Memo

To: Cynthia Roach, Ed.S.
Senior Director of Accountability and Assessment, Indiana State Board of Education Staff
Michele Walker, Ed.D.
Director, Office of Student Assessment, Indiana Department of Education
From: Karla Egan, Ph. D., Karen Barton, Ph.D., and Edward Roeber, Ph.D.
CC: Brian Murphy
Date: October 13, 2015

Re: ISTEP+ Standard Setting

The purpose of this memorandum is to evaluate the appropriateness of the procedures used and the quality of implementation of the standard setting process for the 2015 ISTEP+ assessments in grades 3 through 8 in English Language/Arts (ELA) and mathematics. Three external experts, Dr. Karla Egan, Dr. Karen Barton, and Dr. Edward Roeber, served as the Technical Advisory Committee (TAC). All members of the TAC attended the standard setting, observed the standard setting process, and reviewed materials used during the standard setting procedure. Dr. Roeber provided a separate report summarizing his observations of the standard setting to the Indiana Department of Education.

This memorandum (1) summarizes the way in which panelists were chosen for the standard setting; (2) overviews the implementation of the standard setting; and (3) evaluates the implementation of standard setting processes. The vendor for a standard setting activity will also produce a step-by-step technical report of the standard setting process for the Department. This standard setting technical report should summarize the panelist round-by-round recommendations and panelist readiness surveys. These detailed analyses were not yet available and thus are not referenced in this memorandum. DRC|CTB provided IDOE, ISBE, and the independent evaluators with the panelist evaluations of the standard setting and the articulation process immediately following the workshop. Throughout this memo, panelist evaluations are provided when relevant. The full results of the evaluations may be found in the DRC|CTB standard setting technical report.

Panelist Selection

For both content areas, the Indiana Department of Education (IDOE) purposefully selected panelists to reflect three factors: geographic region, school type (urban, suburban, rural), and poverty level. The IDOE provided a summary of the panelists' demographics. Table 1 shows the distribution of standard setting panelists by

geographic region, Table 2 shows the distribution of standard setting panelists by school type, and Table 3 shows the distribution of standard setting panelists by poverty level. The evidence in these tables shows that the panelists represented diverse backgrounds that reflect the factors deemed important by IDOE.

Table 1. Distribution of Standard Setting Panelists by Geographic Region

Geographic Region	All Panelists
North	38.0%
Central	42.0%
South	20.0%

Table 2. Distribution of Standard Setting Panelists by School Type

School Type	All Panelists
Urban	31.0%
Suburban	29.0%
Rural	40.0%

Table 3. Distribution of Standard Setting Panelists by Poverty Level

Poverty Level	All Panelists
Low	77.0%
High	23.0%

The selected panelists appeared to be knowledgeable of the content area and of students as demonstrated by the types of conversation observed throughout the standard setting. Almost half of the panelists in each standard setting had 16 or more years of experience (see Table 4).

Table 4. Distribution of Standard Setting Panelists by Years of Experience

Years of Experience	ELA (n=56)	Mathematics (n=54)
1-5	12.5%	14.8%
6-10	14.3%	22.2%
11-15	17.9%	13.0%
16-20	19.6%	18.5%
20+	35.7%	31.5%

Standard Setting Procedure

The bookmark standard setting procedure was implemented for the ISTEP+ grades 3 through 8 ELA and mathematics assessments during the week of October 5-8, 2015 at the Wyndham Hotel in west Indianapolis. This process was also used for the ISTEP+ College and Career Readiness Assessment and Indiana National Center and State Collaborative (NCSC) standard settings.

Bookmark is a content-based process that utilizes an ordered item booklet (OIB), in which the test questions are ordered from easiest to most difficult. Guided by preliminary performance level descriptors (PLDs), which were written by IDOE content-area specialists, panelists study the ordered test questions and place a cut score that separates the content students should know to enter a performance level (i.e., Does Not Pass, Pass, Pass+) from the content that is more than enough. Panelists engage in three rounds of activities during a bookmark standard setting. At a high-level, the following occurs in each round of activity:

- Round 1: Panelists review, discuss, and edit the PLDs; take the test; review the OIB, and recommend bookmarks. (Bookmarks are translated to cut scores by the vendor staff.)
- Round 2: Within the table groups and led by table leaders, panelists discuss the range of individual bookmark placements and recommend bookmarks.
- Round 3: As a large group and led by room facilitators, panelists review the range of bookmark placements for their grade/content area, review impact data (the percentage of Indiana students in each performance level) based on the Round 2 median bookmarks, and recommend final bookmarks.

Implementation of the ISTEP+ Grades 3 – 8 ELA and Mathematics Assessment Standard Setting

This section provides an overview of the implementation of the ISTEP+ bookmark process, including the configuration of the panelists, opening session, bookmark training, closing session, and articulation processes. When available, panelist evaluations of events are included.

Panel Configuration

The 110 panelists were separated into six groups based on their experience:

- Grade 3-4 ELA
- Grade 5-6 ELA
- Grade 7-8 ELA
- Grade 3-4 mathematics
- Grade 5-6 mathematics
- Grade 7-8 mathematics

Within each group, panelists recommended cut scores for the lower grade followed by the upper grade in each grade pair. As described above, IDOE recruited a diverse sample of Indiana educators to make recommendations about the content-based cut scores.

Each group was sub-divided into four groups of four to five panelists to facilitate active engagement of all panelists. Three or four panelists from each grade group were selected to serve as table leaders for the process. In this role, they facilitated the small group discussions that occurred during Rounds 1 and 2. These panelists received additional training over the lunch hour that immediately followed the initial training.

Opening Session

Dr. Michele Walker welcomed panelists to the process and overviewed Indiana's test development process. Dr. Walker also explained that the panelists would make recommendations that will be approved by the Indiana State Board of Education. Mr. Ricardo Mercado, DRC|CTB, provided an overview of the standard setting process and preliminary training on the Bookmark procedure.

Table 5 shows the panelist feedback about the opening session. Most panelists agreed that the opening session provided a clear overview of the cut score process and a clear explanation of the development of the tests.

Table 5. Panelist Evaluations of the Opening Session*

	ELA (n=56)		Mathematics (n=54)	
	Disagree	Agree	Disagree	Agree
The opening session provided a clear overview of the cut score process.	1.8%	92.8%	3.7%	92.6%
The opening session provided a clear explanation of the development of the tests.	7.2%	76.8%	5.6%	83.3%

*The percent selecting the neutral category is not included here.

Round 1

Following the opening session, the panelists went to their breakout rooms to engage in the Round 1 activities. Panelists studied the PLDs and discussed the knowledge, skills, and abilities of the target student. The target student is defined as the student just entering a performance level. Panelists discussed their ordered item booklets. Prior to placing bookmarks, they were trained on how to place a bookmark.

Bookmark Training

In-depth training on how to place a bookmark was conducted just prior to the time when panelists made cut score recommendations. Mr. Mercado spent about an hour training panelists on the mechanics of bookmark placement as well as the relationship between items and students.

Table 6 shows the panelist feedback about bookmark training. Over 90% of panelists in both content areas agreed that bookmark training helped them understand the task and that they were prepared to compete the task.

Table 6. Panelist Evaluations of Bookmark Training*

	ELA (n=56)		Mathematics (n=54)	
	Disagree	Agree	Disagree	Agree
The training on bookmark placement helped me understand what we were preparing to do.	5.4%	92.9%	1.9%	90.7%
After the training, I felt confident I was prepared to complete the cut score setting task.	7.2%	91.1%	1.9%	90.8%
I understood how to place my bookmarks.	3.6%	94.6%	1.9%	94.4%

*The percent selecting the neutral category is not included here.

Rounds 2 and 3

During Round 2, panelists discussed the range of bookmark placements within their tables. Following discussion, they recommend Round 2 bookmarks. During Round 3, the room facilitator led panelists through a discussion of their Round 2 bookmarks. Panelists were also shown impact data based on their Round 2 recommendations. Mr. Mercado presented the impact data, and Dr. Walker answered process questions related to the impact data.

Closing Session

At the end of the workshop after final bookmark recommendations, all panelists again gathered into a single room where Mr. Mercado presented the across-grade cut scores and impact data for each content area. Upon returning to their rooms, the panelists discussed the across-grade results with the table leaders. The panelists provided a range in which their table leaders could adjust cut scores in order to promote the cross-grade coherence of results and reflect their content-based recommendations. Typically, panelists are asked to indicate their level of support for the final cut scores. Unfortunately, this question was not on the panelist evaluation.

Articulation Process

On Day 4 of the standard setting workshop, Dr. Juan d'Brot led the 22 table leaders through a vertical articulation process for each content area. During the vertical articulation process, the table leaders discussed the coherence of the results across grades. Within each content area, table leaders were shown the results across all grades, first in ELA and then in mathematics. If table leaders wanted to examine an area of disarticulation in the across-grade results, then Dr. d'Brot directed the table leaders to consider the content of the OIB and PLDs, as well as the range of cut scores provided by the panelists to their table leaders at the conclusion of the standard setting activity the previous day. This process ensured any recommended changes remained tied to content.

Table 7 shows the table leader evaluation of the articulation process. ELA table leaders unanimously agreed that they understood the benefits of well-articulated performance standards, the final recommendations represent the work of the standard setting committee, the recommendations are reasonable, and the pattern of impact data is explainable. The mathematics table leaders also agreed with these statements, but not to the same extent as the ELA table leaders.

Table 7. Table Leader Evaluations of Articulation Process*

	ELA (n=12)		Mathematics (n=10)	
	Disagree	**Agree**	**Disagree**	**Agree**
I understood the benefits of well-articulated performance standards	0%	100%	0%	100%
The final recommendations represent the work of the standard setting committee.	0%	100%	30%	70%
I feel the recommendations that resulted from this process are reasonable.	0%	100%	30%	70%
In general, the impact data form an explainable pattern across grades.	0%	100%	10%	90%

*The percent selecting the neutral category is not included here.

Review by the Technical Advisory Committee

Immediately following the vertical articulation process, staff members from ISBE and IDOE met with members of the TAC to discuss results. The TAC considered the coherence of the system of cut scores and the conversations of the table leaders. The TAC recommended a few adjustments in the cut scores to address a couple of areas of remaining disarticulation. These adjustments were within one combined standard error of the

panelist-set cut scores. The combined standard error accounts for the standard error of the assessment and of the Bookmark process.

Expert Evaluation of the Standard Setting Processes

DRC|CTB implemented the standard setting procedure with fidelity to the original design, and the process adhered to best practices and AERA/APA/NCME standards. Table 8 describes each procedure's adherence to best practices in the field of standard setting. It also notes weaknesses in the process.

The panelists appeared to be knowledgeable of the content and diligent in their cut score recommendations. The panelists provided content-related rationales for placement of Bookmarks and did not appear to have a preconceived idea about the placement of Bookmarks. The training processes were particularly strong at this standard setting. The content-based standard setting activities, overall, were conducted in a manner consistent with sound psychometric practices.

The evaluators observed that panelists actively participated throughout the standard setting. Multiple panelists are invited to standard setting in order for a cross-section of opinion to be reflected in the final recommended cut scores; therefore, it is important to watch for panelists who are not participating or for panelists who are dominating discussion during the standard setting. These types of panelists were not noted in any of the rooms. The few issues that arose were comparatively minor and did not substantially affect the validity of the results.

In addition to Mr. Mercado, DRC|CTB provided skilled facilitators for each breakout room who were able to deal with issues as they arose and two psychometricians (Dr. Christie Plackner and Dr. Juan d'Brot) who monitored the implementation of the standard setting, assisting when needed. In addition, the project psychometrician, Dr. Dong-In Kim, attended the standard setting and discussed the comparability of the computer and pencil/paper forms and answered test score related questions in each room. Throughout the process, the evaluators observed that DRC|CTB facilitators and psychometricians routinely guiding panelists to use the PLDs and to think of all students (not just the students in their classroom) when recommending cut scores.

On the negative side, the panelists sometimes seemed rushed for time during the standard setting and during the articulation process. DRC|CTB did not show group-level recommendations following Round 1 of standard setting. It is typical to provide panelists with some sort of feedback after recommendations gathered in Round 1, but this did not occur until following Round 2. Even though this particular workshop was a strong implementation of the Bookmark standard setting procedure, the DRC|CTB process would have benefitted from the use of detailed facilitator scripts. This would have ensured that each facilitator implemented the process in the same manner.

Also, the panelists experienced a good deal of downtime (sometimes over an hour) while Dr. Walker and Mr. Mercado presented impact data to each room. While the intention was to ensure consistent discussion and time for questioning by participants of the IDOE, this meant that the next round of rating or work on the second grade began before the results of the prior round or grade level were presented and discussed. This may have

minimized discussion of the results of ratings from the prior round or grade level; and it served to interrupt work on the next round of ratings or second grade level. During future standard settings, it is recommended that multiple staff members are available to present impact data and to help facilitate discussion about it so that all panelists can adhere to the same schedule.

For the articulation process, it is recommended that the content groups be separated so that each can use the entire time to engage in conversation. The mathematics group was rushed to complete this phase of the process due to the length of the ELA discussions that preceded it.

In addition to best practices, there are professional standards (AERA/APA/NCME, 2015) in the measurement field related to standard setting. Table 9 shows how each procedure adhered to the AERA/APA/NCME standards. In both cases, the content-based standard setting process met the criteria represented in the *Standards for Educational and Psychological Testing.*

Limitations

There are limitations for interpreting this evaluation report. Only the procedural evidence of validity of the standard setting process was evaluated. Further, this is only one piece of information that should be collected when gathering validity evidence to support the proposed cut score. Procedural evidence is important, and it provides support that the process used to establish cut scores was reasonable and implemented with fidelity to professional standards. While procedural evidence is necessary, however, it is not sufficient in establishing validity evidence for a proposed cut score. The DRC|CTB technical report is another important source of validity evidence, and this was not available to the TAC for its evaluation and conclusions about the standard setting prior to preparing this report for review. As with any assessment system, additional studies should be planned to examine the internal and external validity evidence to support the interpretations and use of the ISTEP+ grades 3 – 8 ELA and mathematics assessments.

Conclusions

Based on observations and review of standard setting materials, it is the opinion of Dr. Egan, Dr. Barton, and Dr. Roeber that the standard setting process implemented by DRC|CTB for the ISTEP+ grades 3 – 8 ELA and mathematics assessments was sufficiently executed in accordance with best practices and industry standards in the field of psychometrics. The TAC recommends the adoption of the standards that were set for grades 3 – 8 ELA and mathematics.

Table 8. Adherence of the DRC|CTB Standard Setting Process to Best Practices

	Best Practice	ISTEP+ Standard Setting Evaluation
Panels	Panels should be recruited so that they are representative of important demographic groups, and they should be knowledgeable of the content area and of students. Panels should also be sufficiently large.	Serious attention was given to create panels that were representative of Indiana based on three factors: geographic region, school type (urban, suburban, rural), and poverty level. The six panels consisted of approximately 20 panelists divided into four groups. Each group consisted of four to six panelists. This provides a mechanism for checking generalizability of the performance standards (Hambleton, Pitoniak, & Copella, 2012). Observations confirmed that all of the panelists were knowledgeable of the content and were diligent in setting the standards.
Method	The standard setting method should be appropriate for the type of test administered and the understandability of the judgment task.	The Bookmark method was appropriate for use with the ISTEP+, which was a mixture of item types. DRC|CTB was diligent in their training for the judgment task, spending an hour on this training. They also checked for understanding by administering check sets. The DRC|CTB facilitators and psychometricians regularly checked with panelists to ensure understanding.
Implementation	There are various aspects of implementation that must be considered when evaluating a standard setting. These include: (a) training; (b) using PLDs, (c) taking the test; (d) using an iterative process; (e) providing opportunity for discussion; and (f) presenting impact data. In addition, the	The purpose of the assessment and the uses of the test scores were explained to panelists during the opening session. Panelists were exposed to the assessment and how it was scored. The panelists engaged in an iterative process and used the descriptions of the performance levels effectively. They were shown impact data following the second round and

	Best Practice	ISTEP+ Standard Setting Evaluation
	method should be efficient, allow transparency in the computation of cut scores, and provide time for evaluations.	again following the final round. The method was implemented efficiently, and panelists completed evaluations. Following the standard setting, an articulation committee comprised of the 24 table leaders and the TAC met separately to examine the coherence of the system of cut scores. This is an important component of modern standard setting where cut scores are set in contiguous grades. This provides panelists with an opportunity to examine the consistency of recommendations across grades. While the standard setting process followed best practices in standard setting implementation, there is room for improvement in future standard settings. It is suggested that panelists be provided feedback following each round. In addition, multiple teams should be available to present impact data so that panelists do not have unnecessary downtime and all panels carry out their tasks in a timely manner.

Table 9. Adherence of the DRC|CTB Standard Setting Process to AERA/APA/NCME Standards

Standard	Text of Standard	ISTEP+ Standard Setting Evaluation
5.21	When proposed score interpretation involves one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.	Standard 5.21 was fulfilled through DRC|CTB standard setting design 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.	As explained in the previous section, the Bookmark procedure provided a reasonable means for panelists to share their knowledge and experience through group discussions and to make judgments in an intuitive manner. Almost all of the panelists agreed that they understood how to place their bookmarks.
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 was presented to panelists based on Round 2 recommendations. This data was based on the Spring 2015 implementation of the ISTEP+. Panelists were again shown impact data based on their final cut scores.