

Systematic Approaches to Training Raters for Performance Assessments

Troy Cox, PhD

Pause Point

What's your purpose?



Baltimore Hotels

- **Hilton Baltimore**
- **Hyatt Regency Baltimore**
- **Sheraton Inner Harbor Hotel**
- **Baltimore Marriott Inner Harbor**
- **Holiday Inn Inner Harbor**
- **Lord Baltimore**
- **Hampton Inn Baltimore**
- **Monaco Baltimore**

Selection from Expedia Hotel Ratings

Stars	3	3.5	4	4.5	5
Rest.	1	1+	1+	1+ (with awards)	Gourmet Dining
Staff	Baggage Assistance	Uniformed	Turndown by request Valet Parking 24-Hour Room Service	Automatic Turndown 24-Hour Room Service (Hot food items)	Address guests by name Personal Trainers, Floor Butlers
Lobby	Small with fake plants	Upscale décor and live plants	Multiple conversation areas	Fresh Flowers	Fresh flowers in rooms
Pool	-	Small	Larger with poolside food servers	Larger with poolside food servers	Full Service Health Facility
Spas	-	-	Small	Full-feature	
Artwork	-	-	Basic	Original	Original
Bedding	-	-	Pillow-top mattress	Luxurious linen with triple sheeting	?

Hotels

Hotel	Rating
Hilton Baltimore	3.5
Hyatt Regency Baltimore	4
Sheraton Inner Harbor Hotel	3
Baltimore Marriott Inner Harbor	3
Holiday Inn Inner Harbor	3
Lord Baltimore	3.5
Hampton Inn Baltimore	2.5
Monaco Baltimore	4

Let's try this again

Hotel	Rating	~Price	~Distance
Hilton Baltimore	3.5	\$299 \$229	0.1
Hyatt Regency Baltimore	4	\$399 \$229	0.4
Sheraton Inner Harbor Hotel	3	\$519 \$364	0.3
Baltimore Marriott Inner Harbor	3	\$399 \$299	0.2
Holiday Inn Inner Harbor	3	\$123 \$90	0.1
Lord Baltimore	3.5	\$139 \$118	0.4
Hampton Inn Baltimore	2.5	\$249 \$129	0.3
Monaco Baltimore	4	\$369 \$159	0.5

Pause Point

Describe your experience.

Rating Principles

Judge not that ye be not judged.

—Matthew 7:1

Movies

Stores

Professors

Pervasive

Music

Books

Cars

Restaurants

Round 1: Rating vs. Ranking



Similarities?

Differences?

Examples?

Round 2: Formal vs. Informal

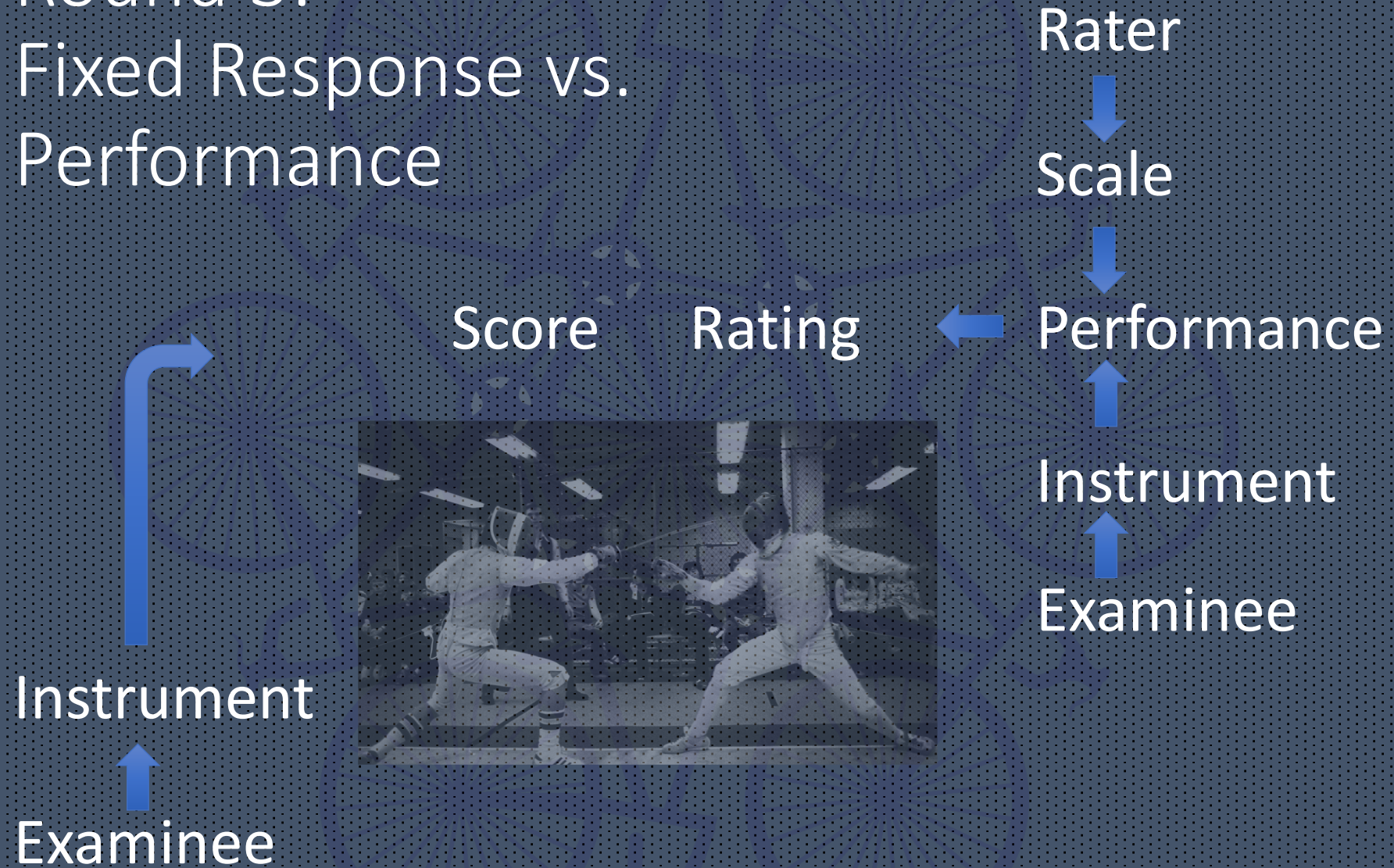


Similarities?

Differences?

Examples?

Round 3: Fixed Response vs. Performance



Observed Score = True Score + Error

Reliability increases as error decreases.

What are the sources of error?

Fixed Response

- Instrument

Performance

- Scale
- Instrument
- Rater

- Definition: Rating is a human process that involves estimating the degree to which a person, object or event possesses a particular trait or characteristic that can be identified and described.
- Types:
 - Quantitative (estimates of frequency, amount or duration)
 - Qualitative (estimates of what kind? or how good?)
- The rater is the real **instrument**—not the rating scale.

Example: Huot (1990)

- 300 Freshmen English Papers
- 2 Topics
- 9 Point Scale
- 53 Readers

**Be an instrument not a
tool!**

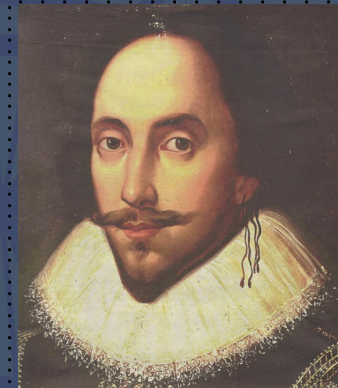
- What percent of the papers had at least 7 different scores?
- What was the minimum number of scores any paper had?

94%

5

Characteristics of Rubrics

"To thine own self be true."



Rating Scales/Scoring Rubrics

Set of generic descriptions of performance which can be used to assign scores to an individual's performance.

Purpose: Get raters out of their heads.

Preliminary Thinking - Rubrics

- Do you know what you are looking for?
- Can you define good performance and talk about it?
- Can you defend your judgment to a peer?

Characteristics of a good rubric/rating scale

1. Mutually Exclusive Divisions
2. Mutually Exhaustive Divisions
3. Defined clearly enough that they can be readily understood by different
 1. raters
 2. subjects

Rubrics/Rating scales should help raters...

1. Focus their attention on the specific features being rated.
2. Provide a common frame of reference for making comparisons.
3. Provide a convenient form on which to record their ratings.

Three Types of Rubrics / Rating Scales

Holistic Scale/Scoring

Performance is judged in terms of its overall quality.

A single rating is given in the form of a grade or number. While a holistic rubric/rating scale might have multiple components for a given rating, these components will be viewed in combination not in isolation.

Holistic scoring assumes all relevant performance features of a given level occur at the same rate.

Holistic Scale/Scoring

Pros:

- Faster
- Requires fewer resources.
- May occur more naturally for raters.

Cons:

- Does not provide specific information about different aspects about performance.
- May not discriminate.
 - Two students receive same score for different reasons.
- Mask performance deficiencies.
- May group language characteristics that are not highly related.

Analytic Scale/Scoring

Performance is divided into its composite parts.

Each part receives a score/rating that is independent of the other parts. These sub-scores may then be averaged or combined.

Analytic scoring assumes that different aspects of performance develop at different rates.

Analytic Scale/Scoring

Pros:

- Provides more abundant and useful diagnostic/descriptive information.
- Gives greater control over rater behavior. Rater's interpretation of construct is more transparent and can be adjusted for.
- May be more reliable particularly for new or less trained users.

Cons:

- Resource intensive to develop.
- Resource intensive to maintain/train.
 - Presents greater cognitive load.

Objective Scale/Scoring

Performance is judged in terms of quantifiable constructs.

i.e. accuracy in the form of error-free clauses / T-units; instances of particular language features.

Cautions:

- Objectivity is a heavy burden.
- Often 'objectivity' just isn't.

Pause Point

Describe a rubric you have used; are using; or think you may need to use to accomplish your institutions current assessment needs.

Creating A Rubric

Developing a new scale

- Analytic vs Holistic

- If holistic, how many levels?

Range of ability; purpose; who will use it; rater reliability

- If analytic, how many criteria?

One subscale per component of the construct

4-5 start to strain raters; 7 max.

Are criteria separate?

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. New York: Cambridge.

Determining Criteria

- Intuitive rating scale development
 - One or more 'experts' frame expectations based on their experience (instructors, theory, syllabus, curriculum)
- Empirical rating scale development
 - Start with a sampling of the target performance
 - Divide into groups based on the level of their performance.
 - Norm referenced: based on relative performance within the sampling.
 - Criterion referenced: based on relation to preexisting levels.
 - Describe characteristics of the responses at each level.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. New York: Cambridge.

Writing Descriptors

- Descriptors should reference performance and not 'levels of division.'
 - Needs of both 'novice' and 'experienced' users.
 - Descriptors that reference performance can provide diagnostic information.
- Highs & Lows
 - Zero assigned to reference the sample did not contain evidence appropriate for the construct.
 - Highest reflects mastery of the construct for a given component.

Writing the Descriptors

- Best fit vs. Lowest sustained.
 - Best fit/Most fit: which band most resembles the sample.
 - Direct users to be consistent in border judgments.
- Lowest sustained: which is the highest band where each component is met fully.
 - Generally seen as the less appropriate method.

Writing the Descriptors

- Focus on what learner can do.
- Be concise.
- Strive for parallelism
- Avoid ambiguity

Pause Point

3 – Points you'd relate to a colleague interested in what we've discussed about rubrics.

2 – Ideas/Questions/Sparks that you are going to keep thinking about

1 – Action you are going to implement.

Adapting A Rubric

Adapt or Build?

What is the purpose of the assessment?

Placement, Proficiency, Achievement, Research

Who will use the rubric? How will they be trained?

Teachers, Students, Administrators

What is being measured?

Define the construct;

Aspects of Performance; which are most important

Adapting rubrics

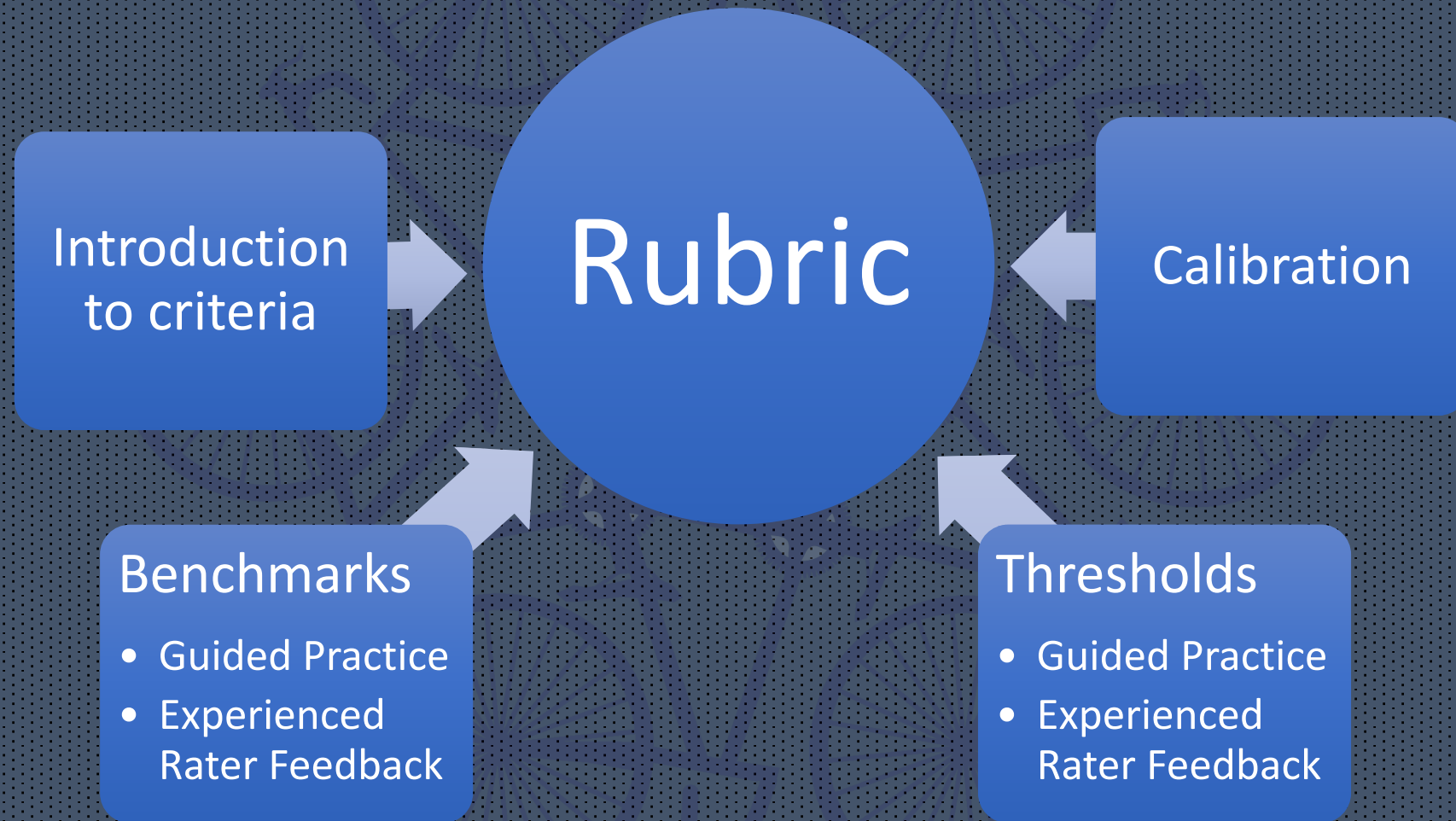
- Be familiar with what is out there.
 - Examples of common rubrics.
 - TOEFL, IELTS, ACTFL, Institution specific
- Cautions
 - Understand instrument differences behind what is being rated.
 - Are 'those' being rated sufficiently equivalent.
 - Proficiency scale – too many, too few levels?

Pause Point

Of the sample rubrics provided, which seems to be closest fit to your needs? Which characteristics work well? Which would need to be modified?

A Rater Training Model

Rater Training Model



Rubric

- You can't reliably rate what hasn't been defined!

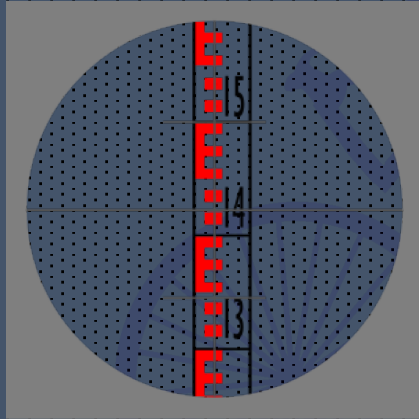
THE HAMBURGER WRITING RUBRIC			
1	2	3	4
- Explanation is <u>not</u> complete; there is <u>no evidence</u> to support or prove their answer. - Question <u>not</u> answered correctly. - Work is <u>not</u> neat and clear. - Work has <u>many</u> spelling and grammatical errors.	- Student uses evidence in the response, but the <u>explanation is weak and confusing</u>. - Question is <u>partially</u> answered. - Work is unorganized. - Work has <u>some</u> spelling and grammatical errors.	- Question is restated and answered in a <u>thoughtful, coherent</u> way. - Response is <u>based on evidence</u> from text or from class activities. - Question is answered correctly. - Work is organized. <u>Very few</u> spelling and grammatical errors. - Demonstrates <u>proficiency</u>.	Supersize Me! - Student demonstrates understanding and <u>performance beyond proficiency</u>. - Question is answered in a clear, logical, and coherent way. - Work is neat and organized. - No spelling and grammatical errors. - Student went <u>above and beyond</u> expectations!
"You get a hot dog instead of the hamburger you ordered."	"You get the hamburger on time, but it is barely warm, missing it's top bun, and does not have the ketchup and mustard you requested."	"You get the hamburger you requested hot off the grill."	"You get the hamburger you requested hot off the grill and you get upgraded to a double-stack with all the toppings – for free!"
			
THE "HOT DOG"	THE "PATTY"	THE "AVERAGE JOE"	THE "KITCHEN SINK"

Introduction to Criteria

- What's included/excluded?
- How is it mutually exhaustive?
- What should raters do when characteristics span categories?
 - Compensatory Scoring
 - Prone to false positive errors
 - Fairer to examinees
 - Conjunctive Scoring
 - Prone to false negative errors
 - Fairer to institutions
 - Hybrid Scoring Models
 - Minimizes (but does not eliminate) categorization errors

Benchmarks

What and Why?

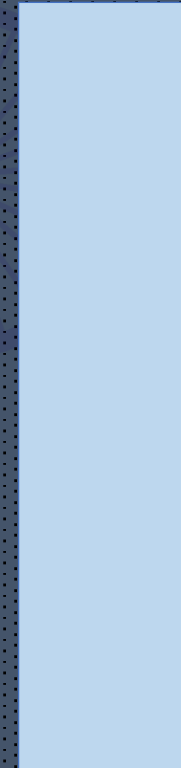
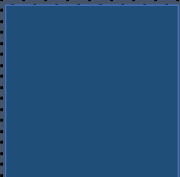


The term benchmark originates from the chiseled horizontal marks that surveyors cut in a wall, pillar, or building and used bench to place a leveling rod to measure altitude.



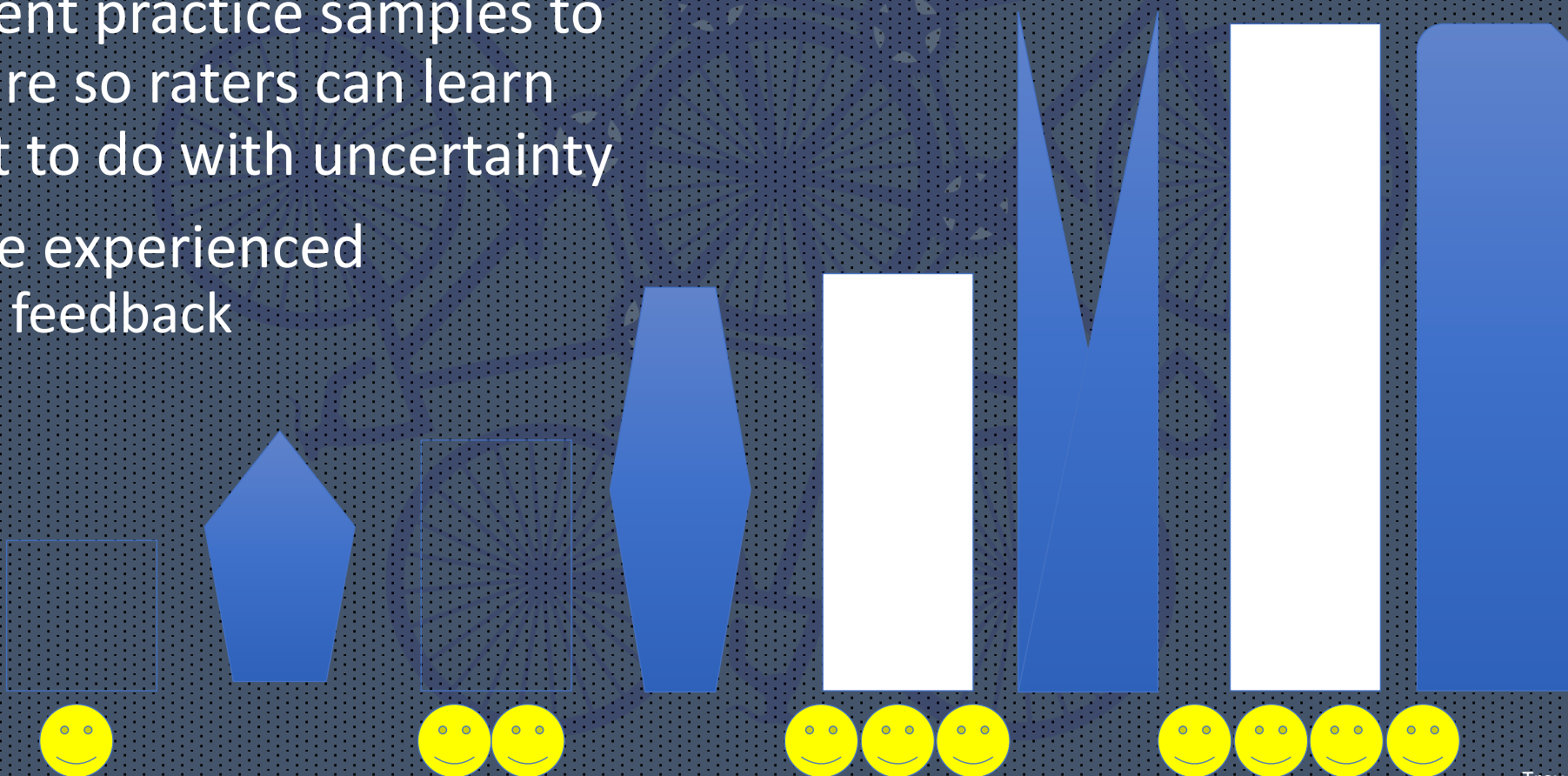
Benchmark Examples

- Provide clear examples of rubric categories in which all the criteria align.
- Present examples that can be used as references
- Present practice samples to ensure raters understand aligned rubric



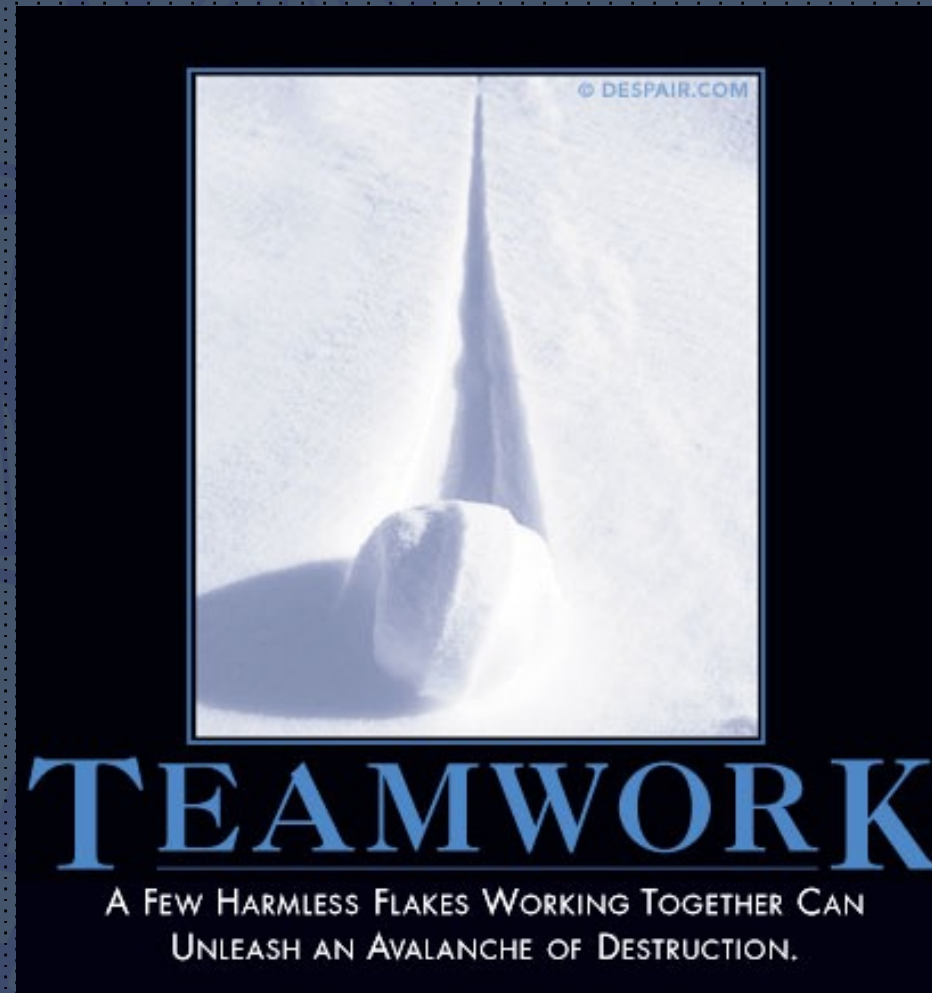
Threshold Examples

- Provide examples in which the criteria do NOT align.
- Present practice samples to ensure so raters can learn what to do with uncertainty
- Share experienced rater feedback



Calibration

- Take some time before each rating session to make sure everyone remembers how to use the rubric
- A little self-retrospection is good for the soul



Break Time

20 minutes– run free!

Context and Timeline

- BYU's English Language Center
 - Intensive English Language Program
 - Lab School (staffed in large part by MA TESOL students)
 - Non-matriculated ESL students (230-260; 34 countries; True beginners to University/Graduate Studies ready).
 - Foundations Program (1-3)
 - Prep Program (0)
 - Academic Program (4-6)
 - Prep program (-4)
 - University prep program.

Context and Timeline

- Full Program Testing and Rating before each semester (1st semester students); end of semester (all enrolled students).
- Rating is staffed by all teachers who have a teaching assignment. 3 hours is added to their contract for rating participation per class assigned. Teacher gets one rating assignment for each class taught.
 - Calibration practice
 - Calibration meeting
 - Individual rating assignments (20 speaking/ 40 writing per rating assignment).
- Rating Calibration distributed and collected by Assessment Coordinator.
- Calibration meeting conducted by Assessment Coordinator or Skill Area Supervisor.
- Individual ratings the day following computerized testing.

Context and Timeline

- True novice (pre-teaching) raters trained on benchmarks and thresholds in practicum courses
- Raters assigned to rate speaking or writing (usually not both). Rater design is constructed. All students are blind rated twice (boundary students are rated 3 times)
- Raters receive rating calibration practice sheet and samples 12 days before testing begins and have 1 week to complete them.
- Raters' identities are anonymized and put into a table that will allow for discussion of rater errors and biases.

Context and Timeline

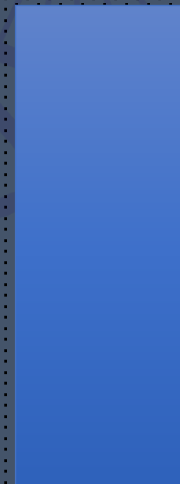
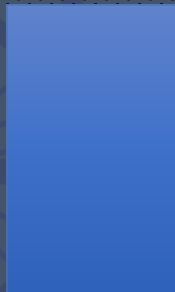
- Student Samples are collected.
- Raters meet for calibration. See how they compare. Listen/Read student samples discuss rating tendencies.
- Raters rate performance.
- Assessment coordinator uses MFRM to determine a FairAverage Score for students. Occasional 3rd Rating is required.

Speed Rating (not Speed Dating)

A Rater Training Case Study

Benchmark Examples

- Let's read a clear example of a 1 (pages 1-2 of Timed Writing Rater Training packet)
 - Why is it a 1?
- Now a clear example of a 7 (pages 13-14)
 - How do you know it's a 7?



Let's Practice!

Benchmark Samples

- Pages 50, 42, 34, 38 & 41 *individually* using your rubric as reference
- Then discuss your ratings and ideas with your table
- What's the consensus? Why?

Discussion Questions

- How long and organized is the text?
- How does the student function with the language and vocabulary?
- How complex is the student's grammar? How clear is the student's meaning?
- Was this sample difficult for you to rate or choose?

Let's Practice!

Threshold Samples

- Rate Pages 32, 33, 37, 45 & 46 individually using your rubric as reference
- Look at the experienced rater feedback
- Discuss with your table

Discussion Questions

- Were these samples easier or more difficult to rate?
- What did you do when trying to decide between two different scores?
- How did the experienced rater feedback affect you?

Rating Reliably

Employ Sound Rating Practices



- **Refer to the rubric**

- Rate blindly
- Avoid comparing examinees to each other
- Rate one prompt at a time
- Be aware of time and setting when you rate so that all students are treated fairly
- Change the order of your ratings
- Use multiple raters for any high stakes testing

Avoid Common Rater Errors

1. Generosity
2. Severity
3. Central Tendency (Scale interpretation and timidity)
4. Halo (Rater and item, rater and examinee)

Generosity Error

Definition: A tendency to systematically overrate examinees.

Potential Causes: An unwillingness or inability to recognize poor performance.

Effect: Examinees receive higher ratings on their performance than their peers who have the same ability level.

How to Detect: In a fully-crossed design (each examinee is rated by each rater), look at the average ratings of each rater. The rater with the highest average might have a tendency toward this error.

Severity Error

Definition: A tendency to systematically underrate examinees.

Potential Causes: An unwillingness or inability to recognize good performance.

Effect: Examinees receive lower ratings on their performance than their peers who have the same ability level.

How to Detect: In a fully-crossed design, look at the average ratings of each rater. The rater with the lowest average might have a tendency toward this error.

Central Tendency Error

Definition: A tendency to limit ratings to the middle range of the scale.

Potential Causes: An inability to discriminate between good and bad performance or a lack of confidence on the part of the rater due to a desire to play-it-safe.

Effect: Examinee ratings cluster around the middle of the rubric and thus it is difficult to differentiate one examinee from another.

How to Detect: In a fully-crossed design, look at the standard deviation of each rater. The rater with the lowest standard deviation might have a tendency toward this error.

Rater/Examinee Interaction Error

Definition: A tendency to have some extraneous characteristic (native language, gender, age, etc.) influence the rating of examinees with that characteristic.

Potential Causes: An inability to prevent some previous experience with a group from affecting the rating given to an examinee of that group. The experience could be positive or negative and the bias could be conscious or subconscious.

Example 1: A rater who is fluent in Spanish may systematically rate Spanish examinees higher or lower than they rate other examinees due to being more familiar with the types of errors Spanish learners make.

Rater/Examinee Interaction Error (cont)

Effect: Examinees that have the characteristic receive higher or lower scores than their peers who have the same ability level.

How to Detect: In a fully-crossed design, compute the average ratings of each rater for each group that shares that characteristic. Compare the averages between the groups. Raters with the highest differences between groups might have some kind of interaction error.

Halo Error

Rater/Task-Trait Interaction Error

Definition: A tendency for a global rating to be over-influenced by a single trait/item **or** for the rating of a specific trait/item to be over-influenced by the global rating or other items.

Potential Causes: An inability to distinguish between the specific traits being rated or rate different items independently.

Halo Error

Rater/Task-Trait Interaction Error (cont.)

- *Example 1: With a speaking test that has more than one item, the performance on the first question might influence the rating of the second question.*
- *Example 2: When rating overall writing ability, letting a single trait such as grammar over-influence the rating that is awarded.*
- *Example 3: When rating the specific trait of written linguistic accuracy, letting the overall impression of the essay (length, vocabulary, rhetoric, etc.) over-influence the linguistic accuracy rating.*

Halo Error

Rater/Task-Trait Interaction Error (cont.)

Effect: Examinees have similar ratings across all the traits/items that should have been independently assessed and therefore it is more difficult to diagnose/assess differential development.

How to Detect: This can only be detected if each trait/item is rated by multiple raters.

Pause Point

- As a table, complete the Rater Errors and Biases Worksheet.

When Raters Disagree...

- Decide beforehand what will happen.
- Average
 - Add both ratings together and divide by two
 - If ratings are NOT adjacent, get a third rater to arbitrate
- Arbitrate
 - Get a third rater
 - Each raters rating is counted as a vote. The most votes win.
- Adjust
 - Use Multi-Facet Rasch Measurement (MFRM) to adjust the ratings to a Fair Average (more details to come later).

Rating Schedules

Fully-Crossed

Every examinee rated by every rater.

	R1	R2	R3	R4
1	X	X	X	X
2	X	X	X	X
3	X	X	X	X
4	X	X	X	X
5	X	X	X	X
6	X	X	X	X
7	X	X	X	X
8	X	X	X	X
9	X	X	X	X
10	X	X	X	X

Disconnected Subsets

Every examinee double-rated, but no overlap between raters.

	R1	R2	R3	R4
1	X	X		
2	X	X		
3	X	X		
4	X	X		
5	X	X		
6			X	X
7			X	X
8			X	X
9			X	X
10			X	X

Rating Schedules (Continued)

Incomplete, Connected Subsets

All raters rate a shared subset of examinees and are then paired for the rest.

	R1	R2	R3	R4
1	X	X	X	X
2	X	X	X	X
3	X	X		
4	X	X		
5	X	X		
6	X	X		
7			X	X
8			X	X
9			X	X
10			X	X

Incomplete, Connected Spiral Subsets

Each rater is paired with every other rater.

	R1	R2	R3	R4
1	X	X	X	X
2	X	X		
3	X	X		
4	X	X		
5		X	X	
6		X	X	
7			X	X
8			X	X
9	X			X
10	X			X

Case Study

Calibration Meeting

Pause Point

- As a table, determine raters who exhibit the errors listed above.

Introduction to Many FACETs

Rasch Measurement

What is a valid rater?

Two Approaches to Rating

- Rating machines
 - Raters give the “correct” rating every time.
 - Agree with each other on the exact rating
 - Agree on which performances are better and which are worse
 - Use Classical Test Theory/Generalizability Theory
- Dependable Independent Experts
 - Agree on overall competence (expert)
 - Disagree on details (independent)
 - Maintain own standards of leniency and severity (dependable)
 - Use MFRM

Rasch Measurement

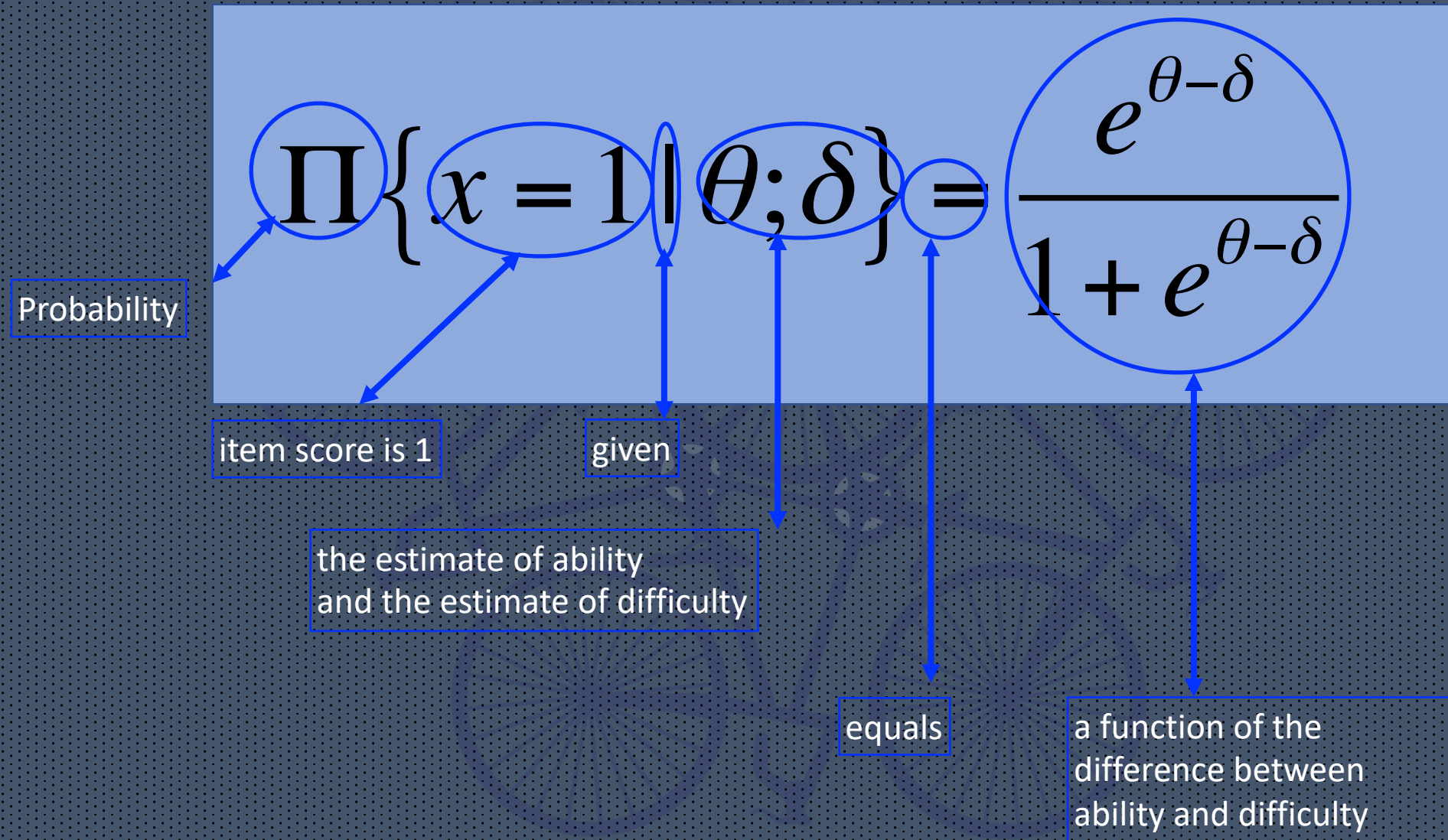
- Is person-independent
- Is item-independent
- Converts raw scores of items and persons to a shared scale (logits)
- Is defined as...

Probability of Success

=

Function (Person Ability-Item Difficulty)

Rasch Formula-Dichotomous Data



Examinees and items are on the same linear scale allowing for easy comparison at the item/person level.

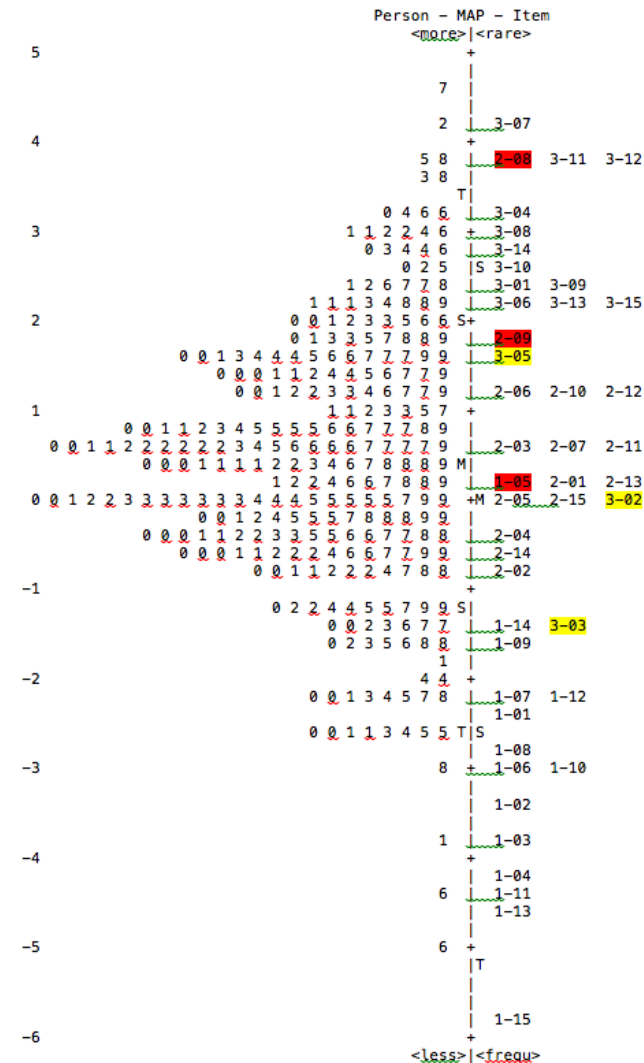
TABLE 1.0 English Proficiency Screener

ZOU452WS.TXT

May 28 14:22 2013

INPUT: 296 Person 45 Item MEASURED: 296 Person 45 Item 2 CATS WINSTEPS 3.70.0.3

Harder than Expected, Easier than expected



Many Facet Rasch Measurement

- Probabilistic model that converts raw scores to a linear scale
- Facet: any factor, variable or component that is assumed to affect scores in a systematic way (Eckes, 2011)
- What facets could contribute to the variance of the performance ratings?
 - The ability of a particular PARTICIPANT
 - The difficulty of a particular TOPIC
 - The severity of a particular RATER
 - The SCALE that is used

MFRM-Polytomous Data

$$\ln \left[\frac{P_{nljk}}{P_{nljk-1}} \right] = \theta_n - \delta_i - \alpha_j - \tau_k$$

- θ_n = ability of participant n;
- δ_i = difficulty of topic/prompt i;
- α_j = severity of rater j;
- and
- τ_k = scale being used

Examinees, raters, and rubric are on the same linear scale allowing for easy comparison at the facet level.

Furthermore, we can see which individuals do not “fit” the model, giving us insight to intra-rating reliability

Winter 2013 Timed Writing Calibration 4/17/2013 9:55:34 AM
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1A,2A,S) Yardstick (columns lines low high extreme)= 0,3,-5,8,End

Measr	+Raters	+Student	Essays	Scale
8 +				+ (6)
7 +		490		+ ---
6 +		120		+ 5
5 +		611 884		+ ---
4 +				+
3 +	Admiral Marcus	767		+ 4
2 +	Bones	513		+
1 +	Kirk	863		+ ---
0 *	* Checkov * Captain Abbott * Christopher Pike	* 718 * 588		* 3 *
-1 +	Sulu Bar Patron Cadet			+
-2 +	Uhura Transport Officer Khan			+ ---
-3 +				+
-4 +				+ 2
-5 +		813		+ (0)
Measr	+Raters	+Student	Essays	Scale

MFRM...

- Provides a mathematical framework to compare data at the facet level.
- Converts raw scores (dichotomous and polytomous) to a linear scale.
- Has the quality of invariant measurement.
- Allows us to adjust student scores based on the raters leniency/severity.
- Allows diagnostics of raters and rubrics
- Quite useful, but requires specialized training. Most appropriate for high stakes testing.