Setting Performance Standards

Texas Assessment of Knowledge & Skills
(TAKS)

August - September, 2002

Session Overview - Day 1

I. What is "standard setting" ?

II. Describe the 3 performance "categories;" refine the performance descriptors

III. Review & discuss actual TAKS test; map test content onto the categories

IV. "Item Mapping" procedure & practice

V. Round 1 of independent judgments

Session Overview -- Day 2

I. Review Day 1 -- questions & issues

II. Feedback & discussion of Round 1; student performance data

III. Round 2 ratings -- reconsider Round 1

IV. Round 2 feedback & discussion; statewide implications of the cuts

V. Final ratings & debriefing questionnaire

Session Overview -- Day 2

I. Review Day 1 -- questions & issues

II. Feedback & discussion of Round 1; student performance data

III. Round 2 ratings -- reconsider Round 1

IV. Round 2 feedback & discussion; examine statewide implications of cuts

V. Final ratings & debriefing questionnaire

VI. Review 2nd test; Round 1 ratings on the *2nd* test

Session Overview -- Day 3

I. Review *final* recommendations for Test 1

II. Review first round results for Test 2

III. Discuss Round 1 Ratings; student data

IV. Make Round 2 of Ratings

V. Discuss Round 2 Ratings; impact data

VI. Final recommendations & evaluation

Setting Performance Standards

Who's involved? TEA & contractor roles

Why BETA? Who's moderating? Role ?
Not content experts, but facilitators

Why you? -- individually & collectively:

You are the ***experts***.

You ***represent*** various groups.

You are ***judges***, not psychometricians.

You are ***advisors***, not policy makers.

Groundrules

CONFIDENTIALITY

+

NO DISCUSSIONS **about the TAKS program**

OR

- **why the state is setting standards**

- **the philosophy of educational assessment**

- **the TEKS curriculum standards**

All discussions should be *as a group.*

What *IS* Standard Setting?

- **another frame of reference for interpreting test scores ("how good is *good*"?)**
- **for teachers, a routine, daily activity**

- **true "criterion-referencing"**
- **a semi-quantitative, semi-standardized, socio-political judgment process**

- **NOT a science!**

4 Keys to Being a Great Judge:

1. **Judgments** ***vs.*** **Data**

2. **"Should" *vs.* "Will"**

3. **Consider *ALL* Texas students who take the TAKS**

4. **Think of *threshold* students, *not* all who met the standard**

“Competence”

Low **High**

Low **?????** **High**

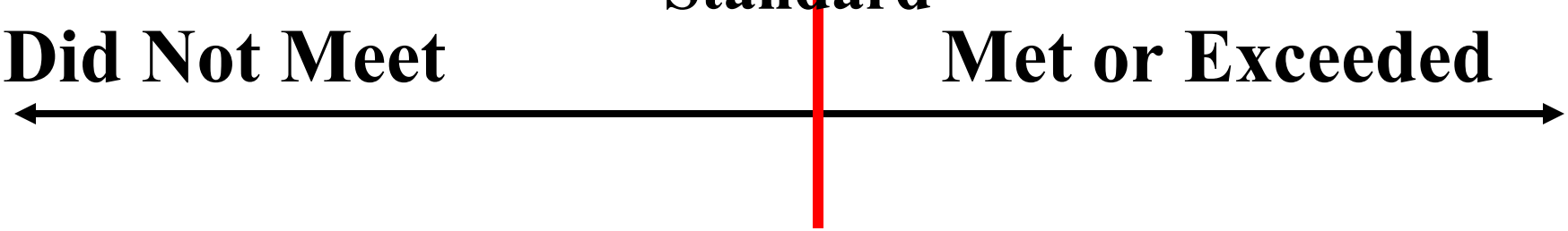
"Met the Standard" on TAKS

Below **? ? ? ? ? ?** **Above**

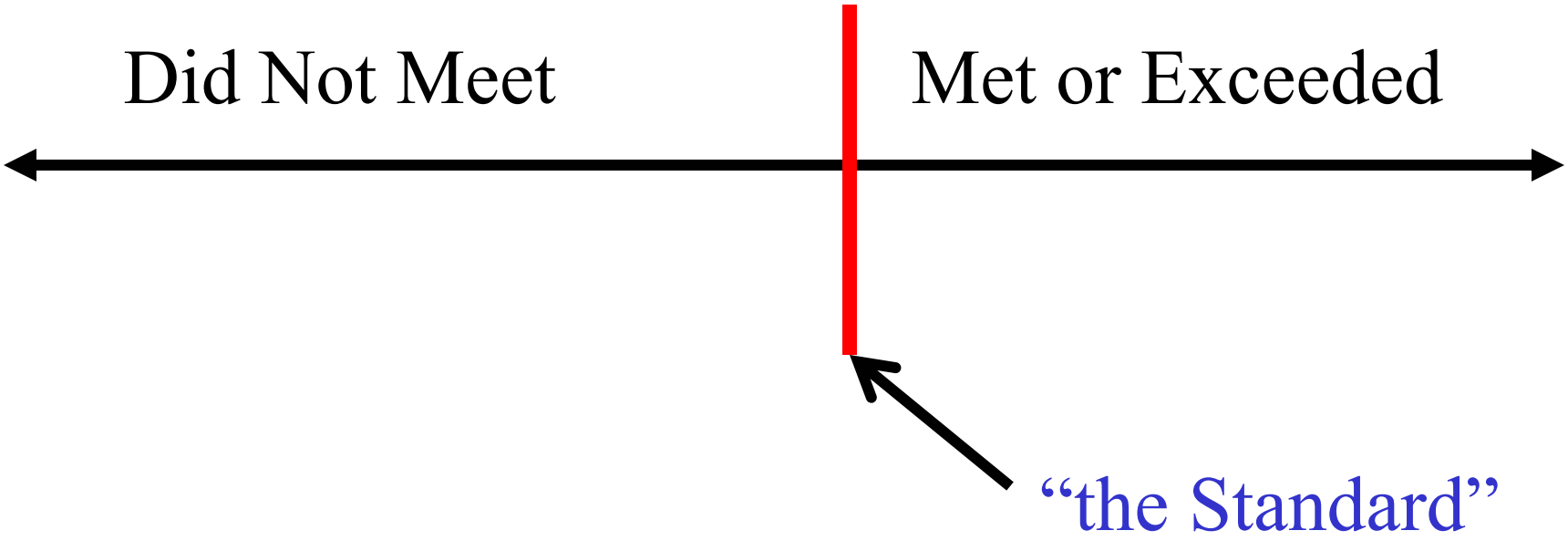
Your Task:

Standard

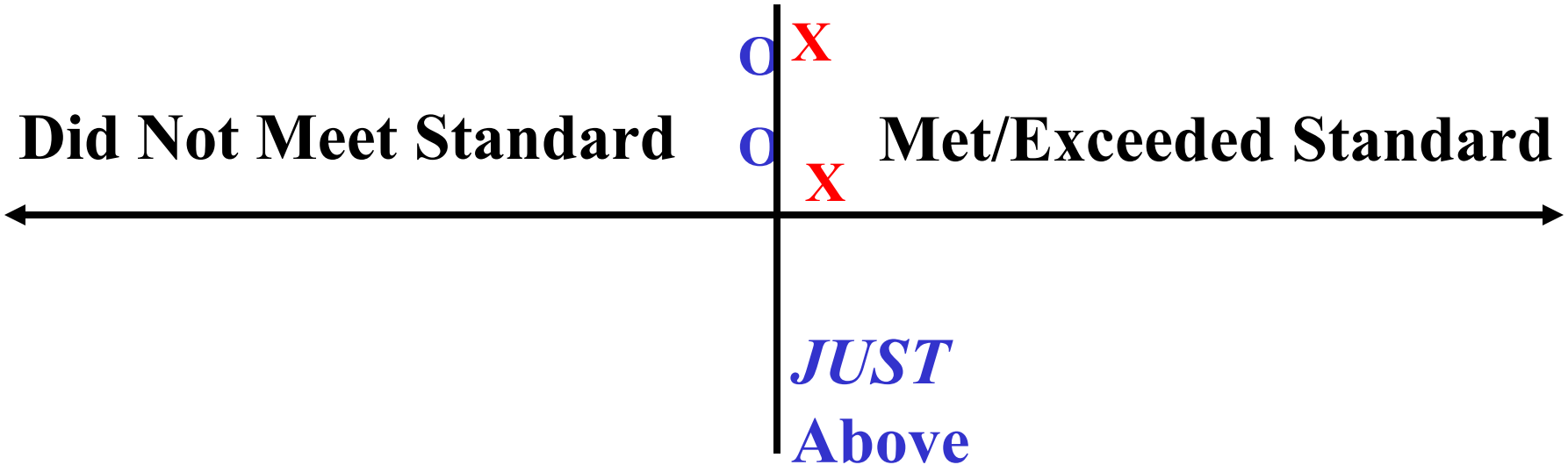
Did Not Meet **Met or Exceeded**



TAKS Performance Standard



“Met the Standard”



Are the Xs *really* better than the Os ??

Your Task:

From:

Competence

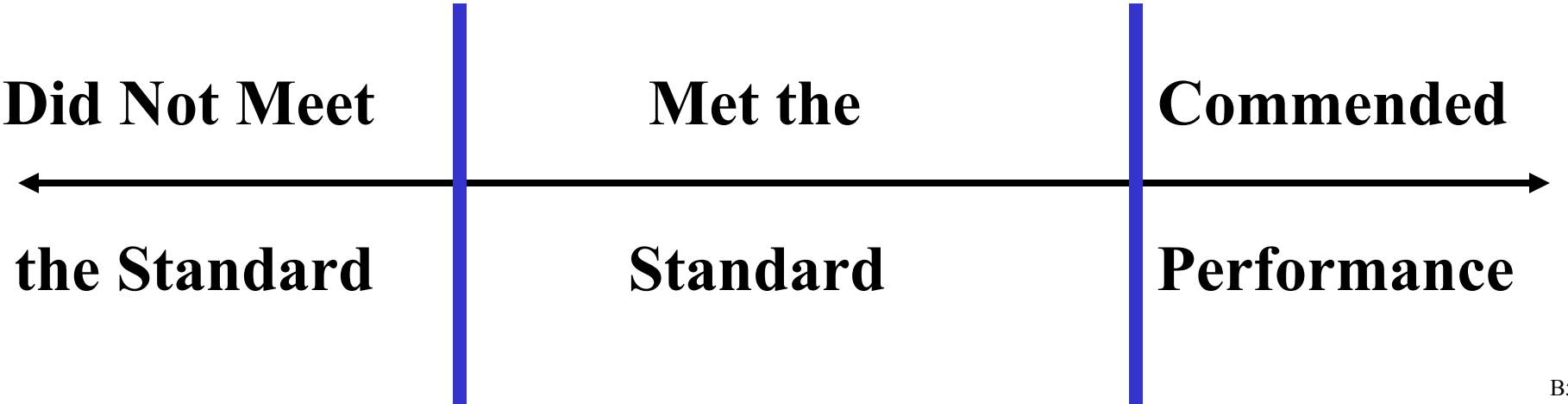
Low **High**

To:

Did Not Meet the Standard

Met the Standard

Commended Performance



Problem:

What do these *general* performance descriptions mean *concretely* in YOUR subject-matter area & at your grade ?

Labels Used for TAKS

1. **Did Not Meet** the Standard
2. **Met** the Standard
3. **Commended** Performance

Did Not Meet the Performance Standard

- ***Unsatisfactory*** **academic achievement**

- **Students performed at a level that was *below* the state passing standard.**

- **Students demonstrated an *insufficient* understanding of the knowledge and skills measured at this grade.**

Met the Performance Standard

- ***Satisfactory*** **academic achievement**

- **Student performed at a level that was *at or somewhat above* the state passing standard.**

- **Students demonstrated a *sufficient* understanding of the knowledge and skills measured at this grade.**

Commended Performance

- ***High*** **academic achievement**

- **Students performed at a level that was *considerably above* the state passing standard.**

- **Students demonstrated a *thorough* understanding of the knowledge and skills measured at this grade.**

Summary of the "Labels"

Did Not Meet the Standard --

"*Unsatisfactory* achievement" **"*Insufficient* understanding"**

Met the Performance Standard --

"*Satisfactory* achievement" **"*Sufficient* understanding"**

Commended Performance --

"*High* achievement" **"*Thorough* understanding"**

Focus on the two "cuts"!

Describe Concretely Students Who "Met the Standard" & "Commended Performance" Students

- **What can they *do*? *Not* do?**
- **What TEKS *skills* do they possess?**
- **What do they *know*?**
- **What *behaviors* demonstrate that they "met standard" or were "commendable"?**

"Experience" the Test !

Why? Standards are being set on the TAKS, not in general

What to do? "Be" a student
Think about each question

Think about: Skill(s) / behaviors being tapped
"Met the Standard" / "Commended"
"Threshold" students

ASK: **SHOULD a student who JUST**
Met the Standard answer this correctly?

"Experience" the Test !

Why? Standards are for TAKS, not in general

What to do? "Be" a student
Think about each question
Rough out answers to open-ended Qs

Think about: Skill(s) / behaviors being tapped
"Met the Standard" / "Commended"
"Threshold" students

ASK: **SHOULD students who JUST Met the Standard answer this correctly (answer this well) ?**

"Item Mapping" Procedure

- **"Invented" as the *Bookmark Method***
- **Has been used in over 25 states ("validity by application")**
- **Has both positive and negative features**
- **Just *another* way to quantify judgments**

Item Mapping - How Does It Work?

- **Arrange all test items in a "book" from the easiest to most-difficult.**

- **Consider the book of items from front to back. Ask for each item: Should threshold students who *minimally* Met the Standard be able to answer this question correctly?**

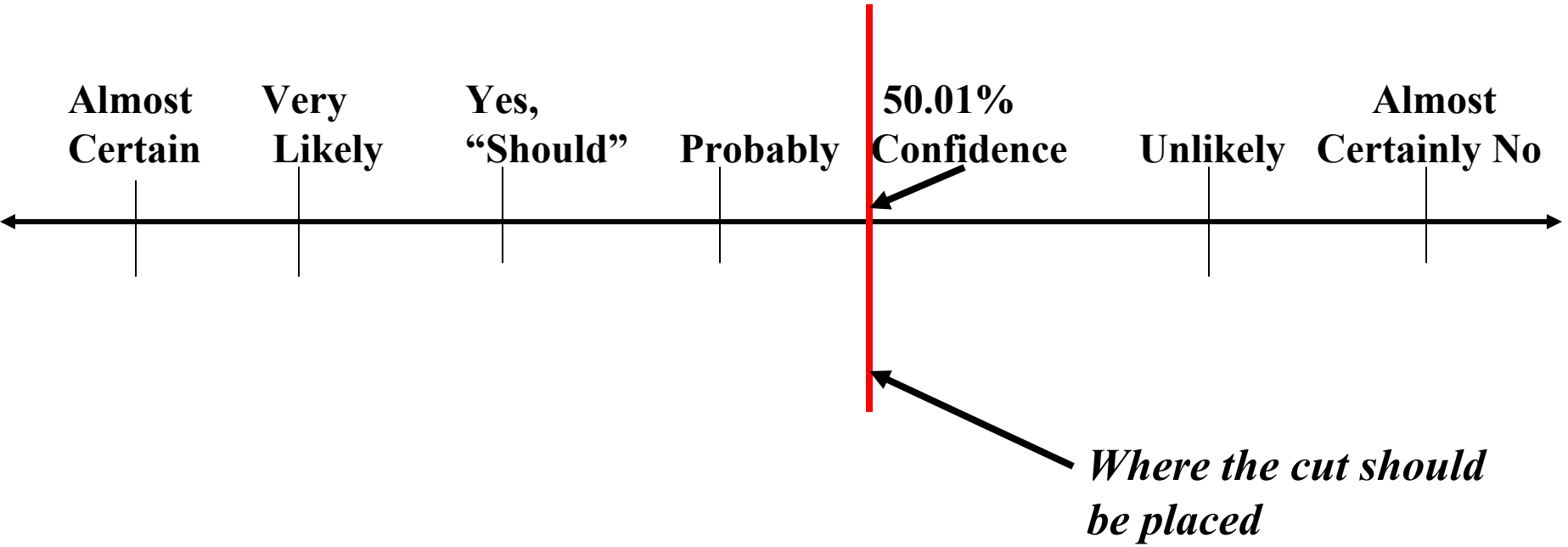
- **"Bookmark" the item prior to the first item for which you answer "no."**

- **Read on to the point separating *Met the Standard* from *Commended*; place another bookmark.**

- **The bookmarks define the two cuts you recommend.**

Degree of Confidence that a "Threshold" Student Should Answer an Item Correctly

Test Booklet Item Position:



Open-ended ("Multipoint") Items

Each such item counts multiple points. Since each page of the book = 1 point :

- **These items appear in the book as many times as they have score points.**

- **An item scored 0 - 4 appears 4 times -- once for each point it contributes to the total score.**

Item Mapping - What Judges Do

- **Read each page. Consider the content assessed. Think about the answer.**

- **Decide: *Should* threshold students who minimally Met the Standard answer this question correctly?**

    - **If YES, read on; if ??, slow up; if NO, "bookmark" the *preceding* item.**

- **Move to the Commended threshold; place a 2nd bookmark.**

- ***Suggestion*: Mark off "zones" first; then "revisit the neighborhoods" to set the cuts**

Maybe an Item Seems "Misplaced"

Difficulty -- **Too hard or Too easy**
What does this judgment mean?

Appropriateness -- **fit the TEKS**
fit what you teach or think *should* be taught.

What to <u>DO?</u> Work holistically

Don't place a bookmark based on one item.

Practice Activity

Think about:

- the **item** - What it measures, intentional or not
- the **curriculum** - Is this taught? Will it be?

- the descriptions of the 3 categories

- **THRESHOLD students; ALL students; "SHOULD"**

Jot down notes, impressions, questions, reactions.

Practice Exercise: How to Do

Consider each item -- what's tested, how hard, *should* students be able to answer correctly?

Bookmark two points:

M = *first* item that students who just MET the standard should probably answer correctly, but those *just below* should not

C = *first* item that threshold COMMENDED students should probably answer correctly, but those *just below* Commended should not

Advice on Setting Goals

- **Set demanding, but attainable standards**
- **What "should be" probably shouldn't disregard what "is"**
- **Focus on the concrete -- students, behaviors, instruction**
- **Item difficulty doesn't reside in the item "stem"**

Use your best judgment !!

Who was the 7th president of the United States?

A. Andrew Jackson

B. Davy Crockett

C. George Bush

D. Elvis Presley

Who was the 7th president of the United States?

A. Andrew Jackson

B. George Washington

C. John F. Kennedy

D. Lyndon B. Johnson

Who was the 7th president of the United States?

A. Andrew Jackson

B. John Quincy Adams

C. Abraham Lincoln

D. James Madison

"Rules" for Ratings

- **Anonymity**
- **Independence**
- **Consider the *answer options***
- **Don't perseverate -- Make a best guess**
- **Find the "neighborhoods," then refine cuts**

"Rules" for Ratings

- **Anonymity**
- **Independence**

- ***Mult-Choice*: consider the answer options**
- ***Open-Ended*: consider responses, not the Q**

- **Don't perseverate -- Make a best guess**

- **Find the "neighborhoods," then refine cuts**

Keep in Mind :

- **Skill(s) being assessed**
- **The TEKS curriculum & actual instruction -- now & over time**
- **How *SHOULD* students perform?**

ISSUES:

Should or Ought, not Will

What separates

"Did Not Meet the Standard" from

"Met the Standard" from

"Commended Performance" ?

Threshold Students **All Students**

Student Performance Data

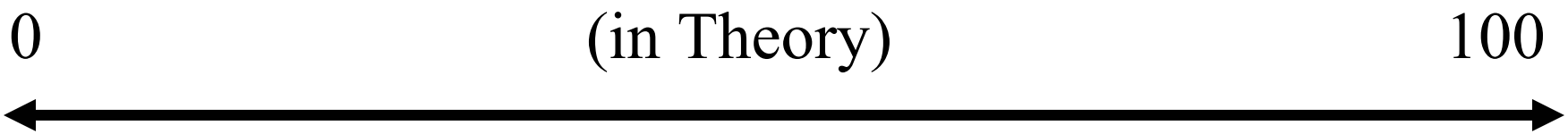
- **Item difficulty = "*p* values" (% correct)**
- **Data tell how students *DID* perform**
- **Data CANNOT tell how students SHOULD perform *nor* even how those who Met the Standard performed**

Issues with the Student Data

- Collected in Spring, ‘02 tryout research.
- Large and representative samples, but most students knew their work “didn’t count.”
- Effect of lack of motivation unknown, but probably not trivial.

Item Difficulty Values

(for multiple-choice items)



Where to put the cuts???

Why Reratings ?

- **You are now a *different* judge**
- **Consider judgments & views of your <u>peers</u>**
- **Consider student performance data**
- **Goal: NOT "consensus," but *reflection***

YOU ARE NOW A *BETTER* JUDGE, because you are a better-informed judge.

Reratings -- What to Do ??

1. **Reflect on earlier ratings -- yours & peers**

2. **Reflect on the discussions we have had**

3. **Consider expanding the "zones" around your earlier cuts**

4. **Reconsider each page in the "zone"**

5. **Choose the point that best defines the *threshold* of each category**

F5/mc

Reratings -- What to Do ??

1. **Reflect on earlier ratings -- yours & peers**

2. **Reflect on the discussions we have had**

3. **Consider expanding the "zones" around your earlier cuts**

4. **Reconsider each item / score point in the "zone"**

5. **Choose the point that best defines the *threshold* of each category**

Keep in Mind:

1. **Skills(s) Assessed**

2. **Curriculum & Instruction - now & later**

3. **How SHOULD students perform?**

4. ***Data: Difficulty & Implications***

Discussion of Preliminary Ratings

- **WHY ????**

- Hearing from your peers helps you to:

  - become more **comfortable** with your judgments -- both the *how* and *where*

  - **reconsider** your earlier judgments

"How do I know if I'm *right*?"

- **There is no "right"**

- **Did you keep in mind:**

  "*should*" ?

  the *threshold* ?

  ***all* Texas students taking TAKS ?**

  the discussions we've had ?