Using the Results of Content Alignment Analyses to Inform Ongoing Item-Level Improvements to an Assessment Program:

A Guide for State Departments of Education and for Assessment Vendors

Alignment is the degree to which learning expectations, curriculum, assessments (and other parts of the education system, e.g., teacher education, instructional practices, professional development, etc.) are in agreement and serve in conjunction with one another to guide the system toward students learning what is expected (Webb, 2007). Considered an ethical imperative in the field of measurement, the Standards for Educational and Psychological Testing call out evidence of content alignment between learning expectations and assessments as a core aspect of a validity argument for the interpretation and use of test scores (AERA/APA/NCME, 2014). The United States Department of Education requires states to submit third-party evidence of the degree of alignment of statewide summative assessments with corresponding academic content standards for Assessment Peer Review. Many State Departments of Education also require assessment vendors to support claims with third-party content alignment evidence. Aside from compliance with state and federal requirements, assessment vendors may also conduct internal alignment analyses and/or solicit third-party alignment analyses as part of a high-quality assessment development and evaluation process.

State Departments of Education and assessment vendors can use the summary results and item-level data from these alignment studies to inform ongoing improvements to assessment programs. ***Importantly, item-level issues may exist even when alignment expectations are met. Therefore, item-level data can be used to inform improvements to all assessment programs, including those that met alignment expectations.***

A recent review of over 100 alignment analyses from states across the country found that all used some version of the alignment methodology develop by Dr. Norman Webb (1997), commonly known as the Webb alignment methodology (Traynor et al., 2020). The original tools and methodology developed by Dr. Webb are freely available and have been reinterpreted by many others for use in alignment analyses. Dr. Webb's work is continued and extended through the WebbAlign program of the non-profit Wisconsin Center for Education Products and Services (WCEPS), affiliated with the University of Wisconsin Center for Education Research (WCER). This guide is written from the WebbAlign perspective but can be applied to the results of other alignment approaches as well.

Consideration 1: Do the item-level data identify any items for removal from the test form or item pool?

Check through any Source of Challenge comments that panelists recorded. A Source of Challenge is a technical or content problem with an item that might cause a student to give a wrong answer for the wrong reason or give the right answer for the wrong reason. If the alignment study is completed as part of the test development process, any items flagged can be removed or revised before the first administration of the assessment. Although items on an operational test will have already gone through editorial passes as well as bias and sensitivity review(s), additional issues are sometimes identified during the course of a content alignment analysis. Items flagged by a majority of panelists are likely to need attention, although all Source of Challenge comments should be examined, as one panelist may have noticed an important problem that no one else did. After stakeholder discussion of any items flagged for Source of Challenge, these items can be removed from or retained in the test form or item bank as appropriate. If removed from a robust item bank, removal will resolve issues and have essentially no negative consequences to the overall assessment program. (Some professional judgement is required in the review of these data, as a general comment or note may sometimes be reported as a Source of Challenge if accidentally entered into the incorrect data input space. It is also possible for an issue to be perceived but inaccurate. For example, a panelist might perceive a feature of an online assessment item to be inoperable when it was simply an artifact of the viewer interface used.)

Consideration 2: To what extent are the ratings of independent panelists for standard (or other assessment target) consistent with the internally assigned metadata?

WebbAlign alignment analyses have typically expected that a vast majority of items, defined as at least 75%, should have internal metadata that identifies the standard (or other assessment target) consistent with the independent majority coding for standard. In other words, it is expected that independent reviewers generally find that items are targeting the intended assessment targets. Some leeway is given to account for reasonable differences in professional opinion, as well as legitimate overlap between and among the content within standards. In other words an independent panel may decide an item is a better fit for one standard even if they would agree that the item reasonably addressed another standard. A comparison of internal metadata with the independent coding can reveal any discrepancies in coding. Items with discrepant coding can be reviewed by stakeholders to identify the difference(s) in interpretation of the content alignment between assessment target and assessment item/task. In some cases, discrepancies may simply be due to internal metadata errors and items are viable after corrections to the metadata.

In some cases, a confirmatory analysis of internal metadata may be conducted instead of (or in addition to) an independent analysis. If a confirmatory analysis is conducted, then at least 90% of items are expected to be considered acceptably coded per internal metadata. A confirmatory analysis expects panelists to consider whether the internal metadata are reasonable. The expectation is that essentially all internal coding should be reasonable, therefore, a 90% cutoff is used to allow for some differences in professional opinion. Depending on study structure, a confirmatory analysis may rate items according to the degree to which they address a core aspect of the targeted standard. In these cases, any items identified as only weakly (or partially) addressing the intended assessment target can be pulled for review to determine if revisions to the item and/or to the metadata are warranted.

Consideration 3: To what extent are the ratings of independent panelists for DOK consistent with the internally assigned metadata?

Depth of Knowledge (DOK) is an evaluative tool used to determine the complexity of engagement (often called "cognitive complexity") required by learning expectations, tasks, questions, prompts, and other units of analysis. Depth of Knowledge Consistency between standards and an assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards. As such, appropriate DOK of items is relative to the DOK of the assessment targets. Suggestions provided here assume stakeholder consensus on blueprint specifications that include consideration for degree of DOK consistency between item/task and assessment target. Checking item-level consistency between internal and independent coding can yield information about the degree to which the items are targeting the intended levels of complexity. Similar to expectations for standard coding, WebbAlign alignment analyses have typically expected that the vast majority of items, defined as at least 75%, should have internal metadata that identifies DOK consistent with the independent average coding for DOK. If a confirmatory analysis is conducted, then at least 90% of items are expected to be considered acceptably coded per internal metadata. A comparison of internal metadata with the independent coding can reveal any discrepancies. Items with discrepant coding can be reviewed by stakeholders to identify the difference(s) in interpretation of the complexity of the assessment item/task. High interrater reliability in DOK coding is expected after appropriate training and calibration, but some difference in professional judgement is reasonable, as some tasks may fall between the defined levels. To compare codings, average the reviewer coding, rounding up or down as appropriate. Note that Webb's DOK has been reinterpreted in various ways, in some cases in ways that fall outside of the intended use. In general, however, we recommend that the same interpretation should be used for standards analysis as for item analysis, allowing for an "apples-to-apples" analysis.

Consideration 4: Do item-level comments identify items that might warrant qualitative editorial improvements?

Panelists' item-level comments are rich sources of actionable suggestions. A review of item-level comments may yield different categories of issues that can be routed through different pathways. For example, some items may benefit from minor editorial adjustments and can be routed accordingly. Other categories of comments may warrant more extensive review before deciding whether to and/or how to address the identified issue(s).

Although not the primary focus of the evaluative processes used in a content alignment analysis, educator panels may identify bias and sensitivity issues in their close examination of items. For example, when looking at items developed for use across states, educators may identify language or contexts that are not appropriate for their region. Because these items do not contain technical errors, panelists may not flag them with Source of Challenge and instead include general comments or notes.

WebbAlign™

Get in touch! If you'd like help interpreting the results of your content alignment analysis, reach out to Sara at sara.christopherson@wceps.org

WCEPS
WISCONSIN CENTER for EDUCATION PRODUCTS & SERVICES