Review of Contractor Responses to the Indiana Department of Education Technical Advisory Committee Recommendations Regarding the Indiana CORE Assessments for Educator Licensure

Richard M. Luecht, Ph.D.

University of North Carolina at Greensboro

November 2017

This review follows a request by the Indiana State Board of Education and the Indiana Department of Education to appraise documentation prepared by Pearson Evaluation Systems in response to six recommendations offered by the Indiana Department of Education (IDOE) Technical Advisory Committee (TAC) about the *Indiana CORE Assessments for Educator Licensure* (CORE Assessments). The CORE Assessments are comprised of three clusters of assessments: (1) three *Core Academic Skills Assessments* (CASA) in Reading, Mathematics and Writing; (2) four tests assessment pedagogical knowledge and skills (Early Childhood Education, Elementary Education, Secondary Education, and P-12 Education); and (3) 52 content-specific examinations in areas ranging from Business to World Languages. Pearson Evaluation Systems develops, delivers, analyzes and generates scores and pass/fail decisions for all CORE Assessments.

From an operational testing perspective, the volume of research, preparatory and test development work, test-delivery logistics, analysis, and processing needed to develop, implement and maintain the CORE Assessments, as a system, cannot be trivialized. There are 59 unique score scales that need to be maintained over time along with defensible pass/fail decisions. New test items and test forms also need to be generated on an annual basis to deal with item exposures. Considering that Pearson Evaluation Systems was awarded the contract to develop the CORE Assessments in Fall 2011 and operationally implemented the first sets of test forms in July 2013, it should not be surprising that there are some flaws in the system. No testing program is without flaws and all test-scores are fallible. The question is, does the testing agency and its contractors appear to be making reasonable progress toward continually improving the quality of every test form and the score scales over time—that is, *are there tangible efforts that suggest positive modifications to the item development and test design/assembly practices-ideally modifications aimed at maximizing the precision of scores for the intended licensure purposes, demonstrable evidence that Pearson Evaluation Systems is applying strong, proven quality-control mechanisms to systematically reduce any the myriad of sources of measurement errors, and ultimately that the testing agency and its contractor are taking steps to continually evaluate the fairness and validity of the test items, the processing procedures, and the scoring practices used*?

1.0 The TAC's Six Recommendations

The IDOE TAC offered six important recommendations about the CORE Assessments. Those are listed below.

1. *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.*
2. *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 blueprint 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.*
3. *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 provide 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.)*
4. *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 the can occur for very low volume fields about why target values may not be achieved.*
5. *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).*
6. *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.*

Recommendations #1 to #3 relate to documentation of the validity framework and test development practices that undergird the CORE Assessments. Recommendation #4 focuses on documenting various types of measurement and decision errors associated with the 59 score scales. Recommendations #5 and #6 refer to standard setting practices and evidentiary requirements to demonstrate validity and fair application of the cut scores within the state.

2.0 Contractor's Response to the IDOE TAC Recommendations

Pearson Evaluation Systems' response to the six TAC recommendations was to generate a 345-page report[1] (including appendices containing background reports and numerous tables). The body of the report addressed each of the six TAC recommendations, referring as needed to the appendices for supporting detail or as documented evidence of their assertions. At a somewhat abstract level, Pearson Evaluation Systems therefore appears to have at least addressed all the TAC's recommendations. However, a more directed and in-depth evaluation aimed at the *quality* of the evidence suggests room for improvement.

Most of the documentation in the Pearson Evaluation Systems report falls into two categories: (i) content validity evidence concerning the alignment between current practices and "standards" and (ii) empirical evidence—largely statistical in nature—related to the quality of the scores and cut scores used for making teacher and educator licensing decisions. Sections 2.1 and 2.2 evaluate the quality of these two classes of evidence and highlight some potential deficits in Pearson Evaluation Systems' response.

2.1 Content Validity Evidence and the Alignment of Standards to Test Scores

In professional licensure contexts, *content validity* refers to the correspondence between the test items and the knowledge and skills deemed necessary for initial license to practice. However, providing solid evidence of the alignment that documents in a coherent way the linkages between test item content and the knowledge, skills and abilities (KSAs) needed for professional practice often depends on a nebulous chain of inferential reasoning. Figure 1 shows the developmental components for a typical professional licensure test.

The inference chain begins with information gathering to describe the relevant content knowledge and performance-related behaviors in the professional domain of interest. There are multiple sources that can be used in this information-gathering phase including the authoritative professional literature, consultations with higher education faculty and other subject-matter preceptors, and discussions with supervisors and recently licensed practitioners. A key next step is to carry out a *practice analysis*. A practice analysis may involve convening focus groups comprised of subject-matter experts (SMEs) who work with psychologists and/or test development experts to craft a *survey* that articulates a wide range of discrete, specific task statements that describe what practitioners might do in their jobs. Content knowledge requirements can also be included in the survey.

[1] See Pearson Evaluation Systems (2017).

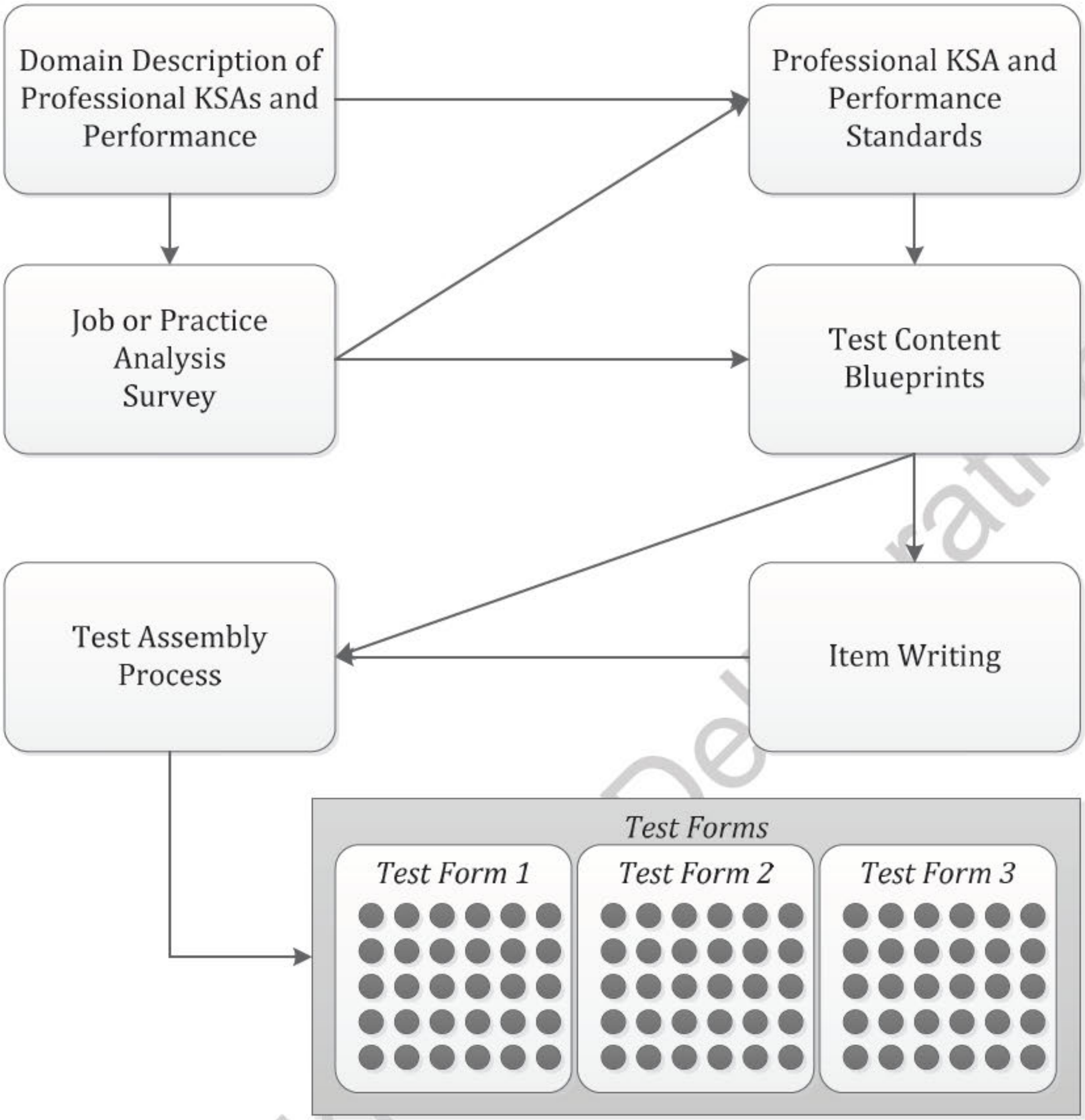

Domain Description of Professional KSAs and Performance
Professional KSA and Performance Standards
Job or Practice Analysis Survey
Test Content Blueprints
Test Assembly Process
Item Writing
Test Forms
Test Form 1
Test Form 2
Test Form 3

Figure 1. The Inference Chain from Domain Descriptors to Test Forms

The survey is then ideally administered to a broad and representative sample of newly licensed practitioners in the field of interest. Formal statistical sampling can be used to ensure that the results generalize to the larger target population. The survey participants rate each task or knowledge element in terms of three factors: (a) whether or not they do the task (or need to understand the knowledge element) ; (b) how frequently they perform the task/use the knowledge; and (c) their perception of how important the task/knowledge is in their job. Most practice analyses ask the respondents to focus on their experiences in practice, not on hypothetical or vaguely "typical" practitioners.

The survey results are analyzed and the most frequent and/or most important tasks/knowledge elements are used to generate KSA and performance standards for professional practice. The professional practice KSA and performance standards may be further refined by using follow-up focus groups and/or additional surveys administered to targeted expert samples. The next step is to generate formal test content specifications: the *blueprint*. Test content blueprints can be lists or outlines of key topics, descriptors of the depth of knowledge required (Webb, 2005), and/or skill-focused brief statements of performance. The test content blueprints also include the weights, item counts (exact quantities or ranges of acceptable counts), or raw-score points that must be included on every test form. The practice analysis results may be incorporated into the required content blueprint allocations as indicated in Figure 1. Finally, the test content blueprints are used in a test assembly process[2] to generate content- and [hopefully] statistically parallel test forms.

> Most (well-designed) test-form assembly specifications include three types of requirements: (i) the test length; (ii) the content blueprint that indicates all content-specific elements to be covered by the test and associated relative weights or prescribed [exact or ranges of] counts of items or points per element; and (iii) statistical specifications that can range from a target difficulty for each test form to more elaborate specifications that impact the properties of the score scale (e.g., minimum reliability specifications). Auxiliary requirements such as the cognitive distribution of items in terms of depth of knowledge may also be incorporated into a test content blueprint. The reason these test specifications are essential is that they concretely describe the content and key statistical properties for building every test form.

If the process depicted earlier in Figure 1 HAD BEEN employed for the Indiana CORE Assessments, the IDOE TAC likely would not have questioned the content validity evidence for the 59 tests (see Recommendations #1 to #3). A streamlined process was instead used; it by-passed the formal practice analysis survey altogether. That procedural omission in the development process—whether driven by budgetary, logistical or contractual constraints—generates serious questions about the remaining test design and development steps in the inference chain depicted in Figure 1. However, to better understand the evidence deficits in alignment of items and test content to professional standards, it is important to understand what Pearson Evaluation Systems did.

After being awarded an initial contract to develop the Indiana educator standards, Pearson Evaluation Systems staff independently collected, collated and organized a substantial amount of information and online documentation from various teacher-educator practice and content knowledge sources. "Development specialists" and subject-matter experts working for the contractor apparently used this information to prepare draft standards. The draft standards were next reviewed twice by select Indiana educators

[2] Test assembly practices range from manual generation of test forms using simple item queries and sorts to select the test items for each form to sophisticated automated test assembly algorithms that can create hundreds or thousands of content-parallel and statistically isomorphic test forms in milliseconds.

(IDOE participants, subject-matter experts, practicing teachers and higher education faculty involved in teacher education). The first review asked a limited number of *invited* Indiana educators to review the draft standards and provide feedback/suggestions for improvement. For the most part, each set of draft standards was reviewed by three individuals. The feedback was used to modify the draft standards—in some cases dropping irrelevant standards, but also clarifying ambiguous language and providing exemplars where needed. Those modified standards where then reviewed by a second group of 287 Indiana educators—spread across almost 60 teacher licensure categories. That second group again provided feedback on the essential knowledge elements for each standard as well as the perceived alignment to supporting documentation for each standard. Although the second group of reviewers was slightly larger than the first group, it again comprised only *invited* educators. Furthermore, each set of draft standards was only reviewed by 2 to 5 individuals (with some notable exceptions like "Elementary Generalist" and "School Setting Development Standards: Middle School"—both of which had more than 20 reviewers looking at the applicable standards). The National Council on Teacher Quality and the Indiana Association of Colleges for Teacher Education were also invited to review and comment on the draft standards.

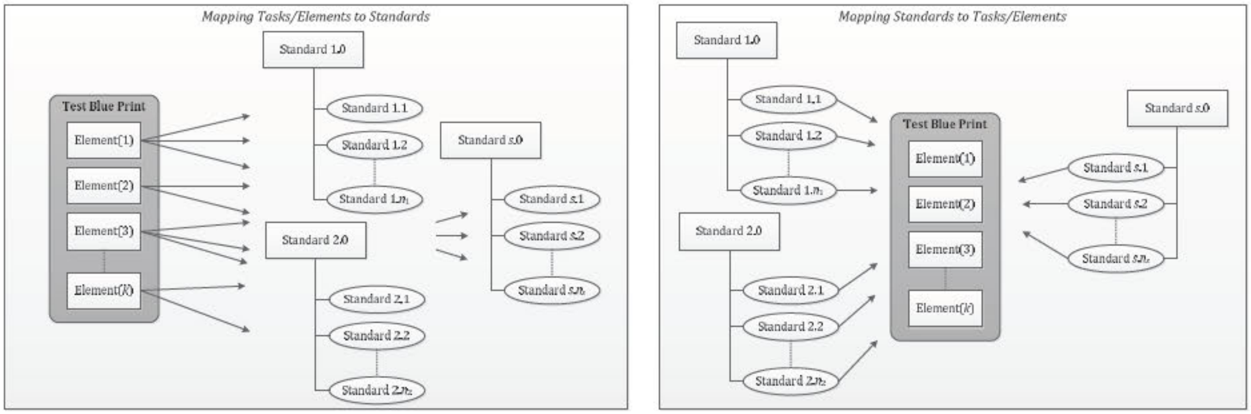
The Pearson Evaluation Systems' report refers to a follow-up "standards correlation[3] study" that apparently provided the formal descriptors, essential knowledge elements, and alignment of each Indiana teacher and educator standard to state and national standards documents. It is not clear who was included in this final alignment review (other than a rather vague allusion to "content experts"). The results of this final alignment study were passed onto the Indiana Advisory Board to the Division of Professional Studies (ABDPS). The ABDPS approved the *Indiana Developmental and Content Standards for Educators* in December 2010.

There are multiple "standards" referenced in Pearson Evaluation Systems' report, including but not limited to: (a) the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014); (b) the *Rules for Educator Preparation and Accountability* (REPA) standards; (c) the *Indiana Academic Standards* (doe.in.gov/standards); (d) the *Indiana Core Standards* (doe.in.gov/standards); (e) the *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects* (CCSS.org); (f) the *Common Core State Standards for Mathematics* (CCSS.org); (g) the *International Society for Technology in Education Standards* (ISTE.org);(h) the *InTASC Model Core Teaching Standards* (CCSSO.org); (i) the *National Council of Teachers of English* (NCTE.org); and (j) *the National Council of Teachers of Mathematics* (NCTE.org).

[3] The term "correlation" was used to denote logical connection/alignment between the Indiana educator standards and national standards. It does not imply a statistical correlation or covariance.

Pearson was subsequently awarded a second contract to develop the CORE Assessments. It is essential to understand that the approved *Indiana Developmental and Content Standards for Educators* are NOT test content blueprints (test-form assembly specifications) for any of the 59 CORE Assessments. Rather, Pearson Evaluation Systems apparently started with those standards and then crafted the 59 blueprints, with input from the IDOE.

In addition to the lack of a larger-scale practice analysis with appropriate and representative sampling (see Figure 1, earlier), the lack of details about the connections between the test-form blueprints for the 59 CORE Assessments and the official *Indiana Developmental and Content Standards for Educators* should be questioned—which the TAC did. Taking inherent *ambiguity* as a given in most content alignment studies, the actual alignment process can be carried out in two ways. One approach is to map individual knowledge, skill or performance *standards* to specific job-related tasks, test items, or content blueprint elements. The other approach is to map the job-related tasks, test items or elements to the standards. Figure 2 shows a conceptualization these two approaches.



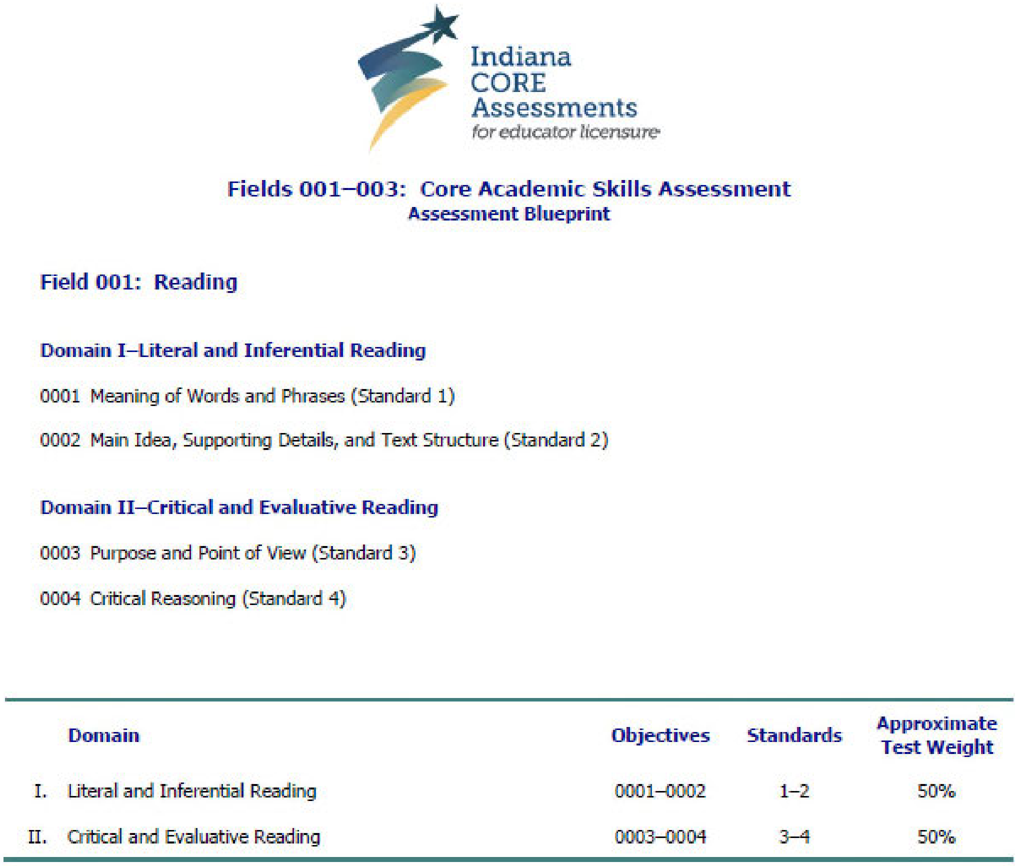
Mapping Tasks/Elements to Standards

Mapping Standards to Tasks/Elements

Figure 2. Two Approaches for Aligning Test Content to Standards

It is not clear whether Pearson Evaluation Systems used either of the two approaches shown in Figure 2 to build and justify the CORE Assessment blueprints. The blueprints are, fortunately, available online[4]. Each blueprint document contains a list of knowledge or skill elements and one or more standards (by number). Logical "domain" clusters of elements associated with multiple standards are then listed in a table embedded in the blueprint document along with percentages denoting the constrained representation of content on the test forms—at least for most of the examinations. Figure 3 provides an example for the CASA in Reading.

[4] See http://www.in.nesinc.com/PageView.aspx?f=HTML_FRAG/GENRB_PrepFramework.html



Indiana
CORE
Assessments
for educator licensure

Fields 001–003: Core Academic Skills Assessment
Assessment Blueprint

Field 001: Reading

Domain I–Literal and Inferential Reading

0001 Meaning of Words and Phrases (Standard 1)

0002 Main Idea, Supporting Details, and Text Structure (Standard 2)

Domain II–Critical and Evaluative Reading

0003 Purpose and Point of View (Standard 3)

0004 Critical Reasoning (Standard 4)

	Domain	Objectives	Standards	Approximate Test Weight
I.	Literal and Inferential Reading	0001–0002	1–2	50%
II.	Critical and Evaluative Reading	0003–0004	3–4	50%

Figure 3. A Sample Blueprint for CASA Reading

As noted above, having the blueprint documents online is good; however, the procedures and evidence for linking the elements to standards and the process for determining the content blueprint item-count/points percentage constraints—termed "weights" in the Pearson Evaluation Systems' report—remains somewhat of a mystery.

Clearly, the 59 test content blueprints were developed BEFORE any items were written and BEFORE the Content Advisory Committees (CACs) and Bias Review Committees (BRCs) reviewed any test items. Pearson Evaluation Systems also employed professional item writers (current and former teachers) and provided the necessary item-writing training. In short, the contractor appears to have independently handled all of the interim steps with little outside input (or technical oversight).

The items were then reviewed by the BRCs and the CACs. The primary role of the BRC was to flag items that exhibited potential bias with respect to content choices, language usage, offensiveness, stereotypes, fairness and diversity. The BRC recommendations were passed on to the CACs. The CACs reviewed every test item for content accuracy and appropriateness, with the options to accept, revise or delete the items. They also individually provided "valid" or "not valid" ratings for every item, after any final revisions[5].

In their response to the TAC, Pearson summarized the basic demographics composition of the six BRCs and the fifty-eight CACs[6] and described the basic procedural training and directives issued to the committees. The CACs met at six different "conferences" between 2012 and 2015. Each committee had 10 to 11 members on average (the range of participants, however, was 4 to 22). Members of select CACs (the CASA, the School Administrator examinations, and six World Language tests) were also invited to participate in rubric refinement and range-finding for constructed-response items included on those tests. These range-finding committees were comprised of four to eight participants who were also SMEs for the corresponding licensure area. In addition to scoring pilot-test papers, the range-finding meetings also produced initial rubric-point exemplars used for operational scoring (i.e., termed "anchors" in the Pearson Evaluation Systems report).

In sum, Pearson Evaluation Systems clearly provided most of the requested background information in response to the TAC's Recommendations #1 to #3. However, there were evidence deficits that could lead to challenges to the content validity of the 59 examinations (see Section 3.0 for a discussion of risks).

2.2 Empirical Evidence About the Quality of the 59 Score Scales/Cut Scores

The empirical [statistical] evidence in the Pearson Evaluation Systems report was definitely more concrete than the content validity and alignment evidence, but also tended to be sparse for various reasons[7]. The contractor carried out correlational analyses aimed at quantifying the accuracy of the scores (reliability) and did additional analyses to produce classification consistency results for most of the examinations, but could have produced a more coherent set of arguments showing that the reported scores and classification decisions uniformly meet industry standards. That statement does NOT imply that any of the 59 CORE Assessments fail to meet standards—merely that the documentation lacks coherence.

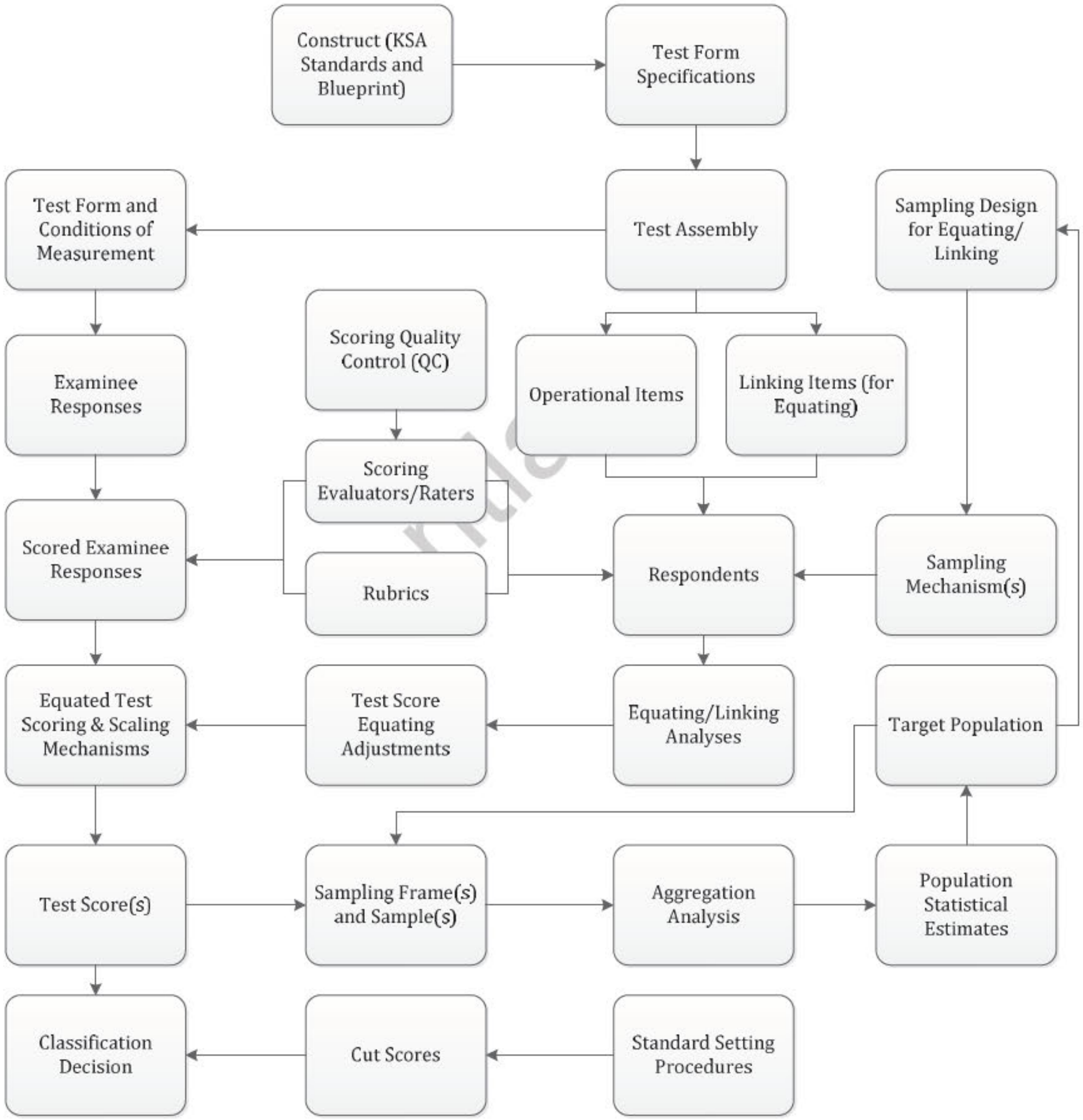
Consider Figure 4 as an example. This figure shows in reasonable detail the multiple procedures, analyses, data components, and data exchanges that occur behind the scenes in an operational testing program. The leftmost column of rounded rectangles reflects how

[5] The item validity classifications were used as a final flagging mechanism used by Pearson Evaluation Systems to retain, revise or delete items from the item banks.

[6] Some of the CACs reviewed items for multiple exams (e.g., there was a single CAC for all three sections of the CASA).

[7] Small sample sizes for some of the content-specific CORE Assessments cited as a reason for suppressing some statistical results (i.e., the inherent instability of statistics estimated with a limited number of data points). However, as noted further on in this report, there is an inconsistency in reasoning with that logic insofar as other uses of the same [unstable] data for equating, scaling and standard setting.

an individual candidate receives his or her score on a given test form, including any follow-up classifications (e.g., a pass/fail result for the test). The centermost columns denote how test forms are generated and how test score equating is carried out. Note that the resulting statistical equating functions are applied to individual test scores so that ALL examinees' scores are on a common scale. Equating also ensures that the approved cut scores derived from standard setting can be uniformly and fairly applied to scores from every test form. For example, if two examinees taking different test forms both correctly answered 76% of the items, and the cut score was 75%, we might conclude that both examinees should pass. But if one or both test forms were significantly easier than the test form used in standard setting, that "pass" decision could be wrong. Statistical equating ensures that the scores are on the same common scale as the cut scores.



Construct (KSA Standards and Blueprint)
Test Form Specifications
Test Form and Conditions of Measurement
Test Assembly
Sampling Design for Equating/ Linking
Scoring Quality Control (QC)
Examinee Responses
Operational Items
Linking Items (for Equating)
Scoring Evaluators/Raters
Scored Examinee Responses
Rubrics
Respondents
Sampling Mechanism(s)
Equated Test Scoring & Scaling Mechanisms
Test Score Equating Adjustments
Equating/Linking Analyses
Target Population
Test Score(s)
Sampling Frame(s) and Sample(s)
Aggregation Analysis
Population Statistical Estimates
Classification Decision
Cut Scores
Standard Setting Procedures

Figure 4. Overview of the Process of Equating, Test Scoring and Aggregating Test Scores

Finally, the rightmost column in Figure 4 shows how we might want to aggregate the equated test scores or classification decisions to report population-based estimates of relevant statistics like means, standard deviations, or correlations.

What is perhaps most important about Figure 4—other than showing that test assembly, scoring, score equating and statistical aggregation are quite complex—is that *every arrow in the figure denotes a potential source of error.* Some of those errors can be processing related, others are related to statistical sample, and still others are psychometric in nature (i.e., relating to threats to score validity, score imprecision, and decision inaccuracies). Standard errors associated with reported scores are typically reported to denote the imprecision along the score scale; however, some of the other errors implied by Figure 4 can filter down to the score for an individual examinee. Arguably, the primary job of any assessment contractor should be to demonstrate systematic and continual reductions of error (threats to reliability and validity) over time.

Appendices V, W and W1 of the Pearson Evaluation Systems report provide tables that contain eight types of statistics per test form: (i) test-taker counts; (ii) mean scale scores; (iii) standard errors of measurement[8] (SEM); (iv) decision consistency (see Livingston & Lewis, 1995 for technical details); (v) an adjusted score scale reliability coefficient; (vi) the number of selected-response (SR) items; (vii) a KR20 reliability coefficient for only the SR-based raw scores; and (viii) generalizability theory "dependability" coefficients (see Brennan, 2001) for those CORE Assessments that employ constructed-response (CR) items scored by human raters. These statistics were apparently based on census samples of test takers by year (2013-14 and 2014-15) and then combined 2013-16.

Note: the multi-year (2013-16) test statistics in Appendices W and W1 appeared to be inconsistent with the 2013-14 and 2014-15 results in Appendix V. Either the forms letters are misspecified or the counts were not properly accumulated across forms. For example, in 2013-14, N=2,033 candidates took Form A of the examination, *001 Core Academic Skills Assessment – Reading*. In 2014-15, N=409 took the examination (per Appendix X). However, in Appendices W and W1 (filtered for minimum counts of "N≤100" and "N≤30", respectively), the count for Form A of that examination is N=2,033. So, either the form letters denote different forms across years within examination title or the counts and—assumedly the other reported statistics in those tables—are not properly aggregated across the 2013-16 timeframe. Results reported here therefore only reflect the disaggregated results for 2013-14 and 2014-15.

[8] The standard error of measurement is computed as $SEM = Bs_x(1 - r_{xx'})^{1/2}$ where B is the scale score slope, s_x is the standard deviation of the estimated or equated observed scores and $r_{xx'}$ is a reliability coefficient.

The Pearson Evaluation Systems' response to the IDOE TAC states that the CORE Assessments annual technical manual includes ample evidence of reliability, scorer consistency (for CR items) and decision accuracy "*…to appraise the degree of dependability of the CORE Assessment scores for their intended use*." This an important point because it suggests that the CORE Assessments comply with basic technical reporting recommendations from the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). However, meeting reporting requirements is NOT the same as evaluating the quality of those reported statistics "*…for the intended use* [of scores and decisions]."

Figure 5 shows boxplots of the KR-20 reliability coefficients (SR items only). Reliability coefficients reflect the average precision across the score scale. Values of 0.85 or higher are typically considered to be acceptable to most higher stakes assessments. The Pearson Evaluation Systems report suggests that lower values might be acceptable for teacher licensure tests like the Indiana CORE Assessments. That might be true for the initial administrations of any testing program, but some improvements in the score scale reliability coefficients might also be expected as the testing program matures. Except for the Pedagogy examinations, there was not demonstrable improvement in the reliability coefficients from 2013-14 to 2014-15.

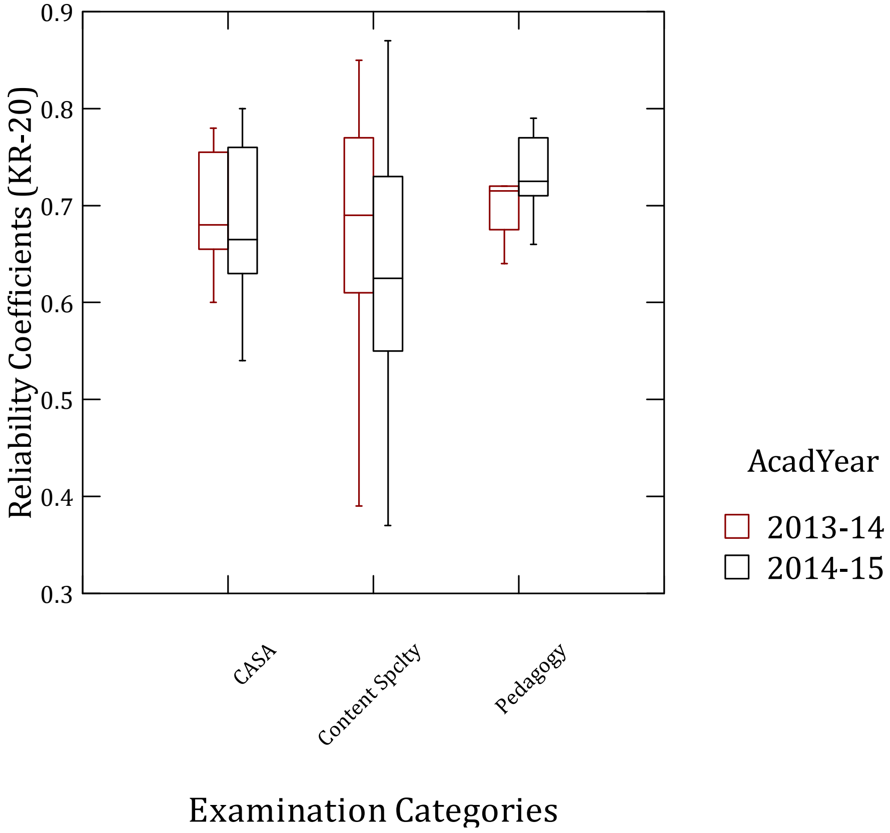


Figure 5. Box Plots of the Distributions of Reliability Coefficients (KR-20) by Exam Type

In any case, a reliability coefficient is not necessary useful to report for certification and licensure tests—especially for tests with relatively low or relatively high passing rates because: (a) that type of coefficient does not reflect the precision of scores in the region of the pass/fail cut score and (b) the *intentional* design of a test to optimize the precision of the pass/fail decisions can restrict the overall score variance due to reduced precision elsewhere along the score scale. Both factors can lower a reliability coefficient. However, it is not acceptable to merely assume that an observed lower reliability coefficient (e.g., less than 0.8) stems from either of these two causes. That is, it is also entirely possible that marginal item quality, multidimensionality and other types of nuisance variance may contribute more "noise" than "signal" to the score scale.

Figure 6 shows the decision consistency that results from raw test scores and the associated cut scores. That is, if the examinees were administered two tests of equal difficulty, scored, and then classified as passers or failers, how much agreement would there be? Livingston and Lewis (1995) developed a statistically elegant approach to estimating decision consistency from a single test form. However, the index does make strong assumptions about the psychometric and construct equivalence of the test-form halves, as well as the distributions of "true" scores. The procedure also depends on a having an estimated reliability coefficient (see discussion above).

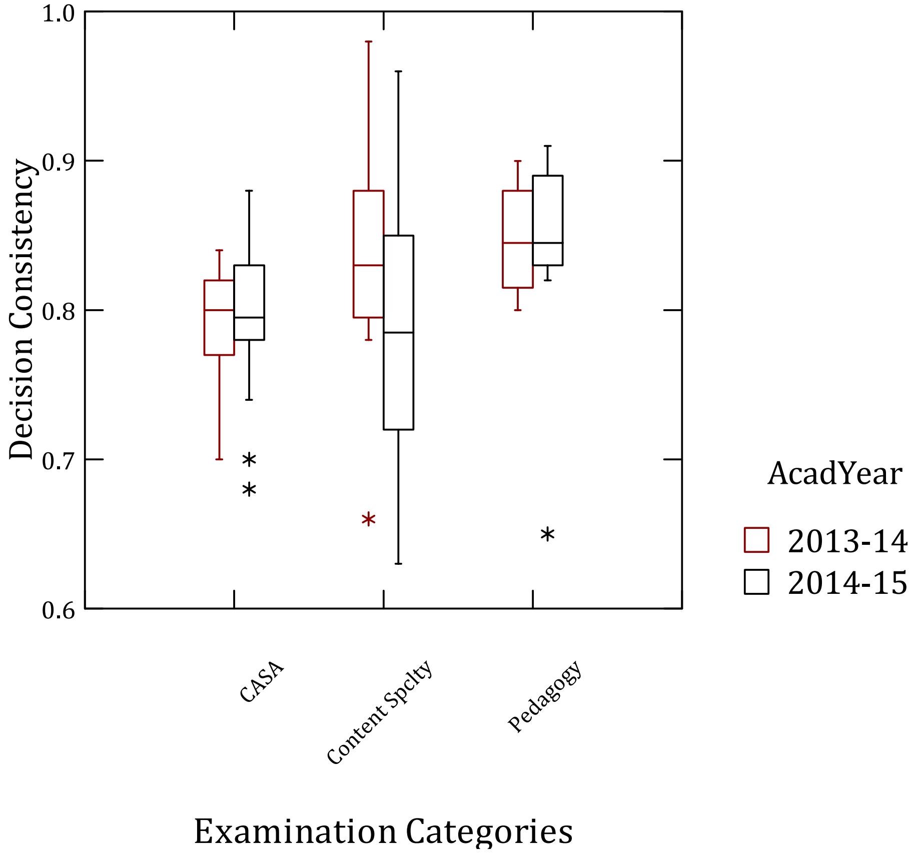


Figure 6. Box Plots of the Distributions of Decision Consistency Statistics by Exam Type

The decision consistency results perhaps suggest better outcomes (i.e., higher numbers than the reliability results). But again, there is not strong evidence of "improvement" from 2013-14 to 2014-15 in the test design, item writing or overall psychometric quality of the tests for their intended purpose.

Pearson Evaluation Systems appears to clearly recognize that there are some technical issues with at least some of the examinations. For example, they state:

> "*Our recommended approach is to strike a balance between actions to improve test form performance and the sufficiency of data available to corroborate the positive effects of the improvement efforts on test reliability and decision consistency. The continued practice of formulating test redevelopment plans with phases to prioritize improvement decisions of the test/forms performance will be part of the CORE Assessments program.* [Bulleted] *For example, Phase I of the plan will identify tests with the most need for improvement on reliability and decision consistency. Such determination would be made by considering the historical size of the estimates and the number of test forms showing low estimates. For example, Field 063 Elem. Ed. Gen. Sub. 4, which consistently has showed the lowest test form reliability estimates, would be included in Phase I of the improvement plan. The plan will include a total of three phases, with forms identified for improvement and a detailed schedule for each Phase*." (p. 34)

It would seem prudent for the Indiana DOE and the TAC to hold the contractor accountable to demonstrate improvement in line with their plans.

A fundamental problem with the technical, empirical evidence and future reporting plans provided in Pearson Evaluation Systems' report is the absence of context or relevant psychometric information about the *intended* test designs, the test assembly process and outcomes, and the linking/equating designs and quality of the linking[9]. At the very least, it would be useful to understand how "test forms" are identified in the tables. For example, is form "A" always the same collection of items, even across years? And how the equating was performed? What were the nature of the link functions across forms (e.g., random groups supported by randomly spiraled forms or random assignment, common-item links)? Finally, what evidence is there as to the quality of the equating by test form (e.g., standard errors of equating)? Consider a simple thought exercise. Appendices V, W and W1 show mean scale scores reported for virtually all the CORE Assessments, but we have no information to suggest whether examinees taking a particular form can be legitimately compared to others taking different test forms. While details about the test designs, item analyses, test assembly and equating (see Figure 4, shown earlier) are no doubt included in

[9] Pearson Education, Inc. (2016) indicates that: (a) single-group equating designs were used for the CORE Assessments, 2013-15; (b) forms were "pre-equated"; and (c) mean-sigma linear equating was used as the equating method of choice. Unfortunately, those vague methodological descriptions do little to clarify exactly what was done and how scores on two or more forms are reasonably linked to a common scale.

the CORE Assessment annual technical manuals, some of those relevant details would have been useful for Pearson Evaluation Systems to provide in their response to the TAC.

Another potential problem relates to the Pearson Evaluation Systems proposal to prepare multi-year technical reports, but to suppress reporting for examinations with small sample sizes—as they did in their current report. One can certainly can follow the "statistically responsible" logic of not reporting test-score results with extremely low sample counts (N≤30 or N≤100) in certain contexts (e.g., for disaggregated results by teacher training programs or subgroups with limited participation). However, how does Pearson Evaluation Systems rationalize, equating/linking the score scales across test forms, reporting item statistics and potential target population impact during standard setting, and computing scale scores with those SAME small sample sizes? There is a growing body of psychometric and statistical research that suggests that combining data across multiple time points leads to more robust estimation than computing (annual) results and then trying to link together the various statistics.

From a certain sampling perspective, small census samples for some of the CORE Assessments could be considered "finite" populations. In those cases, finite-population corrections that can be applied to adjust, for example, the variance estimates that would be used in computing reliability coefficients. Consider that the scale reliability coefficients and other distributional properties WERE used to report ±1 and ±2 standard error of measurement (SEM) adjustment options during the 2013 standard setting for those SAME examinations identified here as having too few candidates to justify reporting reliability and decision-consistency statistics.

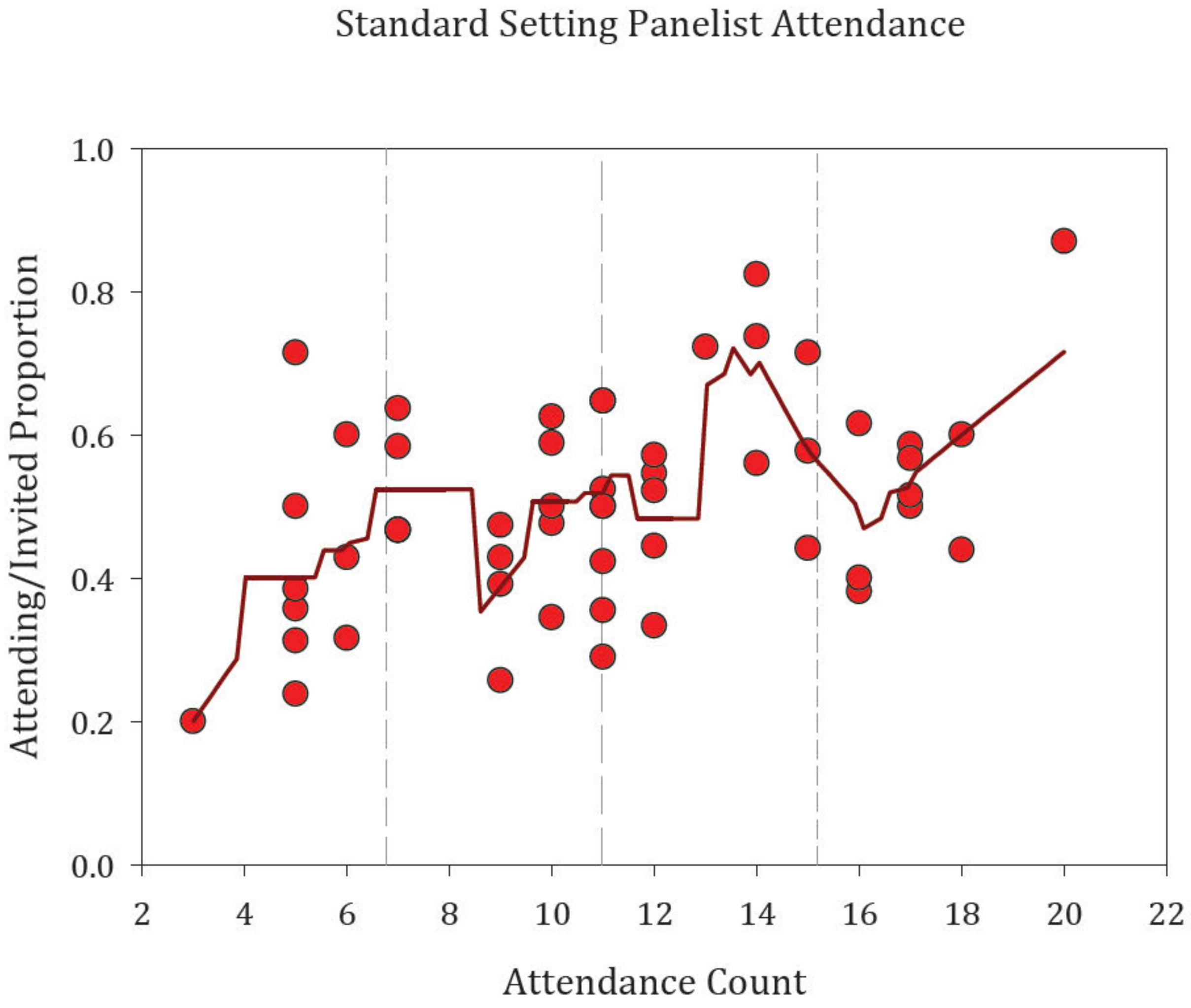
2.3 Standard Setting

Standard setting was the final point of concern by the IDOE TAC (see Recommendations #5 and #6). Here, Pearson Evaluation Systems seemed to "get it right"—at least in terms of meeting industry standards. The modified Angoff method is highly popular in certification and licensure testing and seems to consistently lead to legally defensible standards and cut scores. Nothing in the contractor's description of their panelist selections or procedures suggested serious logistical, operational or analytical problems. Additional details about the nomination process for identifying the panelist "sampling frames" would be useful for the contractor to provide in any future documentation.

Figure 7 shows a scatterplot[10] of the attendance rates (vertical axis) by panelist counts (horizontal axis) across all standard-setting events. The average number of panelists was 11.6 (11 or 12 panelists). The average attendance rate was 50.2%. Note that

[10] A smoothed trend line—here computed using a "negative exponential" smoothing function—is also shown in Figure 7. The sole purpose of that line is to suggest a possible non-linear tread in the functional relationship between the final standard setting panelist counts and the attendance rates (relative to the number of invited panelists).

higher attendance rates coupled with lower panelist counts may suggest that Pearson Evaluation Systems and the IDOE had limited pools of subject-matter experts (SMEs) from which to draw. It appears that the standard-setting panelists were reasonable well motivated to attend. There were only four examinations with acceptance rates lower than 30%.



3.0 Conclusions

As noted in the introductory section of this review, no testing program is perfect and the testing agency and contractors are cautioned to show coherent evidence that documents efforts aimed at continual improvement of each examination program in terms of test design, item writing and test development. Processing steps and psychometrics also need to be examined as often and by as many technical reviewers as possible. The intent is not to go through the motions of quality control and then claim that the procedures are compliant with industry standards. Rather the intent should be to ensure that the best possible scores and decisions are being produced for EVERY examinee—especially in high-stakes settings like teacher licensure. As noted in this review, there is room for improvement for the CORE Assessments. It will be up to the IDOE TAC to provide the scrutiny needed to ensure that progress is made and documented in timely manner.

A rather interesting point on which to end is the topic of legal risk. Licensure tests seems to fall into a gray area between certification and employment. Although the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) somewhat distinguish between them, those are NOT legally definitive distinctions. That is, on one hand, *licensure is not employment, per se*. However, it can be a barrier to employment in many professions such as medicine and allied health, insurance, finance/accounting, and, of course, teaching. But neither are licensure tests merely certifying knowledge and skills.

A [perhaps] good piece of advice when evaluating legal risk in employment and licensure testing is, "*you are safe until you are sued*." That is, it does not matter whether a particular legal question of fairness or disparate impact has never before been tested in the courts. All it takes is one case to snowball into another and another, etc.. For example, a civil lawsuit current making its way through the appellate courts in Minnesota was initiated by 18 teachers who alleged that the Minnesota Board of Teaching made arbitrary and inconsistent decisions when rejecting applications for teaching licenses and further ignored the Legislature's request to streamline the licensing process for educators trained out-of-state and in alternative programs (see Megan, 2016). The appellate court upheld that the district courts DO have jurisdiction over the Board's administrative decisions and "quasi-legislative actions" of the Minnesota Board of Teaching. While this case is still under litigation, it could set an interesting legal precedent for claims that policy decisions related to teacher licensure are entirely under the purview of an elected state board of education or the state department of education.

In line with that same caution about risk of lawsuits, it seems tenuous at best to argue that teacher licensure is wholly distinct from employment testing and therefore exempt from legal precedence, legislation, administrative decisions by the U.S. Department of Labor, or Presidential executive orders. For example, if licensure practices can be empirically shown to violate the well-known ""80% rule[11]", a group of teacher candidates as plaintiffs could argue "disparate impact" and open up the CORE Assessments to intense scrutiny in the courts. One way to avoid that problem is to do the best job possible, document technical quality over time, and improve the processes and tests to reduce as much error as possible.

[11] The 80% test rule was published by the State of California Fair Employment Practice Commission in 1972 (see *State of California Guidelines on Employee Selection Procedures*). The rule states that if the ratio of the hiring proportion of minority-identified candidates relative to the hiring proportion of non-minority candidates drops below .8 (or 80%), there is prima facie evidence of adverse impact. While not proof of discrimination, when applied, this rule can serve as sufficient grounds for extensive data collection and validation research to demonstration that an employment test and/or hiring practices are required, are directly job related, and are NOT discriminatory. This can become a non-trivial "burden of proof".

4.0 References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32(2)*, 179-197.

Megan, C. (2016). Appeals panel denies state's request to dismiss teacher licensing lawsuit. *Pioneer Press*. (Retrieved from http://www.twincities.com/2016/08/09/appeals-panel-denies-states-request-to-dismiss-teacher-licensing-lawsuit/, 01-November 2017)

Pearson Education, Inc. (2016). *Indiana CORE Assessments of Educator Licensure: Technical Manual, 2013-2015*. [Author].

Pearson Evaluation Systems (2017). *Response to Indiana Technical Advisory Committee (TAC) August 10, 2017 Technical Memo*, Dated: September 6 2017. [Author].

Webb, N. L. (2005). *Alignment, depth of knowledge, & change*. Paper presented at the Annual Meeting of the Florida Educational Research Association.