South Carolina End of Course Examination Program (EOCEP) Biology
Standard Setting Report

Introduction

Pearson conducted a standard setting for the End of Course Examination Program (EOCEP) Biology test in Columbia, SC, July 27-28, 2009. Two achievement level cut scores (*Met* and *Exemplary*) were recommended by the standard setting panel through three rounds of the Item Mapping method. Panelists worked through a 90-item ordered item booklet, which represented 150% of an actual test form. Panelists placed recommendations for the *Met* and *Exemplary* achievement level cut scores based on the progression of item difficulty. These cuts (booklet page numbers) were mapped onto item difficulty (Rasch) estimates that were adjusted using the response probability criterion of .67 (RP67).

The Biology items used for standard setting were field tested in two different administrations. In Spring 2008, a random matrix sample field test was conducted in which 176 items were spread evenly across three forms. Fifteen of these items served as links between the three forms. There were approximately 2,000 students tested per form. In Spring 2009, Biology was field tested through a census field test, with 270 items spread evenly across six forms. Additionally, there were 15 common items on each form that served as linking items. Students enrolled in Biology 1 or Applied Biology 1 and eligible for end of course testing participated in the Biology field test for a total of approximately 30,000 students, or about 5,000 students per form. Items deemed viable for operational use from these two field test years were eligible for inclusion in the 150-item ordered item booklet.

Standard Setting Panel

A total of 17 individuals participated for a day and a half in providing recommendations for the *Met* and *Exemplary* achievement level cut scores. These panelists were recruited to represent various stakeholders within the state of South Carolina. Although the majority of the participants were science educators - including a Biology professor from the University of South Carolina - there was an English for Speakers of Other Languages (ESOL) teacher, a former Mathematics teacher, and one person employed outside of education. It should be pointed out the former Mathematics teacher is currently serving as a coordinator for Mathematics, Science, Physical Education, and Health. A summary of the panel's demographic information is provided in Table 1. For the panelists currently involved with Science instruction, the average number of years experience was 22 years, with as few as 5 years and as many as 36 years.

Table 1. Summary of Panel Demographic Information

Gender		Ethnicity		Curriculum Background		
Male	Female	White	Black	Science	Non-Science	None
5	12	14	3	14	2	1

The panelists were seated among three tables, each of which had an appropriate mix of Science curriculum experience and gender and ethnicity representation.

Standard Setting

Prior to the standard setting process, participants worked through the ordered item booklet (OIB) as if they were students taking the test. During this task, the panelists could develop an understanding of the difficulty of the items and how the difficulty of each item might compare to those before and after it as presented in the OIB.

Following this exercise, the panelists reviewed and discussed the definitions of the achievement levels *(Not Met, Met,* and *Exemplary*), indicating the skills that distinguished each achievement level from the others. From this discussion, the panelists were asked to draft achievement level definitions for *threshold* students for both the *Met* and *Exemplary* achievement levels – students who would be classified in the particular achievement level, but minimally (*see Appendix A*). This task was critical for the standard setting process as these definitions were used by the panelists during the process of indicating achievement level cuts.

Prior to the first round of standard setting, a practice round was facilitated to ensure that the panelists understood the process of making achievement level cuts, using the definitions of *threshold* students. As with the actual standard setting rounds, the panelists were asked the following question: *What is the last item that 67 out of 100 "Just Met" (or "Just Exemplary") students will answer correctly?* The panelists were instructed to place their cut on that item. Once the practice round was completed and the panelists felt comfortable with the process, the first round of standard setting began. Prior to each of the three rounds, the panelists answered questions from a 'readiness survey' to indicate they were ready for the standard setting round. This was to ensure that all panelists felt comfortable and were ready to recommend achievement level cuts.

After Round 1, each table of panelists was provided the cut scores and summary statistics of those cuts specified by each panelist of that table. From this, the panelists discussed their cuts, providing rationales according to the definitions of the *threshold* students. Once the panelists felt comfortable with their table discussions, they proceeded to Round 2 of standard setting, making achievement level cuts recommendation as they did in Round 1, but with knowledge gained from the table discussion.

For Round 2, the descriptive statistics of the panel's OIB page numbers were presented to the entire panel along with how the recommended cuts (median OIB page numbers) impacted the distribution of students within each achievement level (*Not Met, Met,* and *Exemplary*), using the Spring 2009 test population. The impact data was shown for the

overall test population as well as by gender and ethnicity (White vs. African American). The standard setting panel discussed the Round 2 results, sharing rationales for individual cuts based on the definitions of the *threshold* students.

Round 3 occurred in the same manner as Round 2 – panels provided another set of cuts using knowledge gained from the panel's discussion of the Round 2 results. The Round 3 cut scores and resulting impact data were shown to the panel as part of the final debriefing of the standard setting meeting. The median of the page numbers, specified as cuts by the panelists, were mapped to the corresponding theta estimate from the thetas used for the OIB. These mapped theta estimates were used as the recommended cut scores provided to the South Carolina Department of Education (SCDE). The results from Round 3 are discussed in the next section.

Table 2 shows the descriptive statistics for the OIB page numbers from each of the standard setting round.

Table 2. OIB Page Numbers by Round

Round	Statistic	Achievement Level Cut	
		Met	*Exemplary*
Round 1	Mean	18.65	61.12
	Median	16.00	64.00
	Minimum	5.00	31.00
	Maximum	39.00	77.00
Round 2	Mean	14.00	59.12
	Median	11.00	62.00
	Minimum	5.00	33.00
	Maximum	38.00	77.00
Round 3	Mean	11.47	60.06
	Median	11.00	62.00
	Minimum	6.00	40.00
	Maximum	17.00	77.00

Analyses

The Biology field-test items were concurrently calibrated across all forms using an anchor item design, placing all items on a common metric. Ninety items (150% of the operational test length) were chosen for the OIB to allow for a thorough spread of theta estimates for the impact data. The items were chosen considering the test blueprint as well as matching the average Rasch difficulty of the overall item bank and the average Rasch value of the item bank *by standard*. The original item parameters and the theta values with RP67 are presented in Table A in the Appendix B.

Recommended Cut Scores

The recommended theta cut score for *Met* from Round 3 was -0.12 (from a median OIB page number of 11), while the recommended theta cut score for *Exemplary* was 0.92 (from a median OIB page number of 62). Table 3 shows the impact data associated with these recommended cut scores.

Table 3. Impact Data from Round 3 Recommended Cut Scores

Subgroup	Achievement Levels			Total
	Below Met	Met	Exemplary	
Overall	50.51%	39.62%	9.86%	100.00%
Gender				
Female	51.00%	40.45%	8.55%	100.00%
Male	49.97%	38.74%	11.29%	100.00%
Ethnicity				
African-American	69.88%	27.62%	2.50%	100.00%
White	37.77%	47.69%	14.54%	100.00%

Panelist Variability

Estimation of panelist variability can be used to evaluate the stability of the cut score recommendations, considering that the standard setting could be replicated using a different collection of panelists. In order to estimate and describe the variability in panelist's judgments, a Generalizability Theory (G-Theory) study was conducted (Lee & Lewis, 2001). For this investigation, the sources of variability of interest were panelists and rounds. For each performance level, the variance associated with each of these sources was estimated using the maximum likelihood (SAS VARCOMP) procedure. After estimation of the variance components, the G-Theory provides a mechanism for describing the variability associated with panelist's judgments. This is important for determining how similar the cut scores might be if a different set of panelists were asked to recommend cut scores. The result is an estimate of the standard error of the cuts cores for this set of panelists' data.

In this analysis, the number of rounds was treated as a fixed factor, meaning that if the meeting were held again, the same number of rounds would be used. Therefore, the three rounds of cut scores were used.

The G-Theory standard error was computed using the formula below, and the standard error estimates were adjusted by 1.253 to account for the use of the median.

$$SE_{cut} = \sqrt{\frac{\sigma^2_{Judges}}{N_{Judges}} + \frac{\sigma^2_{Error}}{3 \bullet N_{Judges}}}$$

It is common for policy-makers to consider the total error associated with cut scores prior to making final decisions, taking into account the uncertainty associated with the recommended cut scores. Total error in this case is conceptualized as the sum of the measurement error associated with the instrument and the error associated with the cut score procedures described above. The total error was calculated as follows:

$$SE_{Total} = \sqrt{(CSEM_{Cut})^2 + (SE_{Cut})^2} \ ,$$

where *CSEM* is the conditional standard error of measurement for the theta cut, and *SE* is the standard error computed using the G-theory. For this analysis, Winsteps (Linacre, 2006) was used to generate a raw-score-to-theta conversion table based on the 90 items presented in the OIB. From this, the standard error of measurement values for the recommended theta cut scores (theta estimates closest to, but not higher than the recommended theta cuts) were obtained for computing the total standard error with the above equation.

Table 4 summarizes the panelist recommendations from the standard setting meeting including the standard error associated with these recommended cut scores.

Table 4. Standard Error Indices with Recommended Theta Cut Scores

Cut	Recommended Cut Score	SE_{cut}	$CSEM_{cut}$	SE_{total}
Met	-0.12	0.07	0.22	0.23
Exemplary	0.92	0.05	0.24	0.25

Note: For Met, SE = 0.069027, CSEM = 0.2186, and total SE = 0.2292; for Exemplary, SE = 0.053231, CSEM = 0.2418, and total SE = 0.2476. All standard error values have been rounded for presentation.

References

Lee, G. & Lewis, D. M. (2001). *A generalizability theory approach toward estimating standard errors of cutscores set using the bookmark standard setting procedure*. Paper presented at the annual meeting of the national council on measurement in education, Seattle, WA.

Linacre, J.M. (2006). *WINSTEPS* Rasch measurement computer program. Chicago: Winsteps.com.

APPENDIX A
Achievement Level Descriptors for Threshold Students

Minimally Met Students…

- are aware of multiple components, but not sure how to connect them (e.g., photosynthesis)
- can interpret basic facts from graphs, but cannot make inferences/conclusions from them.
- have trouble with gaps/holes of knowledge within complete processes.
- can state the scientific method but cannot apply it successfully/consistently; the knowledge breaks down when the process is analyzed piece-meal.
- have some organizational skills (e.g., notebook, lab, note-taking), but they are not refined.

Minimally Exemplary Students…

- know that multiple details interact with each other within scientific processes, but do not know the products.
- are more consistent with their applications of scientific knowledge.
- can compare one process with another, synthesizing similarities.
- can analyze what things would work and what things will not work in particular situations.
- begin to take classroom instruction (knowledge) and apply it outside of the classroom (i.e., seeing the "why" and "what if" instead of just the "how").
- begin to critique and evaluate processes.

APPENDIX B
Ordered Item Book Map

Table A. Biology EOCEP Ordered Item Map

Item Map Order	Original Item Parameter	Theta Parameter with RP67	
1	-1.6400	-0.9320	
2	-1.3665	-0.6585	
3	-1.3550	-0.6470	
4	-1.2872	-0.5792	
5	-1.1990	-0.4910	
6	-1.1305	-0.4225	
7	-0.9740	-0.2660	
8	-0.8860	-0.1780	
9	-0.8790	-0.1710	
10	-0.8390	-0.1310	
11	-0.8300	-0.1220	**<=Level 2 Cut**
12	-0.7589	-0.0509	
13	-0.7560	-0.0480	
14	-0.6740	0.0340	
15	-0.5840	0.1240	
16	-0.5730	0.1350	
17	-0.5650	0.1430	
18	-0.5536	0.1544	
19	-0.5430	0.1650	
20	-0.5370	0.1710	
21	-0.5180	0.1900	
22	-0.5160	0.1920	
23	-0.5000	0.2080	
24	-0.4880	0.2200	
25	-0.4680	0.2400	
26	-0.4610	0.2470	
27	-0.4530	0.2550	
28	-0.4170	0.2910	
29	-0.3940	0.3140	
30	-0.3910	0.3170	
31	-0.3890	0.3190	
32	-0.3820	0.3260	
33	-0.3672	0.3408	
34	-0.3400	0.3680	
35	-0.3194	0.3886	
36	-0.2750	0.4330	
37	-0.2670	0.4410	
38	-0.2370	0.4710	
39	-0.2241	0.4839	
40	-0.1790	0.5290	
41	-0.1263	0.5817	
42	-0.0880	0.6200	

Table A. Biology EOCEP Ordered Item Map (Cont.)

Item Map Order	Original Item Parameter	Theta Parameter with RP67	
43	-0.0810	0.6270	
44	-0.0780	0.6300	
45	-0.0740	0.6340	
46	-0.0681	0.6399	
47	-0.0510	0.6570	
48	-0.0243	0.6837	
49	0.0216	0.7296	
50	0.0320	0.7400	
51	0.0400	0.7480	
52	0.0450	0.7530	
53	0.0530	0.7610	
54	0.0760	0.7840	
55	0.0840	0.7920	
56	0.1030	0.8110	
57	0.1050	0.8130	
58	0.1321	0.8401	
59	0.1350	0.8430	
60	0.1510	0.8590	
61	0.2122	0.9202	
62	0.2156	0.9236	**<=Level 3 Cut**
63	0.2280	0.9360	
64	0.2300	0.9380	
65	0.2420	0.9500	
66	0.2535	0.9615	
67	0.3110	1.0190	
68	0.3280	1.0360	
69	0.3288	1.0368	
70	0.3690	1.0770	
71	0.3700	1.0780	
72	0.3960	1.1040	
73	0.4090	1.1170	
74	0.4460	1.1540	
75	0.4510	1.1590	
76	0.4750	1.1830	
77	0.4781	1.1861	
78	0.4831	1.1911	
79	0.4850	1.1930	
80	0.4950	1.2030	
81	0.5458	1.2538	
82	0.5470	1.2550	
83	0.6330	1.3410	
84	0.6380	1.3460	
85	0.6400	1.3480	
86	0.6790	1.3870	
87	0.7110	1.4190	

Table A. Biology EOCEP Ordered Item Map (Cont.)

Item Map Order	Original Item Parameter	Theta Parameter with RP67
88	0.7420	1.4500
89	0.8130	1.5210
90	0.9840	1.6920