is simply the percent of occasions the judges agreed with each other regarding their ratings. This can be expressed in many ways but the most often used is: Percent Perfect Agreement, Percent Adjacent Agreement, and Percent Non-Adjacent Agreement. In the current example, using the before resolution data, we can see that half of the time the judges provided the same score for the students, 40 percent of the time the ratings were within one-point of each other, with one score (10 percent) beyond one score point. Certainly, these percentages improved when the resolution scores were used.

Summary

This section provided information regarding how to score and collect agreement data when human judgments are involved in the scoring process. Typically this is used when rating student responses to writing essays, performance assessments or other non-machine scorable tasks. The collection of such data is not trivial. Three different pieces of information can be collected to document the consistency (and inconsistency or error present) from a performance scoring. These include an accounting of the number of resolutions required to reach adjacent agreement, the correlation coefficient between readings/raters, and the percent of perfect and adjacent agreement. Note that the example used only two judges and that the logistical work of scoring and tracking judges' responses increases substantially when more judges are used. However, the procedures outlined in this section will generalize to cases requiring more than two judges.

In addition, this section supposed that the very important pieces of the performance assessments or essays were available from a publisher (such as the scoring guides, scoring rubrics, anchor papers and other training materials). These pieces are crucial parts to a successful scoring and should not be taken lightly. For districts who wish to develop their own performance assessments or writing essays, as much if not more attention should be paid to the scoring process as to the stimuli or prompts themselves.

II. Validity

Introduction

Assessment results must show evidence of reliability for the purpose for which they were intended before they can show evidence of validity. Hence, this chapter on validity is presented after the chapter on reliability. This does not mean that validity is less important. In fact, the Standards for Educational and Psychological Testing (hereafter referred to as "the standards") state that validity is "...the most important consideration in test evaluation" (APA, 1985, p.9). The main concern is that the collection of evidence on the accuracy of the interpretations of the scores resulting from a measurement is mostly a judgmental process. Reliability, for the most part, lends itself well to estimation through statistical means, though judgment is required in trying to reduce the error components influencing a measure. Validity evidence, on the other hand, especially regarding the establishment of content validity evidence, is judgmental. This is especially true in the area of educational achievement where information regarding how much content knowledge a student possesses is the paramount interpretation desired from assessment results.

In the history of academic assessment, a distinction was made between different types of "validity." For example, the terms "content validity," "construct validity" and "criterion-referenced validity" were used. This generated more confusion than was necessary:

> *Validity is, and always was, a "unitary concept" (APA, 1985; Linn and Gronlund, 1995, p.49)*

The way evidence was collected to demonstrate valid use and interpretations of assessment results took on many different and specific classifications. While this distinction between validity as a unitary concept and the different types of validity evidence might seem trivial, it is important to remember that all types of validity evidence should be collected when possible. This evidence should document and demonstrate that the interpretations being made from the results of the assessment are appropriate. The standards provide the following (APA, 1985, p.9):

> *"An ideal validation includes several types of evidence, which span all three traditional categories (content-related, criterion-related, construct-related). Other things being equal, more sources of evidence are better than fewer.*

However, the quality of the evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable quality."

The standards go on to state that professional judgment should determine which information should be collected and documented as evidence of valid score use and interpretation. The standards also state that resources should be expended to obtain the evidence that "optimally reflects the value of a test for an intended purpose" (APA, 1985, p. 1). Because a district-wide standards-referenced assessment system will encompass many different purposes (reporting progress toward meeting annual improvement goals to the community, providing feedback to teachers in order to improve instruction and fulfilling statewide accountability requirements), many different types of validity evidence will need to be collected.

The primary concern for all districts will be the match between the assessment components and the content-standards.

Clearly it will do little good to have a very consistent (reliable) measure of well documented content when that content is not what is being taught or even what is seen as important by the local community. Hence, the collection of content validity evidence will be what most districts will spend their resources on. Aligning assessment frameworks to content-standards and the documentation and publication of such alignments will be one important piece of validity evidence. The procedures used to demonstrate validity evidence are presented in the paragraphs that follow.

Content Validity Evidence

It may not seem like it, but there has probably been a lot of work done already within your district to demonstrate the content validity of a particular assessment. Each district has probably considered in great detail the content-standards and benchmarks adopted as current curriculum within the district. This activity almost assuredly did not take place independently of on-going instruction and assessment. In fact, perceived gaps between aspects of the former curriculum and the newly desired content-standards were probably discussed in great detail. In addition, gaps in what was being assessed or how it was being assessed were also likely considered. Unfortunately, these perceptions were probably not collected in a systematic way, so that while many people involved with selecting the new content-standards for a

district have ideas about how well current assessments "fit" these standards, they have not been formally recorded. The content-standards define key aspects of the curriculum that are important. The assessments currently in place presumably measure this content to varying degrees. However, "how much" and "where" the assessments currently in place measure the content-standards are probably not known.

> *Nonetheless, it is critically important that the assessments (both new and old) accurately measure the adopted content-standards.*

Inferences are made about how much content is learned based upon how students score on the assessments. If the assessment does not measure the content-standards being taught then these inferences (i.e., the interpretations) made from the assessment results will be misleading. It is therefore very important to ensure that the assessment, be it a commercially available assessment or one developed by the district, matches the content being taught.

There are many different ways to ensure that the content of the assessments is aligned with the content-standards, but perhaps the most often used procedure relies on expert judgment (APA, 1985, p. 10). Presumably teachers in the content areas as well as other school personnel were involved in establishing the district-wide content-standards. These same teachers could serve as judges in assigning components of the various assessments to the content-standards. In the case of new assessments, these same teachers could form a review panel that would select or develop new assessments based on their match to the content-standards. Such matches provide evidence that the assessments do indeed measure the content-standards. In addition, such matches highlight which assessments are best measuring various pieces of the content-standards such that there will be no "holes" in what is being assessed across the entire assessment system. A simple table showing the number of assessment components, the tasks or items matching each content-standard or benchmark and how the match was decided could serve as this documentation.

Once there is a map showing how the assessments are aligned with the content-standards, additional judgments will be required regarding how well the assessments

measure the range of content. For example, perhaps there is little, if any, match between a particular content-standard regarding student writing skill and some existing or selected measure. In such a situation, it is doubtful that appropriate interpretations of the assessment results can be made regarding the content-standards because so few measures (if any) are linked to these content-standards. In fact, this might point to a need for additional assessment components to be added to the district-wide assessment.

> *Typical assessments are simply single-point in time samples of all possible tasks that might be constructed for a given content-standard.*

Teachers in the district (as well as other district personnel) acting as judges must determine if the particular sample of items or tasks on the selected assessments fairly represents all possible content to be assessed under a particular content-standard or benchmark. In addition, judgments must be made regarding the format and environment in which the student responses are collected. For example, if the content-standards state that the student is to construct multiple solutions to a problem in a particular sub-domain of mathematics but the assessments offer only multiple-choice items, this should be documented in the match. It is doubtful that valid score interpretations are being made regarding student generated solutions on the assessment if students do not actually generate their responses. Therefore, in the district-wide standards-referenced assessment system, it is most likely that a variety of assessment tasks will be needed to cover the range of content-standards as well as formats required.

> *While reliability is manifested statistically, content validity is manifested judgmentally.*

Unlike the reliability estimates presented in the previous chapter, there are no general mathematical indices to establish evidence of content validity. As Mehrens and Lehmann (1987) point out, careful consideration of the match between the content-standards and each test item is probably the best way. While this may be the best way, it is nonetheless subjective. While it may be possible to develop some sort of scoring rubric and generate some inter-judge agreement data regarding the match between content-standards and various assessments, it is probably better to collect judgments from as many experts as possible and to allow for group discussion and /

or consensus building. Additional steps can be taken to minimize the work involved if the selected assessment is commercially available. Test publishers usually provide very detailed classifications of their assessments. These include detailed explanations of what the items measured, as grouped into sub-objective, "content cluster" or skill levels. The district should obtain such classifications prior to matching the items to the content-standards if possible. Such an existing detailed classification of assessment items will provide a structure for the district to follow in matching the assessment to the content-standards. Additionally, publishers of assessments have been known to provide "on demand" matches, particularly for larger districts, in which they will match their test to the district's content-standards.

> *There are procedures that can be used to collect evidence of content-related validity other than expert judgments (Mehrens and Lehmann, 1987, p. 78)*

The problem with these other procedures is that they are complicated and may require measurement expertise beyond that typically found in the district or even the area education agency. Generalizability theory, as presented in the DSRAS Manual (Iowa Department of Education, 1998), is one potential tool that could be used to investigate issues of content validity. For example, Mehrens and Lehmann present a context they call "content reliability" inspired by the research of Robert Ebel (1975, 1983). They suggest that one could construct two tests from the same pool of items (i.e., matching the same content-standards), give both tests to students and correlate the results. Correcting for the unreliability of error, the correlation between scores by these students on these two test forms would provide a "validity coefficient" per se. Again, these alternatives are possible but judgmental procedures will probably still be required.

> *Documentation of the judgments regarding the match between content-standards and the various assessment components is evidence of content validity.*

The steps a district must follow in order to document these judgments are fairly straightforward. First, the content-standards must be known and agreed upon. Second, judgments regarding the degree of match between the components of assessments currently being used and the content-standards and benchmarks should be compiled. These judgments should be summarized in a chart or table showing the

relationship between the currently used assessments and the content-standards. A full explanation of how the judgments were derived should also be included. Third, gaps or holes in the match should be highlighted. These will be used to select additional components of the assessment system. Fourth, district personnel should seek out additional assessment components to cover all other content-standards and as many benchmarks as possible. Finally, district personnel should evaluate the extent of the coverage of the content-standards from all components of the assessment overall. The purpose of this last step is to determine if the content-standards and benchmarks are being measured with enough depth (i.e., with enough items and tasks) such that accurate and reliable estimates of student performance are likely to be obtained.

> *While content-related evidence of valid score use and interpretation will probably be most important, other types of evidence are also allowed and desired. All such validity evidence should be collected if it is relevant.*

Criterion-Related Validity Evidence

Criterion-related validity evidence, according to the standards (APA, 1985, p. 11), attempts to answer the following question:

> *"How accurately can criterion performance be predicted from scores on the test?"*

The standards go on to state that the key to criterion-related validity evidence is the degree of relationship between the assessment items or tasks and an outcome criterion (APA, 1985, p. 11). Furthermore, this relationship must be systematic and predictable. Often, measurement experts trying to collect evidence of appropriate criterion-related score interpretations are confronted with several problems. First, the outcome criteria are usually determined by the purpose of the assessment. For example, often the degree of relationship between "end-of-course" exam results and final course grade (criterion) is disappointingly poor. However, is this the fault of the assessment results or the criterion?

> *Because the results of a district-wide standards-referenced assessment will be used for so many different purposes, the criterion measures will be difficult to determine and may often be in competition with each other.*

Secondly, if the criterion is performance on, say, the ACT Assessment (Ziomek and Svec, 1995), how does the district account for the fact that not all examinees will

necessarily participate on the criterion measure? Clearly, careful consideration regarding the purpose of the assessment and the definition of the criterion must be made. Mehrens and Lehmann (1987, p. 80) state:

- *"One of the most difficult tasks in a study of criterion-related validity is to obtain adequate criterion data. Gathering such data is often a more troublesome measurement problem than constructing the test..."*
- *"Criterion measures, like all other measures, must have certain characteristics if they are to be considered adequate. First of all, they should be relevant. A second desired characteristic of a criterion is that it be reliable."*

Independent of a clearly defined criterion, we would still like to see the results of, say, a science performance assessment agree, for the most part, with the results of a standardized science assessment. Hence, a district could correlate scores on both assessments and provide this as criterion-related validity evidence. In other words, if the inferences about student performance based on the science performance assessment were valid we would expect them to be in general agreement with the results from other measures of science content.

> *Unlike content-related validity evidence, which is essentially judgmental, criterion-related validity evidence is typically demonstrated via a correlation with a criterion measure.*

People collecting criterion-related validity evidence often cite two types of evidence: concurrent and predictive. The only difference between the procedures for collecting these two types of validity evidence is timing. Typically, concurrent evidence is collected from both the assessment and the criterion at the same time. An example might be seen when trying to relate the scores from a district-wide assessment to the ACT assessment. In this example, results from the district-wide assessment and the ACT assessment would be collected in the same semester of the school year (i.e., concurrently). Predictive evidence is usually collected at different times. For example, if the ACT assessment results were used to predict success in the first year of college, the ACT results would be obtained in the junior or senior year of high school whereas the criterion (say college grade-point average) would not be available until the following year. Hence, the correlation generated concurrently shows the relationship to the district-wide assessment as it exists now, whereas the correlation between ACT scores and college GPA is used to predict future college GPA.

When collecting criterion-related validity evidence, the district should be concerned not only with the reliability of the results, but must also consider the reliability of the criterion.

The criterion must be selected carefully and in light of the purpose for which the various assessments are designed. Many different pieces of information regarding the relationship between the assessments and the criterion are possible, but only those that are relevant will yield evidence of valid score use.

Construct-Related Validity Evidence

Collecting construct-related evidence of valid score use is probably the most difficult and misunderstood type of validation evidence typically reported. This might be due to the misunderstood and often misused term "construct" itself. Simply stated (Linn and Gronlund, 1995, p. 67):

"A construct is an individual characteristic that we assume exists in order to explain some aspect of behavior."

Linn and Gronlund explain that when we infer a particular individual characteristic from the assessment results, we are generalizing or making an interpretation in terms of some construct. For example, problem solving is a construct. When we infer that students who master the mathematical reasoning portion of an assessment are "good problem-solvers" we are interpreting the results of the assessment in terms of a construct. As such, we will need to demonstrate that this is a reasonable and valid use of the results.

The standards (APA, 1985, p. 10) suggest that construct-related validity evidence can come from many sources:

- *High inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain or construct;*
- *Substantial relationships between the assessment results and other measures of the same defined construct;*
- *Little or no relationship between the assessment results and other measures which are clearly not of the defined construct;*
- *Substantial relationships between different methods of measurement regarding the same defined construct;*
- *Relationships to non-assessment measures of the same defined construct.*

Using these guidelines, a district-wide standards-referenced assessment system should provide ample opportunity to collect construct-related validity evidence. Additionally, to the extent that the district has a good alignment between its content-standards and assessment components, the evidence should be strong. For example, if a particular content-standard is measured by several assessment components, then a correlation of the items and tasks from across these components should show a strong correlation. Furthermore, if these items and tasks do measure a particular content-standard, then the correlation among them should be higher than the correlation with other content-standards. This should be true despite the format of the items or tasks (i.e., it should not matter if they are multiple-choice items or open-ended tasks). Linn and Gronlund (1995, pp. 68-70) more explicitly define the process of collecting construct-related validity evidence. They say that there are three general steps in the process of construct validation:

- *Identifying and describing the meaning of the construct;*
- *Deriving hypotheses about the performance on an assessment from the theory underlying the construct;*
- *Verifying the hypotheses by empirical and logical means.*

Like most other validity evidence, the collection of construct-related evidence is a continuous and on-going process. It is paramount that the assessments be constructed in light of research regarding the construct being assessed. In addition, the underlying theory regarding how the construct is defined and how it is typically measured must be well understood. Finally, Linn and Gronlund (1995, pp. 69-70) provide the following guidelines:

- *Define the domain or tasks to be measured: well defined assessment specifications will aid in the understanding of the construct being measured;*
- *Analyze the mental process required by the assessment tasks: provide a "field test" or "pilot test" in which students describe how they answered the items. Builders of the assessment can then judge if the students are doing what they desired or if the items are evoking measures about the desired construct;*
- *Compare the scores from known groups of students: a simple comparison of the assessment results for a group of instructed and uninstructed students will reveal the degree to which the construct is being measured.*

- *Compare scores before and after some learning activity: we would clearly like to see continued improvement in the construct being measured as more learning takes place;*
- *Correlate the scores with other measures: the results of the current assessment purportedly measuring a defined construct should correlate highly with the results from another measure of the same construct.*

Before construct-related validity evidence can be obtained, steps must be taken to ensure the assessment measures the desired construct. This begins with research at the time the assessment is selected or constructed and is not independent of the alignment between the content-standards and the assessment components. Use the content-standards to provide an operational definition of the construct being measured. The construct must be meaningful and clearly defined. In addition to collecting correlational information regarding the degree of association among measures of the same construct as well as the lack of association to other constructs, districts could also collect information from currently instructed and uninstructed groups of students. A comparison of scores between these groups provides evidence that the instructed group "has something" the uninstructed group does not, namely the construct. For example, if students instructed in Algebra (i.e., those who have learned Algebra content) received similar scores as those without instruction, it is doubtful the test is measuring the "construct" of Algebra.

Consequential Validity Evidence

It is the opinion, of at least this author, that the goal of all education is improved student learning. It is also the author's opinion that the results of assessment should facilitate improved student learning.

> *Improved student learning will ultimately lead to a windfall of associated consequences (benefits).*

It should therefore be quite common to think about how a district-wide assessment system impacts not only student learning but teaching, as well as any other unanticipated consequences (such as drop out rate, student and teacher anxiety level, length of the work day, stress, etc.) Messick (1989, p. 20, Table 2.1) refers to such impacts as the general consequential basis of test score use and interpretation.

> *If an assessment system has been put into place to ultimately generate good consequences, to what extent has it fulfilled its mission?*

Linn and Gronlund (1995, p. 72) point out that considerations regarding the consequences of assessment score use and interpretation are clearly evident in the move toward "authentic" performance-based assessments. This includes both the intended use of the assessment (i.e., a better look at actual student performance) as well as the unintended consequences (such as delayed reporting time due to the judgmental scoring process required). Linn and Gronlund (1995, p. 73) suggest that the consequences of an assessment be considered in light of the following:

- *Do the assessment tasks address key learning objectives or content-standards? Emphasis on important and not secondary aspects of the content-standards is a desirable consequence.*
- *Is there reason to believe that students study harder in preparation for the assessment? Increased student motivation is a desirable consequence.*
- *Does the assessment artificially constrain the focus of the student's study? The narrowing of the content-standards through over-emphasis or "drill and practice" activities is an undesired consequence.*
- *Does the assessment encourage creative modes of expression? Students exploring new ideas is a desirable consequence.*

Systematically generating a list of questions like those presented by Linn and Gronlund (1995) could help in documenting the consequential aspects of an assessment. Districts should ensure that only the most relevant issues are addressed and that it is in the best interests of the students to participate in and succeed on the assessment. Attitudinal surveys could capture other unanticipated consequences. Additionally, attitudes of the teachers and community and changes in those attitudes over time would provide additional consequential-validity evidence. Finally such varied things as attendance, enrollment in extra-curricular activities, participation in science fairs, and membership in academic clubs could all be evidence of the consequences of an aggressive district-wide assessment system aimed at improving student learning.

Summary

This chapter has defined validity as the evidence generated in support of appropriate use and interpretation of the results, of an assessment. The results of an assessment, and the use or interpretation of those results are not independent from the purpose of

the assessment. While validity was discussed as a unitary concept, four different types of validity evidence were discussed.

> *First, content-related validity evidence questions the degree to which the assessment results are interpreted appropriately regarding the content-standards and what aspects of the content-standards are to be assessed.*
>
> *Second, criterion-related validity evidence asks the question of how well results of the assessment agree with or could be used to predict a criterion outcome measure.*
>
> *Third, construct-validity evidence relates the results of the assessment to individual student characteristics and does so in a way that is clearly understood.*
>
> *Finally, for consequential validity evidence, districts should document the consequences of the assessment, both intended and unintended.*

Districts should further strive to develop assessment systems that provide for the best consequences. While the concepts of both reliability and validity are related, this section has clearly documented the differences between the two. Reliability is primarily a mathematical concept that is empirical in nature and is a necessary condition for validation. The concept of validity, on the other hand, is primarily judgmental and a unitary concept, requiring different types of evidence.

III. Glossary of Terms

The following glossary of terms as used in this document is provided to assist the reader regarding language that may not be familiar. Where possible, terms were taken from existing documentation from the Iowa Department of Education.

Alternate Forms

Two tests constructed to measure the same thing from the same table of specifications and to the same required psychometric and statistical properties. These forms are not strictly parallel due, primarily, to differences in the statistical properties of the two forms.

Annual Improvement Goals

Goals which describe the district's desired rate of improvement for students.

Benchmarks (Major Milestones)

Major milestones, which specify skill or performance levels a student needs to accomplish.

Content-Standards

Content-standards describe the goals for individual student achievement. Content-standards specify what students should know and be able to do in identified disciplines or subject areas.

Consequential Validity

Evidence that the implemented assessment or assessment system results in the planned and desired consequences and that unanticipated consequences do not detract from the goal of the assessment.

Construct Validity

Evidence that performance on the assessment tasks and the individual student behavior that is inferred from the assessment shows strong agreement, and that this agreement is not attributable to other aspects of the individual or assessment.

District-wide Assessment

A large-scale, academic achievement assessment.

General Curriculum

A description of the content-standards and benchmarks adopted by an LEA or schools within an LEA that applies to all children. It is the basis of planning instruction for all students.

Generalizability Theory

A procedure for the study and classification of the components of error.

Internal Consistency Reliability Estimate

A statistic, which represents the correlation between scores obtained from one measure when compared to scores obtained from the same measure on another occasion. Typically this estimate comes as a correlation between different halves of the same test (split-half method), thereby requiring only one test administration.

Inter-Judge Agreement

Consistency statistics describing the relationship or degree of agreement between two or more judges scoring an open-ended assessment.

Inter-rater (Inter-reader) Reliability

Consistency statistics describing the relationship between scores on an open-ended assessment assigned by more that one judge. Typically these statistics are a simple correlation between judges, but other more sophisticated estimates are possible.

Inter-Judge Consistency

See inter-rater reliability and inter-judge agreement.

Parallel Forms

Two tests constructed to measure the same thing from the same table of specifications with the same psychometric and statistical properties. True parallel test forms are not likely ever to be found. Most attempts to construct parallel forms result in alternate test forms.

Reliability Coefficient

A mathematical index of consistency of results between two measures expressed as a ratio of true-score to observed-score. As reliability increases, this coefficient approaches unity.

Standard

A clear statement that expresses what students are expected to know and be able to do. In Iowa, local school districts and communities are responsible for setting high quality standards.

Standard Error of Measurement

Statistic which expresses the unreliability of a particular measure in terms of the reporting metric. Often used to place score-bands or error-bands around individual student scores.

Test / Retest Reliability Estimate

A correlation between scores obtained from one measure with scores obtained from the same or parallel measure on another occasion.

True Score

That piece of an observed student score that is not influenced by error of measurement. The true-score is used for convenience in explaining the concept of reliability and is unknown in actual assessments.

Validity

A psychometric concept associated with the use of assessment results and the appropriateness or soundness of the interpretations regarding those results.

IV. References

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole Publishing.

American Psychological Association (1985). *Standards for educational and psychological testing.* Washington, D. C.: American Psychological Corporation.

Berk, R. A. (1984). *A guide to criterion-referenced test construction* (Ed.) Baltimore: Johns Hopkins University Press.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Programs.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cronbach, L. J. (1951). Coefficeint alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: Massachusetts Institute of Technology.

Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practices, 2*, 7-10.

Ebel, R. L. (1975). *Prediction? Validation? Construct Validity*? Mimeograph.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement*, 3rd Edition. Washington D. C.: American Council on Education.

Frisbie, D. A. (1991). *Essentials of educational measurement (5th ed.)* Prentice Hall: Englewood Cliffs, NJ.

Harcourt Brace Educational Measurement (1997*). Stanford Achievement Test Series, Ninth Edition: Technical data report* (pp. 207-220). San Antonio, TX: Harcourt Brace Educational Measurement.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Bray, G. B., Lewis, J. C., & Qualls, A. L. (1996). *Norms and conversion tables with technical information, ITBS Form M Complete Battery, Levels 5-14*. Chicago: Riverside Publishing.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 254-264.

Iowa Department of Education (1998). Implementing a district-wide standards-referenced assessment system (DSRAS). Iowa City: National Computer Systems.

Iowa Testing Programs (ITP 1997-1998 Revision). *Interpretive supplement for the achievement levels report*. The University of Iowa, Iowa City.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed*.), Educational measurement*, 3rd Edition, pp. 485-514. New York: American Council on Education / Macmillan.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement in assessment and teaching*, 7th edition. New Jersey: Prentice-Hill.

Mehrens, W. A., & Lehmann, I. J. (1987). *Using standardized tests in education*, 4th Edition. New York: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* Washington D.C.: American Council on Education.

Nitko, A. J. (1989). Desigining tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement*, 3rd Edition, pp 447-483. New York: American Council on Education / Macmillan.

SAS© (1990). *SAS© Institute Inc., SAS© Procedures Guide, Version 6, Third Edition.* Cary, NC: SAS© Institute (SAS© is a registered trademark of the SAS© Institute).

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, NJ: SAGE Publications.

SPSS for Windows© (1995). SPSS© Inc., SPSS for Windows©, Standard Version, Release 6.1.3. Chicago: SPSS© Inc (SPSS© is a reistered trademark of SPSS© Inc.).

Traub, R. E. (1994). *Reliability for the social sciences: Theory and application.* Thousand Oaks, CA: SAGE Publications.

Ziomek, R. L., & Svec, J. C. (1995). *High school grades and achievement: Evidence of grade inflation.* ACT Research Report Series, 95-3. Iowa City: The American College Testing Program.

V. Index