CLEAR
Exam Review

VOLUME XXX, NUMBER 2 | WINTER 2020

A Journal

CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 108 Wind Haven Drive, Suite A, Nicholasville, KY 40356.

Design and composition of this journal have been underwritten by Prometric, which specializes in the design, development, and full-service operation of high-quality licensing, certification and other adult examination programs.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $30 per year to others. Contact CLEAR at (859) 269-1289 or cer@clearhq.org for membership and subscription information.

Advertisements and Classified (e.g., position vacancies) for CER may be reserved by contacting CLEAR at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page, and cost $100 or $200, respectively, per issue.

CLEAR Exam Review

VOLUME XXX, NUMBER 2 | WINTER 2020

CONTENTS

 ISSN 1076-8025

Advancing Alignment Arguments in Supporting Scoring Interpretation and Use Claims of Credentialing Exams

LISA M. ABRAMS, Ph.D., associate professor of research, assessment and evaluation,
Virginia Commonwealth University
KATHERINE A. REYNOLDS, Ph.D., assistant research director, questionnaire development and policy research,
TIMSS & PIRLS International Study Center, Boston College
MICHAEL O'LEARY, Ph.D., Prometric chair in assessment, Centre for Assessment Research, Policy and Practice in Education,
Dublin City University

Correspondence concerning this article should be addressed to Lisa Abrams, lmabrams@vcu.edu.

Introduction

Determining the inferences that can be drawn from test scores and ascertaining appropriate uses for those scores are essential considerations in test design and specification. Articulation of these ideas is commonly referred to as the interpretation and use argument (IUA) and involves providing documentary evidence to demonstrate the validity of the intended inferences, conclusions, and/or decisions made based on test scores. The *Standards for Educational and Psychological Testing* (AERA et al., 2014), hereafter referred to as the *Standards,* are especially relevant: "Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores" (p. 125). Ferrara and Lai (2016) extend this expectation by including evidence required over time and across all stages of test specification, design, development, and implementation. Credentialing exam developers can draw on alignment models dominant in educational (K-12 in particular) testing to meet increased evidentiary expectations at the test specification stage in ways that support the industry's *Standards for the Accreditation of Certification Programs* (NCCA, 2014).

The initial test specification stage requires ". . . documentation of the purpose and intended uses of the test, as well as detailed decisions about test content, format, test length, psychometric characteristics of the items and test, delivery mode, administration, scoring and score reporting" (AERA et al., 2014, p. 76). Test specifications identify the content domains measured in relation to the test's purpose. Credentialing tests are intended to measure the extent to which examinees have acquired the necessary knowledge, skills, and dispositions to engage in safe and effective practice in the workplace. In contrast, achievement tests aim to measure mastery of specific content and skills and are most commonly used in educational settings. Credentialing tests focus on job or occupation responsibilities, while achievement tests center on academic content and skills. In both cases, considerable emphasis is placed initially on defining test content with precision and clarity so that items and tasks can be developed and subsequently used to make valid inferences about examinees (Webb, 2006). However, as Ferrara and Lai (2016) argue, procedures for routine documentation of test content validity evidence are more established in achievement testing than they are in credentialing examinations.

Practices implemented in educational achievement testing can benefit those working in the field of credentialing. Even though the specifics of the test development procedures may differ across these two broad fields, the fundamental principles are similar. Buckendahl (2017) explains that the lines distinguishing

educational assessments, credentialing exams, and tests for employment have become increasingly blurred as the use of test results has expanded to serve a range of purposes and stakeholder needs, thus affording an opportunity for credentialing exam developers to consider the applicability and benefits of well-established practices in achievement testing. With this in mind, we first consider current test content specification practices in the credentialing field and then discuss how these can be strengthened by drawing on best practices in achievement testing.

Content Specification in Credentialing Exams

Credentialing exams, including licensure and certification tests, are administered to assess readiness for a professional role or to ensure that workers can carry out job-related responsibilities safely and in accordance with professional guidelines and standards. Specifying the test content, usually in the form of professional knowledge, skills, and judgments (KSJs), prior to item and task development is fundamental to test quality. The IUA for credentialing exams suggests that passing scores demonstrate a level of mastery of the KSJs that denote a degree of competence or readiness to perform a professional practice or activity. A key underlying assumption for the validity argument is that test content and tasks measure the KSJ domains necessary for effective practice. Examining the alignment between test content, tasks, and the KSJ domains is essential to demonstrating that several of these assumptions have been met and the IUA is credible, thus constituting an important source of validity evidence. As described in the *Standards* (AERA et al., 2014), "When test content is a primary source of validity evidence in support of the interpretation for the use of a test for employment decisions or credentialing, a close link between test content and the job or professional/occupational requirements should be demonstrated" (p. 178). Accordingly, content specifications for credentialing exams are based on empirical job or practice analysis to determine the knowledge, skills, and attitudes needed to perform workplace responsibilities effectively (Raymond, 2016; Raymond & Neustel, 2006).

Practice analysis requires systematic and thorough methods of questionnaire design, data collection, and analysis in order to arrive at accurate indicators of job duties and successful performance. Approaches to practice analyses vary. They include developing a working theory of the profession, identifying behaviors that enable or constrain performance, and conducting a task inventory (Raymond & Neustel, 2006). Completed practice analysis studies result in process- and content-based specifications, which are similar to blueprints or frameworks that specify the broader content domains measured by achievement tests. A content-process matrix is used to create a framework for item development for credentialing exams. The content-process matrix is one of the more common ways to integrate the work-related knowledge needed to apply this information to authentic tasks or processes (Raymond & Neustel, 2006). Documentation of the content specification process is one source of evidence to support validity and use arguments.

Emerging Practices in Test Documentation

As noted earlier, the *Standards* (AERA et al., 2014) provide guidelines for best practices in assessment documentation. In addition, the National Commission for Certifying Agencies' *Standards for the Accreditation of Certification Programs* (NCCA, 2014) sets out the practices, policies, and procedures programs should have with respect to (1) purpose, governance, and resources; (2) responsibilities to stakeholders; (3) assessment instruments; (4) recertification; and (5) maintaining accreditation (p. 6). However, reviews of documentation and

technical reporting for individual credentialing exams have found several persistent limitations, including a lack of comprehensive and publicly available documentation for certification and licensure testing programs (Ferrara & Lai, 2016). Ferrara and Lai found that reports based on different types of validity evidence were published at different times and for different audiences. Although general information for test-takers and candidates was often readily accessible, a single comprehensive program document was rarely available and often difficult to obtain. Large-scale and national testing programs had more complete information available (in technical and lay forms), compared to smaller and regional programs. They concluded that most documentation presented technical evidence and information for individual programs in discrete or disconnected ways; a coherent argument or chain of evidence to support claims about interpretations and uses of test scores was often missing.

In an effort to advance the science of test design, Ferrara and Lai (2016) recommend a new framework for test documentation based on the interpretation/use argument, which they call the *IUA Report.* They describe the report as "a collection of technical, procedural, and other evidence that, taken together, provides a comprehensive and coherent collection of evidence to support claims about how test scores can be interpreted and used" (p. 613). They argue that focusing on capturing evidence at each stage of the test design, development, and implementation process ensures a systematic focus on validity. The *IUA Report* expands current practices, especially in the credentialing field, and offers new ways of thinking about collecting and presenting evidence in a single document that evaluates purported interpretation and use claims at every stage of the process. Ferrara and Lai's exemplar *IUA Report* includes developing claims and providing evidence for the following steps: "(1) determination of testing program policies and articulation of intended interpretations and use of test scores; (2) test design and development; (3) test implementation; (4) response scoring; (5) technical analyses; (6) delivery of scores and other feedback to examinees, candidates, and other test users . . . ; and (7) interpretation of score reports to guide decisions and take other actions" (p. 615-616). They suggest claims at the test design and development stage demonstrate that "item development procedures produce items that elicit evidence of targeted content knowledge and skills" and that evidence for this claim may include ". . . research evidence that the items elicit targeted knowledge and skills" (p. 615). They go on to explain that this claim could be supported by independent reviews as well as depth-of-knowledge judgments by subject matter experts who can determine the extent to which the cognitive process needed to respond to a test item aligns with the required knowledge, skill, or ability (p. 619). The Webb model of alignment, commonly used in educational testing, provides an example of one method to document claims as part of test design and development.

Alignment Studies in Achievement Testing

A series of federal education reform efforts in the United States in early 2000 substantially expanded educational testing programs and required a peer review process to provide different kinds of evidence to support the use of test scores. Alignment between state-developed content standards and their corresponding assessments was required; as a result, empirical studies of alignment became common. Among several models, Webb's model emerged as the dominant one and was regarded as providing the strongest quantitative information to evaluate alignment based on several criteria (Martone & Sireci, 2009).

According to Webb (1997; 2006), alignment is defined as the degree to which assessment and learning objectives agree. Alignment can be measured using four criteria: (1) *categorical concurrence*—the

sufficiency of the item sample from the content domains measured by the test, (2) *depth-of-knowledge consistency*—the extent to which test items meet/exceed the cognitive demand expressed in the associated content standard, (3) *range-of-knowledge correspondence*—the diversity of content and extent to which different content standards are represented in the assessment, and (4) *balance of representation*—an index representing the relative emphasis placed on individual standards measured by the assessment (Webb, 2007). Each criterion has different thresholds that can be used to indicate the extent or degree of alignment. Tests are not in alignment or out of alignment—rather, a series of criteria can be used to collectively determine the degree of coherence among standards, assessments, and instruction (Resnick, Rothman, Slattery & Vranek, 2004; Roach, Elliot & Webb, 2003). Webb's model is consistent with several of the NCCA standards guiding instrumentation development, reporting, and interpretation. For example, the alignment criteria speak directly to one of the essential elements of NCCA's *Standard 15: Examination Specifications,* which states that "the plan for weighting sections of an examination must be based on a job analysis; the plan must provide precise direction regarding the weighting structure for each section" (p. 22). Drawing on Webb's approach allows for clear documentation and evidence of the NCCA standards.

The Webb Model of Alignment—Implications for Test Documentation

Implementation of Webb's model requires two main phases. The first involves making determinations about the level of cognitive demand or depth-of-knowledge (DOK) evident in the content standard.[1] This phase requires that content area experts work as a review committee to determine the DOK levels for the content standards by group consensus. Webb developed distinct DOK categories to reflect different levels of cognition: *Level 1—recall and reproduction, Level 2—skills and concepts, Level 3—strategic thinking, and Level 4—extended thinking.* These categories are similar to the concepts in Bloom's Taxonomy—*recall knowledge, comprehension, application, analysis, synthesis, and evaluation.* The second phase of the alignment study requires the content experts to closely examine each test item to determine (1) the appropriate content standard/domain the test item is designed to measure and (2) the DOK level of the corresponding item. The reviewers independently code each test item. The reliability of reviewers' ratings is an important factor in this phase of the alignment review process. The information obtained during phases 1 and 2 can be used to produce and examine the degree of alignment according to the four criteria:

1. **Categorical Concurrence:** This criterion addresses questions about the extent to which the content measured on the test is the same as the content expressed in the content standards or learning objectives. Webb's evaluation criteria suggest that at least six test items measuring content from a reporting category are needed for a reasonably reliable estimate of students' content mastery on a subscale or standard (Webb, Alt & Ely, 2005, p. 110). This alignment criterion is closely related to NCCA's Standards 15 and 20 that address construct specification and score reliability.
2. **Depth-of-Knowledge Consistency:** This criterion is focused on the match or agreement between the cognitive process or DOK expressed in the learning objectives and the aligned test item. Webb et al. (2005) indicates that for depth-of-knowledge consistency to exist between the assessment and the reporting category, at least 50% of targeted objectives should be "hit" by items of the appropriate complexity. Webb's cut-point is based on the "assumption that a minimal passing score for any one objective/domain of greater than 50% would require the student to successfully answer at least some

[1] This is a very general description of the Webb alignment model process. For more about Webb's alignment methodology, see the following 2009 publication: https://www.nagb.gov/content/nagb/assets/documents/publications/design-document-final.pdf

items at or above the depth-of-knowledge level of the corresponding standard" (Webb et al., 2005, p. 111). Here, this criterion is consistent with essential elements of NCCA's Standard 15, where the cognitive or performance task required to respond to the test item is clearly associated with "what the examination is intended to measure" such as knowledge, skills, and competency (p. 22).

3. **Range-of-Knowledge Correspondence:** This criterion addresses item sampling and the sufficiency of content coverage. According to Webb et al., 50% of the standards should have at least one related test item in order to determine the range-of-correspondence criterion "acceptable." Similarly, alignment on this criterion is determined "weakly" met if 41%- 49% of the objectives for a reporting category had a corresponding test item and "not met" if less than 41% of the standards had at least one corresponding test item. The ideas of this alignment criterion support essential elements of NCCA's *Standard 16: Examination Development*, where "the sampling plan for the examination must correspond to the examination specifications" (p. 23).
4. **Balance of Representation:** This criterion relates to range-of-knowledge and addresses the extent to which content is emphasized or distributed on a test. A BOR index[2] is "computed by considering the difference in the proportion of objectives and the proportion of related assessment items for that objective. An index of 1 indicates perfect balance and is obtained if the corresponding items related to a standard are equally distributed among the objectives for a given standard" (Webb et al., 2005, p. 112). Webb suggests that a BOR index of .70 or greater suggests that the criterion has been met.

Table 1 provides a hypothetical example of how the criteria might be applied and could be included as part of an IUA report.

Table 1. Alignment Criteria Evaluation Categories

Alignment Criteria	Strength of Alignment Evidence: *Were the Alignment Criteria Met?*		
	YES	WEAK	NO
Categorical Concurrence	6 or more test items	4-5 test items	less than 4 test items
Depth-of-Knowledge Consistency	greater than 50% agreement	41%-50% agreement	agreement of 40% or less
Range-of-Knowledge Correspondence	50% or greater	41%-49%	40% or less
Balance of Representation	Values of .70 or greater	Values between .60 and .69	Values less than .60

[2] The following formula is used to compute the BOR index: $1-\Sigma\left(\left|I_o-I_H\right|\right)/2$
I_k = the number of test items corresponding to standard K
O = the total number of standards with corresponding test items within a reporting category
H = the total number of test items corresponding to a reporting category

Advancing Alignment Arguments in Supporting Scoring Interpretation and Use Claims of Credentialing Exams

Alignment in Professional Exams

Formal alignment studies appear to be less common in the areas of licensure and certification, perhaps because job analyses and other content domain definition activities are not necessarily explicated through formal standards or policy in the same way as K-12 education. It is also possible that such studies are simply not available to the public. However, the Webb model could be adapted for use in licensure or certification examinations. This section highlights themes from two content alignment studies conducted for credentialing exams for teachers and principals, with particular attention to how the procedures used compare to the Webb model. A summary of these comparisons is presented in Table 2.

In the first of these studies, Reese, Tannenbaum, and Kuku (2015) examined the correspondence between the Praxis Performance Assessment for Teachers (PPAT) and the Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards. The PPAT was explicitly designed to measure the nine InTASC Standards applicable to teacher-candidates that "could be demonstrated during the candidate's preservice teaching assignment, and could be effectively assessed with a structured performance assessment" (p. 2). In the second study, Swigget (2019) presents the results of a distance-based alignment study concerning the Performance Assessment for School Leaders (PASL) and the Professional Standards for Educational Leaders (PSEL). It is worth noting that both studies concern performance assessments, where assessment tasks differ considerably from the multiple-choice or constructed response items typically used in educational achievement tests.

Both Reese et al. (2015) and Swigget (2019) solicited alignment-related judgments from relevant experts—similar to Webb's model. Specifically, both studies involved expert judgments to identify the professional standards most associated with each performance task. Reese et al. (2015) had panelists use a 5-point scale (1 = not measured, 5 = directly measured) to indicate whether or not each InTASC Standard was captured by a particular step within a given PPAT assessment task. Standards receiving a 4 or 5 from at least seven panelists were considered aligned to a task step. These results were compared to the test developers' framework and intended association of the InTASC standard measures by each task step. These procedures are most comparable to the categorical concurrence criterion in the Webb model.

Swigget (2019) prompted review panelists to indicate whether each of three broad PASL tasks were aligned to each of the PSEL standards. PASL tasks include the compilation of "written responses, supporting instructional materials, and artifacts (e.g., student work)" (Swigget, 2019, p. 2). If an overarching PASL task was identified as aligned to a standard, panelists further indicated to which PSEL supporting elements the task was aligned. This approach allowed for calculation of how many PSEL standards were represented by the assessment, as well as the percentage of supporting elements assessed. Again, this might be considered comparable to the Webb model's categorical concurrence criterion.

Reese et al. (2015) also asked panelists to evaluate the applicability of performance indicators for completion of assessment tasks; scoring rubric adequacy; and the relevance, importance, and authenticity of each PPAT task. Scoring rubric adequacy might be considered comparable to Webb's DOK consistency criterion, as panelists were prompted to render dichotomous judgments of whether performance descriptions at the highest level for each task step reflected the performance indicators accompanying the standards. However, panelists did not formally classify these indicators along any sort of DOK continuum. These additional judgments imply a

broader view of "alignment" than the Webb model, invoking not only alignment among assessment tasks and standards but also scoring materials and tasks, as well as assessment content and the tasks routinely performed in the teaching profession. These are examples of additional validity evidence that can be provided in an IUA Report, but they are distinct from content alignment as conceptualized by Webb. Both alignment studies could be strengthened through incorporation of additional Webb model indices, which would yield evidence not only of task-standard alignment but also of cognitive level alignment and overall representativeness of the assessments with respect to the standards.

The lack of publicly available content alignment studies in professional testing represents an opportunity for those working in credentialing and licensure to capitalize on a powerful source of content validity evidence. Normalizing the publication of these kinds of studies might serve to increase public confidence in the appropriateness of credentialing exam content, especially considering the relatively straightforward results obtained from content alignment studies. Use of well-established alignment study models, such as the Webb model, also lends credence and procedural validity to the process.

Table 2. Summary Comparison of the Webb Model, Reese et al. (2015), and Swigget (2019) Alignment Procedures

	Categorical Concurrence	Depth-of-Knowledge Consistency	Range-of-Knowledge Correspondence	Balance of Representation
Webb Model	Panelists classify items to standards	Panelists classify standards and items by DOK levels	Calculation of the percentage of standards represented by assessment items according to panelist classification	Examines the balance of standard representation across an assessment
Reese et al. (2015)	Panelists use a Likert scale to indicate if a standard is measured by a task	Not directly addressed—tangentially addressed by panelists' judgments of scoring rubrics compared to performance indicators	Calculation of total number of InTASC standards represented across the PPAT tasks	Not directly addressed
Swigget (2019)	Panelists provide yes/no judgments about task alignment to each standard	Not directly addressed	Calculation of the percentage of standards/supporting elements represented by assessment items according to panelist dichotomous judgments	Not directly addressed

Conclusions

This paper describes how credentialing test developers can respond to increased calls to enhance technical documentation and reporting in ways that support an interpretation and use validity argument at all stages of

the testing program. Specific attention to the content specification stage is essential for designing high quality assessments. The use of systematic methods, such as implementing a performance or job analysis, heightens the validity of the inferences made based on credentialing exam scores. Especially in cases where these inferences can lead to high-stakes consequences for examinees and require strong assurances of validity evidence, Webb's model provides an approach to bolster technical documentation. We have argued that while alignment studies are not currently a common feature of the credentialing exam development process, there is much to be learned from such studies in educational K-12 testing, where content domains and standards are more clearly explicated. Webb's model provides test developers with a clear set of quantitative indicators that can be used to inform the initial stages of test development and demonstrate validity evidence for test content. In helping to mitigate validity challenges, the model has implications for how legal defensibility in testing might be conceptualized and for how quality assurance during important stages in the test development process can be actioned.

Reference List

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.

Buckendahl, C. W. (2017). Credentialing: A continuum of measurement theories, policies and practices. In S. Davis-Becker & C. W. Buckendahl (Eds.) *Testing in the professions: Credentialing policies and practice* (pp. 1-20). New York, NY: Routledge.

Ferrara, S., & Lai, E. (2016). Documentation to support test score interpretation and use. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development (2nd ed.)* (pp. 603-623). New York, NY: Routledge.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research,* 79(4), 1332-1361.

National Commission for Certifying Agencies. (2014). *Standards for the accreditation of certification programs.* Washington, DC: Institute for Credentialing Excellence.

Raymond, M. (2016). Job analysis, practice analysis and the content of credentialing examinations. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development (2nd ed.)* (pp. 144-164). New York, NY: Routledge.

Raymond, M., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook for test development* (pp. 181-224). New York, NY: Routledge.

Reese, C. M., Tannenbuam, R. J., & Kuku, B. (2015). Alignment between the Praxis performance assessment for teachers (PPAT) and the interstate teacher assessment and support consortium (InTASC) model core teaching standards. ETS Research Memorandum-15-10. https://www.ets.org/research/policy_research_reports/publications/report/2015/jvgo

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2004). Benchmarking and alignment of standards and testing. *Educational Assessment,* 9 (1-2), 1-27. https://doi.org/10.1080/10627197.2004.9652957

Roach, A. T., Elliott, S. N., & Webb, N. L. (2003). Alignment analysis and content validity of the Wisconsin alternate assessments for students with disabilities. *Wisconsin Center for Education Research.* https://eric.ed.gov/?id=ED497575

Swigget, W. D. (2019). Alignment of the Performance Assessment for School Leaders to the Professional Standards for Educational Leaders: A distance-based approach. ETS Research Report Series ISSN 2330-8516. https://doi.org/10.1002/ets2.12259

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Educa*tion, 20(1), 7-25. https://www.cehd.umn.edu/edpsych/C-BAS-R/Docs/Webb2007.pdf

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook for test development* (pp. 155-180). New York, NY: Routledge.

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005). *Web alignment tool (WAT): Training manual version 1.1.* Wisconsin Center for Education Research, University of Wisconsin. Retrieved from http://watv2.wceruw.org

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. *National Institute for Science Education.* Retrieved from https://eric.ed.gov/?id=ED414305