Review of Indiana CORE Assessments for Educator Licensure

- TAC Recommendations
- Contractor Responses
- Evaluation

S. E. Phillips, PhD, JD
Assessment Law Consultant

Report prepared for the
Indiana State Board of Education

Table of Contents

Tables

Executive Summary

The purpose of this report is to memorialize recommendations presented at the November 8-9, 2017 Technical Advisory Committee (TAC) meeting. The recommendations were formulated after review of a set of TAC recommendations to the contractor for the CORE Assessment program, the contractor's responses, and other program data and information.

Several program strengths were identified by the evaluator. These included the appropriateness of the data and information provided by the contractor, proposed additional studies, training of educator review panels, research and subject matter expert input supporting the content validity of the examinations, and well-documented descriptions of passing score determinations consistent with best practices.

For the evaluation, the TAC recommendations were organized into the three psychometric categories of validity, reliability and standard setting. Following brief discussions of program data and information relevant to each TAC recommendation and its corresponding contractor response, short term and longer term evaluative recommendations were presented. These evaluative recommendations are summarized below.

Validity

TAC Recommendation #1 – Validity evidence aligned with test-taker volume

Contractor's Response – Responsive on content validity for all fields

Evaluator's Recommendations

- Field volume differences need more discussion
- Responsible authority (SBE) should consider formally adopting the content standards for each licensure examination
- Greater Content Advisory Committee (CAC) diversity or more authority for the Bias Review Committee (BRC) is needed to strengthen the item sensitivity reviews
- Creation/reporting of crosswalk tables of educator and student content standards would strengthen the validity evidence for the relevant examinations

TAC Recommendation #2 – Job analysis survey of practitioners by field

Contractor's Response – Responsive with proposed practitioner surveys

Evaluator's Recommendations

- Derive survey tasks from educator standards
- Tailor sampling plans to field volumes and ensure diversity
- Rate *importance* and *frequency*
- Revise educator standards and examination frameworks based on survey results
- Align items with revised standards and frameworks

TAC Recommendation #3 – Documentation of CAC & BRC demographics, activities and results

Contractor's Response – Mostly responsive with proposed DIF studies

Evaluator's Recommendations

- Calculate DIF statistics for subgroups with sufficient sample sizes
- Improve CAC diversity or increase BRC authority to review and evaluate item sensitivity and DIF results
- Document review/decision criteria for evaluating DIF results

Reliability

TAC Recommendation #4 – Report decision consistency and scoring consistency (CR items); Develop corrective action plans and/or explanations for reliability estimates that need improvement

Contractor's Response – Mostly responsive with recommended additions

Evaluator's Recommendations

- Report conditional SEMs at the passing scores
- Report rater agreement data for constructed response items
- Develop tailored action plans for examinations with very low reliability estimates after
  1) researching possible explanations (e.g., mixed content and short tests, as in the 32-item Social Studies/Fine Arts subtest, contribute to low KR_{20}s and may require state/contractor cooperation to improve), and
  2) considering studies to estimate alternate forms decision consistency reliabilities and/or judgmental split half reliabilities (with Spearman-Brown corrections for test length)

Standard Setting

TAC Recommendation #5 – Policy for periodic review of passing scores

Contractor's Response – Responsive

Evaluator's Recommendations

- Passing scores have already been appropriately reviewed and adjusted for some fields with low passing rates across multiple administrations
- Develop and document systematic and replicable criteria and procedures for continued periodic review and possible adjustment of passing scores

TAC Recommendation #6 – External validity evidence for reviewing passing scores for reasonableness

Contractor's Response – Partially responsive

Evaluator's Recommendations

- Instructional validity evidence is not applicable to licensure tests but may be useful for other purposes (e.g., institutional program evaluation). Licensure test content should be aligned to the minimum skills necessary for effective practice that the test is intended to assess. Using external information to review cut scores may create pressure for unwarranted changes.

Task Assignment

The General Counsel of the Indiana State Board of Education (SBE) requested expert feedback and recommendations related to the Indiana CORE Assessments for Educator Licensure program. Specifically, the following tasks were requested: (1) review written recommendations provided to the program contractor by the Technical Advisory Committee (TAC) and the contractor's responses, (2) attend the next TAC meeting to present an evaluation with recommendations, and (3) prepare a summary report. The program areas addressed by the evaluation and recommendations included:

- assessment procedures
- methodologies
- psychometrics
- professional standards and best testing practices
- psychometric and legal defensibility

specifically related to the topics of validity, reliability and standard setting addressed by the TAC recommendations and contractor's responses.

Materials Reviewed

Extensive program materials were made available for the review. These documents and related information included:

- Contractor's responses to TAC recommendations with multiple appendices of test information and data
- Technical Manual (updated)
- Correspondence related to TAC requests for information
- State and contractor CORE Assessment program websites
- Sample Items (CORE Academic Skills Assessment (CASA), English language arts, mathematics, life science, history, and secondary education were reviewed).

Further, review of the following additional materials was suggested to provide a more comprehensive evaluation of the psychometric and legal defensibility of the licensure testing program:

- TAC memoranda to the SBE
- Sample Test Administration Manual
- Accommodations Policy
- Test Security Policy
- Disparate Impact Data
- Item data for ONE large-N form of an examination (e.g., difficulty, discrimination, distractor analysis, alignment, bias and content reviews).

These matters were tabled for consideration at a later date.

Program Strengths

Based on review of the materials described above, the following program strengths were identified:

- Contractor responded with specific information related to the TAC recommendations
- Contractor provided relevant, accessible supporting data in the appendices to the response
- Contractor proposed additional data collection studies to strengthen evidence of adherence to professional standards
- Good training for educator panels was provided
- Thorough research and subject matter expert input supported the content validity of the licensure exams
- Well-documented descriptions of the standard setting activities were consistent with best practices.

Overview of Evaluation of Contractor's Responses

The state employs a Technical Advisory Committee (TAC) of educational and psychometric professionals to provide advice about its educator licensure testing program. Prior to this review, the TAC had provided six written recommendations to the contractor (Memorandum, 8/10/17) and the contractor had provided written responses (Report, 9/6/17). These TAC recommendations and contractor responses have been grouped into three psychometric categories for evaluation as follows:

- **Validity** TAC Recommendations 1, 2 & 3
- **Reliability** TAC Recommendation 4
- **Setting Passing Standards** TAC Recommendations 5 & 6.

The next sections of this report provide a professional evaluation of the contractor's responses and recommendations for each of the three psychometric categories (validity, reliability, setting passing standards) based on the program materials and information listed above that were reviewed. The recommendations were informed by the 2014 *Standards for Educational and Psychological Measurement* (*Test Standards*),[1] best practices in the industry, and the reviewer's psychometric and legal training and experience.

Validity

General Considerations

In a power point presentation by the Joint Committee responsible for revising the *Test Standards*, the authors answered two major questions about validity evidence for tests with the following professional advice:

> ***What makes a test valid?***
> A test is not valid in itself, but specific interpretations of the test scores are valid for particular uses. Evidence required to demonstrate validity varies with interpretation and use. For example, evidence of the alignment of test content to content standards is essential if test scores are interpreted as indicating mastery of the targeted content.
>
> ***How much validity evidence is required?***
> Validation is an open-ended process. Validity evidence should be collected prior to initial test use and further data analyzed as the test continues in operational use. Higher levels of evidence are required when test use has important consequences for individuals or for society. Professional judgment is needed to balance evidence supporting or contradicting test score interpretations for particular uses.[2]

The purpose of an educator licensure examination is to protect the public from unqualified practitioners by requiring educators seeking licenses to demonstrate minimum levels of relevant knowledge and skills for effective

[1] American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Testing,* Washington DC: AERA.

[2] Joint Committee, *Standards for Educational and Psychological Testing: Essential Guidance and Key Developments in a New Era of Testing.* (Sept. 12, 2014). Presentation announcing the release of the 2014 *Test Standards*, Washington DC.

practice in an entry-level position. The state uses licensure examinations with specified passing scores for this purpose. As the professional advice quoted above indicates, higher levels of validity evidence are required for licensure examinations because these tests have important consequences for licensure candidates who must achieve specified passing scores before being granted an educator license by the state.

The constructs being measured by educator licensure examinations are the minimum knowledge and skills for effective entry-level practice in specific fields, so content validity evidence is the primary and most appropriate validity evidence for these examinations. Educator licensure examinations are not designed to predict future performance, to distinguish levels of performance, or to match any particular instructional program. Therefore, educator licensure examinations are not employment tests and neither predictive nor instructional validity evidence is required.

The next sections separately address the three TAC recommendations related to validity: (#1) validity evidence aligned with test-taker volume, (#2) job analysis surveys of practitioners by field, and (#3) documentation of content advisory committee (CAC) and bias review committee (BRC) demographics, activities, and results. In each case, relevant 2014 *Test Standards,* summary program data and information, and specific recommendations are discussed.

TAC Recommendation #1

Validity Evidence Aligned with Test-taker Volume

> Contractor plans for addressing sources of validity evidence should reflect strategies that are appropriately aligned with the volume of test takers in a respective program.

The contractor responded to TAC Recommendation #1 by providing content validity evidence for the CORE Assessment educator licensure examinations. This evidence focused on the content and development standards for educators on which the licensure examinations were based. The procedures used and types of validity evidence collected were similar for all educator standards and examinations. Differentiation between high and low volume fields primarily involved the number of participating reviewers and field test sample sizes.

Indiana Developmental and Content Standards for Educators

The contractor provided content validity evidence for two types of educator standards and examinations: developmental and content. Standards were developed for 3 core academic skills, 4 developmental areas and 52 specialty fields. Relevant research and existing policy materials were reviewed, Indiana educators provided input (2-5 for developmental standards; 2-8 for content standards), subject matter experts and national organizations were consulted, and correlation studies were conducted. The Advisory Board adopted final versions of each set of standards in December 2010.

Assessment Blueprints and Items

Information about the examination blueprints and item development offered by the contractor in response to TAC Recommendation #2 also provided relevant validity evidence. The contractor provided an overview of how the test blueprints and items were developed to align with the content and developmental standards for educators adopted by the Advisory Committee.

In addition to evidence related to examination blueprint construction, the contractor also described how test items were developed and reviewed. Prior to field testing, all items were reviewed for appropriateness by a Content Advisory Committee (CAC) and a Bias Review Committee (BRC). Field testing samples varied in size based on whether the examination represented a low or high volume administration. Differential performance (DIF) statistics based on field test data should be calculated and evaluated for examinations with sufficient sample sizes.

Selected *Test Standards* (2014) for Validity Evidence

Standard 11.1

> *… a clear statement of the intended interpretations of test scores for specified uses should be made. The subsequent validation effort should be designed to determine how well this has been achieved for **all relevant subgroups**.* (emphasis added)

Standard 1.11 Comment

> *… The match of test content to the targeted domain in terms of **cognitive complexity** and the **accessibility** of the test content to all members of the intended population are also important considerations.* (emphasis added)

CORE Assessment Data

The number of educators who reviewed selected content/developmental educator standards by CORE Assessment field is summarized in Table 1. The data in Table 1 include the number of teachers and higher education reviewers for each of two rounds of review plus the combined total number of reviewers for each field. The first nine rows represent selected high and low volume fields with a smaller number of reviewers. For comparison, the last two rows of Table 1 present fields with a larger number of reviewers.

As indicated in Table 1, the number of reviewers was relatively small, even for the fields with larger numbers of reviewers. In addition, some fields had none or only 1 teacher reviewer (i.e., Reading, Early Childhood Generalist, School Setting: Elementary Education), or none or only 1 higher education reviewer (i.e., School Setting: P-12, Career & Technical Education: Marketing, English language arts). For most of the fields listed in Table 1, there appear to have been too few teachers and/or higher education reviewers to adequately represent the diversity of views of these subgroups of educators. However, such limited review would be acceptable if job analysis survey data were also obtained from larger, representative, diverse samples of educators.

TAC Recommendation #2

Job Analysis Survey of Practitioners by Field

> Documentation of the draft task statements or competencies; the tasks or competencies that characterize a given domain; the sampling plan for a survey of practitioners for each domain; and the data analyses, results, decision rules, and how the results were translated into the test blueprints that are used to develop CORE Assessment forms should be provided by the contractor. Such documentation would, necessarily, be provided separately for each field for which a CORE Assessment is administered.

In addition to aligning examination frameworks and items to the relevant content or developmental standards for educators, the educator standards themselves should be aligned to the results of appropriate job analysis surveys. Usually, the job analysis surveys are completed before examination frameworks and items are developed to ensure that the frameworks and items accurately reflect any content revisions or weighting adjustments indicated by the job analysis surveys. In addition, revisions to the content and developmental standards for educators based on the results

of the job analysis surveys would typically be made prior to adoption of the standards by the Advisory Committee.

TABLE 1
Numbers of Educators Reviewing Content/Developmental Standards by Field

FIELD	ROUND 1 (Draft)		ROUND 2 (Revised)		TOTAL*
	Teachers	Higher Educ	Teachers	Higher Educ	
Career & Technical Educ: Marketing	1	0	2	1	4
Early Child Generalist	1	1	0	3	5
Soc Stud: Govt & Citizen	1	0	1	2	4
School Setting: Elem Educ	1	2	0	2	5
School Setting: P-12	2	0	0	0	2
Life Science	1	2	1	0	4
Mathematics	2	1	3	1	7
Reading	0	1	0	1	2
ELA	3	0	2	1	6
School Librarian	1	3	4	3	11
Exceptional Needs Mild	1	2	4	2	9

SOURCE: Contractor's Response to TAC Recommendations, Sept 6, 2017, Appendices E & G
*An unspecified number of educators provided reviews for both rounds.

However, in this case, the examination blueprints and items had already been developed and aligned to the content and developmental standards for educators adopted by the Advisory Committee without the benefit of job analysis survey data. Consequently, the contractor proposed to conduct revalidation job analysis studies to confirm the appropriateness of the existing educator standards and examination frameworks.

Selected *Test Standards* (2014) for Job Analysis

Standard 11.3

When test content is a primary source of validity evidence … a close link between test content and the job … should be demonstrated.

Comment: … It is often valuable to also include information about the relative frequency, importance, or criticality of the [tasks].

Standard 11.13

The content domain to be covered by a credentialing test should be defined clearly and justified in terms of the importance of the content for credential-worthy performance …

Comment: Typically, some form of job … analysis provides the primary basis for defining the content domain. If the same examination is used in the credentialing of people employed in a variety of settings and specialties, a number of different job settings may need to be analyzed. … In tests used for licensure, knowledge and skills that may be important to success but are not directly related to the purpose of licensure (e.g., protecting the public) should not be included.

CORE Assessment Data

As indicated above, the data in Table 1 demonstrate that the samples of teachers and higher education reviewers for the CORE Assessment Content and Developmental Standards for Educators were too small and unrepresentative to adequately capture the diversity of educators' views. These data were also insufficient to provide adequate job analysis data as specified by the 2014 *Test Standards* for licensure examinations.

TAC Recommendation #3

Documentation of CAC and BRC Demographics, Activities & Results

In addition to the evidence from the job analysis, a detailed written summary of the Content Advisory Committee (CAC) and Bias Review Committee (BRC) plans and procedures should be provided to document the people, process, results, and decision rules used for each CORE Assessment field. (NB: A generic description for the collection of CORE Assessments is not acceptable.)

Bias Review Committee (BRC)

BRC members reviewed items for possible bias prior to field testing. They considered the content, language, possible offensiveness, stereotypes, fairness and diversity of the items for all subgroups of test takers.

For all fields (except Theater Arts, Health/PE, and Elementary Education Generalist), BRC recommendations were provided to the CAC for consideration. For the three exceptions, the BRC provided recommendations 10 months to 2^+ years after the CAC reviews were completed. It appeared from the documentation (Appendix L) that the BRC provided consensus recommendations but it was not clear if that meant 100% or majority agreement. The documentation also did not specify any decision rules for BRC minority dissents or for CAC consensus. In addition, the decision rules that were provided for CAC item validity votes lacked adequate justification.

The proportion of female to male representatives on the BRC was 5+/1 which may have mirrored the educator demographics of the state. Based on the list of participants provided in the documentation, BRC Hispanic representation was insufficient.

Content Advisory Committee (CAC)

CAC members reviewed items for match to objective, accuracy, freedom from bias, and job-relatedness. Specific criteria and decision rules were not included in the documentation. Based on the list of participants provided in the documentation, many CACs exhibited little or no diversity. For example, selected demographic information for CAC members included:

- Ethnic Representation: 24 fields (≈ 46%) all White (or unspecified)
  - agriculture, journalism, computer education, economics, mathematics, MS ELA, school librarian, vocal music, family & consumer science, marketing, early childhood generalist, blind & low vision, mild intervention, general music, health & physical education, MS math, MS science, MS social studies, school counselor, life science, historical perspectives, psychology, virtual instruction, sociology (N=2)
- Gender Representation – (female, male)
  - Predominately male: engineering (1, 6); economics (2, 6); instrumental music (2, 8)
  - Exclusively female: business (10, 0); early childhood (11, 0); intense intervention (10, 0); MS ELA (7, 0); school librarian (7, 0), family & consumer science (9, 0); blind/low vision (3, 0); mild intervention reading (4, 0); sociology (2, 0); virtual instruction (11, 0).

Some observers might question whether a committee composed of only majority members can adequately determine if an item is free from bias against unrepresented minority groups.

Selected *Test Standards* (2014) for Item/Test Reviewers

Standard 1.9

> *When a validation rests in part on the opinions or decisions of expert judges, … procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures … should indicate whether participants reached their decisions independently, and should report the level of agreement reached. …*
>
> *Comment: … The basis for specifying certain types of individuals … as appropriate experts for the judgment or rating task should be articulated. …*

Standard 3.3

> *Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test.*
>
> *Comment: … If sample sizes permit, it is often valuable to carry out separate analyses for relevant subgroups of the population. When it is not possible … test results may be accumulated and used to conduct such analyses …*

Validity Recommendations

Advice related to the TAC Recommendations for validity (#1, #2 & #3) and the corresponding responses of the contractor is presented in two categories: short term and longer term. Short term recommendations are the most critical and should be addressed as soon as possible. Longer term recommendations should be addressed when time and resources permit.

Short Term

Short term recommendations for strengthening the validity evidence for the CORE Assessments include conducting job analysis surveys, adjusting the decision-making procedures for evaluating potential differential item performance, and developing crosswalks between the educator standards and state content standards for students.

Job Analysis Surveys. Conduct job analysis surveys prioritized by field volume and criticality (i.e., where differences are most likely or educator reviews were least adequate). Recommended procedures for conducting the job analysis surveys and using the results are as follows:

- Develop task lists, conduct surveys, revise standards and examination frameworks, complete alignment studies
- Derive tasks from existing Content/Developmental Standards for Educators using existing items as exemplars to illustrate knowledge and skill expectations
- Create sampling plans that reflect field volumes and diversity
- Ask participating educators to rate the *importance* and *frequency* of each task
- Use the job analysis survey results to revise the Content and Developmental Standards for Educators where indicated and to appropriately weight the knowledge and skill categories specified in the examination frameworks
- Check the alignment of examination items with the revised standards and the revised frameworks.

Differential Item Performance. The 2014 *Test Standards* dictate that all examination items be appropriately screened for fairness to historically disadvantaged groups. There are two aspects to this activity: item sensitivity reviews prior to field testing and statistical evaluation of items based on field test data when sample sizes are sufficient.

Currently, sensitivity reviews are conducted by the Bias Review Committee (BRC) and their recommendations are considered by the Content Advisory Committee (CAC) before items are field tested. The contractor has proposed the calculation of differential item performance (DIF) statistics in the future using pooled data. Either the CAC or BRC should review these data. However, as indicated above, the CAC may not be sufficiently diverse for defensible decision-making based on DIF statistics. To remedy this weakness, it is recommended that CAC diversity be substantially increased or that the BRC be given additional responsibility for reviewing DIF statistics. In either case, the specific procedures and decision criteria utilized by the committee should be documented.

Crosswalks. For the moderate volume English language arts, mathematics and science examinations corresponding to subjects for which student standards and assessments are required by federal legislation,

construct (report) crosswalk tables relating Indiana educator and student standards to provide additional validity evidence for the content inclusion and weighting decisions used to generate the examination frameworks.

Longer Term

Longer term recommendations for strengthening the validity evidence for the CORE Assessments include prioritizing the most salient evidence, considering pre-equating of examinations, and using alternative statistical methods for evaluating DIF for low volume fields.

Prioritizing Validity Evidence. Neither instructional validity (alignment of tested content to preparation program instructional content) nor predictive validity (predicting college course grades) is a requirement for licensure examinations. The purpose of a licensure examination is to protect the public by measuring critical skills for effective, entry-level practice.[3]

Entry-year, educator ratings of effective practice provided by supervising administrators may constitute relevant criterion validity evidence for a licensure examination. However, such evidence may be limited by attenuation (only data for licensed, employed candidates can be collected), contamination (when test results have been used to determine licensure), availability (missing data), and unevenness in the quality and reliability of the administrators' ratings.

Pre-equating Examinations. For large volume fields where multiple forms of an examination are utilized, pre- and/or post-equating methodologies are necessary for establishing the statistical parallelism of the forms and accurately aligning passing scores across administrations. Such equating data provide supporting validity evidence for licensure tests.

Item response theory (IRT) models are particularly useful for test construction and equating, and among the commonly used IRT models, Rasch model analyses are most useful when field volumes are moderate (about 300-2000 test takers).[4] It is recommended that consideration be given to the use of Rasch model analyses for CORE Assessments in fields with sufficient test-taker volumes.

---

[3] Note: CASA Assessments measure entry level basic skills in reading, mathematics and writing for educator preparation programs.

[4] When the number of test takers is at least 2000, either the Rasch model or the three-parameter IRT model can be supported.

Alternative DIF Statistics for Low Volume Fields. There are many fields where the field volume is too small for the calculation of traditional DIF statistics. In such cases, it is recommended that the contractor consider whether the evaluation of items could be strengthened by using nonparametric procedures for estimating DIF. (These alternative statistical procedures may also be helpful for the five fields (world languages) for which no reliability data is currently available.)

Reliability

The reliability of the CORE Assessments was addressed by TAC Recommendation #4.

TAC Recommendation #4

Report Decision Consistency & Scoring Consistency (CR items); Develop Corrective Action Plans and/or Explanations for Reliability Estimates That Need Improvement

> In addition to traditional internal consistency reliability, evidence of decision consistency in addition to estimates of scorer or rater error from the scoring process for constructed response questions should be provided for each CORE Assessment. When reliability evidence is presented in technical reports or manuals for the program, additional narrative discussion is needed when the values do not support assertions of the reliability of the scores, scorers, or decisions. This discussion would include corrective action plans to improve the evidence or explanation for instances that can occur for very low volume fields about why target values may not be achieved.

Internal consistency (KR_{20}) reliabilities were available for most fields. These estimates were affected by form length which was 80 items for single CORE Assessment tests and 32 items for fields for which multiple subtests were required (e.g., elementary education generalist). The contractor also reported decision consistency estimates by field and generalizability estimates for the constructed response items on the CASA writing subtest.

Examinations for which reliability estimates were relatively low may have been affected by test length, heterogeneous content, and/or the distribution of candidate test scores, especially when passing rates were low and score distributions markedly skewed. Nonetheless, the collective effects of these factors, taken together, are complex. Consider the following two examples:

- **ELEMENTARY EDUCATION GENERALIST: SOCIAL STUDIES/FINE ARTS**
  - Content: multiple topics in social studies and fine arts
  - 32 items
  - **3-yr KR_{20} = .38 – .49**
  - Passing rates
    - 2013 = 36%
    - 2015 = 77% (adjusted passing score)

- **MATHEMATICS**
  - Content: number, algebra, measurement/geometry, statistics/probability; functions/trigonometry, calculus, discrete math; mathematics instruction and assessment
  - 80 items
  - **3-yr KR_{20} = .83 – .86**
  - Passing rates
    - 2013 = 21%
    - 2015 = 35% (adjusted passing score)

The social studies/fine arts test had fewer items, more heterogeneous content and a substantial change in passing rates while its KR_{20} reliability remained relatively low across three years. On the other hand, the mathematics test was longer, contained more homogeneous content, and had relatively low passing rates while its reliability remained relatively high across three years. In these examples, it appears that the most influential factors affecting internal consistency reliability estimates were test length and content heterogeneity.

Selected *Test Standards* (2014) for Reliability

Standard 11.14 (Credentialing)

> *Estimates of the [decision] consistency of test-based credentialing decisions should be provided in addition to other sources of reliability evidence*. [*See also* Standard 2.16 (General)].
>
> *Comment: … the consistency of decisions on whether to certify is of primary importance.*

Standard 2.7

> *When subjective judgment enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements. …*

Standard 2.14

> *… Where cut scores are specified for … classification, the standard errors of measurement should be reported in the vicinity of each cut score.*

CORE Assessment Data

Reliability estimates by program Year and field are summarized in Table 2 for selected fields. Three-year cumulative statistics are also reported. For each time period and selected field, the number of test takers (N), number of test items (n), decision consistency reliability (DC) and KR_{20} reliability (KR_{20}) are presented. Both reported reliabilities are on a scale from zero to one. Note that retests are permitted allowing for measurement errors to be corrected and remediation to produce successful outcomes.

The data in Table 2 demonstrate significant variability in reliability estimates across fields. For example, for the elementary education generalist subtests, social studies/fine arts (32 items) had a 3-year cumulative KR_{20} reliability of .38 while the corresponding value for mathematics (40 items) was .70. Both subtests had approximately 900-1000+ test takers per year. Passing rates varied significantly from 2013 to 2015 for both tests due to adjusted passing scores (34%-56% for social studies/fine arts; 25%-68% for mathematics – see Table 3). The reliability estimates for each subtest were relatively stable across years. The significant difference in KR_{20} reliability estimates for these subtests was likely attributable, at least in part, to differences in the heterogeneity of the tested content and differences in test length.

In contrast, the specialty area tests in English language arts and mathematics were statistically more similar but also demonstrated significant variability in reliability estimates. Both examinations had moderate volumes (585 and 353, respectively). Both examinations had the same larger number of items (80) and assessed relatively homogeneous content. But the mathematics examination had a much higher 3-year cumulative KR_{20} reliability estimate (.86) than English language arts (.64). Passing rates were adjusted for mathematics but not English language arts yet the KR_{20} reliability estimates for each examination remained consistent across the three years. In this case, the explanation for the difference in reliability estimates is unclear.

As indicated in Table 2, some fields with larger numbers of items and/or test takers have relatively low reliability estimates. Examinations with more than 50 items but unexpectedly low reliability estimates warrant further investigation. In particular, it is recommended that examinations with cumulative KR_{20} estimates less than .70 and cumulative decision consistency estimates less than .80 be prioritized for additional scrutiny.

TABLE 2
Reliability Estimates by Program Year and Field (Form A)

FIELD	2013-14 N	2013-14 n	2013-14 DC	2013-14 KR_{20}	2014-15*‡ N	2014-15*‡ n	2014-15*‡ DC	2014-15*‡ KR_{20}	3-Year Cumulative 2013-16* N	3-Year Cumulative 2013-16* n	3-Year Cumulative 2013-16* DC	3-Year Cumulative 2013-16* KR_{20}
CASA Reading	2033	32	.70	.67	842	32	.78	.63	1035	32	.85	.66
Mathematics	1975	32	.84	.78	1031	32	.78	.78	1251	32	.79	.79
**Writing	1770	33	.80	.66	936	33	.80	.65	1163	33	.81	.65
Engl Lang Arts	145	80	.66	.66	440	80	.74	.63	585	80	.74	.64
Mathematics	107	80	.94	.85	246	80	.88	.86	353	80	.89	.86
ELEM ED GEN Engl Lang Arts	990	40	.81	.56	984	40	.77	.55	1974	40	.71	.56
Mathematics	1043	40	.90	.71	1111	40	.87	.68	2154	40	.81	.70
Science & H/PE	930	32	.83	.53	909	32	.72	.52	1839	32	.66	.52
Social Std & FA	960	32	.86	.39	965	32	.82	.37	1925	32	.68	.38
Life Science	116	80	.83	.83	318	80	.83	.79	487	80	.81	.80
History	200	56	.78	.69	246	56	.84	.62	446	56	.72	.65
Elem Educ	2048	80	.83	.64	864	80	.83	.69	2284	80	.88	.65
Second Educ	1133	80	.90	.72	722	80	.91	.79	1228	80	.94	.72

SOURCE: Contractor's Response to TAC Recommendations, Sept 6, 2017, Appendices V & W

‡ Elementary & Secondary Education Form C

*CASA Form K

** Generalizability statistics for the writing CR section were .84, .90 and .90 for 2013-14, 2014-15 and 3-yr 2013-16, respectively.

Examinations with less than 50 items and low cumulative reliability estimates should improve if test length can be increased. If not, alternative corrective actions may be beneficial. For example, items with very low or very high difficulties may be contributing little information. Replacing such items with equally-content-valid items of more moderate difficulty may improve reliability. Replacing relatively low discriminating items with more highly discriminating items may also be helpful.

In addition, consider whether the tested item content is central to the assessed developmental/content standard or instead represents specialized content that is more ambiguous, arcane, inconsequential, or rarely needed. Replacing items testing less important and/or infrequently used knowledge and skills with items testing more important and/or frequently used knowledge and skills that also align with the assessed developmental or content standard may improve reliability estimates. Remind subject matter experts who serve as item writers and item reviewers that licensure examinations are intended to assess the *minimum* knowledge and skills *essential* for effective, *entry-level* practice.

Reliability Recommendations

The next sections present short term and longer term recommendations for collecting reliability evidence for the CORE Assessments.

Short Term

- Report rater agreement statistics for constructed response items
- Report conditional standard errors (SEMs) at the passing scores
- Develop corrective action plans for examinations with low internal consistency and decision consistency estimates after exploring alternatives and possible explanations for the observed results

Longer Term

- Consider designing studies to estimate alternate forms decision consistency reliabilities and/or judgmental split halves reliabilities (with Spearman-Brown corrections) for selected fields with low cumulative reliability estimates. Starting with larger volume fields that have adequate score variability and test taker diversity may be beneficial. Such studies should increase the accuracy of the reliability estimates

by estimating them directly and basing them on more parallel collections of items.

- For low volume fields where reliability estimates are not currently being reported, consider EAP Bayesian or nonparametric reliability estimates.

Setting Passing Standards

Passing standards for the CORE Assessments are addressed in TAC Recommendations #5 and #6.

TAC Recommendation #5

Policy for Periodic Review of Passing Scores

> A policy establishing periodic review of passing scores for examinations should be adopted for the CORE Assessments. These policies will often define the criteria associated with when to revisit the passing scores and will frequently correspond with the development or redevelopment of a given field (e.g., every 5 to 7 years, when there are significant changes to the content, significant changes in the pool of candidates, significant changes in curriculum or instructional practices at state educator preparation programs, or significant changes in credentialing policy in the state).

The state has already begun to review and adjust the passing standards[5] for selected fields with low passing rates. Standard setting panels[6] were convened to review and possibly adjust the passing scores for 10 examinations in December 2014 and an additional 10 examinations in June 2015.[7] In addition, in December 2014 the time limits for the mathematics, middle school mathematics and physics examinations were extended by 30 minutes.

The panels of educators who convened to review and consider adjusting the passing scores for the selected fields consisted of 6 to 16 members. Approximately 67% of panel members were teachers and 33% were higher education faculty. Approximately 25% were repeat participants.

---

[5] Depending on context, the term *standard* may refer to test content, professional practice, or test performance.

[6] The term *standard setting panels* is commonly used by psychometricians, but this evaluator prefers the term *standard recommending panels* because the final decision is typically made by a policy-making body (or official) with panel input.

[7] The passing standards for all 10 examinations reviewed in December 2014 and 5 of the 10 examinations reviewed in June 2015 were subsequently adjusted.

A modified Angoff methodology was used by panelists in Rounds 1 and 2 to make recommendations for readjusting the passing scores. In 2013, no impact data was shared with panel members but the 2014-15 panels were given impact data prior to providing a third round of ratings. In Round 3, panelists provided holistic, test-based recommendations. For 15 of the 20 examinations reviewed, passing scores were lowered and the new passing scores were effective six months later.[8] The original passing scores for the other 5 fields remained unchanged.

Note that, if sample sizes are sufficient, it is desirable to disaggregate impact data by subgroup (e.g., African-American, Hispanic) when passing standards are being finalized. Also note that the effective passing standard is lower with multiple retests.[9]

Finally, beyond consideration of the qualifications and demographics of panel members, it is also important to investigate potential response bias. For example, in 2013, only 47% of the invited panelists participated. Were the invitees who participated qualitatively different than the nonparticipants?

CORE Assessment Data

Original and adjusted passing scores for the 15 examinations that were reviewed in 2014-15 are summarized in Table 3. For each passing score, its percent correct (Pct), corresponding raw score (RS) and passing rate (Pass %) are reported. The number of items (n) and test takers (N) is also reported. Examinations for which the percent correct at the passing score is less than 60% are printed in bold type[10] and passing rates that remained relatively low after adjustment are highlighted.

At the November 8, 2017 TAC meeting, a Department of Education (DOE) staff member stated that most of the licensure candidates who take middle school content examinations already hold elementary licenses and are seeking an additional endorsement for which no additional preparation or coursework is required. If so, lower passing rates for these tests may

[8] Note that the Angoff methodology, commonly used by licensure testing programs, tends to set higher passing standards than the Bookmark method, which has been widely used by student testing programs. Both are professionally acceptable, consistent with the 2014 *Test Standards,* and produce defensible results with appropriate training and implementation.

[9] Millman, J. (Aug. – Sept. 1989). *If at First You Don't Succeed,* 18(6) EDUC. RESEARCHER, 5-9.

[10] Note that differences in the difficulty of the examinations for different fields may support passing standards corresponding to lower or higher percent correct values.

appropriately reflect lower levels of required content knowledge and skills among less-prepared test takers.

TABLE 3
Adjusted Passing Standards

FIELD		n	N	2013 CUT Pct	2013 CUT RS	2013 CUT Pass %	2015 CUT Pct	2015 CUT RS	2015 CUT Pass %
English Learners		80	54	71%	57	35%	63%	50	61%
Middle School ELA		80	68	74%	59	7%	68%	54	35%
Middle School Mathematics		80	186	71%	57	2%	***56%***	45	19%
Mathematics		80	160	64%	51	21%	***58%***	46	32%
Middle School Science		80	75	66%	53	15%	63%	50	23%
Physical Science		80	18	71%	57	28%	***59%***	47	67%
ELEM EDUC GEN	Read/ELA	40	1340	73%	29	38%	65%	26	66%
	Mathematics	40	1329	75%	30	25%	60%	24	68%
	Science/Health/PE	32	1312	75%	24	41%	66%	21	73%
	Social Studies/Fine Arts	32	1300	69%	22	34%	63%	20	56%
Early Child Gen:	Read/ELA	40	229	75%	30	24%	60%	24	80%
	Mathematics	40	214	73%	29	54%	63%	25	85%
	Science/Health/PE	32	226	72%	23	50%	63%	20	81%
	Social Studies/Fine Arts	32	224	72%	23	36%	***59%***	19	77%
Life Science		80	238	65%	52	38%	60%	48	53%

SOURCE: Contractor Response to TAC Recommendations, Sept 6, 2017, Appendix Y

Selected *Test Standards* (2014) for Setting Passing Standards

Standard 5.22

> *When cut scores defining pass-fail … are based on direct judgments about the adequacy of … test performances, the judgmental process should be designed so that the [panelists] can bring their knowledge and experience to bear in a reasonable way.*

Comment: … feedback on the pass rates entailed by provisional proficiency standards, and other forms of information may be beneficial in helping participants to reach sound and principled decisions.

TAC Recommendation #6

External Validity Evidence for Reviewing Passing Scores for Reasonableness

> The cut scores used for CORE Assessments should be reviewed in combination with external validity evidence that may include performance disaggregated by institutions, course taking patterns of candidates, and candidate subgroups (e.g., traditional pathway, alternative certification). These additional analyses may be useful to policymakers in their evaluation of the reasonableness of the passing standard when comparing expectations to observed performance.

Accurately matching test data with external data may be challenging. Self-reported demographic information collected at test registration may not exactly match information from external sources. Data obtained from external sources may also be incomplete. Although external validity evidence is not directly relevant to the primary licensure testing purpose of protecting the public via the establishment of minimum knowledge and skill requirements for effective entry-level practice, such information may provide useful diagnostics for institutions evaluating their programs.

Impact data is a good source of reasonableness when setting passing scores for licensure tests. When collecting validity evidence for licensure tests, content validity is paramount.

Selected *Test Standards* (2014) for External Validity Evidence

Standard 5.23

> *When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.*
>
> *Comment: … With many [tests] used in credentialing, suitable criterion groups (e.g., successful versus unsuccessful practitioners) are often unavailable. Nevertheless,*

when appropriate and feasible, the test developer should investigate and report the relation between test scores and performance in relevant practical settings. …

Chapter 11: Workplace Testing and Credentialing

Adjusting the cut score periodically … implies that standards are set higher in some years than in others, a practice that is difficult to justify on the grounds of quality of performance.

Standard Setting Recommendations

The next sections present short term and longer term recommendations for setting passing standards for the CORE Assessments.

Short Term

- Provide disaggregated impact data by subgroup (e.g., African-American, Hispanic) when passing standards are being finalized if sample sizes are sufficient
- Investigate potential response bias in panel acceptances
- Develop a policy plan for systematic re-validation of passing scores.

Longer Term

- Investigate other standard setting methods (e.g., the Bookmark procedure) that may be more intuitive for panelists and provide more direct feedback to assist panelists in clearly communicating their holistic recommendations
- Exercise caution when collecting external validity evidence for passing standard reviews
  - Evidence of differences in test performance by instructional program or institutional demographics has limited relevance when adjusting passing standards for licensure examinations but may assist institutions to appropriately revise their curricula and instruction
  - Potentially relevant criteria for effective practice may be unavailable or of limited variability and reliability

Summary Evaluation of Contractor's Responses

The following list summarizes the evaluation of the contractor's responses to the TAC recommendations:

- **TAC RECOMMENDATION #1** – RESPONSIVE ON CONTENT VALIDITY FOR ALL FIELDS
  - Field volume differences need more discussion
  - Responsible authority (SBE) should consider formally adopting the content standards for each licensure examination
  - Greater CAC diversity or more authority for the BRC is needed to strengthen the item sensitivity reviews
  - Creation/reporting of crosswalk tables of educator and student content standards would strengthen the validity evidence for the relevant examinations
- **TAC RECOMMENDATION #2 –** RESPONSIVE WITH PROPOSED PRACTITIONER SURVEYS
  - Derive survey tasks from educator standards
  - Tailor sampling plans to field volumes and ensure diversity
  - Rate *importance* and *frequency*
  - Revise educator standards and examination frameworks based on survey results
  - Align items with revised standards and frameworks
- **TAC RECOMMENDATION #3 –** MOSTLY RESPONSIVE WITH PROPOSED DIF STUDIES
  - Calculate DIF statistics for subgroups with sufficient sample sizes
  - Improve CAC diversity or increase BRC authority to review and evaluate item sensitivity and DIF results
  - Document review/decision criteria for evaluating DIF results
- **TAC RECOMMENDATION #4 –** MOSTLY RESPONSIVE WITH RECOMMENDED ADDITIONS
  - Report conditional SEMs at the passing scores
  - Report rater agreement data for constructed response items

  - Develop tailored action plans for examinations with very low reliability estimates after
    1) researching possible explanations (e.g., mixed content and short tests, as in the 32-item Social Studies/Fine Arts subtest, contribute to low KR_{20}s and may require state/contractor cooperation to improve), and
    2) considering studies to estimate alternate forms decision consistency reliabilities and/or judgmental split half reliabilities (with Spearman-Brown corrections for test length)

- **TAC RECOMMENDATION #5 –** RESPONSIVE
  - Passing scores have already been appropriately reviewed and adjusted in some fields with low passing rates across multiple administrations
  - Develop and document systematic and replicable criteria and procedures for continued periodic review and possible adjustment of passing scores
- **TAC RECOMMENDATION #6 –** PARTIALLY RESPONSIVE
  - Instructional validity evidence is not applicable to licensure tests but may be useful for other purposes (e.g., institutional program evaluation). Licensure test content should be aligned to the minimum skills necessary for effective practice that the test is intended to assess. Using external information to review cut scores may create pressure for unwarranted changes.

Additional information on these topics can be found in Phillips, S.E. (2017). Legal Issues for Credentialing Examination Programs, Invited chapter in Davis-Becker, S. & Buckendahl, C.W. (Eds.), *Testing in the Professions: Credentialing Policies and Practice,* New York: Routledge.