CSAI Report

Evaluating Content Alignment in the Context of Computer-Adaptive Testing:

Guidance for State Education Agencies

Carole Gallagher, Ph.D.

June 2016

The work reported herein was supported by grant number #S283B050022A between the U.S. Department of Education and WestEd with a subcontract to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The findings and opinions expressed in this publication are those of the authors and do not necessarily reflect the positions or poli-cies of CRESST, WestEd, or the U.S. Department of Education.

WestEd
WestEd.org

WestEd is a nonpartisan, nonprofit research, development, and service agency that works with education and other communities throughout the United States and abroad to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has more than a dozen offices nationwide, from Massachusetts, Vermont and Georgia, to Illinois, Arizona and California, with headquarters in San Francisco. For more information about WestEd, visit WestEd.org; call 415.565.3000 or, toll-free, (877) 4-WestEd; or write: WestEd / 730 Harrison Street / San Francisco, CA 94107-1242.

Evaluating Content Alignment in the Context of Computer-Adaptive Testing: Guidance for State Education Agencies[1]

Introduction

Evidence that a test and the content domain it is intended to measure are aligned is critical for supporting claims about the validity of inferences drawn from results and defending test use for a given purpose. Findings from an alignment evaluation provide the test developer and the test user with valuable information about the degree to which a test measures what it was intended to measure. The importance of conducting independent, third-party alignment studies has long been recognized, but with the increased use of computer-adaptive tests (CATs), we must reconsider how to go about evaluating alignment in this unique context.

Most current approaches to evaluating alignment between tests and standards were designed for traditional end-of-grade or end-of-course assessments in which all students in a particular grade or course are assessed using the same test items. In the context of this fixed-form testing, alignment is defined as the degree of agreement between the test content and the standards it is intended to measure, in terms of both breadth (i.e., coverage of all eligible elements of the standards) and depth (i.e., coverage at the same level of cognitive complexity as represented in the standards). With a CAT, this traditional definition of alignment no longer suffices.

In a CAT, a set of items are administered via a computer that is customized for each student *as* he or she moves through the test, with a well-designed algorithm selecting and delivering upcoming items in accordance with his or her responses to prior items in the test. The result is an assessment that is optimally suited for each student and, as a result, promotes engagement rather than the frustration that can arise when students are faced with items that are too easy or too difficult. In addition, because each item provides maximally useful information about what a student knows and can do, the testing process is more efficient and the results are more precise for students at all achievement levels.

A CAT may be conducted through full or partial customization, depending on the extent to which the student's testing experience is individualized. A CAT is partially customized when all students see one or more common blocks of items, with additional items delivered to each student based on his or her prior pattern of responses. A CAT also may be item- or stage-adaptive, depending on whether one item or a

[1] Recommendations presented in this article were developed in collaboration with Drs. Stephen Wise and Gage Kingsbury. Please see their research article for background information: Wise, S.L., Kingsbury, G.G., & Webb, N.L. (2015). Evaluating content alignment in computer adaptive testing. *Educational Measurement: Issues and Practice, 34*(4), 41-48.

set of items is delivered to each student based on his or her prior pattern of responses.[2] With all of these types of CATs, all items are drawn from a common pool, but each student's assessment, or test event, is individualized to some extent.

Given the customization that characterizes a CAT, traditional alignment approaches would not provide a complete picture of the degree to which the knowledge and skills that are tested correspond with the standards the test is intended to measure. Wise, Kingsbury, and Webb (2015) concluded the following about evaluating alignment in an adaptive context:

> *Content alignment for an adaptive test event needs to consider an additional frame of reference beyond the content standards (which are common to all test takers). It also must consider the test content that is optimally appropriate for the particular individual taking the test... If a test event is made up of items that perfectly represent the content standards but [the event] fails to provide the appropriate level of challenge for the test taker, this test event is not content-aligned for this test taker. At the same time, if a test event is appropriately challenging for a test taker but fails to measure the content standards well, it also is not content-aligned. (p. 1)*

This is to say, whether a test is traditional or computer adaptive, an alignment evaluation must consider the degree to which the items administered to each student represent the intended depth and breadth of the standards the test is intended to measure (i.e., are content-aligned). But for a CAT, alignment evaluators must go further; specifically, they must *also* consider the degree to which the items delivered to each student provide optimal challenge for that student (i.e., are difficulty-aligned).[3] Collection of this additional type of evidence provides the state and its stakeholders with assurance that students are administered a customized set of items, as intended. In addition, by closely examining the degree to which each student's test event actually matches the test plan, the "black box" underlying the item selection and delivery system becomes more transparent to all constituents.

This report describes a research-supported approach to conducting a study of alignment between a set of standards and a pool of test items that are delivered adaptively for purposes such as instructional planning, measuring end-of-grade or end-of-course achievement, and/or school-, district-, or state-level accountability. The information in this report is intended to support state education agencies (SEAs) that seek to

- Interpret evidence presented by a CAT developer (e.g., a test publisher or vendor) about the alignment of its item pool to a set of content standards;

---

[2] In the multistage approach, sets of items that may be administered are determined to be broadly representative in terms of content and difficulty using information from prior administrations.

[3] It is important to note that the concept of having items "difficulty-aligned" is not the same as having them representative of the level of depth, or cognitive complexity, inherent in the standards intended to be measured. The latter (e.g., Depth of Knowledge rating) is an item characteristic that describes the level of complexity at which an item measures a particular standard, while the former refers to the effectiveness of the item delivery system in ensuring that each subsequent item or set of items on a CAT is delivered in accordance with the student's pattern of prior responses.

- Develop an RFP for a contractor to conduct an independent analysis of alignment between the state standards and a computer adaptive item pool; and/or
- Conduct internal quality control checks on the alignment between the items in an existing adaptive test pool and a new (e.g., Common Core State Standards, Next Generation Science Standards) or revised set of standards.

The sections that follow describe in greater detail an alignment evaluation approach that is well suited for computer adaptive testing. The authors propose a series of steps for carrying out this approach and, for each step, offer a set of guiding questions that states may find useful for executing such a study.

A New Alignment Approach for States Using a CAT

The potential advantages of a CAT are widely appreciated, and use of CATs by states and multi-state consortia has grown steadily over the last two decades (Davey, 2011; Reckase, 2010).[4] But to achieve the potential of adaptive testing with precision and efficiency while also engaging students, intentional activities must be conducted during all phases of item development and use. These activities provide the opportunity to collect evidence to support test use from the earliest stages of test design throughout key stages of development of the item pool, test blueprints, and the item-selection algorithm. Alignment evaluations are more comprehensive and useful when they include reviews of information collected during these stages, such as the content specifications for the item pool, how the content representation and optimal challenge guidelines will be achieved, and how items are selected and exposure rates monitored. Equally critical are analyses of data collected during post-administration activities that examine the degree to which individual test events (i.e., items delivered to a given student) are aligned with the requirements for student-level testing that are documented in the test plan or blueprints.

Dr. Wise and his colleagues (2015) provide specific recommendations for evaluating item-to-standards alignment when the assessments are delivered adaptively. Their approach prescribes the steps needed to systematically collect evidence during all stages of test development and administration that can be used to support claims that a CAT is measuring the breadth of the standards as intended with efficiency and precision. Assessment and alignment specialists from the Center on Standards and Assessment Implementation (CSAI) at WestEd have adapted these research-supported recommendations into a user-friendly action plan for states and districts seeking to collect evidence about content alignment when test items are delivered adaptively. That plan includes the following three steps:

1. Examine test planning documentation
2. Evaluate the alignment of items in the pool to a set of standards
3. Evaluate test-event records

---

[4] Davey, T. (2011). *A guide to computer adaptive testing systems*. Washington, DC: Technical Issues in Large-Scale Assessment, Council of Chief State School Officers. Reckase, M. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

Ideally, these steps are completed by the test developer (e.g., the state's test vendor), with documentation of findings available as part of a technical report or validation study that can be reviewed by prospective test users or included as part of the body of evidence submitted for federal peer review. The steps can also be undertaken by an independent contractor as part of a post-administration verification/validation activity. Each step in this process is described in greater detail below.

It is important to note that this approach to evaluating alignment between a CAT and a set of standards seeks to ensure maximum efficiency while yielding trustworthy findings. For this reason, it incorporates strategies for streamlining the process that do not compromise the integrity of results or constrain the types of claims that an SEA can make about the validity of its adaptive system. To this end, as discussed in the steps below, this approach to alignment evaluation relies on the application of research-supported sampling methods by a team of experienced and specially trained content, assessment, and alignment experts.

Step 1: Examine Test Planning Documentation (e.g., statements of purpose, standards on which test is based, grade-level blueprints, item delivery plan)

The first step in this comprehensive alignment approach calls for the alignment evaluator to examine information from the CAT developer that describes the conceptual framework for the test, such as documentation of test purpose, the target population for testing, and the content standards on which the test is based. Alignment evaluators will be looking for information about the ways in which results are intended to be used and the claims a state seeks to make about students, teachers, or schools based on test results; the standards on which the test is based; the frequency with which students are tested; and the type of "stakes," or potential consequences, for students, teachers, schools, districts, or the state based on test results. These data also will guide decision-making about the appropriate unit of analysis for the alignment evaluation (e.g., within the standards hierarchy, will items be aligned with the broad standard statement, more specific objectives, or detailed skill statements?) and will, thus, ensure that the most relevant types of information are collected and reviewed by evaluators.

For example, a common purpose for a test is to measure end-of-grade achievement for state or federal accountability purposes. In such instances, every student's test must assess content from the domain (e.g., mathematics) that is specified as eligible for that grade, so the evaluation of alignment must be focused on that requirement; evaluators will be seeking evidence that the tested content is broadly and meaningfully representative of the content standards for that grade. Another common use of a test is to monitor students' progress toward achieving high-priority instructional milestones or their growth in learning critical concepts or skills. Item pools for this type of interim assessment typically include items designed to probe more deeply into what students know and can do at key points in time in relation to specific content and skills, so in this context evaluators are seeking evidence of clear linkage between items and standards, both within and across grades in key content strands (e.g., reading fluency, mathematical reasoning). Alternatively, if the purpose of a test is to determine students' readiness for a particular program (e.g., gifted and talented) or course (e.g., algebra I), evaluators will be reviewing the connection between items in the pool and a much more constrained set of grade- or course-specific standards.

This review also will include collecting and studying test plans, maps, or blueprints. This information is valuable for determining whether the specifications guiding the testing process enable the development of sufficient numbers of items with particular content and difficulty classifications. Information about content expectations collected during this step will help evaluators understand the specific targets that were set for item development, as well as the prescribed composition of the item pool at each grade (i.e., proportion of the test intended to be dedicated to the measurement of particular standards). Because a CAT test plan or blueprint is integrally linked to the preliminary judgments of a state's item developer about the standards with which each item is intended to align as well as to decisions about the item-selection algorithm and the required composition of student test events, the alignment evaluator will refer to findings from Step 1 during all subsequent steps of the alignment process.

Guiding questions for Step 1 of the evaluation of alignment in a CAT context:

- What is the purpose of this assessment? What does the state intend to measure?
- Who will use the results? In what ways?
- What content standards are assessed at each grade?
- Which item types are allowed?
- For what level of adaptivity was this system designed (e.g., stage-adaptive vs. item-adaptive)? Will all students view one or more common sets of items?
- How many items will be delivered to each student at each grade in each content area?
- What is the minimum and maximum test length?
- Can a student at a particular grade level be exposed to items that measure standards above or below that grade level?
- What are the expectations for each test event in terms of providing optimal challenge during testing?
- What are the plans for regulating item exposure rates?

Documentation also will be collected about the assumptions underlying design of the delivery system. This information helps evaluators better understand the theory of action underlying the assessment (i.e., how it is intended to work) and, also, explore the reasonableness of the assumptions associated with design of the adaptive item-selection process. Ideally, the evaluator will find clear links between the intent of the algorithm used to guide item selection and the grade- and course-specific test plans. It is important to note that this step does *not* require alignment evaluators to assess the technical qualities of the test engine or to examine system specifications. Instead, it calls for evaluators to review any existing documentation that describes the *approach* (e.g., item-level adaptive vs. stage-adaptive) to testing. The goal is to ensure that the item-selection plan is defensible in terms of test purpose and has the capacity to strategically manage content and difficulty constraints so as to ensure that items are administered to students as called for in the test plan.

Resources for answering the guiding questions for Step 1 may be found in a range of documents, including, for example, state or federal regulations mandating assessment, publications posted on the state web site, key state communiqués to key stakeholder groups, and presentations to district administrators. Other resources include publicly available test blueprints for the assessments developed at each grade level as well as any other relevant documentation from the state (e.g., planning documents that describe the numbers of each type of item in each pool, item and test specifications, or relevant sections of a technical report). Ideally, documentation will be available from the test delivery

vendor that describes the state's intent for the adaptive item-selection process. Information collected during this step will be used formatively by the alignment evaluator throughout the course of the study, though the state may also request a copy of the findings from this phase of work to be used for purposes such as federal peer review.

Step 2: Evaluate the Alignment of Items to the Standards

The second step in evaluating the alignment of a CAT to a set of standards requires the alignment evaluator to make independent judgments about the degree to which items in the CAT pool measure the full depth and breadth of the standards.[5] To answer the guiding questions for this step, the evaluator systematically examines each item in a pool, seeking evidence of the strength of correspondence between the item and the standard to which it appears to align and the level of cognitive complexity at which the item assesses that standard (Webb's categorical concurrence and depth-of-knowledge [DOK] ratings).[6] The following information then is documented by the evaluator:

- **Categorical concurrence:** The evaluator documents a judgment of *full, partial,* or *no alignment*. A rating of *full alignment* signifies that the evaluator found evidence that the item assesses a fundamental skill or concept of the standard or taps the central idea of an eligible standard. The item does not need to address *all* of the content embodied by the standard in order to receive this rating. A rating of *partial alignment* signifies that the item assesses the standard in a superficial way or does not address a central or fundamental idea of the standard. A rating of *partial alignment* generally is accompanied a rationale for the rating. A rating of *No Alignment* indicates that the evaluator could find no eligible standard to which the item aligned.
- **Cognitive complexity:** The alignment evaluator also documents a judgment about the level of cognitive complexity at which each item assesses the eligible standard using Webb's DOK rating system. In that system, level 1 is *recall*; items at this level require students to recall a fact, definition, procedure, or piece of information. Level 2 is *basic application*; items at this level require students to use a skill or concept. Items at level 3 of this rating system, *strategic thinking*, require students to demonstrate deep content knowledge and engage in abstract

Guiding questions for Step 2 of the evaluation of alignment in a CAT context:

- To what extent do sampled test items measure the intended breadth of the content standards for each grade?
- To what extent do sampled test items measure the intended depth or level of cognitive complexity of the content standards?
- Overall, to what extent do findings from these analyses suggest that items in the full pool have the potential to meet all specifications called for by the test plan and/or blueprints, including those regulating content and difficulty alignment?

[5] Alternatively, depending on the purpose of the study, an alignment evaluator may be called upon to verify the specific content attributes initially assigned by the item development vendor and subsequently vetted and approved by state stakeholders and SEA staff. These attributes, which are found in the meta-data associated with each item, generally include the developer's judgments about the standard(s) that each item is intended to measure (e.g., content strand, domain, benchmark, and grade-level standard) and its intended level of cognitive complexity (e.g., Depth of Knowledge [DOK] rating).

[6] Webb, N. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four States*. Washington, DC: Council of Chief State School Officers.

thinking, while items at level 4, *extended thinking,* require students to demonstrate complex reasoning processes, higher-order thinking, and deep conceptual understanding. This step is intended to ensure that the item requires students to demonstrate what they know and can do in ways that are aligned to the level of complexity embodied in the standard.

Given the large size of most adaptive item pools at each grade or course in each content area, for this phase of work (i.e., evaluating item alignment) research-supported methods can be used to select a sufficiently large *and* representative sample of items to allow for defensible inferences to be drawn. In most cases, a target sample size is *at least 20 per*cent of the item pool at each grade or a minimum of 100 items, whichever is larger.

The methods used in Step 2 and findings from the analyses are formally presented in a report of findings. Alignment evaluators generally report the findings in tabular format, with explanatory text following each table. Depending on the purpose of the study, findings can be presented at the overall (cross-grade and/or cross-course) pool level in each content area and for each grade level in each content area, with specific strengths or concerns noted.

Step 3: Evaluate Test Event Records

During this final step in the evaluation of content alignment in an adaptive context, the evaluator reviews records that provide information from students' testing experiences. The evaluator explores the degree to which each test event actually measured the intended depth and breadth of the content domain while also providing optimal challenge for each student. This last step is a culminating activity in the alignment evaluation because misalignment at the test event level is a signal to the state that one or more of the key components of its system may not be working as intended by the state, the item-pool developer, or the designer of the item-selection algorithm.

Guiding questions for Step 3 of the evaluation of alignment in a CAT context:

- To what extent does each sampled test event meet the criteria for measuring the breadth and depth of the standards specified in the test plan or blueprint?
- To what extent does each sampled test event demonstrate alignment with the test plan in terms of optimal challenge?
- Overall, does the sample of test events demonstrate alignment to the content standards intended to be measured by the CAT?
- Overall, does the sample of test events demonstrate alignment to the goals for optimal challenge specified in the test plan?

The alignment evaluation team can use research-supported stratified sampling methods to identify a subset of test events that will be representative of the tested population. For example, stratification variables may include content area, grade level, student-level demographic information (e.g., gender, race/ethnicity), and performance level (e.g., by quartiles from lowest to highest performers). The alignment evaluators then will need support from the SEA to access records from the state's administration vendor so they can identify sufficient numbers of test event records to fill the quotas requested for each category (sample cell) of students.

Once the appropriate sample of event records has been identified, data can be transferred to the alignment evaluators for their analyses. For each event, an evaluator will examine (a) a list of the items

delivered, in the order of delivery, (b) the meta-data associated with each item administered during that event, and (c) whether the student answered a given item correctly or incorrectly. An evaluator also will review, if available, each student's provisional ability estimates (i.e., those emerging as a student responds to each item) and final ability estimates.

Using this information for each test event, the evaluator assigns a judgment of *full alignment, partial alignment,* or *no alignment* with the specifications called for in the grade- or course-specific test plans or blueprints. The evaluator will assign two ratings, one to represent the degree to which the guidelines for content alignment were met and one to represent the degree to which the guidelines for optimal challenge were met.

- A rating of *full alignment,* for either content or optimal challenge, signifies that the evaluator found substantial or conclusive evidence to support the claim that the test event met expectations for alignment to the test plan or blueprint for that grade or course. For each event, the evaluator records a rationale for his or her rating (e.g., set of items administered clearly aligned with breadth and depth of content prescribed in the blueprint).
- A rating of *partial alignment* for either content or optimal challenge signifies that the evaluator found only limited evidence to support the claim that the test event met expectations for alignment to the test plan or blueprint for that grade or course. For each event, the evaluator records a rationale for his or her rating (e.g., items were not delivered in ways that promoted optimal challenge).
- A rating of *no alignment* indicates that the evaluator found no evidence to support the claim that the test event met expectations for alignment to the test plan or blueprint for that grade or course. The evaluator may record any particular issues noted in terms of content- or challenge-level issues with the alignment of the test event to the respective guidelines.

This final step yields results at both the test event and item pool (overall) levels. A summary of results for each grade level in each content area can be presented in the report of findings. This evidence is critical to the state, as it can be used to support its claim that the item pool, test plan or blueprint, and item-selection algorithm all work together as intended to ensure that *each* student is assessed on the standards as intended, using items that are specifically selected to promote student engagement and usefulness of results for the specified purpose.

Conclusion

The author of this report argues that information collected through traditional studies of alignment is not sufficient to fully support claims of content validity in the context of computer-adaptive testing. Using recommendations from Wise, Kingsbury, & Webb (2015), alignment experts from the Center on Standards and Assessment Implementation at WestEd have developed a protocol that state and district education agencies may find useful for examining the relationships between items intended to be delivered adaptively and the set of standards that the items are intended to measure. Specifically, the steps in this protocol provide a framework for systematically evaluating the degree to which the CAT's test plan or blueprint, item pool, and item-selection algorithm work synchronously to ensure that all students are tested on the desired depth and breadth of the standards, with optimal challenge during testing.

For additional information, please contact the Center on Standards and Assessment Implementation at the following address: CSAI@wested.org.