Using the Many-Facet Rasch Model to evaluate standard-setting judgments: Setting performance standards for Advanced Placement ® examinations

Pamela Kaliski, The College Board

Stefanie A. Wind & George Engelhard, Jr., Emory University

Deanna Morgan, The College Board

Barbara Plake, University of Nebraska, Faculty Emeriti

Rosemary Reshetar, The College Board

AERA April 15th, 2012, Vancouver, BC

CollegeBoard
inspiring minds

AP® Environmental Science Background

- Course and exam launched in 1998
- 5 AP score categories (requires 4 cut scores)
  - Cut scores initially established with a college comparability study
  - Recently, AP program is moving towards conducting panel-based standard setting in place of college comparability
- In June 2011, first AP standard setting was conducted

CollegeBoard
inspiring minds

Importance of gathering validity evidence for standard setting procedures

- Types of validity evidence
  - Procedural validity
  - Internal validity
  - External validity
- How have judgments from standard settings been evaluated?
  - Utilize g-theory (e.g., Brennan, 1995)
  - Many-facet Rasch model (Engelhard, 2011)

(Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kane, 2001)

Many Facet Rasch Model

- Focus is on standard setting judgments rather than scores a rater-mediated assessment

$$\ln [P_{nijk} / P_{nijk-1}] = \theta_n - \delta_i - \omega_j - \tau_k$$

- Variability in ratings is a function of specified facets (e.g., Panelist severity, Judged item difficulty, Judged average performance level)
- MFR Model provides:
  - Rating quality indices
  - Model-data fit
  - Display of the facets on a variable map

CollegeBoard inspiring minds

Purpose of study

Use the MFR Model to evaluate quality of standard setting judgments from AP Environmental Science standard setting, specifically:

1. What are the locations of the panelists, items, rounds, and performance standards on the construct being measured (i.e., AP Environmental Science)?
2. Do panelists characteristics of gender of level of course taught (high school or college) influence their conceptualization of the underlying construct?

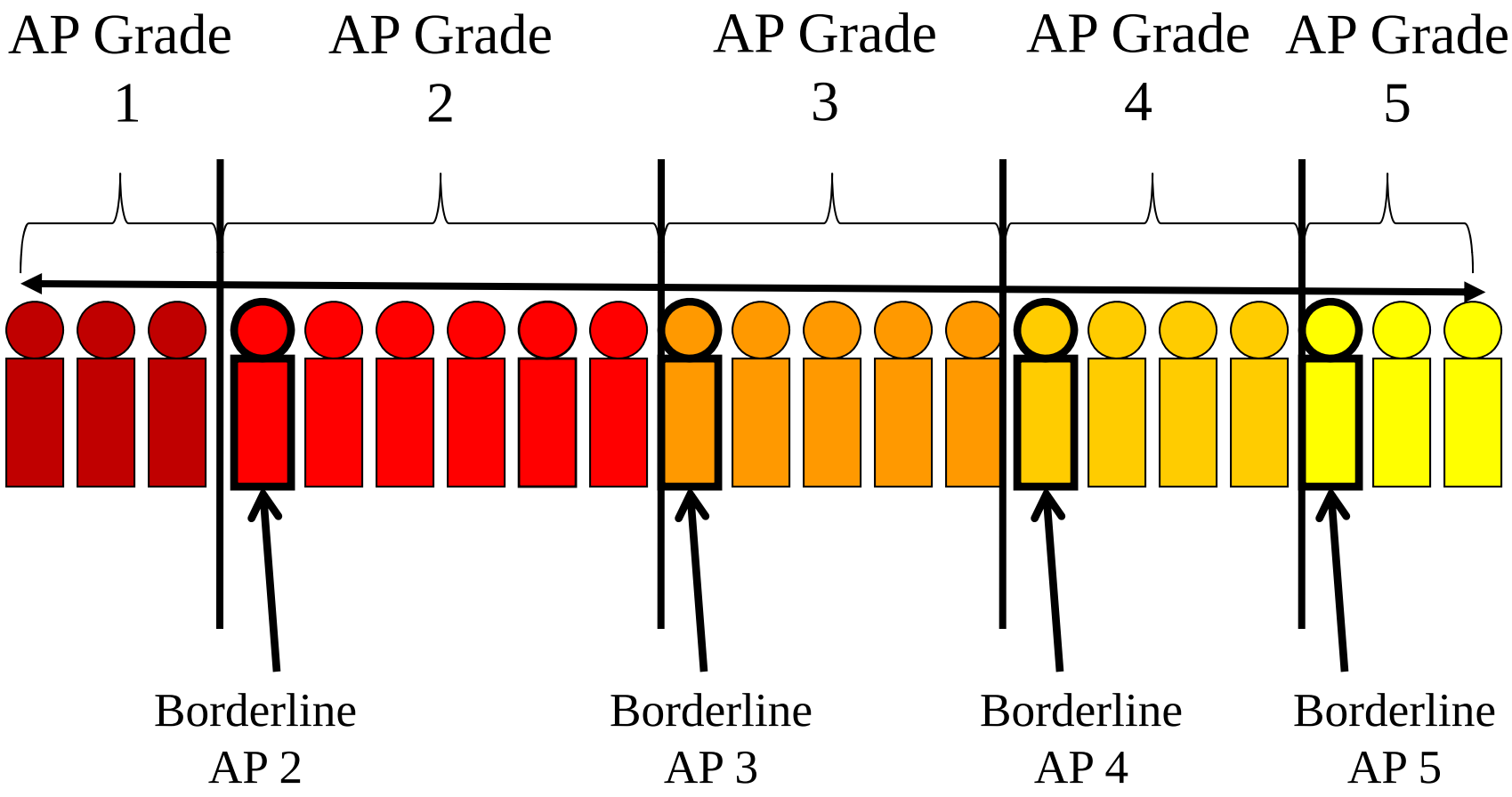
CollegeBoard
inspiring minds

Methods

- 15 panelists were Environmental Science SMEs
- APES exam: 100 MCQs, 4 FRQs
  - We only focused on MCQ item analyses
- Multiple Yes/No standard setting procedure
- Data Analysis
  - Applied MFR Model to analyze panelist judgments

CollegeBoard
inspiring minds

Borderline Examinees for APES



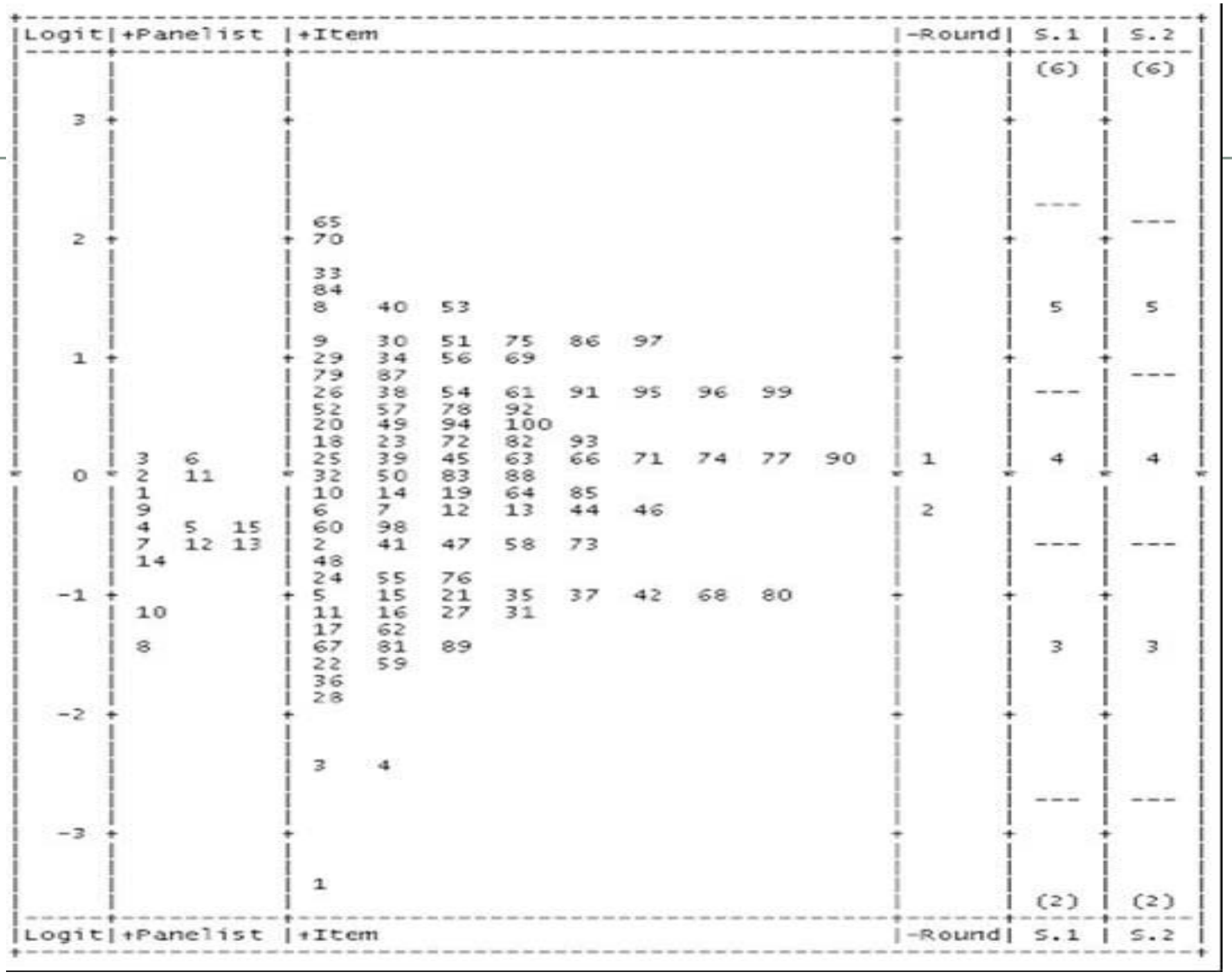
CollegeBoard
inspiring minds

Rating Task for Panelists

- Should the Borderline AP 2 Examinee answer the item correctly?
  - **Yes**, circle 1/2 on the rating form
  - **No**, Read the Borderline PLD for AP 3
- Should the Borderline AP 3 Examinee answer the item correctly?
  - **Yes**, circle 2/3 on the rating form
  - **No**, Read the Borderline PLD for AP 4

- **Question A** 1/2 (2/3) 3/4 4/5 Above 5 Cut

CollegeBoard inspiring minds



Logit	+Panelist	+Item	-Round	S.1	S.2
				(6)	(6)
3					

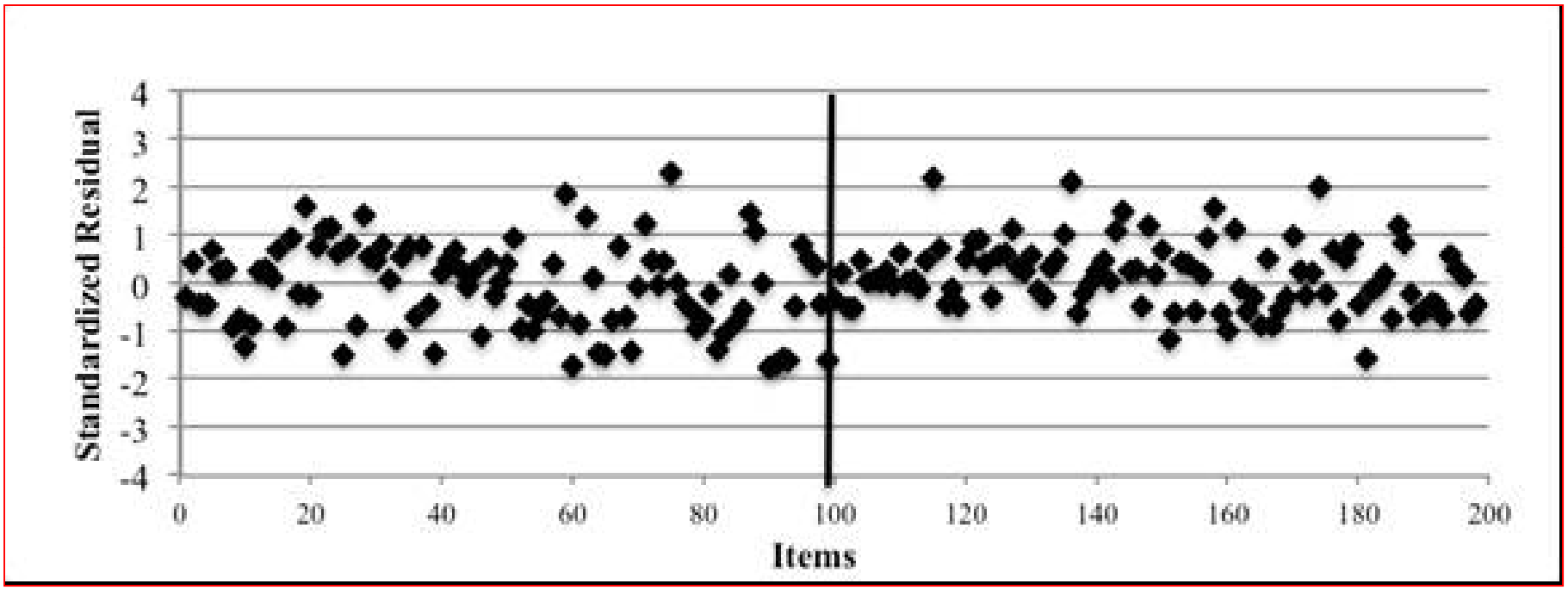
		65			---
2		70			
		33			
		84			
		8 40 53		5	5
		9 30 51 75 86 97			
1		29 34 56 69			---
		79 87			
		26 38 54 61 91 95 96 99		---	
		52 57 78 92			
		20 49 94 100			
		18 23 72 82 93			
	3 6	25 39 45 63 66 71 74 77 90	1	4	4
0	2 11	32 50 83 88			
	1	10 14 19 64 85			
	9	6 7 12 13 44 46	2		
	4 5 15	60 98			
	7 12 13	2 41 47 58 73		---	---
	14	48			
		24 55 76			
-1		5 15 21 35 37 42 68 80			
	10	11 16 27 31			
		17 62			
	8	67 81 89		3	3
		22 59			
		36			
		28			
-2					
		3 4			
				---	---
-3					
		1			
				(2)	(2)
Logit	+Panelist	+Item	-Round	S.1	S.2

Results Research Question 1: Panelists

Panelist	Panelist Severity Measure	Mean Rating	SEM	INFIT	OUTFIT
3	0.17	4.02	0.09	1.07	1.03
6	0.17	4.03	0.09	1.43	1.42
11	0.14	4.00	0.09	1.02	1.00
2	0.09	3.96	0.09	1.07	1.09
1	-0.13	3.82	0.09	0.73	0.71
9	-0.18	3.78	0.09	0.81	0.83
15	-0.29	3.71	0.09	0.74	0.77
4	-0.32	3.69	0.09	1.07	1.08
5	-0.35	3.67	0.09	1.38	1.35
12	-0.47	3.60	0.09	1.31	1.26
7	-0.50	3.58	0.09	0.84	0.84
13	-0.55	3.55	0.09	1.01	1.01
14	-0.58	3.53	0.09	0.73	0.81
10	-1.07	3.24	0.10	0.86	0.83
8	-1.38	3.08	0.10	0.65	0.67

CollegeBoard
inspiring minds

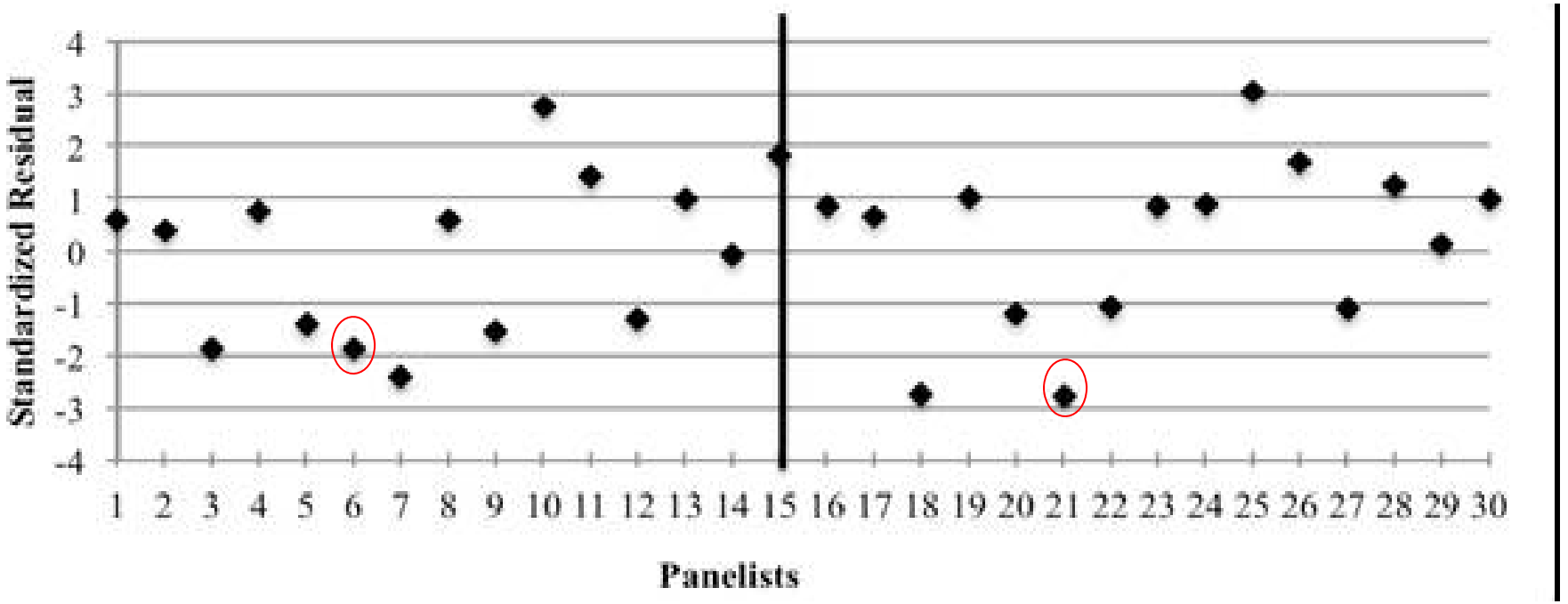
Panelist 8 residual plot



Results Research Question 1: Items

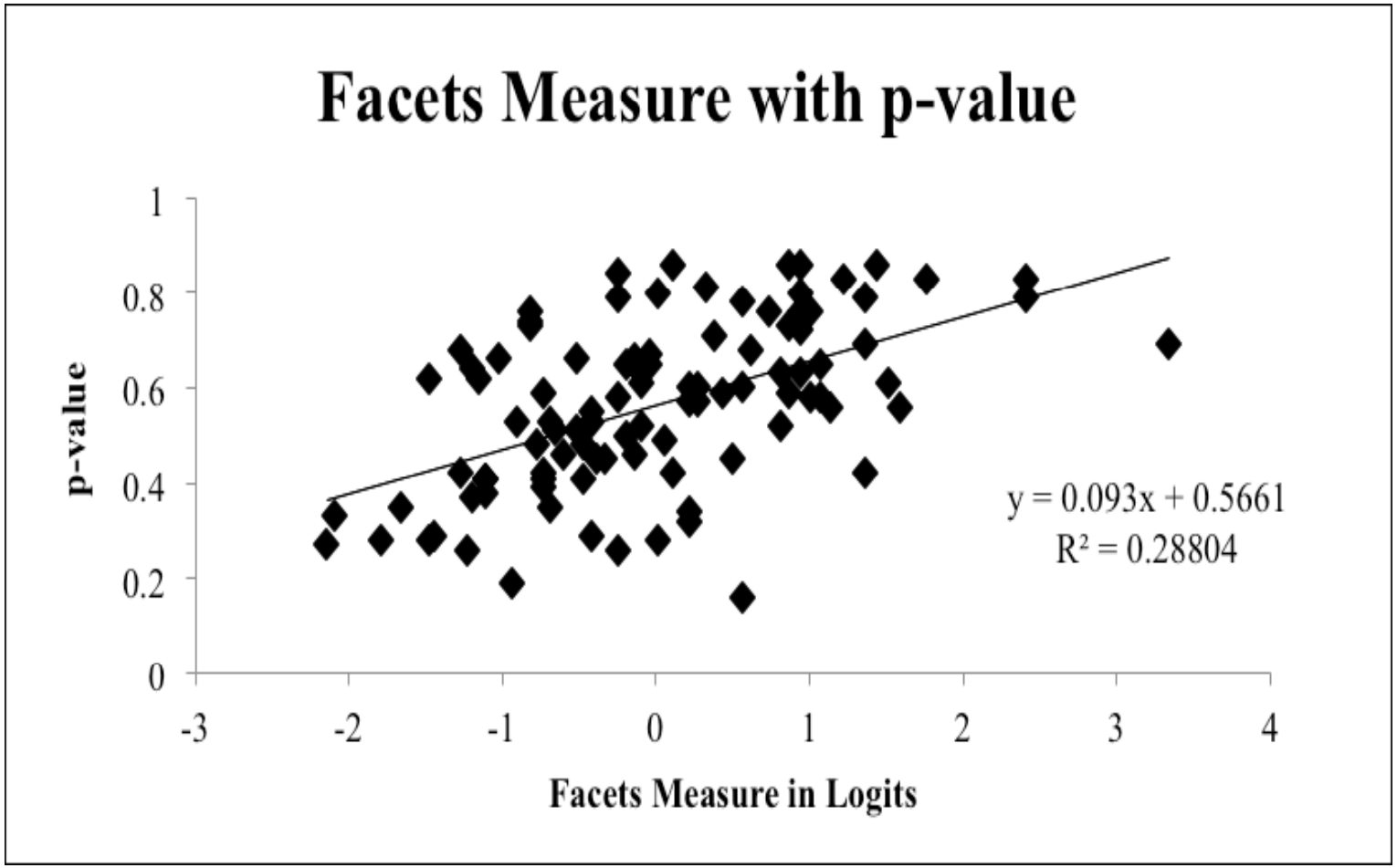
Item	Item Difficulty Measure	Mean Rating	S.E.	Infit MSE	Outfit MSE
65	2.14	5.23	0.23	0.91	0.85
70	2.09	5.19	0.23	0.76	0.75
33	1.79	4.98	0.22	0.91	0.91
84	1.66	4.88	0.21	0.73	0.72
40	1.48	4.74	0.21	0.90	0.92
53	1.48	4.74	0.21	0.57	0.58
8	1.44	4.70	0.21	1.08	1.10
51	1.27	4.56	0.20	0.76	0.74
86	1.27	4.56	0.20	0.30	0.30
75	1.23	4.53	0.20	0.83	0.82
9	1.19	4.49	0.20	0.98	0.97

Item 96 Residual Plot



CollegeBoard
inspiring minds

Relationship between observed and judged item difficulties



$r = 0.54$

CollegeBoard
inspiring minds

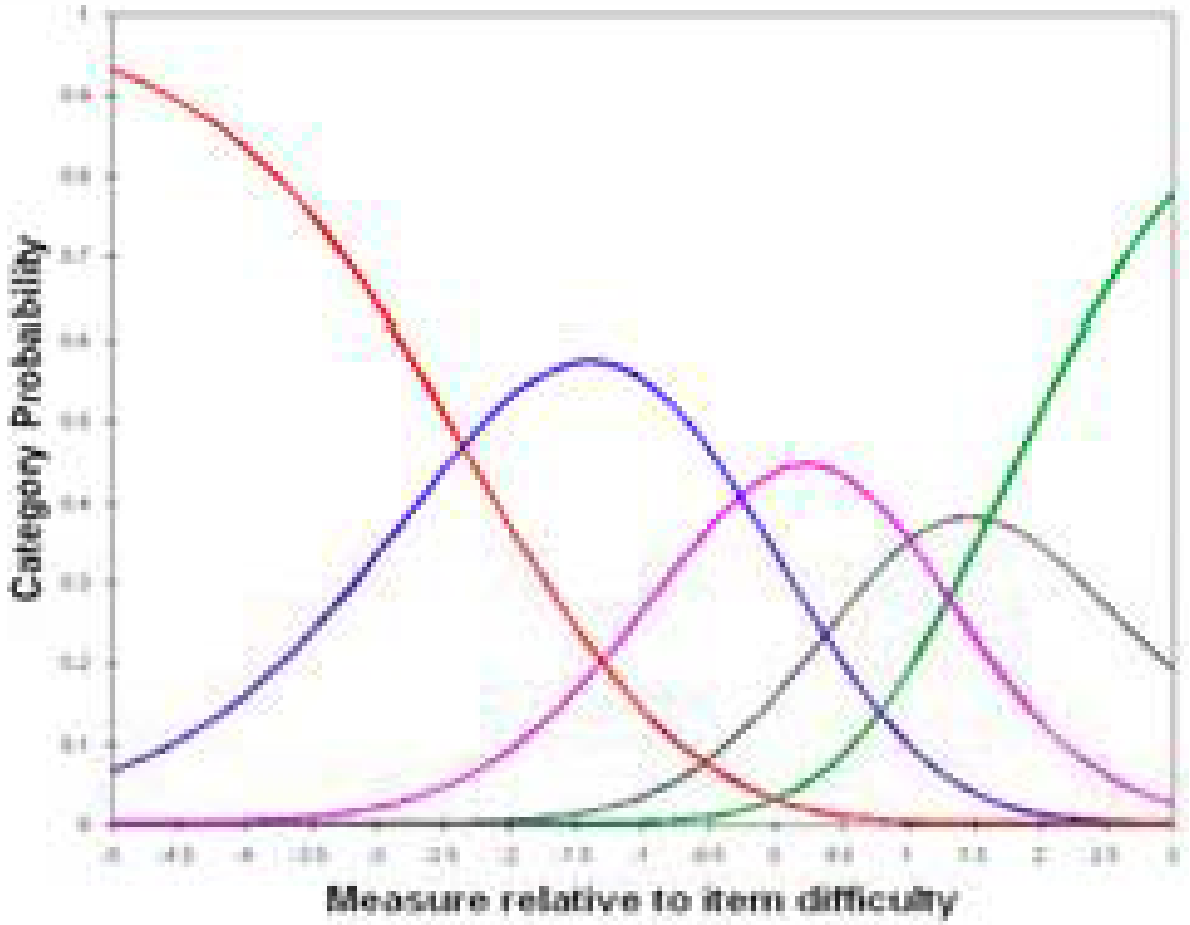
Results Research Question 1: Rounds

Round	Measure	S.E.	Infit MSE	Outfit MSE
1	0.16	0.03	1.07	1.06
2	-0.16	0.03	0.91	0.91
Mean	*0.00*	*0.03*	*0.99*	*0.98*
SD	*0.03*	*0.00*	*0.08*	*0.07*

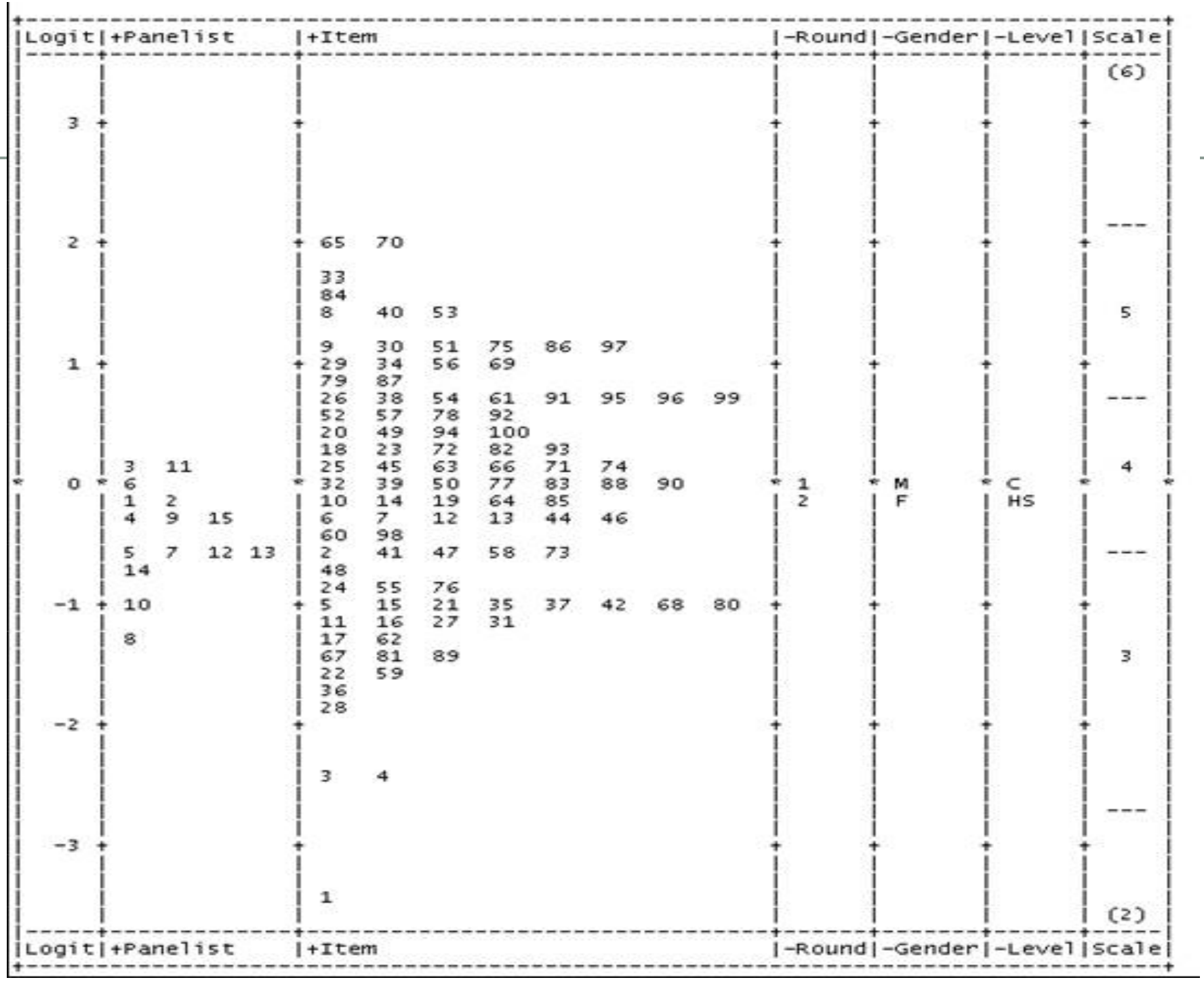
Results Research Question 1: Performance Standards

Category	Count	Percentage	Mean	OUTFIT	Rasch Threshold	S.E.
			Round 1			
Above 5	71	5	0.69	1.20	1.75	0.13
4/5	212	14	0.51	0.80	0.86	0.08
3/4	437	29	-0.22	1.10	-0.24	0.06
2/3	571	38	-0.89	1.00	-2.37	0.09
1/2	194	13	-1.60	1.10		
			Round 2			
Above 5	119	8	1.19	0.90	1.50	0.11
4/5	222	15	0.76	0.70	1.14	0.07
3/4	494	33	-0.01	0.90	-0.29	0.06
2/3	505	34	-0.75	0.90	-2.34	0.10
1/2	145	10	-1.45	1.10		

Category Characteristic Functions



CollegeBoard
inspiring minds


Logit | +Panelist | +Item | -Round | -Gender | -Level | Scale
(6)
3
2
65 70
33
84
8 40 53
5
9 30 51 75 86 97
1
29 34 56 69
79 87
26 38 54 61 91 95 96 99
52 57 78 92
20 49 94 100
18 23 72 82 93
3 11
25 45 63 66 71 74
4
0
6
32 39 50 77 83 88 90
1
M
C
1 2
10 14 19 64 85
2
F
HS
4 9 15
6 7 12 13 44 46
60 98
5 7 12 13
2 41 47 58 73
14
48
24 55 76
-1
10
5 15 21 35 37 42 68 80
11 16 27 31
8
17 62
67 81 89
3
22 59
36
28
-2
3 4
-3
1
(2)
Logit | +Panelist | +Item | -Round | -Gender | -Level | Scale

Research Question 2: Gender Facet

Gender	Measure	SEM	INFIT	OUTFIT
Males	0.05	0.03	0.88	0.88
Females	-0.05	0.04	1.14	1.13
Mean	*0.00*	*0.03*	*1.01*	*1.01*
SD	*0.05*	*0.00*	*0.13*	*0.13*

Research Question 2: Level of Course Taught

Level	Measure	SE	Infit MSE	Outfit MSE
College	0.05	0.03	0.90	0.90
High School	-0.05	0.04	1.11	1.11
Mean	*0.00*	*0.03*	*1.01*	*1.00*
SD	*0.05*	*0.00*	*0.10*	*0.11*

CollegeBoard
inspiring minds

Discussion

- Benefits of utilizing MFR Model for evaluating standard setting ratings:
  - Holistic depiction on variable map
  - Panelist and item-specific residuals
  - Can incorporate explanatory variables
  - Validity evidence, both internal and procedural
- Provided evidence of acceptable quality of ratings

THANK YOU!

CollegeBoard
inspiring minds