Executive Summary

Process Evaluation Report for ISTAR Standard Setting

Prepared by

Lori Nebelsick-Gullett, NG Consulting, SBOE Technical Advisory Committee

August 10, 2017


ISTAR Standard Setting Meetings

English Language Arts and Science

July 19-21, 2017

Mathematics and Social Studies

July 26-28, 2017

Indianapolis, Indiana

Validation of ELA Performance Standards

August 7, 2017

Indianapolis, Indiana

## Executive Summary

Results from two studies are summarized in this Executive Summary. The first section of this document provides an overview of the results of an evaluation of the procedures and processes implemented to establish performance standards for the Indiana Standards Tool for Alternate Reporting (ISTAR) in the subject areas of English/Language Arts (ELA) and mathematics for grades 3-8 and 10; science for grades 4, 6, and 10; and social studies for grades 5 and 7. The second section of this document provides an overview of the results of an evaluation of the procedures and processes implemented for the validation of the performance standards for all ELA assessments. The validation process was conducted post standard setting to address an error in computation of the impact data that was shared with panelists during original standard setting meeting.

## Standard Setting Study

The Item-Descriptor (ID) matching procedure was used to establish recommended performance standards for each ISTAR assessment. All activities during the standard setting meetings were organized and implemented by Questar Assessment, Inc. (Questar). Panelists were recruited by the Indiana Department of Education. Evidence presented in this report is based on a 3rd party, independent evaluator's review of materials, on-site observations, and evaluation of information collected from panelists.

The design of the ID matching procedure was implemented across three days (two days for grade 10) and called for an iterative process to include four rounds of judgments and result in two recommended cut scores for each ISTAR test—a Meeting Proficiency cut score and an Exceeding Proficiency cut score. The process was designed to include the following components:

- General Session. This initial session includes all panelists for a meeting and provides an overview of ISTAR, an introduction to the standard setting process, and a review of logistics such as security procedures and nondisclosure agreements. Panelists then break into grade-band panels and implement the remaining components for each grade-level assessment for which they are recommending performance standards.
- Experience the assessment. Panelists on a grade-band panel can gain insight and understanding of an assessment by taking a form of the assessment under conditions like those experienced by students.
- PLD review and discussion. Panelists independently review and then discuss the performance level descriptors (PLDs) associated with an ISTAR assessment.
- OIB review and discussion. Panelists review and discuss each item in an ordered item booklet (OIB), noting the knowledge, skills, and cognitive processes required to answer an item correctly or achieve a score point on a polytomous item, and discussing why an item was more difficult for students than the previous items.
- ID Matching practice round. For the upper grade level assessment addressed by a grade-band panel, panelists practice the task of matching items to PLDs using a small sample of items arranged in an OIB.

- Rounds of judgments and feedback. Panelists implement an iterative process that includes four rounds of judgments. Feedback is provided between each round that can be used to evaluate and inform their judgments in subsequent rounds.
- Vertical articulation. Table leaders from each panel within a content area serve on a vertical articulation panel to examine the reasonableness of the panels' cut score recommendations given the change in performance expectations across the grades and associated impact data.
- Meeting process evaluation. Each panelist completes a process evaluation survey through which they share their perspectives regarding the training provided, the standard setting process, and the recommended cut scores – including their confidence in the cut scores recommended by their panel. Panelists participating in vertical articulation complete an additional survey focused on that vertical articulation process.

The ID Matching standard setting procedure was implemented as designed—adhering to the intended processes and procedures. Based on formative feedback from IDOE and the process evaluation observer, Questar leadership staff adjusted the guidance provided to facilitators that supported consistency across panels and ensured the process was implemented with fidelity to the intended design and with adherence to standards of best practice. Table 1 provides a summary of the evidence relative to best practices in the field of standard setting and Table 2 provides a summary of the evidence relative to relevant AERA/APA/NCME standards.

*Table 1. Adherence of the Standard Setting Process to Best Practices*

| Process Component | Best Practice | Evaluation |
|---|---|---|
| **Panels** | Panelists should be recruited such that panels have the diversity needed to represent key demographic groups and have sufficiently broad representation.<br><br>Panels should also be sufficiently large. And representative to be judged suitable for setting performance standards.<br><br>Panelists should be knowledgeable of the content area and of students who will take the assessment. | IDOE, in partnership with Quastar, designed and implemented a multistep process to create panels that were representative of the geographic location, school setting, and socio-economic status composition of Indiana schools. Recruitment efforts resulted in panels with diversity regarding these indicators with somewhat lower overall percentage of panelists from the southern region of the state and representing higher SES areas. In addition, there was little diversity of panelists in terms of gender and ethnicity—88% of panelists were white females. IDOE and Questar should investigate if the student groups taught by panelists are representative of the students across the state of Indiana and, more |

| | | specifically, students in this population. |
|---|---|---|
| | | The size of all but one grade-band panel met the expected range of 7-10 panelists. These panels are relatively small when compared to recommendations in the literature, however, panels of this size are in line with those used in past standard setting studies approved through peer review. |
| | | Observations confirmed that all panelists were knowledgeable of the content and most were familiar with the students who took the assessment. (A few general education teachers had less familiarity with the specifics for students in this population.) All panelists attended to the tasks, asked questions, and remained focused throughout the standard setting process. |
| **Method** | The standard setting method should be appropriate for the type of test administered.<br><br>The judgment task should be understandable to those making the judgments. | The ID Matching method was appropriate for use with the ISTAR assessments. The task of matching knowledge, skills, and processes addressed by items to those represented through the PLDs was understood and applied with fidelity by the panelists. |
| **Implementation** | Key aspects of the standard setting process were implemented in accordance with best practices. These include:<br>  1. Facilitator training<br>  2. Panelist training<br>  3. Clarity and use of performance category descriptions<br>  4. Opportunity to experience the test | Overall, the implementation of the ID Matching procedures occurred as designed and met the parameters outlined for best practices. Each are bulleted component in the cell to the left was implemented. The areas for additional comment here include training of facilitators and process evaluation.<br><br>Formative feedback was used early in the process to provide additional |

| | | |
|---|---|---|
| | 5. For an iterative process: <br>    a. Opportunity for discussion; <br>    b. Interpretation and use of feedback; <br> 6. Interpretation and use of impact data (when used) <br> 7. Process conducted efficiently <br> 8. Computation of cut scores, was transparent <br> 9. Panelist completed process evaluations. | support to facilitators and their presentation/implementation of the methodology with grade-band panels. This improved consistency and supported fidelity of implementation. <br><br> As noted, Questar responded immediately to formative feedback from observers and to questions from facilitators, which ensured consistent, faithful implementation of the designed procedures. Given the observed variability in facilitator implementation at the beginning of the first standard setting meeting, additional support and instruction were needed to ensure all facilitators had the depth of understanding needed to implement all component of the design. For future studies, IDOE and Questar should ensure adequate time and materials for training facilitators, including facilitator walk through of all processes and procedures with adequate opportunity for discussion to address inconsistencies in understanding, interpretation, and planned implementation of the designed methodology. <br><br> Other than the input panelists provided through the readiness surveys prior to each round of the iterative process, for week one, panelists had no formal opportunity to provide formative feedback on each component of the procedures as implementation occurred. Panelists completed an evaluation at the end of the standard setting meeting as summative feedback regarding the processes, procedures, and results. |

| | | The results of this survey indicate that having the information formatively may have benefited the facilitators as they could have more effectively addressed panelists' needs.<br><br>For week two, panelists completed a paper evaluation at the end of each day that had two questions focused on what panelists liked and what could be clearer. This information was reviewed by the project leaders.<br><br>For future studies, IDOE and Questar should consider implementing process evaluations throughout implementation of the standard setting study; and use scales that are balanced regarding positive and negative responses. |
|---|---|---|

*Table 2. Adherence of the Standard Setting Process to AERA/APA/NCME Standards*

| Standard | Text of Standard | Evaluation |
|---|---|---|
| **5.21** | When proposed score interpretation involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. | Standard 5.21 was fulfilled through the standard setting design document in which the rationale and procedures were first documented. During the opening session, the rationale and procedures were explained to panelists. |
| **5.22** | When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of an item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way. | The ID Matching procedure provided panelists the opportunity to apply their knowledge, skills, and experiences in a reasonable way. The tasks of identifying and matching knowledge and skills between items and PLDs and discussing those judgments with peers aligns with educators' professional experiences. |

| 5.23 | When feasible and appropriate, cut scores defining categories and distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria. | Empirical data (impact data) was presented to and discussed with panelists using their judgments from Round 3, prior to their final Round 4 judgments. The impact data was based on the Spring 2017 implementation of ISTAR assessments. An error in the impact data was discovered post standard setting for all ELA assessments. See the Validation Study section for a description of how this error was addressed. |
|---|---|---|

**Conclusions based on the Standard Setting Study**

The evidence provided through on-site observations, review of materials, and examination of panelist data provide support for the validity of the outcomes of the standard setting procedures and processes. The process issues identified during the standard setting meetings, were addressed and did not negatively impact the overall fidelity of implementation nor did they deter from the validity of the results. As noted in Table 2, for ELA only, an error in the impact data provided to the ELA panelists at all grade levels was identified post standard setting. The impact data shared with panelists as part of Round 3 and Round 4 feedback, as well as vertical articulation, was based on the wrong response probability. The erroneous data could have impacted panelists interpretation and application of that information during panel discussions as well as their judgments during Round 4 and vertical articulation.

It is the opinion of the independent evaluator that, overall, the iterative standard setting process Questar implemented for the ISTAR assessments was executed in a systematic fashion in accordance with best practices and met the nature of the professional standards identified in the AERA/APA/NCME standards except for the use of the erroneous impact data for the ELA panels. The IDOE and Questar designed and implemented a validation process for all the ELA assessments, bringing a subgroup of each panel together to conduct the validation process described in the next section.

**Validation of ELA Performance Standards**

To ensure the validity of the performance standards recommended for each ELA ISTAR assessment, Questar and the IDOE designed a validation study to present panelists with the correct impact data and to evaluate the following question:

> How different would panelists cut score recommendations be if they had considered the correct impact data?

All panelists who attended the original standard setting in June of 2017 were invited to participate in the validation process. Table 3 shows the number of panelists from each panel who returned for the validation process. At least half of each panel participated in the validation meeting with all panels except grade 10 including at least one table leader from the standard setting panel.

*Table 3. Number of Participants*

| Content Area | Grade(s) | Validation Participants | Returning Table Leaders | Original Panel Size |
|---|---|---|---|---|
| ELA | 3–4 | 6 | 2 | 10 |
| | 5–6 | 4 | 1 | 9 |
| | 7–8 | 4 | 2 | 8 |
| | 10 | 3 | 0 | 5 |

The design of the standards validation procedure was implemented during a half day meeting. All panels were convened in a single room for the meeting. The process was designed to include the following components:
- General Overview. Panelists receive an initial overview focused first on setting the context for panelists, assuring panelists they had implemented the original standard setting process with fidelity, describing the error that occurred, and explaining both how the error was corrected and how the correction was validated. The focus was then intended to shift to describing the standards validation process and sharing the original and corrected impact data.
- Instructions in Standards Validation. Panelists receive step-by-step instructions on the standards validation process. Panelists are instructed to focus on the place they were in the process, and in their thinking, when the impact data was introduced prior to their Round 4 judgments. The validation process was designed to have panelists:
  - Conduct a grade-band panel review of the Round 3 judgements and the corrected impact data, and make recommendations for Round 4 cut scores
  - Conduct a cross-grade vertical articulation and make final recommendations for cut scores at all grade levels
  - Complete a process evaluation

  The process was designed to focus on two key questions: (1) Would the grade-band committee's Round 4 recommendations be different if panelists had seen the corrected Round 3 impact data? and (2) Would the vertical articulation committee's recommendation be different if the corrected impact data were provided?

The standards validation process was implemented as designed—adhering to the intended processes and procedures. The table level discussions for each grade-band panel centered, as intended, on the content of the items and the content of the PLDs, with a specific focus on items around the Round 3 cut scores—which were marked in panelists' OIBs. A member of the Questar team provided support at each table, addressed questions, and ensured the panel maintained the intended focus. Each panel asked questions as needed to ensure they understood the process and the information available. The grade-band panels applied the process with fidelity, implementing

each of the intended steps, incorporating all viewpoints, and ensuring there was an agreed upon rationale for decisions/recommendations from the panel.

For the cross-grade articulation process, panelists were reorganized into three cross-grade-band groups, ensuring that each table had a representative from all grade-band panels. The mixed panel table discussions were focused, as intended, first on the patterns and trends they would expect, based on the PLDs across grade levels, and then on the impact data and how that did/did not align with their expectations. All panelists were then engaged in a discussion about recommendations going forward. Any changes to the Round 4 cut score recommendations were made by consensus. The large group implemented the same process that was used for vertical articulation in the original standard setting study. Panelists were instructed to stay true to the content-based recommendations and to provide cut score recommendations that produce reasonable impact within and across grades.

Panelists completed an evaluation survey at the end of the meeting. The majority of panelists were satisfied with the recommended cut scores and were satisfied the cut scores are based on what students in this population know and are able to do. One or two panelists in each of the elementary and middle school panels reported they were somewhat satisfied and one panelist was dissatisfied. The comments associated with these responses were not focused on the standard setting or the validation processes, rather they focused on the rigor of the expectations as expressed through the PLD, the appropriateness of that rigor for this population of students, and concerns with the impact on teacher morale. Two panelists in each of the elementary and middle school panels reported that the impact data were somewhat accurate in reasonably reflecting the percentages of students in this population who should be classified into each performance level in 2017. One panelist responded that the impact data were not at all accurate.

**Conclusions based on the Standards Validation Study**

The evidence provided through on-site observations, review of materials, and examination of panelist data provide support for the validity of the outcomes of the validation procedures and processes. It is the opinion of the independent evaluator that the standards validation process implemented for the ELA ISTAR assessments was executed in a systematic fashion in accordance with best practices, supports the validity of the outcomes, and ensures the overall process of setting performance standards for the ISTAR assessments met the nature of the professional standards identified in the AERA/APA/NCME standards.