A Checklist Evaluating Credentialing Testing Programs

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

America has been called “the credential society” (Collins, 1979). In the U.S., hundreds of organizations make decisions to award or deny licenses or certification in a range of occupations and professions. The entities responsible for making credentialing decisions are most often concerned with promoting high quality practice within a profession, increasing the status of those in the field, or promoting public protection through processes intended to ensure competent, safe, and effective practice by those who are awarded the credential.

Credentialing decisions are commonly made using *conjunctive models;* that is, the decisions are based on multiple criteria, each of which must be met in order for a candidate to be awarded the credential. Examples of such criteria include required academic degrees, coursework or other academic preparation, time spent in an apprenticeship or supervised practice, grades, recommendations, age, citizenship, or other requirements. Failure to satisfy any one of the criteria typically results in the denial of the credential or refusal to permit the candidate to attempt other steps in the credentialing process.

One of the most common of the additional criteria used in credentialing programs is an assessment of the candidate’s knowledge, skills, and abilities relevant to the occupation or profession. Such assessments may take the form of structured observations of performance, simulations, examinations of work samples, or (most commonly) an objective test.

Purpose

The purpose of this document is to aid in the evaluation of tests used as part of credentialing decisions. The checklist provides a list of practices that credentialing tests should follow; however, it should be noted that this document does not necessarily support quality judgments about a credentialing testing program. That is, a credentialing test may observe all of the checklist criteria while failing to adhere to thresholds of acceptable practices within the measurement field on those same criteria. For example, a credentialing testing 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, credentialing

testing 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 credentialing testing programs.

Intended Audience

The intended audience for this document includes those who develop and administer credentialing examinations and, ultimately, those responsible for licensure and certification programs.

Development of the Checklist

This version of the *Checklist for Evaluating Credentialing Testing Programs* is an update of a 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). The evaluation criteria upon which the checklist is based have been drawn from four broadly accepted sources of guidelines or standards for practice, including the following:

1. Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014);
2. Standards for the Accreditation of Certification Programs (National Organization for Competency Assurance, 2004);
3. Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004); and
4. Rights and Responsibilities of Test Takers (Joint Committee on Testing Practices, 1998).

Following this introduction are sections that contain brief descriptions of five key aspects of credentialing examinations and the associated checklists that can be used to evaluate credentialing examination programs against professionally-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 credentialing examination 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 checklists are intended to be used primarily in a formative manner to enable licensure and certification agencies to identify strengths of their programs and areas for improvement.

The checklists each contain several succinct statements that capture the fine-grain elements that comprise each of the five key aspects of licensure and certification testing programs. When applicable to a particular program, the statements represent the criteria that should be applied when evaluating each

aspect. The checklists also provide a format for indicating whether the 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 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). Those interested in K-12 assessment contexts should refer to *A Checklist for Evaluating K-12 Assessment Programs* (Cizek, Schmid, Kosh, & Germuth, 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 developing a test typically provide the most extensive source of evidence about the quality of the test and are the foundation for instruments that will yield reliable and valid 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 test development process begins and 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.

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

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. Delineation of the test domain can be theory-driven, judgmentally decided based on the input of subject-matter experts (SMEs), or empirically-derived from data generated by a job or practice analysis. Regardless of the source, the overall product from this phase is a delineation of the specific objectives or content standards that are to be measured. In licensure and certification contexts, best practices involve the domain delineation via job analyses that are conducted to identify the knowledge, skills, and abilities (KSAs) necessary for safe and competent practice in a field. Typically, job analyses are conducted on a repeating cycle to ensure the currency of the domain specification.

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, formats that will be used for items or tasks, and acceptable response formats. 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, tasks, prompts, cases, simulated patient profiles, and so on. SMEs are routinely involved in this phase also; to create and select items for a test, they must be instructed in item/task construction, be knowledgeable about the purpose of the test, and be familiar with the characteristics of the intended test takers. When 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 KSAs identified in the job analysis. 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 test population. Test developers should document the procedures and actions of the field testing. They should also ensure that the characteristics of the field test participants match those of the intended test takers 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 is 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 test takers is specified.
☐	☐	☐	TD3 - The appropriate and inappropriate uses of the test are specified.
☐	☐	☐	TD4 - A detailed domain definition and specific objectives or content standards are provided.
☐	☐	☐	TD5 - Domain definition is based on systematic, scientific procedures (e.g., curriculum review, job analysis, practice analyses, etc.).
☐	☐	☐	TD6 - A psychometrically sound method for determining test content specifications is used and the process is documented.
☐	☐	☐	TD7 - 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.
☐	☐	☐	TD8 - The item/task development process is documented.
☐	☐	☐	TD9 - The qualifications of item/task developers are documented.
☐	☐	☐	TD10 - A rigorous item/task review process is followed and documented.
☐	☐	☐	TD11 - The qualifications of item/task reviewers are documented.
☐	☐	☐	TD12 - Test administration procedures, including clear directions and information regarding the rights and responsibilities of test takers should be provided.
☐	☐	☐	TD13 - Computer-based/computer-adaptive testing protocols (e.g., item selection algorithms) are documented.
☐	☐	☐	TD14 - Items and tasks are field tested with relevant populations.

O	N	NA	Evaluation Elements
☐	☐	☐	TD15 - Field test procedures match expected administration conditions.
☐	☐	☐	TD16 - The characteristics of the field test sample and method of selection are described.
☐	☐	☐	TD17 – Field test sample characteristics match the intended test population as closely as possible.
☐	☐	☐	TD18 - Information on the psychometric properties of items/tasks (e.g., alignment, difficulty, discrimination) is reported.
☐	☐	☐	TD19 - Information on the psychometric properties of test forms (e.g., reliability, validity) is reported.
☐	☐	☐	TD20 - Procedures used to translate a test to alternate languages are documented.

Comments:

Section II: Test Administration

A professional credential is intended to provide assurance to examinees about their own preparation for a profession and to the public and other stakeholders regarding the skill and competence of the person to whom the license or certification has been awarded. Examinations are often used as part of the process by which candidates for a credential are evaluated. Those examinations must produce reliable and valid scores and limit the potential for errors in the measurement process that might attenuate those characteristics. Test administration procedures and the test environment are essential components that support interpretations of test results as faithful indicators of what candidates 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 four components of test administration:

1. examination registration procedures;
2. test administration personnel responsibilities;
3. test security procedures; and
4. 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. Registration procedures should allow test takers to register with ease. This may be accomplished by having clear instructions available to candidates before the test date. These clear instructions should explain the purpose of the examination, provide accurate directions to the test site, and describe the expectations and procedures that will be in place when candidates arrive at the testing center. Test takers should be aware of the availability of accommodations before arriving at the test site, as well as ways to best prepare for the examination.

Testing personnel responsibilities include ensuring that all candidates take the test in the same manner unless testing accommodations are required. If so, such accommodations should be provided for all examinees that need them. Test administration personnel are responsible for ensuring that test takers follow all required procedures. Thus, test administrators must be trained to handle a variety of situations, including how to handle emergencies that may occur and how to report testing irregularities.

Security breaches can jeopardize the validity of examination results. Therefore, *test security procedures* should be put in place. These procedures should ensure that the person taking the test is indeed the person registered to take the exam, test materials are handled in a secure manner at all times, and examinees are informed before the exam date about what materials and actions are permissible/not permissible during the testing process.

Finally, the *testing environment and test administration process* should maximize the test taker's ability to accurately demonstrate his or her level of knowledge and skill. Test administrators should be available to assist test takers with procedural matters, to make sure all security policies are followed, and to ensure test takers' needs are met during the testing period. If a test is timed, test takers need the ability to monitor their time. 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. Accommodations for individuals with disabilities and, as appropriate, language accommodations, may be necessary to assist

examinees to demonstrate their knowledge and skill. All test takers should receive comparable and equitable treatment in a comfortable 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 test takers are documented.
☐	☐	☐	TA3 - Test takers are provided with test administration regulations and procedures.
☐	☐	☐	TA4 – Reasonable test preparation materials are made available to test takers free of charge.
☐	☐	☐	TA5 - Test takers are informed of procedures for obtaining technical assistance and/or registering complaints.
☐	☐	☐	TA6 - Test takers are made aware of the availability of appropriate testing accommodations.
☐	☐	☐	TA7 - The eligibility review process for test accommodation requests is documented.
☐	☐	☐	TA8 - Requests for accommodations and supporting documentation are kept confidential.
☐	☐	☐	TA9 - Test adaptation and retake policies and procedures are documented.
☐	☐	☐	TA10 – Qualifications, responsibilities, and training of test administrators are documented.
☐	☐	☐	TA11 - A test administration manual containing standardized procedures for administration is produced.
☐	☐	☐	TA12 - A process for test administrators to record test irregularities is in place.
☐	☐	☐	TA13 - Test takers are informed of score reporting procedures.
☐	☐	☐	TA14 - Procedures for dealing with emergencies are provided to test administrators and test takers.

O	N	NA	Evaluation Elements
☐	☐	☐	TA15 - All test materials are shipped, handled, transmitted and/or maintained appropriately.
☐	☐	☐	TA16 – Procedures to ensure that test takers provide their authorization prior to taking the test are in place.
☐	☐	☐	TA17 - A policy on required and prohibited materials for taking the test is developed and communicated to test takers.
☐	☐	☐	TA18 - Test takers 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 - Test takers are provided with information about their rights and responsibilities.
☐	☐	☐	TA22 - Test takers 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). Thus, reliability is a characteristic of the test data rather than the test itself. It is essential to investigate the reliability of scores to gain information about the extent to which any individual score should be considered to be a dependable estimate of an examinee's 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 method used to estimate reliability is the coefficient of both stability and equivalence. The *coefficient of stability and equivalence* is an estimate of the extent to which forms are parallel and measure a stable characteristic over time.

The most commonly encountered estimates of reliability for credentialing examinations 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 examinees. Put another way, internal consistency reliability estimates gauge how dependable a collection of items is at measuring a common construct of interest. For example, the entity responsible for a licensure or certification examination might find it valuable to know that the total test yields reliable scores related to overall knowledge and skill in some profession, but also that smaller subsets of items (i.e., subtests) produce less dependable measures of a candidate's ability 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 items/task and directions, 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). A conditional SEM is particularly applicable to tests which use cut-scores (i.e., the score which delineates whether candidates' performances fall in either the pass or fail range).

Another perspective on reliability that is particularly appropriate for licensure and certification contexts in which a pass/fail or similar categorical decision is made is *decision consistency*. Indices of decision consistency are an indication of the reliability of decision making.

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. A high degree of interrater reliability may or may not be accompanied by strong interrater agreement. 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, and combined scores 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.
☐	☐	☐	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). Like reliability, validity is not a characteristic of a test, but of a set of scores on 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. 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, it may be appropriate to restate and elaborate on some key points related to validity:

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 test maker, 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) provide 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 every aspect of a testing program bears on the validity of score inferences; the other aspects can provide either supporting evidence for the intended inference or weaken validity claims. Consequently, this section cannot be considered as a stand-alone checklist for evaluating validity evidence. Rather, the entire body of supporting evidence in this report must be considered when evaluating validity claims. However, in an effort to be as parsimonious as possible and avoid overlap with information presented in other sections, the Validity Evidence Checklist is focused 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?
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 use of the test is specified.
☐	☐	☐	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 necessary KSAs of the occupation or qualifications for licensure/certification (e.g., job analysis, role delineation study, etc.).
☐	☐	☐	V8 - Detailed documentation for validity evidence 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 - An integrated evaluative summary of the validity evidence bearing on the intended score inferences is provided.
☐	☐	☐	V10 - When cut scores are used to classify examinees, the process for deriving the cut scores should be documented, including the qualifications of those who participated in the process.

O	N	NA	Evaluation Elements
☐	☐	☐	V11 - Evidence is provided regarding the relationship of cut scores to the purpose of the test. Cut scores should be based on requirements for licensure or certification and not as a mechanism to regulate the number of persons passing a test.
☐	☐	☐	V12 - Cut scores, and the procedures used to establish them are regularly reviewed to assess their validity.
☐	☐	☐	V13 - Errors that affect interpretation of the test scores are promptly corrected.
☐	☐	☐	V14 - Intended and unintended consequences of the use of the test are investigated.

Comments:

Section V: Scoring and Reporting

Following administration of a licensure or certification examination, examinees' scores must be computed accurately; reported in a timely and, where appropriate, confidential manner; and provided in a format that is clear, relevant, and useful to each intended audience. Such audiences would include the individual test taker, but also could include the credentialing entity, education or training programs, and agencies with reciprocity agreements regarding licensure or certification, among 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, or other transformed scores), requirements that prescribe specific reporting timelines, and whether a score equating procedure 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 decisions such as pass/fail, certify/deny certification, or other categorical classifications based on an examinee's performance on a credentialing examination, 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 the examinee regardless of which test form he or she attempted. Additionally, computer-scoring via optical scanning of answer documents may be used, or examinees' performances or responses to items may be evaluated by human raters.

Score reports for examinees may simply provide an indication of passing or failing, but they might also provide information on how the examinee fared on any subtests, and diagnostic information or recommendations for how to improve competence in the area(s). 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: ***Policies and Procedures***
☐	☐	☐	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.

O	N	NA	Evaluation Elements: *Policies and Procedures*
☐	☐	☐	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 test takers and other appropriate audiences.
☐	☐	☐	SR11 - A clear explanation of the intended test purpose or meaning of test scores 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 – Test takers are informed regarding appropriate uses of test results.
☐	☐	☐	SR14 – Test takers should be informed regarding to whom scores will be released and how long scores will be kept on file.
☐	☐	☐	SR15 - Test users are provided with information about the benefits and limitations of the test results.
☐	☐	☐	SR16 - When score comparability is affected, score reports include information on test modifications.
☐	☐	☐	SR17 - Procedures for ensuring confidentiality of test results and protecting test results from unauthorized release and access are documented.

O	N	NA	**Evaluation Elements:** ***Policies and Procedures***
☐	☐	☐	SR18 - Procedures for identifying and correcting scoring errors are documented.
☐	☐	☐	SR19 - Test takers are informed of score appeal procedures.
☐	☐	☐	SR20 – Test takers are informed of procedures and guidelines for retesting.
☐	☐	☐	SR21 – Policies and procedures regarding record keeping (e.g., duration of retention) are documented.
☐	☐	☐	SR22 - 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.
☐	☐	☐	SR23 - Methods used to determine norm-referenced scores are documented.
☐	☐	☐	SR24 - Automatically generated interpretations of test scores include documentation of how those interpretations were developed and their limitations.

Comments:

Section VI: Reference

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.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use, *Psychological Methods, 17*(1)*,* 31-43.

Cizek, G. J., Schmid, L. A., Kosh, A. E., & Germuth, A. A (2015). *A checklist for evaluating K-12 assessment programs.* Kalamazoo, MI: The Evaluation Center, Western Michigan University.

Collins, R. (1979). *The credential society: An historical sociology of education and stratification.* New York: Academic.

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.

National Commission for Certifying Agencies. (2004). *Standards for the accreditation of certification programs.* Washington, DC: Author.

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. Following a brief description of each source is a series of five tables that show the results of a crosswalk of the checklists' elements against the professional guidelines, indicating the specific locations within the sources that relate to each checklist element.

Source I = The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999)

The *Standards* are widely considered to be the 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 *Standards for Educational and Psychological Testing* is subdivided into chapters and specific standards. In Tables A-1 through A-5 below, an entry in Column I of "4.2" would indicate a reference to Chapter 4, and Standard 2 within that chapter. The 1999 version of the *Standards* is the latest edition of guidelines spanning five editions. It is presently under revision and, although it is not known when the next edition will be finalized, the authors of this checklist will endeavor to update the entries from that source accordingly when the next edition of the *Standards* is released.

Source II = The *Standards for the Accreditation of Certification Programs* (National Commission for Certifying Agencies, 2004)

The National Commission for Certifying Agencies (NCCA) accredits certification organizations complying with its standards. The mission of NCCA is to help ensure the health, welfare, and safety of the public through the accreditation of certification programs/organizations that assess professional competence. The NCCA *Standards for the Accreditation of Certification Programs* addresses the structure and governance of the certifying agency; characteristics of the certification program; information required to be available to applicants, certificants, and the public; and recertification initiatives of the credentialing agency. A subset of nine standards for assessment instruments are noted by the NCCA and are provided in Appendix B. Each standard is further divided into one or more "Essential Elements" that are identified by letters. An entry in Column II in one of the following tables of "13-B" would indicate a reference to Standard 13, and Essential Element B under that standard.

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 Test Takers (7 guidelines). An entry of "B-D1" in Column III in one of tables below would indicate a reference to 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 Test Takers* (Joint Committee on Testing Practices, 1998)

This source provides guidance for testing professionals, grouped into guidelines related to the (A) Rights of Test Takers (10 statements) and (B) Responsibilities of Test Takers (10 statements). Some statements are followed by finer-grained elaborations that 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 Section A (Rights of Test Takers), Guideline 4, and elaboration b. An abridged copy of the *Rights and Responsibilities of Test Takers* is included with this report as Appendix D.

Table A-1

Crosswalk of Test Development Checklist Elements and Professional Guidelines

Checklist Element	Professional Guidelines (Sources*)			
	I	II	III	IV
TD1	1.0, 1.1, 4.1, 7.1, 9.2, 9.3, 9.15	11-D	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		A-D1, U1	A-4b, e
TD4	1.1, 4.2		A-D1, U1	
TD5	4.0, 4.1	10-A, B; 11-A, B, C, D		
TD6	4.0, 4.1, 4.2, 4.6, 4.12	10-A; 11-A, B, C, D	A-D2, U2	A-3a
TD7	4.2, 4.6		A-D3, U3	A-4a
TD8	4.7, 4.9, 7.4	11-A		
TD9	4.6, 4.8			
TD10	4.6, 4.7, 4.8	11-C		
TD11	4.6, 4.8			
TD12	4.15, 4.16, 7.8, 7.13	11-A	B-D3,U3	A-1a
TD13	4.3`			
TD14	4.9	11-D		
TD15	4.9, 7.5			
TD16	4.9, 7.5			
TD17	4.9		A-D9, U9	

Checklist Element	Professional Guidelines (Sources*)			
	I	II	III	IV
TD18	4.0, 4.4, 4.10, 7.4, 7.6			
TD19	4.0, 4.4, 4.10, 7.4, 7.6	14-A	A-D5, U5	A-3a
TD20	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

Checklist Element	Professional Guidelines (Sources*)			
	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		A-D8, U8; B-D2, U2	A-3b; 4l
TA7	3.10		B-D2	

Checklist Element	Professional Guidelines (Sources*)			
	I	II	III	IV
TA8				A-9d
TA9	3.10, 9.18		D-D3	A-4f
TA10	7.7			A-6c
TA11	4.15, 4.16, 7.8, 7.13		B-D1, U1	
TA12	6.3		D-D6, U6	
TA13	6.14, 8.2, 8.5	13-A; 18-A	D-D5	A-4h; 8b, e, g
TA14				
TA15	6.6, 6.7	17-A	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
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 Evidence Checklist Elements and Professional Guidelines

Checklist Element	Professional Guidelines (Sources*)			
	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	13-A; 14-A	A-D5, U5, D9, U9; C-D1	A-3a
R2	2.0, 2.11, 2.19, 7.5	14-A	C-D6, U6	A-3a
R3	2.19	14-A		
R4	2.6, 2.11, 2.12			
R5	2.13, 2.14	14-A		
R6	2.13, 2.14	14-A	C-D4	A-8e
R7	4.14			
R8	2.7, 4.20, 4.21			
R9	2.7, 4.20, 4.21	13-B		

*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 Evidence Checklist Elements and Professional Guidelines

Checklist Element	Professional Guidelines (Sources*)			
	I	II	III	IV
V1	1.0, 1.1, 4.1, 7.1, 9.2, 9.3, 9.15, 11.1		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, 11.2, 11.3	10-A; 11–B, E	A-D2, U2	
V7	1.11, 4.6, 4.12, 11.3, 11.13	10-A, B		
V8	1.0, 1.8, 1.9, 1.10, 1.14, 3.6, 3.8, 7.4, 7.5, 7.6			
V9	1.2, 3.3, 3.6, 3.11, 3.12			
V10	5.21, 5.22, 7.4		C-D4, U4	A-8e
V11	1.2, 1.18, 5.21, 5.23, 11.16	12-A, B		
V12			C-D4, U4	
V13	6.13, 9.5		B-D6, U6	
V14		4.1h		

*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

Checklist Element	Professional Guidelines (Sources*)			
	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			A-6c
SR7	5.1, 5.4	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	5.6	C-D1, U1	A-8b,e
SR13	6.10, 8.2, 9.15		C-D1, U1, D3, U3; D-D4	A-4e

Checklist Element	Professional Guidelines (Sources*)			
	I	II	III	IV
SR14	6.14, 8.2, 9.15		C-D5, U5; D-D3, D5	A-9a,b,c
SR15			C- D1, U1, D3, U3, D5, U5; D6, U6; D-D4	
SR16	3.10	6.1.2	C-D2, U2	A-8a
SR17	6.16, 8.5, 8.6, 9.19		D-D5	A-4b; 9a,d
SR18	6.9, 6.13			A-8h,i
SR19	8.11, 8.12, 9.17		D-D7	A-4h, 8d, 10a,b; B-8a, b
SR20	9.18		D-D3	A-4f; B-8b
SR21	6.14, 8.2, 9.15		D-D5	A-4b
SR22	6.8			
SR23	5.8, 5.9, 5.10, 5.11, 7.2			
SR24	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

The Standards for the Accreditation of Certification Programs

(National Commission for Certifying Agencies, 2004)

ASSESSMENT INSTRUMENTS

Standard 10 - *The certification program must analyze, define, and publish performance domains and tasks related to the purpose of the credential, and the knowledge and/or skill associated with the performance domains and tasks, and use them to develop specifications for the assessment instruments.*

Essential Elements:

A. A job/practice analysis must be conducted leading to clearly delineated performance domains and tasks, associated knowledge and/or skills, and sets of content/item specifications to be used as the basis for developing each type of assessment instrument (e.g., multiple-choice, essay, oral examination).

B. A report must be published that links the job/practice analysis to specifications for the assessment instruments.

Standard 11 - *The certification program must employ assessment instruments that are derived from the job/practice analysis and that are consistent with generally accepted psychometric principles.*

Essential Elements:

A. Assessment instruments, including assessment items, exhibits, instructions to examinees, scoring procedures, and training procedures for administration of assessments, must be products of an appropriately designed and documented development process.

B. The content sampling plan for test items or other assessment components must correspond to content as delineated and specified in the job/practice analysis.

C. An ongoing process must exist to ensure that linkage between the assessment instruments and the job/practice analysis is maintained, as assessment components are revised and replaced over time. This linkage between assessment content and job/practice analysis must be documented and available for review by stakeholders.

D. Certification programs must follow a valid development process that is appropriate for assessment instruments.

E. A systematic plan must be created and implemented to minimize the impact of content error and bias on the assessment development process. Assessment content must be reviewed by qualified subject matter experts.

Standard 12 - *The certification program must set the cut score consistent with the purpose of the credential and the established standard of competence for the profession, occupation, role, or skill.*

Essential Elements:

A. Cut scores must be set using information concerning the relationship between assessment performance and relevant criteria based on the standard of competence.

B. A report must be published documenting the methods and procedures used to establish the standard of competence and set the cut score, along with the results of these procedures.

Standard 13 - *The certification program must document the psychometric procedures used to score, interpret, and report assessment results.*

Essential Elements:

A. The certification program must describe procedures for scoring, interpreting, and reporting assessment results.

B. For responses scored by judgment, developers must document training materials and standards for training judges to an acceptable level of valid and reliable performance. Any prerequisite background or experience for selection of judges must also be specified.

C. Candidates must be provided meaningful information on their performance on assessment instruments. Such information must enable failing candidates to benefit from the information and, if psychometrically defensible, understand their strengths and weaknesses as measured by the assessment instruments.

D. Reports of aggregate assessment data in summarized form must be made available to stakeholders without violating confidentiality obligations.

Standard 14 - *The certification program must ensure that reported scores are sufficiently reliable for the intended purposes of the assessment instruments.*

Essential Element:

A. Certification programs must provide information to indicate whether scores (including any subscores) are sufficiently reliable for their intended uses, including estimates of errors of measurement for the reported scores. Information must be provided about reliability or consistency of pass/fail decisions. When appropriate, information should be provided about the standard error of measurement or similar coefficients around the cut score.

Standard 15 - *The certification program must demonstrate that different forms of an assessment instrument assess equivalent content and that candidates are not disadvantaged for taking a form of an assessment instrument that varies in difficulty from another form.*

Essential Elements:

A. Equating or other procedures used to ensure equivalence and fairness must be documented, including a rationale for the procedure used.

B. When assessment instruments are translated or adapted across cultures, certification programs must describe the methods used in determining the adequacy of the translation or adaptation and

demonstrate that information attained from adapted and source versions of the assessment instruments produce comparable test scores and inferences.

Standard 16 - *The certification program must develop and adhere to appropriate, standardized, and secure procedures for the development and administration of the assessment instruments. The fact that such procedures are in force should be published.*

Essential Element:

A. Assessment instruments must be administered securely, using standardized procedures that have been specified by the certification program sponsor.

Standard 17 - *The certification program must establish and document policies and procedures for retaining all information and data required to provide evidence of validity and reliability of the assessment instruments.*

Essential Element:

A. Policies and procedures must ensure that items and forms of the assessment instruments are stored in a medium and method that emphasizes security, while being accessible to authorized personnel. Such policies must not only describe procedures for a secure system but also address actions required of personnel.

Standard 18 - *The certification program must establish and apply policies and procedures for secure retention of assessment results and scores of all candidates.*

Essential Element:

A. Organizational policy must determine the length of time that assessment results will be retained.

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 test takers 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 test takers 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 Test Takers

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], 1999). 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 test takers 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 test takers with disabilities who need special accommodations.

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

D9. Obtain and provide evidence on the performance of test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 Test Takers

TEST DEVELOPERS (D) / TEST USERS (U)

Under some circumstances, test developers have direct communication with the test takers 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 test takers 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 test takers 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 test takers 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 test takers or their parents/guardians with information about rights test takers 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 test takers or their parents/guardians with information about responsibilities test takers 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 test takers 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

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

1. Because test takers 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 test takers of these rights and responsibilities.
2. Because test takers 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 test takers 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 test takers are provided with reasonable access to testing services.
3. Because test takers 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 test takers from varying societal groups.
    b. Advise test takers 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 test takers 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 test takers 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 test takers about the type of feedback and interpretation that is routinely provided, as well as what is available for a fee. Test takers 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 test takers may have about basic test administration procedures.

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

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

5. Because that test takers 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 test takers are expected to pay, test professionals should:

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

    b. Inform test takers 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 test takers 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 test takers at a disadvantage.

7. Because test takers 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 test takers after the test takers 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 test takers have already taken (e.g., such as when applying for employment and a personnel examination is mandated).

    b. Explain to test takers 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 test takers 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 test takers 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 test takers 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 test takers about their options for obtaining a second interpretation of their results. Test takers may select an appropriately trained professional to provide this second opinion.

    e. Provide test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers know that they have specific responsibilities in addition to their rights described above.

1. Testing professionals need to inform test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers about the documentation needed to receive a requested testing accommodation.

b. Inform test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 test takers 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 Kosh is a doctoral student in Learning Sciences and Psychological Studies at the University of North Carolina at Chapel Hill. She is currently a research fellow at MetaMetrics, where she works on developing mathematics assessments and conducting educational measurement research on automatic item generation. Ms. 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. Ms. Kosh holds a M.A.T. in Secondary Mathematics from American University.

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., Germuth, A. A., Kosh, A. E., & Schmid, L. A. (2016). *A checklist evaluating credentialing testing programs*. Retrieved from http://wmich.edu/evaluation/checklists