Validating Performance Level Descriptors (PLDs) for the AP® Environmental Science Exam

Rosemary Reshetar, Pamela Kaliski, Michael Chajewski, Karen Lionberger
The College Board

Barbara Plake
University of Nebraska–Lincoln, Emeritus

ITC Conference, July 5, 2012

CollegeBoard
inspiring minds™

Background: Performance Level Descriptors

- Assessment programs that classify test-takers into performance categories
- PLDs are used to communicate the meaning of scores in each category
- PLDs also serve a role in setting cut scores

Background: Advanced Placement® (AP®) Exams

The 34 AP Courses and Exams

Course	Exam
•College level	•Summative
•Taken in High School	•3 Hour length
•Full year	•Mixed format (MC & CR)
	•Scores 1 – 5

Background: Environmental Science

- Over 100,000 test takers
- Over 4,000 schools

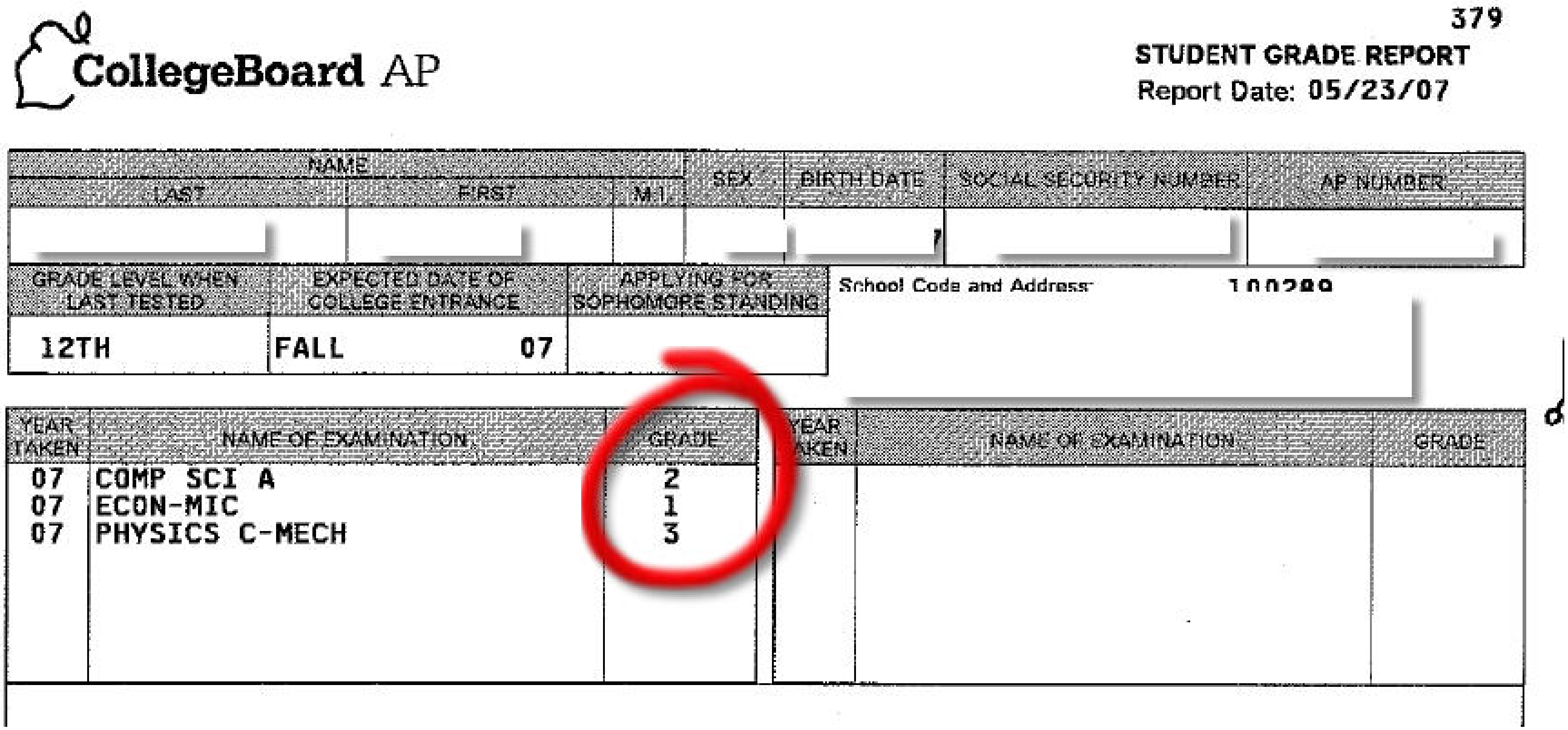
The Exam: Two Forms	
Multiple Choice	**Constructed Response**
100 Items	4 Items, 10 points each
90 Minutes	90 Minutes
50% Score	50% Score

CollegeBoard
inspiring minds™

Focus of this Study

- Explore analytical methods and systematic procedures for PLD validation studies in the AP context
- Challenge: Developing procedures for analytically scored constructed response items
- Apply PLD validation procedures to the AP Environmental Science Exam

AP Scores 1 through 5



CollegeBoard AP

379

STUDENT GRADE REPORT

Report Date: 05/23/07

NAME LAST	 FIRST	 M.I.	SEX	BIRTH DATE	SOCIAL SECURITY NUMBER	AP NUMBER

GRADE LEVEL WHEN LAST TESTED	EXPECTED DATE OF COLLEGE ENTRANCE	APPLYING FOR SOPHOMORE STANDING
12TH	FALL 07	

School Code and Address: 100289

YEAR TAKEN	NAME OF EXAMINATION	GRADE	YEAR TAKEN	NAME OF EXAMINATION	GRADE
07	COMP SCI A	2			
07	ECON-MIC	1			
07	PHYSICS C-MECH	3			

CollegeBoard
inspiring minds™

Definition of Score Categories

- 5: Extremely well qualified;
- 4: Well qualified;
- 3: Qualified;
- 2: Possibly qualified; and
- 1: No recommendation.

What do students at each level know?
What can students at each level do?

CollegeBoard
inspiring minds™

Background of AP Environmental Science PLDs

- Panel of experts created PLDs March 2011
- For each AP score
  - Nine cognitive processes (1–9),
  - Six quantitative skills (10–15), and
  - Three scientific processes (16–18)
  - Example: AP Score Level 4, Cognitive Process 3

    Elaborates about how complex systems function – for example, describing tropospheric ozone formation with some connections to combustion, temperature, and sunlight.

AP Environmental Science Procedures

- A separate panel participated in standard setting June 2011
- Standards set on the 2011 operational form and applied to the other forms
- PLD validation study conducted in March 2012
    - 1 ½ Day working meeting
    - 5 Environmental Science Subject Matter Experts
        - All faculty at 4-year colleges

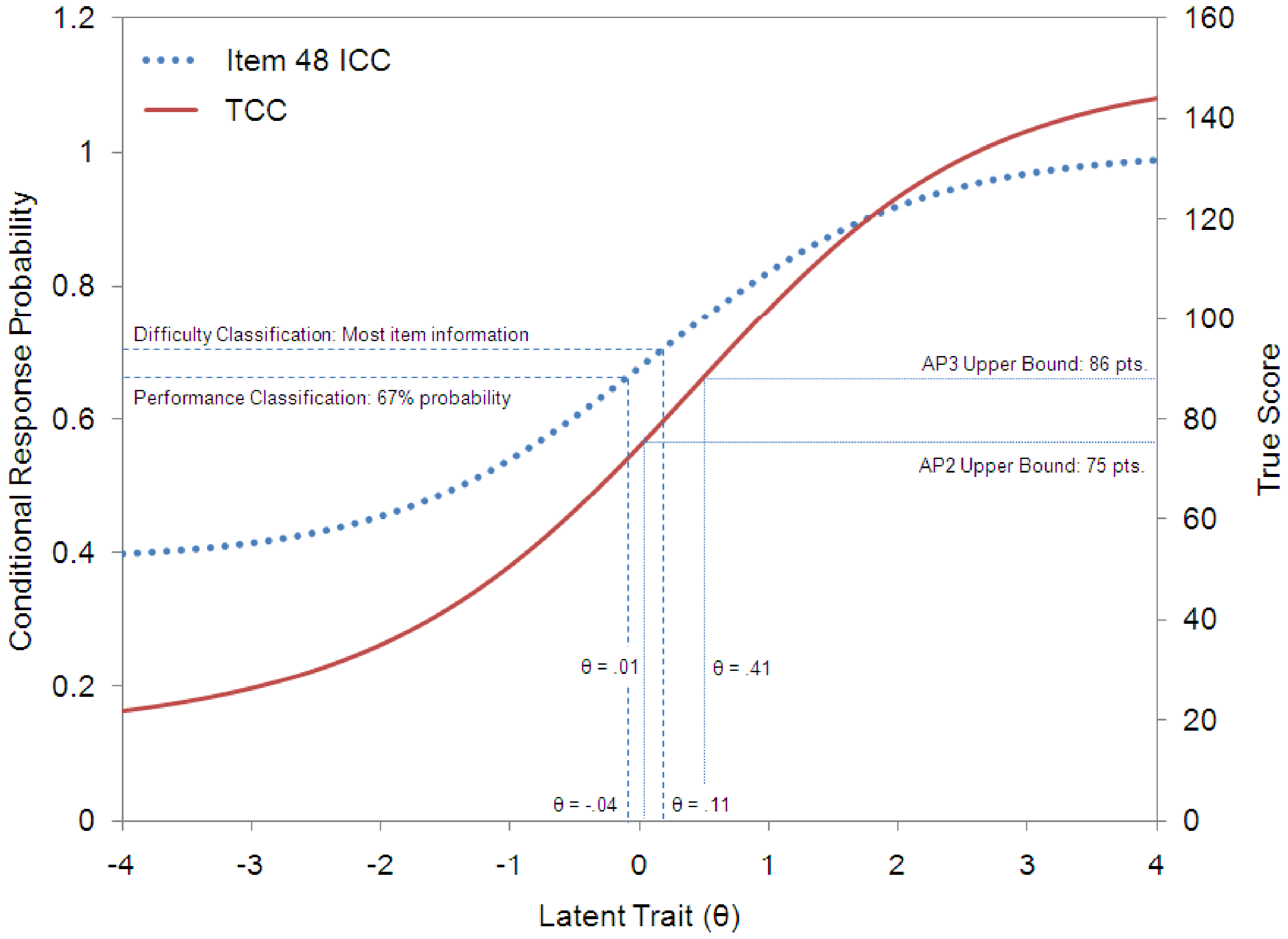
PLD Validation Steps

- Items classified into categories 1 through 5
- Item bundles for each category presented to panelists for review
  - 15 MC items selected per category
  - CR score points & rubrics for 4 CR items
- Panelists noted if item matched PLD level or not
  - If matched, which descriptor(s)?
  - If not, what PLD level is correct?

Methods: Items Classified into Categories

- Theta ranges establish for AP Scores:
  - **1:** -4.00 to -.75
  - **2:** -.74 to .01
  - **3:** .02 to .41
  - **4:** .42 to 1.22
  - **5:** 1.23 to 4.00
- IRT calibrations for items with Multilog
  - 3 P-L for MC items
  - GPC for CR items
- Items parameters from both forms on same scale

MC: IRT Location & Performance 67% Probability

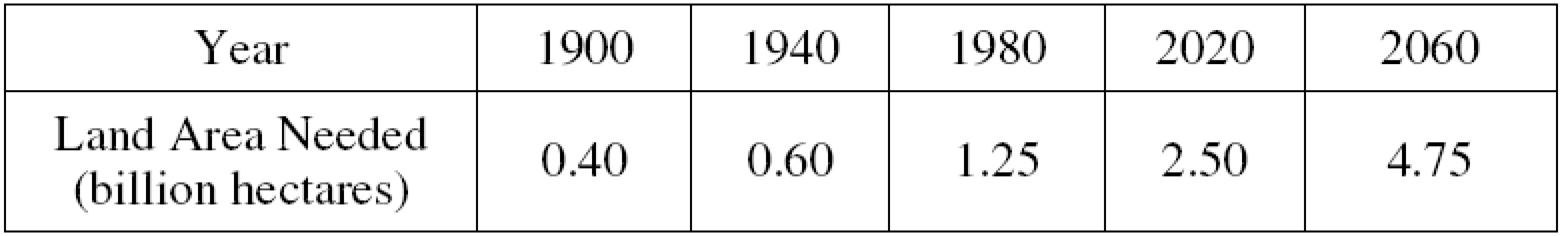


Methods: CR Item points classification

- CR items analytically scored (add example)
- For each level:
  - **Number of points** out of 11 (0 – 10)
  - **Which points** most likely comprise the number of points
- IRT location method for CR items (65% of the item’s value)
- Step function constructed to reflect region across latent trait scale for specific score’s highest response probability

Item 4

As the world’s population increases and availability of new arable land decreases, providing sufficient food for the world’s human population is becoming increasingly difficult. The table below shows the area of land needed to feed the world’s population from 1900 projected to the year 2060.

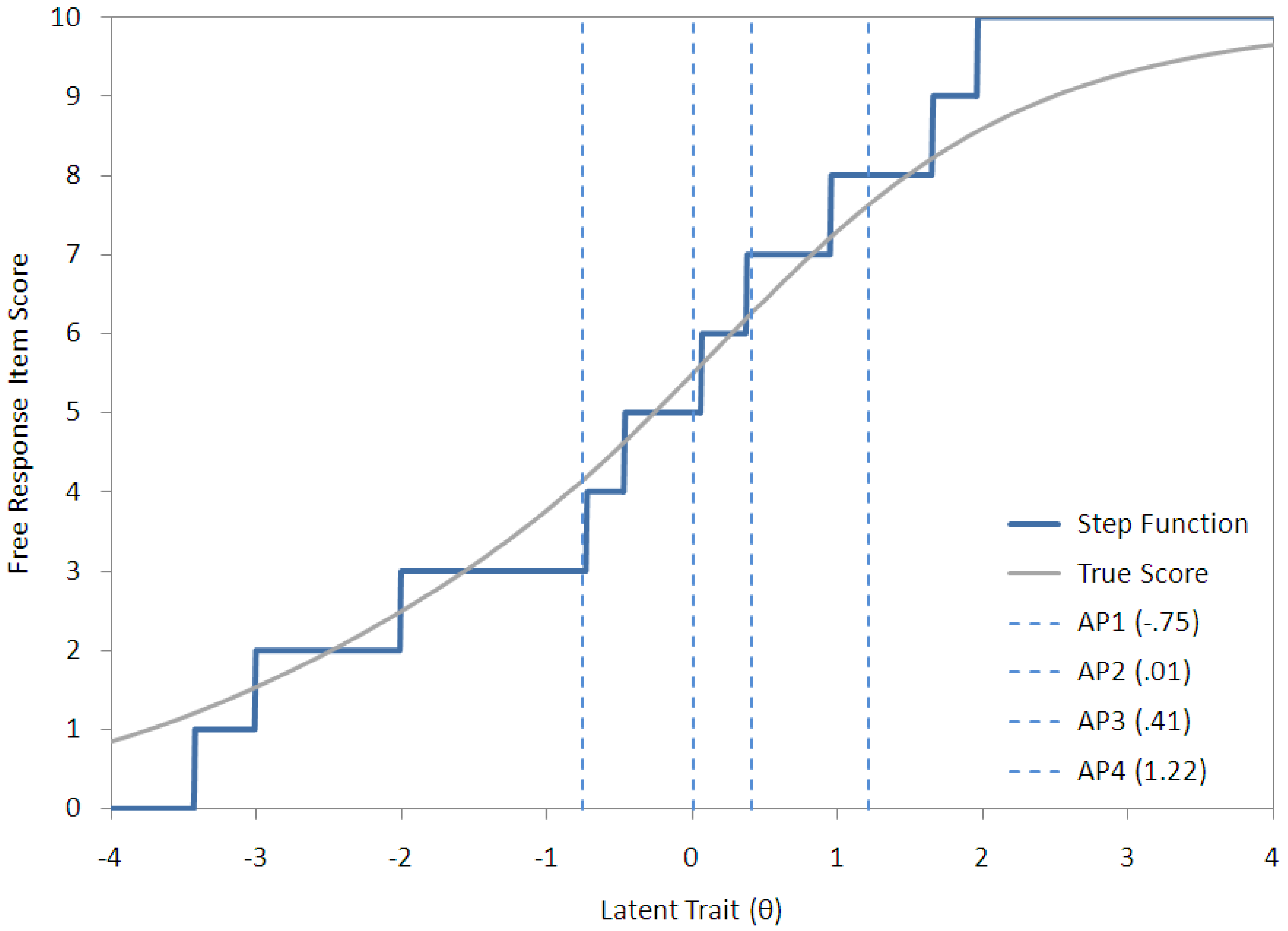


Year	1900	1940	1980	2020	2060
Land Area Needed (billion hectares)	0.40	0.60	1.25	2.50	4.75

Item 4 Scoring Breakdown (Max 11 points)

Segment/Points	Description
4.a – 2 points	Plot data on graph & draw smooth curve.
4.b – 1 point	Determine the year in which the human population is likely to run out of arable land for agriculture.
4.c – 4 points	Identify TWO physical and/or chemical properties of soils and describe the role of each property in determining soil quality.
4.d – 2 points	Describe TWO viable strategies for reducing the amount of land needed for agriculture.
4.e.i – 1 point	Describe how salinization occurs.
4.e.ii – 1 point	Describe one method to prevent or remediate soil salinization.

CR Item 4 Step Function



Methods: CR Item points accumulation

- Which points most likely comprise the number of points?
- Sample of responses for each item rescored
- For each response, which points were received
- Matrix of total score (0-10) by specific points analyzed to develop most likely points
- This merged with step function information to create categorization for panelists to review

Ex. Item 4 Information for Review

AP Level	Rubric Category	Points
5	4.c 4.d 4.e.ii	4th 2nd
4	4.c	3rd
3	4.d	1st
2	4.a 4.c	2nd 1st & 2nd
1	4.a 4.b	1st 1st

Note: 4.e.i, beyond cat 5

Panelists' Review Results

- Majority of cases: Raters believed the item classifications into the PLD categories were accurate
- Raters agreed on the direction of changes

Panelists’ Review Results

MC Items (15 per category)		
Category	**No. of Items “No”**	**Comments**
5	4	Should be 4 or 3
4	1	Should be 3
3	7	Should be 4
2	9	Most should be 3, a couple 4
1	9	Should be 2
“No” if 3 or more raters selected No		

CR Items

•1 segment in PLD 2 and 2 in PLD 1 rated “No”

Results: Content Review

- Difficulty in applying descriptors that included language that assumes an inherent understanding of student performance over time rather than in one instance on one question.
  - E.g., “struggles to” or “inconsistently uses”
- Some PLD statements are difficult to apply to MC questions
  - E.g., “discriminate among a variety of information sources as to their scientific and scholarly validity”.

CollegeBoard
inspiring minds™

Results: Content Review

- Some questions cut across different PLD categories
    - E.g., basic quantitative skill with more complex conceptual topic
- Some panelists may have made ratings based only on the content in the stem and not assessing the difficulty added by the item’s distracters.
    - Some distracters are included to represent common misconceptions

Results: Content Review

- Too many Cognitive PLDs (8) and not differentiable
- Quantitative Skills PLDs 10 and 11
    - Trouble with differentiating levels 3 through 5

2nd Task, Assigning MC Items to PLDs

- 2 items in each PLD category based on stats
- Ratings tended toward the center
  - Items in categories 1 and 2 rated higher
  - Items in categories 3 and 4 rated lower

Observations and Summary

- Having panelists familiar with PLDs was useful
- Panelists had difficulty distinguishing this task from a typical standard setting task that asks, “Could students at this level answer correctly?”
- Need more work with panelists’ blind assignment of items to PLDs
- Challenging to incorporate ratings into PLD modifications; need iterative process
- Group discussion extremely helpful

CollegeBoard
inspiring minds™

Upcoming plans

- Slight modifications to statistical procedures for selection
- Implementation in 2012 with new courses and exams
    - World Language & Culture Exams in French, German and Italian
        - Spoken and written responses
        - Holistic scoring
    - World History
        - Essay items

CollegeBoard
inspiring minds