Checklist for Evaluating K-12 Assessment Programs

Gregory J. Cizek, Lorrie A. Schmid, Audra E. Kosh, and Amy A. Germuth

Since the first tests mandated for high school graduation in the 1970s, student achievement testing has taken on increasing importance in American schools. In the decades following the 1970s, the breadth and stakes associated with performance on student achievement tests have also increased. In 2001, the scope of testing widened further with the requirements of the No Child Left Behind Act (NCLB, 2001) which mandated annual testing of every U.S. student in reading and mathematics in grades three through eight and at least once in grades ten through twelve, as well as annual, every-pupil testing in science at least once in each of three grade spans (3-5, 6-9, and 10-12). Those mandates remain in place with the successor to NCLB, the Every Student Succeeds Act (ESSA, 2015), although at the time this checklist was produced, final regulations had not been released.

The accountability provisions that are often associated with mandated testing in elementary and secondary school contexts have highlighted the consequences that such tests can have for students, school employees, and school systems. As a result, educators, parents, researchers, and policymakers have recognized that such tests must be held to high standards of technical quality. Indeed, in the late 1970s, the Evaluation Center at Western Michigan University coordinated the development and dissemination of a checklist for evaluating large-scale assessment programs (Shepard, 1977). Checklists are appropriate for identifying the factors, components and/or dimensions needed to perform a task. As Scriven (2007) has noted, checklists have great utility for evaluating tests including: reducing the chance of leaving out important components, easy to use and understand, judgment occurs by merit rather than influence, and a large amount of knowledge can be condensed.

Much has changed since the publication of that initial checklist. For example, in the 1970s, many testing programs relied on commercially-developed, off-the-shelf, norm-referenced tests (NRTs). The term standards-referenced had not yet come into use. In the 1970s, the National Assessment of Educational Progress (NAEP) was just being developed, many states did not have statewide every-pupil testing, most did not have state-adopted content standards, and few had accountability systems that relied as heavily on achievement test results as is seen today. In addition, the results of NRTs typically did not have meaningful consequences for students or educators. The ubiquity, prominence,

and stakes associated with current tests are characteristics that illustrate how today's K-12 assessment systems bear little resemblance to previous generations of large-scale testing programs.

The influence of individual legislative initiatives such as NCLB is likely to wax and wane. However, if the trend begun in the 1970s is instructive, it seems likely that large-scale achievement testing will continue to be a prominent feature of mandated assessment and accountability systems. Accordingly, it seems as important as ever—or more so—that assessments used to measure educational achievement in K-12 contexts meet high standards of technical quality. Such assessments should be routinely evaluated so that the information they yield can be confidently, accurately, and dependably interpreted by students, parents, educators, and others who rely on assessment information to inform decisions about educational planning, promotion, selection, graduation, staffing, funding, resource allocation, reform initiatives, and other decisions.

Most of these decisions are commonly made using conjunctive models; that is, the decisions are based on multiple criteria, each of which must be satisfied. For example, high school graduation decisions are often conjunctive to the extent that criteria such as completion of specific coursework, attendance requirements, accumulation of credit hours, grades, and so on are considered, with failure to satisfy any one of the criteria resulting in a negative graduation decision. However, the results of large-scale assessments used in K-12 contexts are often the most visible or simply the last of the conjunctive hurdles and they rightfully attract scrutiny by all concerned that they contribute to sound decisions.

Purpose

The purpose of this document is to aid in the evaluation of assessments used in K-12 contexts to measure student achievement. To be clear, this document focuses only on evaluation of the assessments in an assessment program; that is, the scope of this document is narrower than the earlier checklist (Shepard, 1977), which also included evaluation of cost, program management, and other aspects of an assessment program broadly conceived. We believe that, although such aspects are important, the technical quality of an assessment program—that is, does it permit confident conclusions about student learning—is the ultimate criterion, and that evaluation of this aspect alone is warranted given the evolution and complexity of large-scale student achievement testing in schools.

This checklist provides a list of practices that K-12 assessment programs should follow; however, it should be noted that this document does not necessarily support quality judgments about an assessment program. That is, an assessment program may observe all of the checklist criteria while failing to adhere to thresholds of acceptable practices within the educational measurement field on those same criteria. For example, an assessment program may succeed at reporting reliability estimates for a variety of examinee subgroups, yet those estimates may indicate extremely low reliability of the test. Thus,

assessments programs that observe the elements on this checklist cannot be determined to be successful or not solely by observing all of the elements in the checklist. Nonetheless, this document provides a framework for comprehensively evaluating K-12 assessment programs.

Intended Audience

The intended audience for this document includes those who develop and administer such tests, interested stakeholders such as educators and policymakers, and ultimately, all of those affected by the results of K-12 assessment programs.

Development of the Checklist

This version of the Checklist for Evaluating K-12 Assessment Programs is an update of the version published in 2012; the updates in this version primarily reflect changes from the 1999 to 2014 editions of the Standards for Educational and Psychological Testing (AERA, APA, NCME). To develop this document and the associated Assessment Evaluation Checklists, we reviewed several sources. We began by reviewing the first checklist for large-scale testing programs produced by Shepard (1977). We also reviewed current literature on the topic of professional standards for large-scale testing, including the work of Plake (2002) and Yen and Henderson (2002); policy statements and guidelines (e.g., AERA, 2000; CCSSO & ATP, 2010), and we relied on important handbooks on educational measurement, for example, Educational Measurement (Brennan, 2006).

The specific citations for the evaluation criteria included in the checklists based have been drawn from four broadly accepted sources of standards for best assessment practice, including

I. the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014);
II. the Standards and Assessment Peer Review Guidance (USDOE, 2009);
III. the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004); and
IV. the Rights and Responsibilities of Examinees (Joint Committee on Testing Practices, 1998).

Taken together, these resources provide convergent guidelines that can be used to evaluate any large-scale achievement testing program used in elementary and secondary school settings, including statewide assessment programs, district or local assessment programs, or commercially-produced tests for measuring student achievement. We note also that, in developing this Checklist for Evaluating K-12 Assessment Programs, we have purposefully included only guidelines and standards that transcend specific legislation. For example, although we have included the Standards and Assessment Peer Review Guidance (USDOE, 2009) produced to support the implementation of NCLB (and ESSA), we have cited only the portions of that document that would be relevant to all large-scale assessment systems; we have focused on the key elements of sound testing that would remain relevant in the face of changes that will likely be contained in the final regulations for ESSA or accompany subsequent reauthorizations of the broader Elementary and Secondary Education Act (ESEA).

Following this introductory section are sections that contain of brief descriptions of five key aspects of sound K-12 assessments and Assessment Evaluation Checklists that can be used to gauge the extent to

which a large-scale student achievement testing program meets generally accepted criteria. These five key aspects include

1) test development;
2) test administration;
3) reliability evidence;
4) validity evidence; and
5) scoring and reporting.

Of course, to some extent, all aspects of a K-12 assessment program ultimately support the accurate interpretations of scores yielded by the examinations—that is, validity. For example, the procedures used in test development, the test administration conditions, and so on, are sources of evidence that can either strengthen or weaken support for the validity of scores on an examination. Thus, the five aspects used to organize this document and checklists cannot be examined in isolation; their discrete presentation here is primarily for the convenience and accessibility by the user. Additionally, it should be noted that the Assessment Evaluation Checklists are intended to be used primarily in a formative manner to enable those responsible for K-12 assessment programs to identify strengths of their programs and areas for improvement.

The Assessment Evaluation Checklists each contain several succinct statements that capture the fine-grain elements that comprise each of the five key aspects of sound K-12 testing programs. When applicable to a particular program, the statements represent the criteria that should be applied when evaluating each aspect. The Assessment Evaluation Checklists also provide a format for indicating the extent to which available evidence provides support for each element subsumed by the aspect. Elements within each checklist are accompanied by the following scale:

- ☐ O = Observed
- ☐ N = Not Observed
- ☐ NA = Not Applicable

Using the Checklists

To use one of the Assessment Evaluation Checklists, an evaluator considers the individual element in light of the evidence related to that element, then indicates whether the element was observed (O), not observed (N), or not applicable (NA). An area for elaboration or comments is provided at the bottom of each checklist.

This document concludes with a reference list, a crosswalk between the elements in each checklist and the professional guidelines upon which the elements are based (see Appendix A), copies of the relevant professional guidelines (see Appendices B, C and D), and biographical information for the authors of this document (Appendix E). Finally, a related document has been produced for evaluating licensure and certification testing programs. Those interested in credentialing contexts should refer to A Checklist for Evaluating Licensure and Certification Testing Programs (Cizek, Germuth, Schmid, & Kosh, 2015).

Questions about this document, suggestions, or recommendations should be directed to Gregory J. Cizek (cizek@unc.edu).

Section I: Test Development

The procedures employed in the course of developing a test are the foundation for instruments that will yield reliable and valid scores and these procedures typically provide the most extensive source of evidence about the quality of the test. The development of a test is guided by clear, explicit statements regarding the purpose of the test and the inferences that are intended to be made from the test scores. Test development is the process of creating all aspects of a test—test specifications, items, formats, directions for examinees, practice materials, scoring procedures, test administration procedures and other instructions–and putting them together in a purposeful way designed to accomplish the overall aims of the test.

The Standards for Educational and Psychological Testing (2014) list six steps in the test development process, including:

1) consideration of expected interpretations for intended uses of the scores to be generated by the test;
2) specification of the content and format of the test;
3) specification of test administration and scoring/reporting procedures;
4) development of items or tasks aligned to the specifications;
5) screening of items or tasks according to criteria appropriate to the use of the test; and
6) development and implementation of procedures for scoring responses to the items and tasks, and for reviewing and evaluating the test as a whole. (AERA, APA, NCME, 2014, p. 75).

The six steps can be collapsed into four primary phases. The first phase of test development is delineating the domain to be assessed. Typically, test developers start with a stated need or purpose for the test, an explicit statement of the intended score interpretations and uses, and a description of the characteristics of anticipated test takers. For educational achievement tests, the domains to be delimited are routinely those defined by specific grades and subjects, such as fourth grade English language arts, sixth grade mathematics, or high school biology. Most commonly, those with specific expertise in a content area and knowledge of the intended examinee population as well as other important stakeholders are empaneled for this phase. Participants might include teachers of the relevant grade and subject (primarily), curriculum specialists, administrators, educators with knowledge of English-language learners, special needs students, and, in some cases, persons representing business, policy, or parent groups. The product from this phase is a delineation of the specific objectives, typically referred to as content standards that will be assessed.

The second phase builds on the delineation of the domain and subdomains by establishing test specifications. Test specifications further refine and illustrate the testing domain by detailing content to be covered by the test, the number of items, tasks or questions that will comprise a test form, the formats

that will be used (e.g., multiple-choice, constructed-response, performance), and the acceptable response modes. Specifications also typically include targets for psychometric properties such as overall difficulty, reliability, and other statistical indices. In this phase, relevant item and task administration and plans for how test scores will be reported are also developed.

The third phase of test development is the actual creation and selection of test items, performance tasks, essay prompts, and so on. If developed within a state or local education agency, many of the same perspectives represented in the domain-specification phase would also be tapped for item and task development. If developed by a publisher or testing company, in-house content expertise may be used, although the resulting items and tasks would likely be reviewed by qualified state or local personnel before being accepted for use on an assessment. Developed items, tasks, scoring rubrics and so on are collected into a database commonly referred to as an item bank or item pool.

All items or tasks eligible to be selected for a test must first be screened and evaluated both qualitatively and quantitatively. Each item should be assessed qualitatively to make sure that there is a clear stimulus, that the item is appropriate to the purpose of the test, and that it is well-aligned to the content standards. Items should be reviewed to ensure grammatical correctness, absence of offensive language, and for potential bias. Quantitative evaluations consider statistical indices such as item difficulty, item discrimination, readability, and other factors.

When a new test is being developed, the item review and selection process usually includes field testing the test materials and items or tasks with a sample of the intended student population. The test development process should document the procedures and actions of the field testing. It should also review the characteristics of the field test participants to make sure that they match characteristics of the intended examinees defined in the domain.

The last phase of test development involves assembly and evaluation of the test for operational use and planning for evaluating the performance of items/tasks. In this phase, the test developer identifies and selects the best items or tasks based on test specifications and psychometric properties. Assembly of the test according to specifications should be documented as part of the total test system. Finally, the developer should assess the entire test cycle by looking at the strengths and weaknesses of the entire process and initiating improvements as necessary. Periodic review of the items, scores, and purpose of the test are needed to amend or revise the test development process as conditions change. Table 1 provides the Test Development Checklist.

Table 1 - Test Development Checklist

O = Observed N = Not observed NA = Not applicable

O	N	NA	Evaluation Elements
☐	☐	☐	TD1 - The need/purpose of the test is clearly stated.
☐	☐	☐	TD2 - The appropriate population of intended examinees is specified.
☐	☐	☐	TD3 - The appropriate and inappropriate uses of the test is specified.

O	N	NA	Evaluation Elements
☐	☐	☐	TD4 - Domain definition is based on systematic, scientific procedures (e.g. curriculum review).
☐	☐	☐	TD5 - Content standards have been developed and are documented.
☐	☐	☐	TD6 - Content standards development involved relevant stakeholders.
☐	☐	☐	TD7 - A detailed domain definition is provided and the specific objectives, content standards, and performance levels that are to be measured are documented.
☐	☐	☐	TD8 - A psychometrically sound method for determining test content specifications is used and the process is documented.
☐	☐	☐	TD9 - Performance level labels (PLLs) and performance level descriptions (PLDs) clearly describe student achievement.
☐	☐	☐	TD10 - The type of assessment (e.g., standards-referenced, norm-referenced, augmented), test content, length, time allotments, item/task formats, and any section arrangements are described.
☐	☐	☐	TD11 - The item/task development process is documented.
☐	☐	☐	TD12 - The qualifications of item/task developers is documented.
☐	☐	☐	TD13 - Evidence is gathered to determine that items and tasks are written to the grade level targeted by the test.
☐	☐	☐	TD14 - A rigorous item/task review process is followed and documented.
☐	☐	☐	TD15 - The qualifications of item/task reviewers is documented.
☐	☐	☐	TD16 - Test administration procedures, including clear directions and information regarding the rights and responsibilities of examinees are provided.
☐	☐	☐	TD17 - Computer-based/computer-adaptive testing protocols (e.g., item selection algorithms) is documented.
☐	☐	☐	TD18 - Items and tasks are field-tested with relevant populations.
☐	☐	☐	TD19 - Field-test procedures match expected administration conditions.
☐	☐	☐	TD20 - The characteristics of the field-test sample and method of selection are described.
☐	☐	☐	TD21 - Field-test sample characteristics match the intended test population as closely as possible.

O	N	NA	Evaluation Elements
☐	☐	☐	TD22 - Information on the psychometric properties of items/tasks (e.g., alignment, difficulty, discrimination) are reported.
☐	☐	☐	TD23 - Information on the psychometric properties of test forms (e.g., reliability, validity) is reported.
☐	☐	☐	TD24 - Assessments are designed to yield scores that reflect the full range of student achievement.
☐	☐	☐	TD25 - A system is in place for continuous monitoring and improvement of content standards, assessments, and the assessment program.
☐	☐	☐	TD26 - Procedures used to translate a test to alternate languages are documented.

Comments:

Section II: Test Administration

Educational achievement tests are designed to measure the extent to which a student has mastered prescribed content for a given grade and subject. Where performance level labels (PLLs) and performance level descriptions (PLDs) are used, student achievement is also translated into categories that classify students into hierarchical levels of performance. Students and their parents use test results to gauge their learning; educators use test results to diagnose leaning difficulties, to monitor, plan, and individualize instruction, and to inform decisions about placement, promotion, graduation, and other decisions. Finally, legislators and policymakers use test results to inform resource allocation and to initiate and evaluate educational reforms. Each of these uses demands that the information about student achievement be as reliable and valid as possible, and that the potential for errors in the assessment process be minimized. Test administration procedures and the test environment are essential prerequisites for obtaining sound assessment results that can be interpreted as faithful indicators of what students know and can do. Thus, test administration procedures should be evaluated in order to identify potential sources of undependability or inaccuracy.

This section is structured around three components of test administration, including:

1) test administration personnel responsibilities;

2) test security procedures; and

3) the testing environment and test administration process.

Each of these components consists of a number of elements that are considered important by relevant testing and evaluation standards. Testing personnel include all of those responsible for test administration, which would typically include the classroom teacher, but could also include proctors, test directors, and others involved in the handling and distribution of test materials and the actual administration of a test, whether in paper and pencil or computer-based mode of administration. Testing personnel responsibilities include ensuring that all students take the test in the same manner unless testing accommodations are required. If so, such accommodations should be provided for all students who need them and any alternate or modified assessments should be administered to the students for whom they are intended. Test administration personnel are responsible for ensuring that both students and those who handle test materials follow all required procedures. Thus, test administrators must be provided with training that prepares them to handle a variety of situations, including how to handle emergencies that may occur during testing, and how to report testing irregularities.

Security breaches can compromise inventory of items and tasks available for future testing; they can result in a substantial financial loss to the state or district; they jeopardize the validity of test scores; and, most importantly, they can result in inaccurate information about student learning and incorrect educational decisions Thus, attention to the test administration process requires that specific test security procedures be in place. These procedures should include ensuring that students (or adults, where proscribed) do not have inappropriate prior access to test materials, making certain that test materials are handled in a secure manner at all times, and providing advance information to all of those involved in the testing process (e.g., students, teachers, proctors, etc.) regarding what materials and actions are permissible/impermissible during the testing process.

Finally, the testing environment and test administration process should maximize each student's ability to accurately demonstrate his or her level of knowledge and skill. Test administrators should be available to assist examinees with procedural matters, to answer students' questions, to make sure that all students understand the purpose, directions, and mode of responding to the test, and to ensure that all security policies are followed. If a test is a timed test (i.e., speeded) or if there is a penalty for guessing applied when the test will be scored, students should be informed of these characteristics. Some mechanism should be in place to assist students in monitoring their pacing. For example, if the test administrator or computer interface does not provide this information, then clocks or other time devices should be available and visible to all examinees. Students should be advised of test-taking strategies that will facilitate their best performance when a correction-for-guessing is applied or when simple number-correct scoring is used. Accommodations for individuals with disabilities and, as appropriate, language accommodations, may be necessary to assist examinees to demonstrate their knowledge and skill. All examinees should receive comparable and equitable treatment in a comfortable, safe, environment with minimal distractions. Table 2 provides the Test Administration Checklist.

Table 2 - Test Administration Checklist

O = Observed N = Not observed NA = Not applicable

O	N	NA	Evaluation Elements
☐	☐	☐	TA1 - The purpose and use of the test are communicated to test takers.
☐	☐	☐	TA2 - Eligibility requirements for examinees are documented.
☐	☐	☐	TA3 - Examinees are provided with test administration regulations and procedures.
☐	☐	☐	TA4 - Reasonable test preparation materials are available to examinees free of charge.
☐	☐	☐	TA5 - Examinees are informed of procedures for obtaining technical assistance and/or registering complaints.
☐	☐	☐	TA6 - Examinees are made aware of the availability of appropriate testing accommodations or, where appropriate, alternate assessments for students with limited English proficiency or cognitive disabilities, so that all students are able to participate in the assessment program.
☐	☐	☐	TA7 - The eligibility review process for test accommodations or alternate assessments are conducted by qualified individuals and are documented.
☐	☐	☐	TA8 - Requests for accommodations or alternate assessments and supporting documentation are kept confidential.
☐	☐	☐	TA9 - Test adaptation and retake policies and procedures are documented.
☐	☐	☐	TA10 - Qualifications, responsibilities, and training of test administrators for all assessments are documented.
☐	☐	☐	TA11 - A test administration manual containing standardized procedures for administration of all assessments are produced.
☐	☐	☐	TA12 - A process for test administrators to record test irregularities is in place.
☐	☐	☐	TA13 - Examinees are informed of score reporting procedures.
☐	☐	☐	TA14 - Procedures for dealing with emergencies are provided to test administrators and test takers.
☐	☐	☐	TA15 - All secure test materials are shipped, handled, transmitted and/or maintained appropriately.
☐	☐	☐	TA16 - Procedures to ensure that examinees provide their authorization prior to taking the test are in place.

O	N	NA	Evaluation Elements
☐	☐	☐	TA17 - A policy on required and prohibited materials for taking the test are developed and communicated to test takers.
☐	☐	☐	TA18 - Examinees are informed regarding actions that constitute a security breach.
☐	☐	☐	TA19 - Test administrators are trained in procedures for handling security breaches.
☐	☐	☐	TA20 - Procedures for secure administration of internet or computer-based examinations are established.
☐	☐	☐	TA21 - Examinees are provided with information about their rights and responsibilities.
☐	☐	☐	TA22 - Examinees are monitored during the testing process.
☐	☐	☐	TA23 - The testing environment is comfortable and has minimal distractions.
☐	☐	☐	TA24 – The test is administered in the language that is most appropriate for the intended population and intended purpose of the test. If an interpreter is used, the interpreter is fluent in the language of the test, the student's native language, and the content of the test.

Comments:

Section III: Reliability Evidence

According the Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014), reliability represents the "consistency of scores across replications of a testing procedure" (p. 33). Reliability estimates quantify the degree to which test scores are dependable, reproducible, and consistent measurements of the construct being assessed versus attributable to random errors of measurement. Thus, reliability is a characteristic of the test scores rather than the test itself. It is essential to investigate the reliability of test scores to gain information about the extent to which any individual score should be considered to be a dependable estimate of a student's true level of knowledge, skill, or ability.

Reliability estimates can be obtained in several ways, depending on the nature of reliability evidence desired. One type of reliability estimate is the coefficient of stability, also referred to as test-retest reliability. Test-retest reliability is the correlation between scores from administrations of the same test on separate occasions. It is an indication of how stable measurements are over a period of time. Another method used to estimate reliability is the coefficient of equivalence, or parallel forms reliability. The coefficient of equivalence is an indication of the degree to which scores obtained on two forms of a test can be used interchangeably. A third, and perhaps the most commonly-encountered estimates of reliability for student achievement tests are estimates of internal consistency. Estimates of internal consistency provide information about the extent to which items or tasks within a test provide consistent information about test takers. Put another way, internal consistency reliability estimates gauge how dependable a collection of items is at measuring a common construct of interest, such as computational skill or reading comprehension. For example, a teacher might find it valuable to know that students' total test scores provide reliable information about their knowledge and skill in science, but also that smaller, more homogeneous subsets of items (i.e., subtests) on a science test (e.g., clusters of question on life science, physical science, earth science, etc.) produce less dependable information about students' learning in those more narrowly defined areas. Methods used to estimate internal consistency include KR-20, Cronbach's alpha, and others. Many factors affect the reliability of scores, including the homogeneity of the examinee group, examinee motivation, attention, and effort, the clarity of the test directions, items, tasks, scoring rubrics, the test administration conditions, the effectiveness of proctoring, the objectivity of scoring, the effectiveness of rater training, the extent to which the test is speeded, as well as other factors.

The standard error of measurement (SEM) is related to reliability and can be conceptualized as "the average error of measurement ... estimated over some population" (AERA, APA., NCME, 2014, p. 34). The SEM can be used to create confidence intervals around test scores and gives an idea of how much random error is likely to contribute to a student's score. A conditional standard error of measurement (CSEM) can also be reported and is considered essential when performance standards (i.e., cut scores) are used on a test. For example, when certain scores are established to distinguish between Basic, Proficient, and Advanced performance levels, the CSEM helps to express how much random error likely affects scores at those cut points, allowing users to consider the probability that a student scoring at the Basic performance level but just below the cut point for Proficient, might more correctly be classified at the higher level (or vice versa).

Another way of expressing reliability that is appropriate when cut scores are used to classify student achievement into categories called decision consistency. Indices of decision consistency provide an indication of the dependability of the categorical classifications.

Interrater reliability is a special kind of reliability by which the dependability of rater judgments can be estimated, for example, by examining the correlation (i.e., consistency in rank ordering) between the scores assigned by different raters to essays or performances. Interrater agreement is different from interrater reliability in that interrater agreement is a measure of the extent to which raters actually assign the same scores in their evaluations of examinees' performances (i.e., not just similar rank ordering). A high degree of interrater reliability may or may not be accompanied by high interrater agreement. Interrater reliability and interrater agreement are particularly appropriate for student achievement tests when performance tasks or constructed-response items are included and those items or tasks are scored by human raters according to a scoring guide or rubric. In such cases it is typically of interest to estimate how much confidence can be placed in the rubric score assigned to the students' performances. Table 3 provides the Reliability Evidence Checklist.

Table 3 - Reliability Evidence Checklist

O = Observed N = Not observed NA = Not applicable

O	N	NA	Evaluation Elements
☐	☐	☐	R1 - Reliability estimates for each test version, overall score, subscores, combined scores, and subgroups of test takers are provided.
☐	☐	☐	R2 - The samples used for obtaining reliability evidence are described and the intended population of individuals or groups for whom the reliability evidence is intended to apply is identified.
☐	☐	☐	R3 - The method(s) used to estimate reliability (e.g., test-retest, parallel forms, internal consistency, generalizability theory) are described.
☐	☐	☐	R4 – Differences between reliability estimates are clearly described and different estimates are not used interchangeably.
☐	☐	☐	R5 - Standard errors of measurement are reported for each reported score and subscore, and in the appropriate units (e.g., raw or scaled score units).
☐	☐	☐	R6 - Conditional standard errors of measurement are reported for critical score points (e.g., cut scores).
☐	☐	☐	R7 - Information on the extent, if any, of test speededness is presented.
☐	☐	☐	R8 - Where relevant (e.g., if constructed-response, polytomously-scored items or tasks are included), information on interrater reliability and agreement in scoring is provided.

O	N	NA	Evaluation Elements
☐	☐	☐	R9 - Where relevant, evidence related to task-to-task consistency in scoring and within-examinee consistency in scoring is presented.

Comments:

Section IV: Validity Evidence

According to the Standards for Educational and Psychological Testing, validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA/APA/NCME, 2014, p. 11; for a more elaborated and contemporary explication of validity, see Cizek, 2012). Like reliability, validity is not a characteristic of a test, but of the scores produced by a test. Validity evidence provides support for the intended inferences or meaning to be made from examinees' scores. Simply put, validity is the extent to which differences in the scores examinees obtain on a test can be interpreted to reflect real differences among the examinees in the characteristic measured by the test. For example, strong validity evidence would allow users of scores on a fourth grade mathematics achievement test to be confident that students scoring higher on that test had mastered more of the state's 4th grade mathematics content standards than students scoring lower. Validity evidence cannot be separated completely from reliability evidence; adequate reliability evidence is a necessary prerequisite for examining the validity of test scores.

Before proceeding, some foundational aspects of validity are summarized:

1) Validity pertains to the intended test score inferences; that is, validity concerns the interpretations that the entity responsible for the test asserts can be made from the test scores. Thus, the intended inference(s) should be explicitly stated by the entity responsible for the test, and validity evidence should be considered for each intended inference.

2) Validity is a unitary concept; all validity evidence supports the intended test score inferences. The Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014), provides a list of possible sources that can be examined for evidence pertinent to validating the intended test score inferences.

3) Validity is not an all-or-none consideration, but a matter of degree. Evaluations and judgments about validity should be based on the accumulation of evidence for an intended interpretation.

4) Validation often can involve a joint effort of the test developer and test users to gather and evaluate pertinent data. As such, validation should be considered an ongoing endeavor; validity evidence should be continually gathered and evaluated.

It is not an exaggeration to say that, in essence, every aspect of a testing program bears on the validity of test score inferences. Each aspect of the testing program—from test development to test administration—can provide either supporting evidence for the intended inference or weaken validity claims. Consequently, it would not be wholly appropriate to consider this section as a stand-alone checklist for evaluating validity evidence. Rather, the entire body of supporting evidence in this document must be considered when evaluating validity claims. However, in an effort to be as parsimonious and to avoid overlap with information presented in other sections, this section provides a checklist for evaluating validity evidence by focusing on two questions:

1) Does the test content and analyses of the relationships between test scores and other relevant variables support the intended licensure or certification inference regarding the standing of the examinee related to the knowledge, skills, and abilities deemed appropriate for awarding the credential he or she seeks? and

2) Does the test provide pass/fail or other categorical discriminations that reflect individuals' varying levels of competence?

Finally, we note that even rigorous evidence gathering in support of an intended score inference—that is, validation—does not necessarily mean that a test should be used for a given purpose. Entities responsible for credentialing testing programs are encouraged to gather relevant evidence to justify score use for specific purposes (see Cizek, 2012). Table 4 provides the Validity Evidence Checklist.

Table 4 - Validity Evidence Checklist

O = Observed N = Not observed NA = Not applicable

O	N	NA	Evaluation Elements
☐	☐	☐	V1 - A clear statement regarding the testing purpose or intended test score inferences is provided.
☐	☐	☐	V2 - Appropriate uses for test scores are specified.
☐	☐	☐	V3 - Reasonably anticipated inappropriate interpretations or uses of test scores are identified and avoided.
☐	☐	☐	V4 - Constructs measured by the test are clearly identified and explained.
☐	☐	☐	V5 - The intended population for the test is specified, including applicability of the test to specific subgroups

O	N	NA	Evaluation Elements
☐	☐	☐	V6 - The content domain covered by a test is clearly defined, and an explicit rationale is provided for the content covered by the test.
☐	☐	☐	V7 - An appropriate procedure is documented to demonstrate that the test content is aligned to the content standards on which the test is based.
☐	☐	☐	V8 - Detailed documentation of validity evidence for each test version (including both total scores and any subscores) and validity evidence within subgroups is provided, including the sources of the evidence, the characteristics of the samples from which evidence was gathered, the qualifications of any experts involved in providing judgments about validity, and rationales regarding the relevance of the evidence to the intended inferences.
☐	☐	☐	V9 - As appropriate, evidence is gathered to demonstrate that assessments are measuring the knowledge and skills specified in the content standards and the intended cognitive processes and not other, unintended characteristics.
☐	☐	☐	V10 - When intended to yield scores with the same meaning, evidence is produced regarding the comparability of scores across test forms, versions (e.g., language versions, alternate assessments for students with disabilities), modes of administration (e.g., paper and pencil, computer-based), allowable accommodations or test modifications, and whether test takers received practice or coaching.
☐	☐	☐	V11 - An integrated evaluative summary of the validity evidence bearing on the intended score inferences is provided, including inferences from tests administered with accommodations for students with disabilities and limited English proficiency.
☐	☐	☐	V12 - When cut scores are used to classify examinees or establish achievement levels, the rationale and process for deriving the cut scores should be documented, including the qualifications of those who participated in the process.
☐	☐	☐	V13 - Evidence is provided regarding the relationship of cut scores and test-based decisions to the purpose of the test.
☐	☐	☐	V14 - Intended and unintended consequences of the use of the test are investigated.
☐	☐	☐	V15 - Cut scores, and the procedures used to establish them are regularly reviewed to assess their validity.
☐	☐	☐	V16 - Errors that affect interpretation of the test scores are promptly corrected.

O	N	NA	Evaluation Elements
☐	☐	☐	V17 – Examinees have an opportunity to learn test content and multiple opportunities to demonstrate success before test scores are used to make decisions about the examinee's future.

Comments:

Section V: Scoring and Reporting

Following administration of a large-scale achievement test, students' scores must be computed accurately, reported in a timely and confidential manner, and provided in a format that is clear, relevant, and useful to each intended audience. Such audiences would typically include the individual student and his or her teacher, but also could include school counselors, school psychologists, or others for whom knowledge of the student's performance would assist them in understanding and modifying the educational placement or plan for the student. When aggregated, student scores can also be useful to teacher, building or district-level administrative staff, program evaluators, policymakers, and others.

Specific scoring and reporting procedures differ depending on the mode of test administration (e.g., paper-and-pencil versus computer-administered), the kinds of scores to be reported (e.g., raw, scaled, performance levels, or transformed scores such as percentile ranks), requirements that prescribe specific reporting timelines, and whether test score equating is used. Equating is a statistical procedure used to allow for test scores from different forms to be compared and is necessary because test forms may vary in difficulty. When a cut score is used to make categorical classifications based on a student's performance on a test, equating is often used to place scores obtained on different forms of an examination onto the same scale so that the same performance standard must be met by examinees regardless of which test form they were administered. Additionally, automated computer scoring of students' responses may be used, or students' performances or responses to items may be evaluated by human raters.

Score reports for achievement tests may simply provide an indication of overall performance, such as a total score or performance category, but they might also provide information on how the student fared on any subtests, and diagnostic information or recommendations on area(s) for improvement may be provided. Regardless of the information provided, procedures should be in place to ensure that test

scores are reported only to appropriate audiences and that security and confidentiality of reports is assured. Table 5 provides the Scoring and Reporting Checklist.

Table 5 - Scoring and Reporting Checklist

O = Observed N = Not observed NA = Not applicable

O	N	NA	Evaluation Elements
☐	☐	☐	SR1 - Policies and procedures for score release are documented.
☐	☐	☐	SR2 - Policies and procedures for canceling or withholding scores are documented.
☐	☐	☐	SR3 - Policies and procedures regarding test rescoring are documented.
☐	☐	☐	SR4 – Policies and procedures regarding having test results declared invalid are documented.
☐	☐	☐	SR5 - Information is provided that clearly documents scoring procedures, materials and guidelines, allowing for consistent and standardized use by different test users.
☐	☐	☐	SR6 – The selection, training procedures, and qualifications of scorers are documented.
☐	☐	☐	SR7 - Raw and transformed scores are clearly interpretable, and limitations are described.
☐	☐	☐	SR8 - Criteria used for scoring constructed responses or performances (e.g. scoring rubric) are documented.
☐	☐	☐	SR9 - Procedures and guidelines for monitoring the accuracy and dependability of the scoring process are documented.
☐	☐	☐	SR10 - Score reports are provided promptly to examinees and other appropriate audiences.
☐	☐	☐	SR11 - A clear explanation of the intended test purpose or meaning of test scores and any subscores is provided.
☐	☐	☐	SR12 – Score reports provide test results in a clear format with essential information that is easily understood by the intended audience(s).
☐	☐	☐	SR13 – When change scores are reported, clear statements about the derivation of the scores and their proper interpretation are provided.
☐	☐	☐	SR14 – Examinees are informed regarding appropriate uses of test results.
☐	☐	☐	SR15 – Examinees are informed regarding to whom scores will be released and how long scores will be kept on file.

O	N	NA	Evaluation Elements
☐	☐	☐	SR16 - Test users are provided with information about the benefits and limitations of the test results.
☐	☐	☐	SR17 - When score comparability is affected, score reports include information on test modifications.
☐	☐	☐	SR18 - Procedures for ensuring confidentiality of test results and protecting test results from unauthorized release and access are documented.
☐	☐	☐	SR19 - Procedures for identifying and correcting scoring errors are documented.
☐	☐	☐	SR20 - Examinees are informed of score appeal procedures.
☐	☐	☐	SR21 – Examinees are informed of procedures and guidelines for retesting.
☐	☐	☐	SR22 – Policies and procedures regarding record keeping (e.g., duration of retention) are documented.
☐	☐	☐	SR23- When computer algorithms are used to score constructed responses, descriptions of student responses at each score level are provided. The use of automated scoring algorithms and interpretations is documented.
☐	☐	☐	SR24 - Methods used to determine norm-referenced scores are documented.
☐	☐	☐	SR25 - Automatically generated interpretations of test scores include documentation of how those interpretations were developed and their limitations.

Comments:

Section VI: References

American Educational Research Association. (2000, July). *AERA position statement on high-stakes testing in pre-K–12 education.* Washington, DC: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Brennan, R. (Ed.) (2006). *Educational measurement* (3rd ed.). New York: Praeger.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*, 31–43.

Cizek, G. J., Germuth, A. A., Kosh, A.E., & Schmid, L. A. (2015). *A checklist for evaluating credentialing testing programs.* Retrieved from http://wmich.edu/evaluation/checklists

Council of Chief State School Officers [CCSSO] & Association of Test Publishers [ATP]. (2010*). Operational best practices for statewide large-scale assessment programs.* Washington, DC: CCSSO.

Every Student Succeeds Act. (2015). P. L. 114-95, 20 U.S.C. 6301.

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations.* Washington, DC: American Psychological Association, Joint Committee on Testing Practices.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* Washington, DC: American Psychological Association, Joint Committee on Testing Practices.

No Child Left Behind Act. (2001). P. L. 107-110, 20 U.S.C. 6301.

Plake, B. S. (2002). Evaluating the technical quality of educational tests used for high-stakes decisions. *Measurement and Evaluation in Counseling and Development, 35,* 144-152.

Scriven, M. (2007, December). *The logic and methodology of checklists*. Retrieved from http://www.wmich.edu/evalctr/checklists/

Shepard, L. A. (1997). *A checklist for evaluating large-scale assessment programs* [Occasional Paper No. 9]. Kalamazoo, MI: Western Michigan University, The Evaluation Center.

United States Department of Education, Office of Elementary and Secondary Education. (2009, January 12). *Standards and assessments peer review guidance*. Washington, DC: Author.

Yen, W. M., & Henderson, D. L. (2002). Professional standards related to using large-scale state assessments in decisions for individual students. *Measurement and Evaluation in Counseling and Development, 35,* 144-152.

Section VII: Appendices

APPENDIX A

This appendix documents the professional guidelines used to develop the evaluation criteria in each of the checklists included in this document. The guidelines are summarized using tables in which the specific sources and guidelines are abbreviated using a numerical indicator. Additional characters following the numerical indicator identify the portion of the reference where specific criteria related to the evaluation element can be found. The resources, along with sample interpretive information, are abbreviated as follows:

Source I = The Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014)

In the following tables, source I refers to the Standards for Educational and Psychological Testing. The Standards are widely considered to be a definitive compilation of best practices in assessment, being developed and sponsored by three leading professional associations (the American Educational Research Association, the American Psychological Association, and the National Council of Measurement in Education) and endorsed by dozens of other associations, licensure and certification agencies, boards, and education groups. The 2014 version of the Standards is the latest edition of guidelines spanning six editions. The Standards for Educational and Psychological Testing is subdivided into chapters and specific standards. An entry in Column I in one of the following tables of "4.2" would indicate a reference to Source I (the Standards for Educational and Psychological Testing), Chapter 4 within that reference, and Standard 2 within that chapter.

Source II = The Standards and Assessments Peer Review Guidance (United States Department of Education, Office of Elementary and Secondary Education, 2009)

To assist states in meeting the requirements of the No Child Left Behind Act, the U.S. Department of Education developed and disseminated Standards and Assessments Peer Review Guidance in 2004. In the course of implementing the content and performance standards mandated by NCLB, state testing programs were subject to reviews conducted by education and assessment specialists external to each state, so-called peer reviews. The documentation designed to guide a state's preparation for peer review, as well the peer review process itself, was expanded in 2007 to include information related to what have come to be known as modified achievement standards and modified assessments—content standards and tests intended for use in the population of students with certain cognitive disabilities. The documentation was revised again in 2009 with certain technical edits. A subset of guidelines from the Guidance is referenced here as Source II. The guidelines included have been limited to those related to assessment systems broadly-conceived (see Appendix B); guidelines that apply only to unique aspects of NCLB have been omitted. An entry in Column II in one of the following tables of "2.1b" would indicate a reference to Source II (the Standards and Assessments Peer Review Guidance), Standard 2.1b within that reference.

Source III = The Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004)

This source is subdivided into guidelines intended for test developers ("D") and test users ("U"). Within the Code are four portions: (A) Developing and Selecting Tests (9 guidelines), (B) Administering and Scoring Tests (7 guidelines), (C) Reporting and Interpreting Test Results (8 guidelines), and (D) Informing Examinees (7 guidelines). An entry of "B-D1" in Column III in the following tables would indicate a reference to Source III (the Code of Fair Testing Practices in Education), portion B relating to Administering and Scoring Tests, and Guideline 1 for Test Developers (D1) within that portion. An abridged copy of the Code of Fair Testing Practices in Education is included with this report as Appendix C.

Source IV = The Rights and Responsibilities of Examinees (Joint Committee on Testing Practices, 1998)

This source provides guidance for testing professionals, grouped into guidelines related to the (A) Rights of Examinees (10 statements) and (B) Responsibilities of Examinees (10 statements). Some statements are followed by finer-grained elaborations which are indicated here using lower case letters. An entry in Column IV in one of the following tables of —A-4b‖ would indicate a reference to Source IV (the Rights and Responsibilities of Test Takers), Section A (Rights of Test Takers) within that source, Guideline 4, and elaboration (b). An abridged copy of the Rights and Responsibilities of Examinees is included with this report as Appendix D.

Table A-1

Crosswalk of Test Development Checklist Elements and Professional Guidelines

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
TD1	1.0, 1.1, 4.1, 7.1, 9.2, 9.3, 9.15		A-D1, U1; C-D3, U3; D-D4	A-4a, e
TD2	1.1, 4.1, 7.2		A-D1, U1	
TD3	4.1, 7.1, 12.1		A-D1, U1	A-4b, e
TD4	4.0, 4.1	3.6		
TD5		1.1		
TD6		1.4		
TD7	1.1, 4.2		A-D1, U1	
TD8	4.0, 4.1, 4.2, 4.6, 4.12	5.1, 5.2a, b; 5.3, 5.4	A-D2, U2	A-3a
TD9		2.3.1a, 1b		

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
TD10	4.2, 4.6	3.1a	A-D3, U3	A-4a
TD11	4.7, 4.9, 7.4			
TD12	4.6, 4.8			
TD13	4.12	4.1d		
TD14	4.6, 4.7, 4.8			
TD15	4.6, 4.8			
TD16	4.15, 4.16, 7.8, 7.13		B-D3, U3	A-1a
TD17	4.3			
TD18	4.9			
TD19	4.9, 7.5			
TD20	4.9, 7.5			
TD21	4.9		A-D9, U9	
TD22	4.0, 4.4, 4.10, 7.4, 7.6			
TD23	4.0, 4.4, 4.10, 7.4, 7.6		A-D5, U5	A-3a
TD24		5.5		
TD25		4.5c, 5.7		
TD26	7.6			

*Key to Sources

I = Standards for Educational and Psychological Testing

II = Standards and Assessments Peer Review Guidance

III = Code of Fair Testing Practices in Education

IV = Rights and Responsibilities of Test Takers

Table A-2

Crosswalk of Test Administration Checklist Elements and Professional Guidelines

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
TA1	1.0, 1.1, 4.1, 7.1, 8.1, 9.2, 9.3, 9.15		C-D3, U3; D-D4, U4	A-4a, e
TA2				A-4m
TA3	4.15, 4.16, 7.8, 7.13, 8.2, 9.2, 9.15		B-D3, U3	A-1,4a,i,k,m, 5a,b
TA4	4.16, 6.5, 8.1, 8.2		A-D4, U4	A-2a
TA5	8.2, 8.12		D-D3, U3, D7, U7	A-8d, 10a, b, c
TA6	3.9, 6.2, 8.2, 9.14	3.1b, 4.3a, b, c, 6.1.1, 6.2.1a, 6.3	A-D8, U8; B-D2, U2	A-3b, 4l
TA7	3.10	2.3.2, 6.2.2	B-D2	
TA8				A-9d
TA9	3.10, 9.18		D-D3	A-4f
TA10	7.7	6.2.1b		A-6c
TA11	4.15, 4.16, 7.8, 7.13	4.5a	B-D1, U1	
TA12	6.3		D-D6, U6	
TA13	6.14, 8.2, 8.5		D-D5	A-4h, 8b,e,g
TA14				
TA15	6.6, 6.7		B-D4, U4; D7, U7	A-9a
TA16			B-D4, U4	
TA17				A-4d,k
TA18	8.2, 8.9			
TA19	6.6		B-D4, U4	
TA20	6.6, 6.7			
TA21	9.17		D-D3	A-1a
TA22	6.6			A-6f

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
TA23	6.4			A-6d
TA24	3.13, 3.14			

*Key to Sources

I = Standards for Educational and Psychological Testing

II = Standards and Assessments Peer Review Guidance

III = Code of Fair Testing Practices in Education

IV = Rights and Responsibilities of Test Takers

Table A-3

Crosswalk of Reliability Checklist Elements and Professional Guidelines

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
R1	2.0, 2.3, 2.4, 2.8, 2.9, 2.10, 2.11, 2.12, 3.3, 7.4, 7.6, 12.2	3.2, 3.3, 4.2a, 4.2b	A-D5, U5, D9, U9; C-D1	A-3a
R2	2.0, 2.11, 2.19, 7.5	3.4	C-D6, U6	A-3a
R3	2.19	4.2e, 4.4a		
R4	2.6, 2.11, 2.12			
R5	2.13, 2.14			
R6	2.13, 2.14	4.2c	C-D4	A-8e
R7	4.14			
R8	2.7, 4.20, 4.21			
R9	2.7, 4.20, 4.21			

*Key to Sources

I = Standards for Educational and Psychological Testing

II = Standards and Assessments Peer Review Guidance

III = Code of Fair Testing Practices in Education

IV = Rights and Responsibilities of Test Takers

Table A-4

Crosswalk of Validity Checklist Elements and Professional Guidelines

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
V1	1.0, 1.1, 4.1, 7.1, 9.2, 9.3, 9.15, 12.1	4.1a	A-D1, U1	A-4a, e
V2	1.0, 1.1, 4.1, 7.1		A-D1, U1; C-D3, U3	A-4e
V3	7.1, 9.6, 9.8		C-D3, U3, U8, D8	
V4	1.1, 4.1		A-D1, U1	
V5	1.1, 3.15, 4.1, 7.2		A-D1, U1	
V6	1.1, 4.1		A-D2, U2	
V7	1.11, 4.6, 4.12	2.5		
V8	1.0, 1.8, 1.9, 1.10, 1.14, 3.6, 3.8, 7.4, 7.5, 7.6, 12.1			
V9	1.0, 1.2, 1.7, 1.11, 1.12, 1.16, 1.17, 3.2	4.1b, c, f		
V10	1.7, 5.6, 5.7, 5.12, 5.13, 5.14, 5.17, 5.18, 5.19, 5.20, 7.6	3.5, 4.3d, 4.4b,c,		

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
V11	1.2, 3.3, 3.6, 3.11, 3.12	4.6a, b		
V12	5.21, 5.22, 7.4	2.1a. b, c; 2.3.1c, 2.6	C-D4, U4	A-8e
V13	1.2, 1.18, 5.21, 5.23	4.1g		
V14		4.1h		
V15			C-D4, U4	
V16	6.13, 9.5		B-D6, U6	
V17	12.8, 12.9			

*Key to Sources

I = Standards for Educational and Psychological Testing

II = Standards and Assessments Peer Review Guidance

III = Code of Fair Testing Practices in Education

IV = Rights and Responsibilities of Test Takers

Table A-5

Crosswalk of Scoring and Reporting Checklist Elements and Professional Guidelines

	Professional Guidelines (Sources*)			
Checklist Element	I	II	III	IV
SR1	6.14, 8.5, 8.6, 9.20		C-D7, U7; D-D5	A-9b,c
SR2	8.11		D-D3, D6	A-10b,c; B-8a
SR3	6.9, 6.13, 8.12		B-D6, U6; D-D3	A-8d, h, i
SR4	8.11, 8.12		D-D3	A-10b,c
SR5	4.18, 5.2, 6.8	4,5b	B-D5, U5	A-8b
SR6	4.20, 4.21, 7.7, 12.16			A-6c

	Professional Guidelines (Sources*)			
Checklist Element	**I**	**II**	**III**	**IV**
SR7	5.1, 5.4, 12.17, 12.18	5.6		
SR8	4.18, 6.8, 7.8, 9.15	5.6	B-D5, U5	
SR9	4.20	4.5b	B-D5, U5; C-D4, U4	
SR10	8.8, 9.16		C-D7, U7	A-8
SR11	1.0, 1.1, 1.2, 7.1, 8.7	4.1e	C-D3, U3; D6, U6; D8, U8	A-4a,e, 8a
SR12	5.1, 6.10, 8.7, 12.17, 12.18	5.6	C-D1, U1	A-8b,e
SR13	12.11			
SR14	6.10, 8.2, 9.15		C-D1, U1, D3, U3; D-D4	A-4e
SR15	6.14, 8.2, 9.15		C-D5, U5; D-D3, D5	A-9a,b,c
SR16	12.18		C- D1, U1, D3, U3, D5, U5; D6, U6; D-D4	
SR17	3.10, 12.17	6.1.2	C-D2, U2	A-8a
SR18	6.16, 8.5, 8.6, 9.19		D-D5	A-4b; 9a,d
SR19	6.9, 6.13			A-8h,i
SR20	8.11, 8.12, 9.17		D-D7	A-4h, 8d, 10a,b; B-8a, b
SR21	9.18		D-D3	A-4f; B-8b
SR22	6.14, 8.2, 9.15		D-D5	A-4b
SR23	6.8			
SR24	5.8, 5.9, 5.10, 5.11, 7.2			
SR25	6.11			

*Key to Sources

I = Standards for Educational and Psychological Testing

II = Standards and Assessments Peer Review Guidance

III = Code of Fair Testing Practices in Education

IV = Rights and Responsibilities of Test Takers

APPENDIX B

Standards and Assessments Peer Review Guidance

(U.S. Department of Education, Office of Elementary and Secondary Education, 2009)

Section 1: Content Standards

1.1 Academic content standards have been developed.

1.4 Education stakeholders were involved in the development of academic content standards.

Section 2: Achievement Standards

2.1a A documented and validated standards-setting process was used to set achievement standards for general population assessments.

2.1b Where modified assessments have been developed for students with disabilities, a documented and validated standards-setting process was used to establish achievement standards.

2.1c Where alternate assessments have been developed for students with the most significant cognitive disabilities, a documented and validated standards-setting process was used to establish achievement standards.

2.3.1a Achievement level labels (ALLs) have been established for all assessments to indicate how well students are mastering academic content standards.

2.3.1b Achievement level descriptions (ALDs) have been established for all assessments to describe competencies associated with each performance level.

2.3.1c Achievement standards (i.e., cut scores) have been established that differentiate among the achievement levels; the rationale and procedures used to determine each achievement level are documented.

2.3.2 Where modified or alternate achievement standards exist, guidelines have been developed for IEP teams to use in deciding when an individual student should be assessed on the basis of modified or alternate academic achievement standards.

2.5 Content standards are aligned with achievement standards for all assessments.

2.6 Standard setting procedures should describe the selection of qualified panelists, involvement of diverse stakeholders, methodology, and results.

Section 3: High Quality Assessment System

3.1a The type of assessment is described (e.g., criterion-referenced, norm-referenced, alternative assessment, subject test)

3.1b Where applicable, non-English native language assessments are documented.

3.2 All required assessments (e.g., state, district) meet the same technical quality requirements and, where applicable, may be aggregated.

3.3 Where multiple forms of an assessment are used, the forms are equivalent and yield comparable results.

3.4 Assessments yield consistent and coherent information for each student across grades, subject areas and assessment types.

3.5 If alternative or native language instruments are used, the forms are equivalent and yield comparable results.

3.6 Assessments cover the breadth and depth of the academic content standards.

3.7 Alternate assessments are available for students with disabilities who cannot take a general assessment with accommodations.

Section 4: Technical Quality

4.1a The purposes and appropriate uses and decisions based on test performance are specified.

4.1b Evidence is provided that assessments are measuring the knowledge and skills in academic content standards and not other, unintended characteristics.

4.1c Evidence is provided that assessments are measuring the intended cognitive processes.

4.1d Evidence is provided that the items and tasks are at the appropriate grade level.

4.1e Evidence is provided that scoring and reporting structures are consistent with the structure of the academic content standards.

4.1f Evidence is provided that test performance is related to other variables as intended and not with irrelevant characteristics.

4.1g Evidence is provided that decisions based on test performance are consistent with the test purpose.

4.1h Evidence is provided about intended and unintended consequences of the assessment.

4.2a Reliability evidence is provided for each reported score.

4.2b Reliability evidence is provided for total group and relevant subpopulations.

4.2c Conditional standard errors of measurement are reported at each cut score.

4.2d Classification consistency estimates are reported.

4.2e Relevant reliability estimates are reported including, where appropriate, internal consistency, form equivalence, and interrater consistency.

4.3a Appropriate accommodations are provided for students with disabilities.

4.3b Appropriate linguistic accommodations are provided for students with limited English proficiency

4.3c Appropriate procedures have been followed to ensure fairness and accessibility for all students.

4.3d Accommodations do not alter the meaning of test scores.

4.4a Evidence is provided to document consistency of test forms over time.

4.4b Evidence is provided supporting consistent interpretations of results across different forms and/or formats.

4.4c Evidence is provided supporting comparability of scores when a test form is administered in different modes (e.g., paper-and-pencil, computer-administered).

4.5a Clear criteria are documented for administration of assessments.

4.5b Clear criteria are documented for scoring and reporting assessment results.

4.5c A system is in place for continuous monitoring and improvement of the assessment program.

4.6a The validity of accommodations for students with disabilities has been evaluated.

4.6b The validity of accommodations for students with limited English proficiency has been evaluated.

Section 5: Alignment

5.1 A coherent approach has been followed to ensure alignment between assessments, content standards, and achievement standards.

5.2a Assessments are aligned to the full range of content standards.

5.2b Assessments are aligned to the depth, difficulty, cognitive processes and complexity represented by the content standards.

5.3 Assessments and content standards aligned in terms of both content and process (i.e., both what students should know and be able to do).

5.4 All assessments reflect the same degree and pattern of emphasis embodies in the academic content standards.

5.5 Assessments are designed to yield scores that reflect the full range of student achievement.

5.6 Assessment results are reported in appropriate ways, including achievement levels, scaled scores, percentiles, etc.

5.7 A system is in place to maintain and improve alignment of the content standards and assessments.

Section 6: Inclusion

6.1.1 All students are included in the assessment system.

6.1.2 Reports are produced separately for subgroups (e.g., disability, accommodations).

6.2.1a Accommodations are provided so that, to the extent possible, students with disabilities can participate in the assessment program.

6.2.1b Test administrators are trained in how to administer assessments, including the use of accommodations.

6.2.2 Guidelines have been developed to determine which students are eligible to be assessed using which accommodations.

6.3 Assessments are made available in the appropriate language and form for ELL students to yield reliable information.

APPENDIX C

The Code of Fair Testing Practices in Education

(Joint Committee on Testing Practices, 2004)

Prepared by the Joint Committee on Testing Practices The Code of Fair Testing Practices in Education (Code) is a guide for professionals in fulfilling their obligation to provide and use tests that are fair to all examinees regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation,

linguistic background, or other personal characteristics. Fairness is a primary consideration in all aspects of testing. Careful standardization of tests and administration conditions helps to ensure that all examinees are given a comparable opportunity to demonstrate what they know and how they can perform in the area being tested. Fairness implies that every test taker has the opportunity to prepare for the test and is informed about the general nature and content of the test, as appropriate to the purpose of the test. Fairness also extends to the accurate reporting of individual and group test results. Fairness is not an isolated concept, but must be considered in all aspects of the testing process.

The Code applies broadly to testing in education (admissions, educational assessment, educational diagnosis, and student placement) regardless of the mode of presentation, so it is relevant to conventional paper-and-pencil tests, computer based tests, and performance tests. It is not designed to cover employment testing, licensure or certification testing, or other types of testing outside the field of education. The Code is directed primarily at professionally developed tests used in formally administered testing programs. Although the Code is not intended to cover tests made by teachers for use in their own classrooms, teachers are encouraged to use the guidelines to help improve their testing practices.

The Code addresses the roles of test developers and test users separately. Test developers are people and organizations that construct tests, as well as those that set policies for testing programs. Test users are people and agencies that select tests, administer tests, commission test development services, or make decisions on the basis of test scores. Test developer and test user roles may overlap, for example, when a state or local education agency commissions test development services, sets policies that control the test development process, and makes decisions on the basis of the test scores.

Many of the statements in the Code refer to the selection and use of existing tests. When a new test is developed, when an existing test is modified, or when the administration of a test is modified, the Code is intended to provide guidance for this process.

The Code is not intended to be mandatory, exhaustive, or definitive, and may not be applicable to every situation. Instead, the Code is intended to be aspirational, and is not intended to take precedence over the judgment of those who have competence in the subjects addressed.

The Code provides guidance separately for test developers and test users in four critical areas:

A. Developing and Selecting Appropriate Tests

B. Administering and Scoring Tests

C. Reporting and Interpreting Test Results

D. Informing Examinees

The Code is intended to be consistent with the relevant parts of the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological

Association [APA], and National Council on Measurement in Education [NCME], 2014). The Code is not meant to add new principles over and above those in the Standards or to change their meaning. Rather, the Code is intended to represent the spirit of selected portions of the Standards in a way that is relevant and meaningful to developers and users of tests, as well as to examinees and/or their parents or guardians. States, districts, schools, organizations and individual professionals are encouraged to commit themselves to fairness in testing and safeguarding the rights of test takers. The Code is intended to assist in carrying out such commitments.

The Code has been prepared by the Joint Committee on Testing Practices, a cooperative effort among several professional organizations. The aim of the Joint Committee is to act, in the public interest, to advance the quality of testing practices. Members of the Joint Committee include the American Counseling Association (ACA), the American Educational Research Association (AERA), the American Psychological Association (APA), the American Speech-Language-Hearing Association (ASHA), the National Association of School Psychologists (NASP), the National Association of Test Directors (NATD), and the National Council on Measurement in Education (NCME).

 This edition replaces the first edition of the Code, which was published in 1988. Please cite this document as follows: Code of Fair Testing Practices in Education. (2004). Washington, DC: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, Science Directorate, American Psychological

Association, 750 First Street, NE, Washington, DC 20002-4242; http://www.apa.org/science/jctpweb.html.) Contact APA for additional copies.

A. Developing and Selecting Appropriate Tests

TEST DEVELOPERS (D)/ TEST USERS (U)

Test developers should provide the information and supporting evidence that test users need to select appropriate tests. Test users should select tests that meet the intended purpose and that are appropriate for the intended test takers.

D1. Provide evidence of what the test measures, the recommended uses, the intended test takers, and the strengths and limitations of the test, including the level of precision of the test scores.

U1. Define the purpose for testing, the content and skills to be tested, and the intended test takers. Select and use the most appropriate test based on a thorough review of available information.

D2. Describe how the content and skills to be tested were selected and how the tests were developed.

U2. Review and select tests based on the appropriateness of test content, skills tested, and content coverage for the intended purpose of testing.

D3. Communicate information about a test's characteristics at a level of detail appropriate to the intended test users.

U3. Review materials provided by test developers and select tests for which clear, accurate, and complete information is provided.

D4. Provide guidance on the levels of skills, knowledge, and training necessary for appropriate review, selection, and administration of tests.

U4. Select tests through a process that includes persons with appropriate knowledge, skills, and training.

D5. Provide evidence that the technical quality, including reliability and validity, of the test meets its intended purposes.

U5. Evaluate evidence of the technical quality of the test provided by the test developer and any independent reviewers.

D6. Provide to qualified test users representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports.

U6. Evaluate representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports before selecting a test.

D7. Avoid potentially offensive content or language when developing

test questions and related materials.

U7. Evaluate procedures and materials used by test developers, as well as the resulting test, to ensure that potentially offensive content or language is avoided.

D8. Make appropriately modified forms of tests or administration procedures available for examinees with disabilities who need special accommodations.

U8. Select tests with appropriately modified forms or administration procedures for examinees with disabilities who need special accommodations.

D9. Obtain and provide evidence on the performance of examinees of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed.

U9. Evaluate the available evidence on the performance of examinees of diverse subgroups. Determine to the extent feasible which performance differences may have been caused by factors unrelated to the skills being assessed.

B. Administering and Scoring Tests

TEST DEVELOPERS (D) / TEST USERS (U)

Test developers should explain how to administer and score tests correctly and fairly.

Test users should administer and score tests correctly and fairly.

D1. Provide clear descriptions of detailed procedures for administering tests in a standardized manner.

U1. Follow established procedures for administering tests in a standardized manner.

D2. Provide guidelines on reasonable procedures for assessing persons with disabilities who need special accommodations or those with diverse linguistic backgrounds.

U2. Provide and document appropriate procedures for examinees with disabilities who need special accommodations or those with diverse linguistic backgrounds. Some accommodations may be required by law or regulation.

D3. Provide information to examinees or test users on test question formats and procedures for answering test questions, including information on the use of any needed materials and equipment.

U3. Provide examinees with an opportunity to become familiar with test question formats and any materials or equipment that may be used during testing.

D4. Establish and implement procedures to ensure the security of testing materials during all phases of test development, administration, scoring, and reporting.

U4. Protect the security of test materials, including respecting copyrights and eliminating opportunities for examinees to obtain scores by fraudulent means.

D5. Provide procedures, materials and guidelines for scoring the tests, and for monitoring the accuracy of the scoring process. If scoring the test is the responsibility of the test developer, provide adequate training for scorers.

U5. If test scoring is the responsibility of the test user, provide adequate training to scorers and ensure and monitor the accuracy of the scoring process.

D6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.

U6. Correct errors that affect the interpretation of the scores and communicate the corrected results promptly.

D7. Develop and implement procedures for ensuring the confidentiality of scores.

U7. Develop and implement procedures for ensuring the confidentiality of scores.

C. Reporting and Interpreting Test Results

TEST DEVELOPERS (D) / TEST USERS (U)

Test developers should report test results accurately and provide information to help test users interpret test results correctly. Test users should report and interpret test results accurately and clearly.

D1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.

U1. Interpret the meaning of the test results, taking into account the nature of the content, norms or comparison groups, other technical evidence, and benefits and limitations of test results.

D2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.

U2. Interpret test results from modified test or test administration procedures in view of the impact those modifications may have had on test results.

D3. Specify appropriate uses of test results and warn test users of potential misuses.

U3. Avoid using tests for purposes other than those recommended by the test developer unless there is evidence to support the intended use or interpretation.

D4. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.

U4. Review the procedures for setting performance standards or passing scores. Avoid using stigmatizing labels.

D5. Encourage test users to base decisions about examinees on multiple sources of appropriate information, not on a single test score.

U5. Avoid using a single test score as the sole determinant of decisions about test takers. Interpret test scores in conjunction with other information about individuals.

D6. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.

U6. State the intended interpretation and use of test results for groups of test takers. Avoid grouping test results for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use. Report procedures that were followed in determining who were and who were not included in the groups being compared and describe factors that might influence the interpretation of results.

D7. Provide test results in a timely fashion and in a manner that is understood by the test taker.

U7. Communicate test results in a timely fashion and in a manner that is understood by the test taker.

D8. Provide guidance to test users about how to monitor the extent to which the test is fulfilling its intended purposes.

U8. Develop and implement procedures for monitoring test use, including consistency with the intended purposes of the test.

D. Informing Examinees

TEST DEVELOPERS (D)

Under some circumstances, test developers have direct communication with the examinees and/or control of the tests, testing process, and test results. In other circumstances the test users have these responsibilities.

Test developers or test users should inform examinees about the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores.

D1. Inform examinees in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test takers.

D2. When a test is optional, provide examinees or their parents/guardians with information to help them judge whether a test should be taken—including indications of any consequences that may result from not taking the test (e.g., not being eligible to compete for a particular scholarship) —and whether there is an available alternative to the test.

D3. Provide examinees or their parents/guardians with information about rights examinees may have to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid.

D4. Provide examinees or their parents/guardians with information about responsibilities examinees have, such as being aware of the intended purpose and uses of the test, performing at capacity, following directions, and not disclosing test items or interfering with other test takers.

D5. Inform examinees or their parents/guardians how long scores will be kept on file and indicate to whom, under what circumstances, and in what manner test scores and related information will or will not be released. Protect test scores from unauthorized release and access.

D6. Describe procedures for investigating and resolving circumstances that might result in canceling or withholding scores, such as failure to adhere to specified testing procedures.

D7. Describe procedures that test takers, parents/guardians, and other interested parties may use to obtain more information about the test, register complaints, and have problems resolved.

APPENDIX D

The Rights and Responsibilities of Test Takers

(Joint Committee on Testing Practices, 1998)

A) The Rights of Test Takers: Guidelines for Testing Professionals

Examinees have the rights described below. It is the responsibility of the professionals involved in the testing process to ensure that examinees receive these rights.

1. Because examinees have the right to be informed of their rights and responsibilities as test takers, it is normally the responsibility of the individual who administers a test (or the organization that prepared the test) to inform examinees of these rights and responsibilities.
2. Because examinees have the right to be treated with courtesy, respect, and impartiality, regardless of their age, disability, ethnicity, gender, national origin, race, religion, sexual orientation, or other personal characteristics, testing professionals should:
    a. Make examinees aware of any materials that are available to assist them in test preparation. These materials should be clearly described in test registration and/or test familiarization materials.
    b. See that examinees are provided with reasonable access to testing services.
3. Because examinees have the right to be tested with measures that meet professional standards that are appropriate for the test use and the test taker, given the manner in which the results will be used, testing professionals should:
    a. Take steps to utilize measures that meet professional standards and are reliable, relevant, and useful given the intended purpose and are fair for examinees from varying societal groups.
    b. Advise examinees that they are entitled to request reasonable accommodations in test administration that are likely to increase the validity of their test scores if they have a disability recognized under the Americans with Disabilities Act or other relevant legislation.
4. Because examinees have the right to be informed, prior to testing, about the test's purposes, the nature of the test, whether test results will be reported to the test takers, and the planned use of the results (when not in conflict with the testing purposes), testing professionals should:
    a. Give or provide examinees with access to a brief description about the test purpose (e.g., diagnosis, placement, selection, etc.) and the kind(s) of tests and formats that will be used (e.g., individual/group, multiple-choice/free response/performance, timed/untimed, etc.), unless such information might be detrimental to the objectives of the test.

b. Tell test takers, prior to testing, about the planned use(s) of the test results. Upon request, the test taker should be given information about how long such test scores are typically kept on file and remain available.

c. Provide test takers, if requested, with information about any preventative measures that have been instituted to safeguard the accuracy of test scores. Such information would include any quality control procedures that are employed and some of the steps taken to prevent dishonesty in test performance.

d. Inform test takers, in advance of the testing, about required materials that must be brought to the test site (e.g., pencil, paper) and about any rules that allow or prohibit use of other materials (e.g., calculators).

e. Provide test takers, upon request, with general information about the appropriateness of the test for its intended purpose, to the extent that such information does not involve the release of proprietary information. (For example, the test taker might be told, "Scores on this test are useful in predicting how successful people will be in this kind of work" or "Scores on this test, along with other information, help us to determine if students are likely to benefit from this program.")

f. Provide test takers, upon request, with information about re-testing, including if it is possible to re-take the test or another version of it, and if so, how often, how soon, and under what conditions.

g. Provide test takers, upon request, with information about how the test will be scored and in what detail. On multiple-choice tests, this information might include suggestions for test taking and about the use of a correction for guessing. On tests scored using professional judgment (e.g., essay tests or projective techniques), a general description of the scoring procedures might be provided except when such information is proprietary or would tend to influence test performance inappropriately.

h. Inform examinees about the type of feedback and interpretation that is routinely provided, as well as what is available for a fee. Examinees have the right to request and receive information regarding whether or not they can obtain copies of their test answer sheets or their test materials, if they can have their scores verified, and if they may cancel their test results.

i. Provide test takers, prior to testing, either in the written instructions, in other written documents or orally, with answers to questions that examinees may have about basic test administration procedures.

j. Inform test takers, prior to testing, if questions from examinees will not be permitted during the testing process.

k. Provide examinees with information about the use of computers, calculators, or other equipment, if any, used in the testing and give them an opportunity to practice using

such equipment, unless its unpracticed use is part of the test purpose, or practice would compromise the validity of the results, and to provide a testing accommodation for the use of such equipment, if needed.

l. Inform examinees that, if they have a disability, they have the right to request and receive accommodations or modifications in accordance with the provisions of the Americans with Disabilities Act and other relevant legislation.

m. Provide examinees with information that will be of use in making decisions if examinees have options regarding which tests, test forms or test formats to take.

5. Because that examinees have a right to be informed in advance when the test will be administered, if and when test results will be available, and if there is a fee for testing services that the examinees are expected to pay, test professionals should:

    a. Notify examinees of the alteration in a timely manner if a previously announced testing schedule changes, provide a reasonable explanation for the change, and inform examinees of the new schedule. If there is a change, reasonable alternatives to the original schedule should be provided.

    b. Inform examinees prior to testing about any anticipated fee for the testing process, as well as the fees associated with each component of the process, if the components can be separated.

6. Because examinees have the right to have their tests administered and interpreted by appropriately trained individuals, testing professionals should:

    a. Know how to select the appropriate test for the intended purposes.

    b. When testing persons with documented disabilities and other special characteristics that require special testing conditions and/or interpretation of results, have the skills and knowledge for such testing and interpretation.

    c. Provide reasonable information regarding their qualifications, upon request.

    d. Insure that test conditions, especially if unusual, do not unduly interfere with test performance. Test conditions will normally be similar to those used to standardize the test.

    e. Provide candidates with a reasonable amount of time to complete the test, unless a test has a time limit.

    f. Take reasonable actions to safeguard against fraudulent actions (e.g., cheating) that could place honest examinees at a disadvantage.

7. Because examinees have the right to be informed about why they are being asked to take particular tests, if a test is optional, and what the consequences are should they choose not to complete the test, testing professionals should:

a. Normally only engage in testing activities with examinees after the examinees have provided their informed consent to take a test, except when testing without consent has been mandated by law or governmental regulation, or when consent is implied by an action the examinees have already taken (e.g., such as when applying for employment and a personnel examination is mandated).

b. Explain to examinees why they should consider taking voluntary tests.

c. Explain, if a test taker refuses to take or complete a voluntary test, either orally or in writing, what the negative consequences may be to them for their decision to do so.

d. Promptly inform the test taker if a testing professional decides that there is a need to deviate from the testing services to which the test taker initially agreed (e.g., should the testing professional believe it would be wise to administer an additional test or an alternative test), and provide an explanation for the change.

8. Because examinees have a right to receive a written or oral explanation of their test results within a reasonable amount of time after testing and in commonly understood terms, testing professionals should:

a. Interpret test results in light of one or more additional considerations (e.g., disability, language proficiency), if those considerations are relevant to the purposes of the test and performance on the test, and are in accordance with current laws.

b. Provide, upon request, information to examinees about the sources used in interpreting their test results, including technical manuals, technical reports, norms, and a description of the comparison group, or additional information about the test taker(s).

c. Provide, upon request, recommendations to examinees about how they could improve their performance on the test, should they choose or be required to take the test again.

d. Provide, upon request, information to examinees about their options for obtaining a second interpretation of their results. Examinees may select an appropriately trained professional to provide this second opinion.

e. Provide examinees with the criteria used to determine a passing score, when individual test scores are reported and related to a pass-fail standard.

f. Inform test takers, upon request, how much their scores might change, should they elect to take the test again. Such information would include variation in test performance due to measurement error (e.g., the appropriate standard errors of measurement) and changes in performance over time with or without intervention (e.g., additional training or treatment).

g. Communicate test results to examinees in an appropriate and sensitive manner, without use of negative labels or comments likely to inflame or stigmatize the test taker.

h. Provide corrected test scores to examinees as rapidly as possible, should an error occur in the processing or reporting of scores. The length of time is often dictated by individuals responsible for processing or reporting the scores, rather than the individuals responsible for testing, should the two parties indeed differ.

i. Correct any errors as rapidly as possible if there are errors in the process of developing scores.

9. Because examinees have the right to have the results of tests kept confidential to the extent allowed by law, testing professionals should:

    a. Insure that records of test results (in paper or electronic form) are safeguarded and maintained so that only individuals who have a legitimate right to access them will be able to do so.

    b. Should provide test takers, upon request, with information regarding who has a legitimate right to access their test results (when individually identified) and in what form. Testing professionals should respond appropriately to questions regarding the reasons why such individuals may have access to test results and how they may use the results.

    c. Advise examinees that they are entitled to limit access to their results (when individually identified) to those persons or institutions, and for those purposes, revealed to them prior to testing. Exceptions may occur when test takers, or their guardians, consent to release the test results to others or when testing professionals are authorized by law to release test results.

    d. Keep confidential any requests for testing accommodations and the documentation supporting the request.

10. Because examinees have the right to present concerns about the testing process and to receive information about procedures that will be used to address such concerns, testing professionals should:

    a. Inform examinees how they can question the results of the testing if they do not believe that the test was administered properly or scored correctly, or other such concerns.

    b. Inform examinees of the procedures for appealing decisions that they believe are based in whole or in part on erroneous test results.
    Inform test takers, if their test results are under investigation and may be canceled, invalidated, or not released for normal use. In such an event, that investigation should be performed in a timely manner. The investigation should use all available information that addresses the reason(s) for the investigation, and the test taker should also be informed of the information that he/she may need to provide to assist with the investigation.

    c. Inform the test taker, if that test taker's test results are canceled or not released for normal use, why that action was taken. The test taker is entitled to request and receive

information on the types of evidence and procedures that have been used to make that determination.

B) The Responsibilities of Test Takers: Guidelines for Testing Professionals

Testing Professionals should take steps to ensure that examinees know that they have specific responsibilities in addition to their rights described above.

1. Testing professionals need to inform examinees that they should listen to and/or read their rights and responsibilities as a test taker and ask questions about issues they do not understand.
2. Testing professionals should take steps, as appropriate, to ensure that examinees know that they:
    a. Are responsible for their behavior throughout the entire testing process.
    Should not interfere with the rights of others involved in the testing process.
    b. Should not compromise the integrity of the test and its interpretation in any manner.
3. Testing professionals should remind examinees that it is their responsibility to ask questions prior to testing if they are uncertain about why the test is being given, how it will be given, what they will be asked to do, and what will be done with the results. Testing professionals should:
    a. Advise examinees that it is their responsibility to review materials supplied by test publishers and others as part of the testing process and to ask questions about areas that they feel they should understand better prior to the start of testing.
    b. Inform examinees that it is their responsibility to request more information if they are not satisfied with what they know about how their test results will be used and what will be done with them.
4. Testing professionals should inform examinees that it is their responsibility to read descriptive material they receive in advance of a test and to listen carefully to test instructions. Testing professionals should inform examinees that it is their responsibility to inform an examiner in advance of testing if they wish to receive a testing accommodation or if they have a physical condition or illness that may interfere with their performance. Testing professionals should inform examinees that it is their responsibility to inform an examiner if they have difficulty comprehending the language in which the test is given. Testing professionals should:
    a. Inform examinees that, if they need special testing arrangements, it is their responsibility to request appropriate accommodations and to provide any requested documentation as far in advance of the testing date as possible. Testing professionals should inform examinees about the documentation needed to receive a requested testing accommodation.
    b. Inform examinees that, if they request but do not receive a testing accommodation, they could request information about why their request was denied.

5. Testing professionals should inform examinees when and where the test will be given, and whether payment for the testing is required. Having been so informed, it is the responsibility of the test taker to appear on time with any required materials, pay for testing services and be ready to be tested. Testing professionals should:
    a. Inform examinees that they are responsible for familiarizing themselves with the appropriate materials needed for testing and for requesting information about these materials, if needed.
    b. Inform the test taker, if the testing situation requires that examinees bring materials (e.g., personal identification, pencils, calculators, etc.) to the testing site, of this responsibility to do so.
6. Testing professionals should advise test takers, prior to testing, that it is their responsibility to:
    a. Listen to and/or read the directions given to them.
    b. Follow instructions given by testing professionals.
    Complete the test as directed.
    c. Perform to the best of their ability if they want their score to be a reflection of their best effort.
    d. Behave honestly (e.g., not cheating or assisting others who cheat).
7. Testing professionals should inform examinees about the consequences of not taking a test, should they choose not to take the test. Once so informed, it is the responsibility of the test taker to accept such consequences, and the testing professional should so inform the test takers. If examinees have questions regarding these consequences, it is their responsibility to ask questions of the testing professional, and the testing professional should so inform the test takers.
8. Testing professionals should inform examinees that it is their responsibility to notify appropriate persons, as specified by the testing organization, if they do not understand their results, or if they believe that testing conditions affected the results. Testing professionals should:
    a. Provide information to test takers, upon request, about appropriate procedures for questioning or canceling their test scores or results, if relevant to the purposes of testing.
    b. Provide to test takers, upon request, the procedures for reviewing, re-testing, or canceling their scores or test results, if they believe that testing conditions affected their results and if relevant to the purposes of testing.
    c. Provide documentation to the test taker about known testing conditions that might have affected the results of the testing, if relevant to the purposes of testing.
9. Testing professionals should advise examinees that it is their responsibility to ask questions about the confidentiality of their test results, if this aspect concerns them.

10. Testing professionals should advise examinees that it is their responsibility to present concerns about the testing process in a timely, respectful manner.

NOTE: The complete Rights and Responsibilities of Test Takers: Guidelines and Expectations is available free of charge at : http://www.apa.org/science/programs/testing/rights.aspx

APPENDIX E

Author Biographical Information

Gregory J. Cizek is Professor of Educational Measurement and Evaluation at the University of North Carolina at Chapel Hill, where he teaches courses in applied psychometrics, statistics, and research methods. His research interests include standard setting, testing policy, and classroom assessment. He is the author of over 200 journal articles, book chapters, conference papers, and other publications. His work has been published in journals such as Educational Researcher, Educational Assessment, Review of Educational Research, Journal of Educational Measurement, Educational Measurement: Issues and Practice, Educational Policy, Phi Delta Kappan, Education Week, and elsewhere. He is a contributor to the Handbook of Classroom Assessment (Academic Press, 1998); editor and contributor to the Handbook of Educational Policy (Academic Press, 1999) , Setting Performance Standards: Concepts, Methods, and Perspectives (Erlbaum, 2001), Setting Performance Standards: Foundations, Methods, and Innovations (2012), and the Handbook of Formative Assessment (with H. Andrade, Routledge, 2010); and author of Filling in the Blanks (Fordham, 1999), Cheating on Tests: How to Do It, Detect It, and Prevent It (Erlbaum, 1999), Detecting and Preventing Classroom Cheating (Corwin Press, 2003), Addressing Test Anxiety in a High Stakes Environment (with S. Burg, Corwin Press, 2005), and Standard Setting (with M. Bunch, Sage Publications, 2007). He provides expert consultation at the state and national levels on testing programs and policy.

Dr. Cizek received his Ph.D. in Measurement, Evaluation, and Research Design from Michigan State University. He has managed national licensure and certification testing programs for American College Testing (ACT) in Iowa City, Iowa and served as a test development specialist for the Michigan Educational Assessment Program (MEAP). Previously, he was an elementary school teacher for five years in Michigan and professor of educational research and measurement at the University of Toledo (OH). From 1997-99, he was elected to and served as vice-president of a local board of education in Ohio. In 2012, he was elected President of the National Council on Measurement in Education.

Amy A. Germuth is President of EvalWorks, LLC, an evaluation and survey research firm located in Durham, North Carolina, that specializes in evaluations in the fields of science, technology, engineering, and math (STEM) education. Her interests include evaluation and survey methodology, evaluation of STEM programs, and evaluation of research. She has conducted over 50 evaluations, working with clients such as the Bill and Melinda Gates Foundation, the Pew Charitable Trust, the U.S. Department of Education, the New York State Education Department, Chicago Public Schools, and Westat. Dr. Germuth has taught evaluation and instrument design courses across the US and in Australia, and she regularly teaches as part of Duke University's Non-Profit Management Program. She is the author of several research and evaluation reports; her work has been recognized with distinguished paper awards from the North Carolina Association for Research in Education (2001) and the American Education Research Association (2002). She has received multiple grants including a Spencer Foundation Grant (2007) to support her study of quasi-experimental design and a National Science Foundation sponsorship to participate in the Math, Technology, and Science Project Summer Evaluation Institute at The Evaluation

Center at Western Michigan University (2001). In 2005, she was a Policy Fellow with the Institute for Educational Leadership.

Dr. Germuth received her Ph.D. in Education Psychology, Measurement, and Evaluation from University of North Carolina at Chapel Hill. Previously, she was a high school math teacher and a middle and elementary school assistant principal. She has been a member of the American Evaluation Association since 2000 and has served as the chair of multiple topical interest groups and board committees within that organization.

Audra E. Kosh is a Research Associate at MetaMetrics (Durham, NC) where she works on developing mathematics assessments and conducting educational measurement research on automatic item generation. Dr. Kosh previously taught eighth-grade mathematics in Prince George's County Public Schools, Maryland, and worked as a research analyst for Westat in Rockville, MD. Her research interests include educational measurement, mathematics learning and teaching, and informal learning opportunities. Dr. Kosh holds a M.A.T. in Secondary Mathematics from American University and a Ph.D. in Education (Learning Sciences and Psychological Studies) from the University of North Carolina at Chapel Hill.

Lorrie A. Schmid is the Manager of Data Infrastructure and a Research Scientist with the Social Science Research Institute at Duke University. She received her M. A. and Ph.D. in the Department of Educational Psychology, Measurement and Evaluation in the School of Education at the University of North Carolina at Chapel Hill (UNC-CH). Dr. Schmid has also completed the Certificate Program in Survey Methodology from the Odum Institute at the University of North Carolina. Her interests include survey design issues, research methods, adolescent school adjustment, peer influence, and applied behavioral statistics.

Suggested Citation

Cizek, G. J., Schmid, L. A. Kosh, A. E., & Germuth, A. A., (2016). A checklist for evaluating K-12 assessment programs. Retrieved from http://www.wmich.edu/evaluation/checklists