Intealth™
Advancing the Global Health Workforce

ECFMG FAIMER

Basics of Rater Cognition

Adapted from a workshop originally presented as part of the ACGME Assessment Course

July 2024

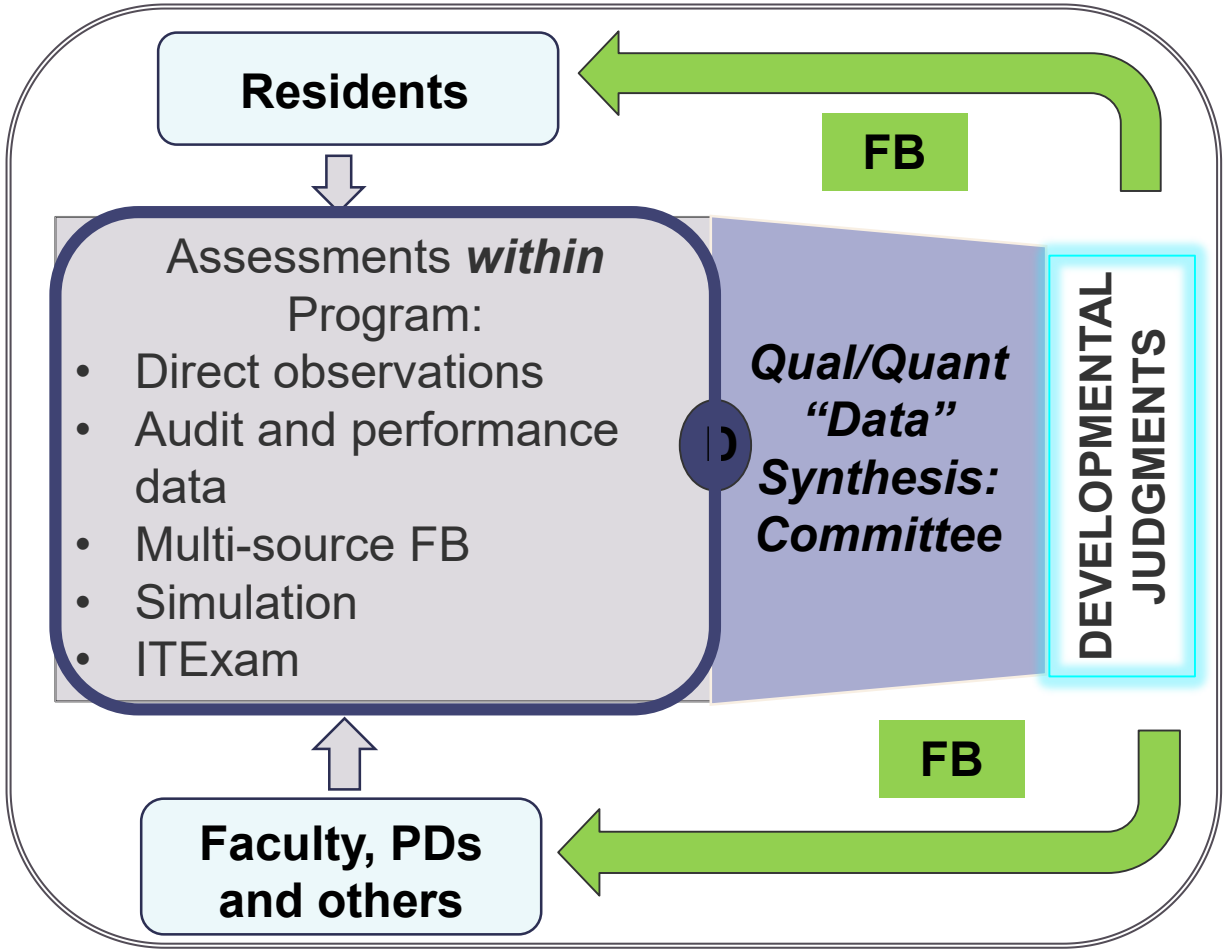
Disclosures

- Eric Holmboe works for Intealth – however, the majority the work presented here was developed during his time at ACGME. He also receives royalties for a textbook on assessment from Elsevier Publishing.

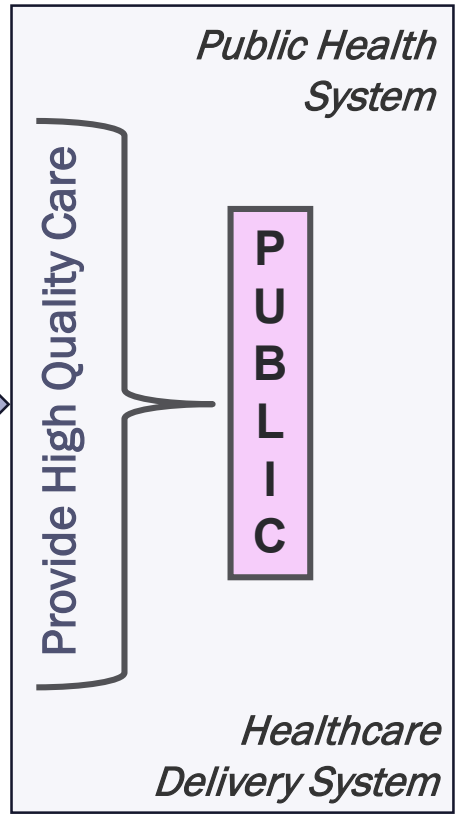
Outline

- Assessment systems and Miller’s pyramid
- Key issues in rater cognition
- Key principles for faculty assessments

A Basic Assessment "System"



Summative Entrustment



The Economist

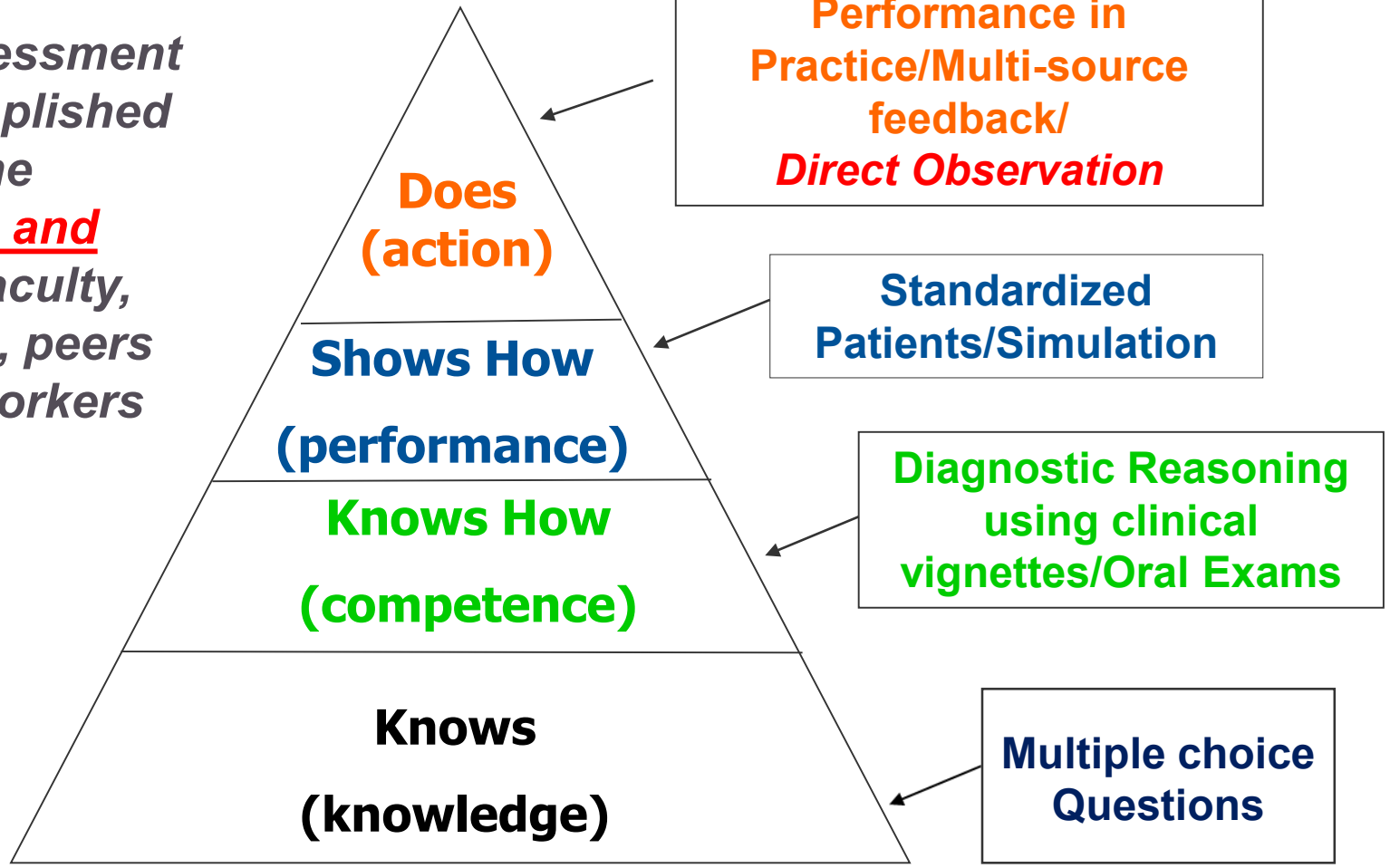
Asia | An Asian model

Why is Thai health care so good?

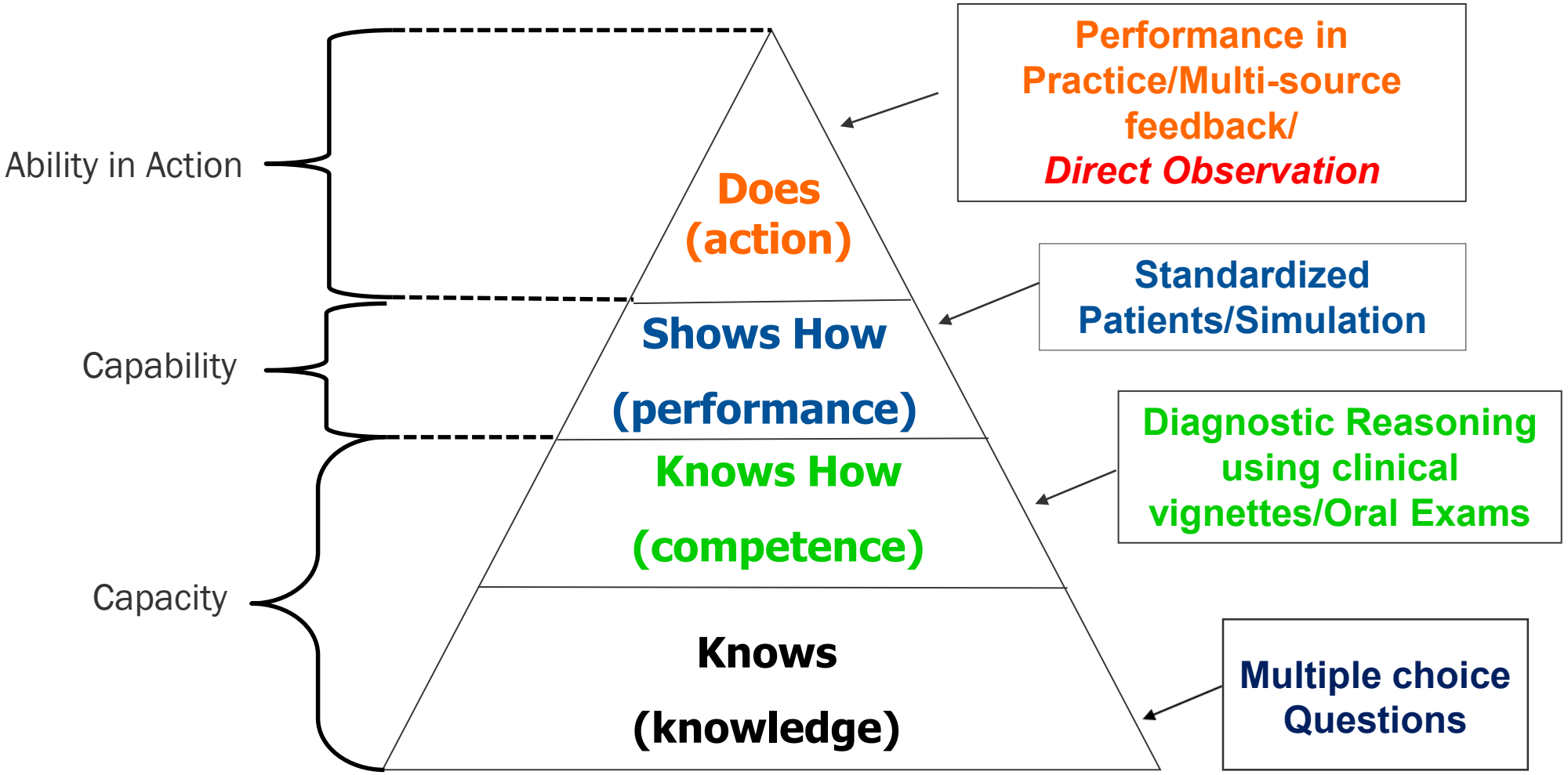
The country could become a model for the region

Assessing for the Desired Outcome

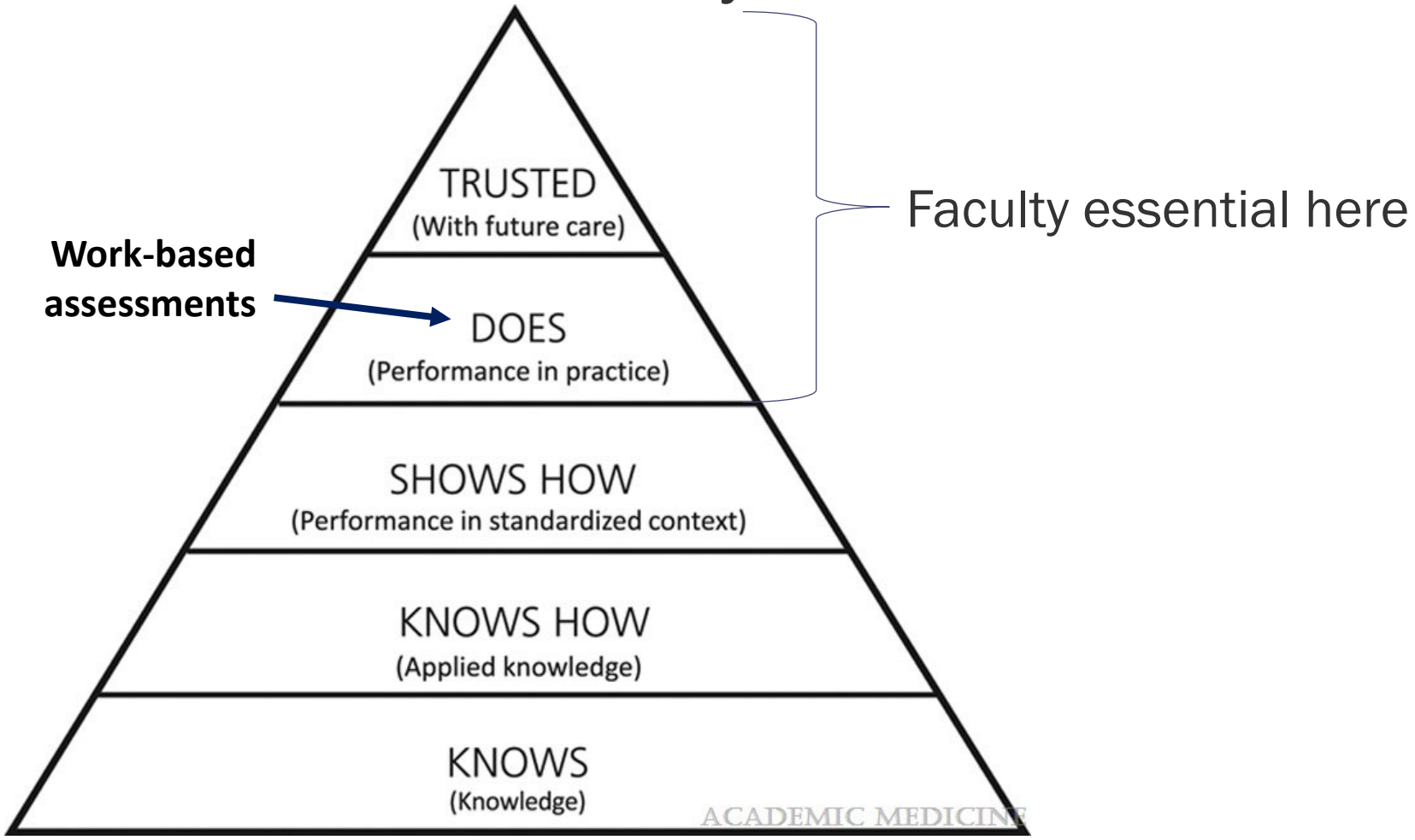
Work-based assessment is mostly accomplished through the observations and questions of faculty, team members, peers and other co-workers



Assessing for the Desired Outcome



“Extended” Miller Pyramid

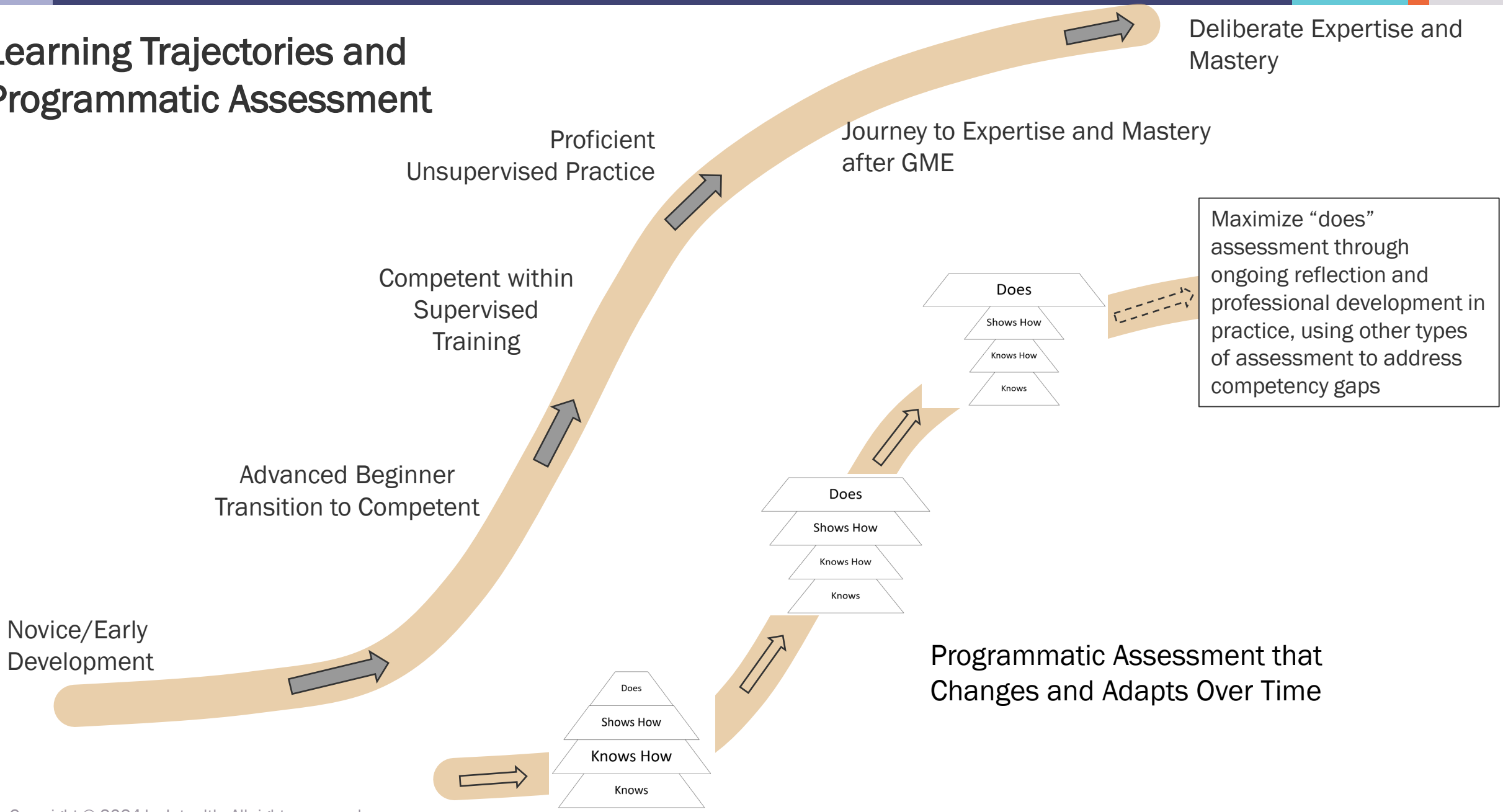


Work-based assessments

Faculty essential here

A new fifth level (“trusted”) reflects the process for reaching the decision to award a learner an attestation of the completion of training, leading to a medical license or specialty registration or certification, that provides permission to act unsupervised and makes the grantors cognizant of the inherent risks.

Learning Trajectories and Programmatic Assessment



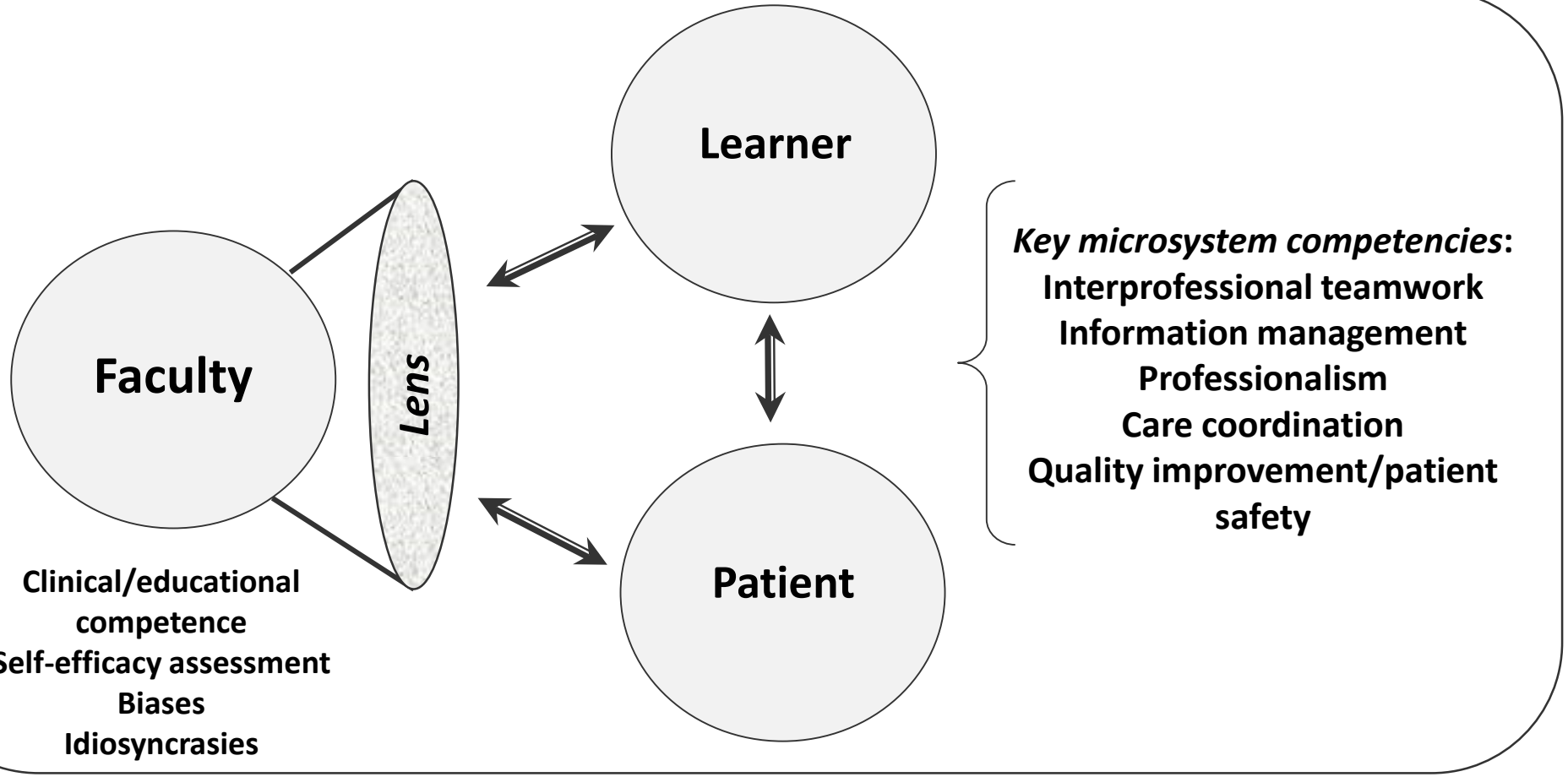
Intealth™

Advancing the Global Health Workforce

ECFMG FAIMER

Faculty Assessments: Issues in “Rater Cognition”

Assessment: Complex and Situated in Context



Microsystems: Clinic, Hospital Ward, Operating Room

Institution and the Clinical Learning Environment

Key Issues: Individual Factors

- Psychometric issues
- Variability among faculty
  - Strengths and weaknesses in
    - Clinical competencies
    - Educational skills
    - Assessment skills
- Human limitations
  - Bias and stereotyping, idiosyncrasy, and inference
  - Cognitive load

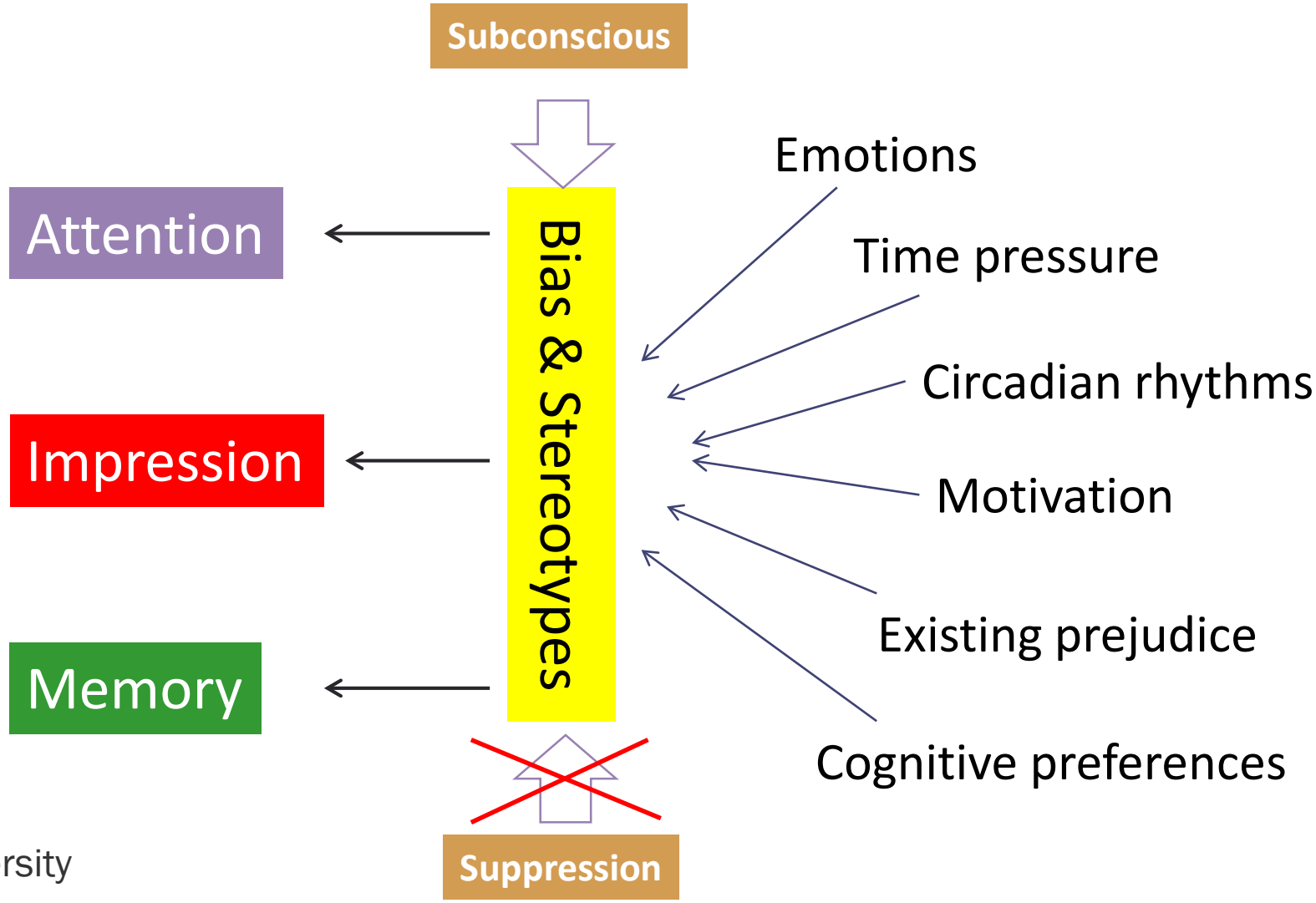
Psychometric Issues

- Multiple studies demonstrating major issues in intra- and inter-rater reliability
  - Usual response – ***change the form or tool that has only modest impact***
- Limited evidence for validity
  - Modest correlations between high-stakes assessments and faculty ratings
- Lack of discrimination among competency domains
  - May not always be a problem – depends on cause
    - Interdependence of competencies

Common Faculty Rating Errors

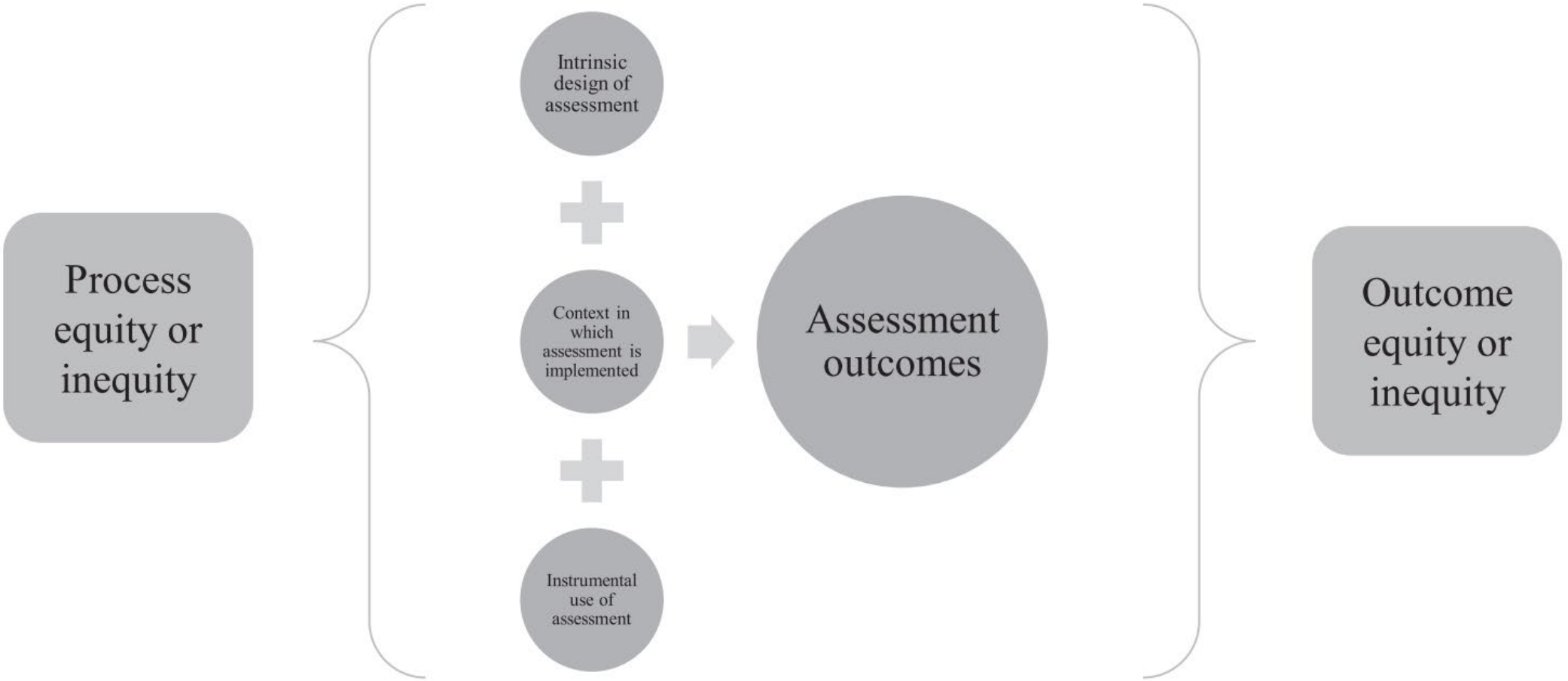
- Correlational errors
  - Halo effect
  - Horn effect
    - Ratings based mostly on *perceived* knowledge and personality
- Distributional errors
  - Leniency error (“Doves”)
  - Severity error (“Hawks”)
  - Central tendency

Bias and Stereotypes



Peter Yeates. Keele University

Components of Equity in Assessment



Lucey, Catherine R. MD; Hauer, Karen E. MD, PhD; Boatright, Dowin MD; Fernandez, Alicia MD. Medical Education's Wicked Problem: Achieving Equity in Assessment for Medical Learners. Academic Medicine 95(12S):p S98-S108, December 2020.

Possible Strategies to Reduce Assessment Bias

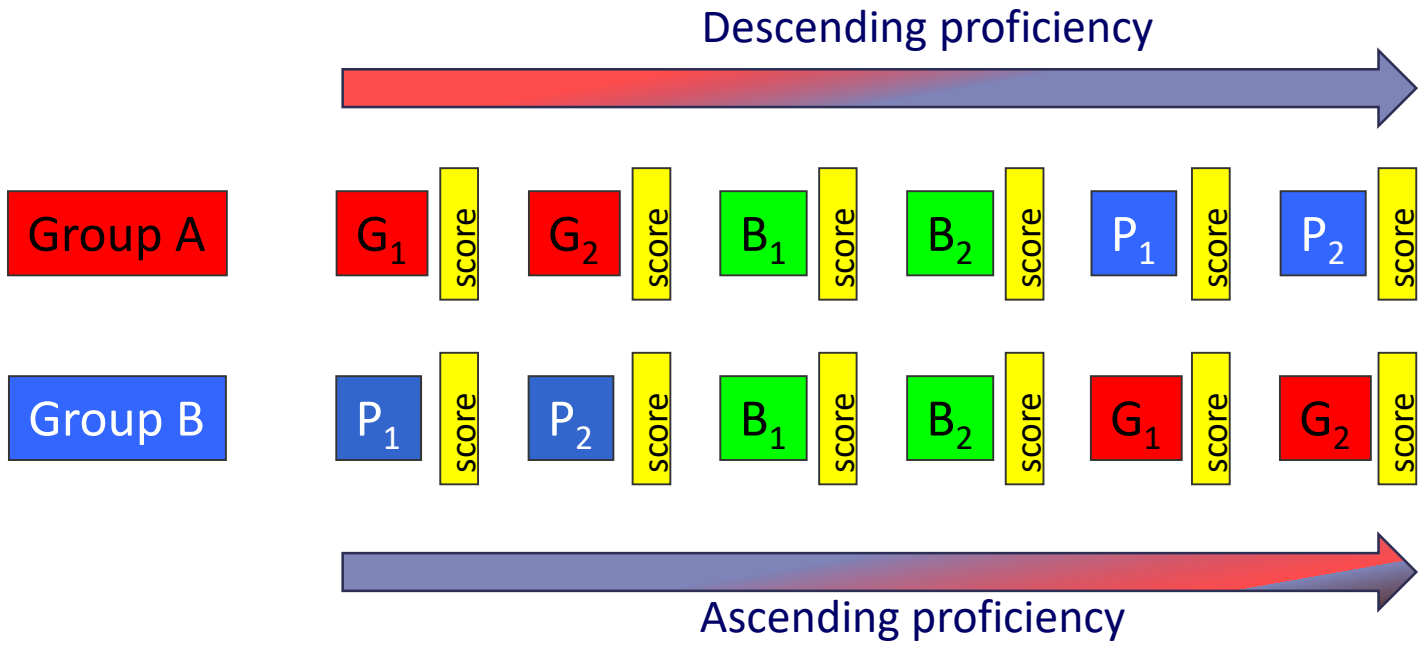
Strategy	Description	Assessment Example
Stereotype replacement	Recognizing when a stereotype has been activated, thinking about why, and then actively substituting non- stereotypical thoughts	When completing a narrative assessment of a female learner, the assessor stops to consider if they may be using gender-laden language or uses an online tool to assess for gender bias. If bias is found, the assessor substitutes evidence-based behavioral skills that are more neutral.
Perspective taking	Considering what it would be like to be a member of the minoritized group	During rounds faculty witness a difficult interaction between a learner from a URiM group with a discriminatory patient. Faculty should ask themselves: What must that be like for the learner? How will I intervene in this situation?
Individuation	Recognizing when you have stereotyped someone according to their group affiliation and instead thinking about what makes them an individual	A faculty member watches a learner from another country struggle to interview a patient with a possible sexually transmitted disease and initially stereotypes the learner as from a group “uncomfortable talking about sex.” Instead, the faculty sees an individual learner struggling and seeks to understand why they are struggling as an individual.

Reflection Question

- What types of assessment bias are you experiencing here in Thailand?

- What can you do to help minimize assessment bias?

Contrast effects



Human Limitations in Assessment

- Limitation in working memory and mental processing
  - The “7 +/- 2” rule for short term memory
  - Cognitive load theory

Human Working Memory

Human Working Memory

Limited working memory capacity

How we organise information and data influences perception

Baddeley, A.D., 1994. *Psychological Review*, 101(2), pp.353–356.

Cognitive Load

- There is a limit as to how much you can ask faculty to observe and capture
    - Clinical units: complex environment
    - Selective attention
- Byrne et. al. (Med Educ 2014)
    - Average cognitive load for faculty judging OSCE stations was higher than anesthesia trainees performing an induction for routine surgery
    - OSCE had 21-22 items in an 8 minute station

Useful Dictum

- The longer the evaluation or rating form (or checklist) and the shorter the “exposure” or observation time, the more likely you are to get less useful ratings and information from the evaluation form
  - Long evaluation forms + short faculty rotations = trouble in assessment land

Idiosyncrasy

- Faculty idiosyncrasy might be useful and the variability associated with faculty idiosyncrasy might not represent "error" (i.e. "warranted variation") ***IF the idiosyncrasy represents expertise & mastery***

- In other words, there ***could*** be situations where there are multiple correct answers from faculty depending on their underlying expertise

Andrea Gingerich and Marjan Govaerts

Summary of Key Lessons in Rater Cognition

Assessment (rater cognition) is a complex process

- Training can be effective, but will not solve “all problems”
  - Broad and longitudinal sampling remains essential
- Clarity on outcomes
- Shared mental models
- Own clinical skills matter
- Multiple raters in multiple setting
- Not all variation is bad, but not limitless
  - Variation is a bounded condition

Intealth™
Advancing the Global Health Workforce

ECFMG FAIMER

Faculty Assessment Tools: Key Principles

What Are Your Faculty Assessing?

- 1:1 patient encounter
  - Admissions
  - Clinic visits
  - Daily rounds
  - Procedures
- Work-day or shift performance
- Rotation-based or longitudinal performance

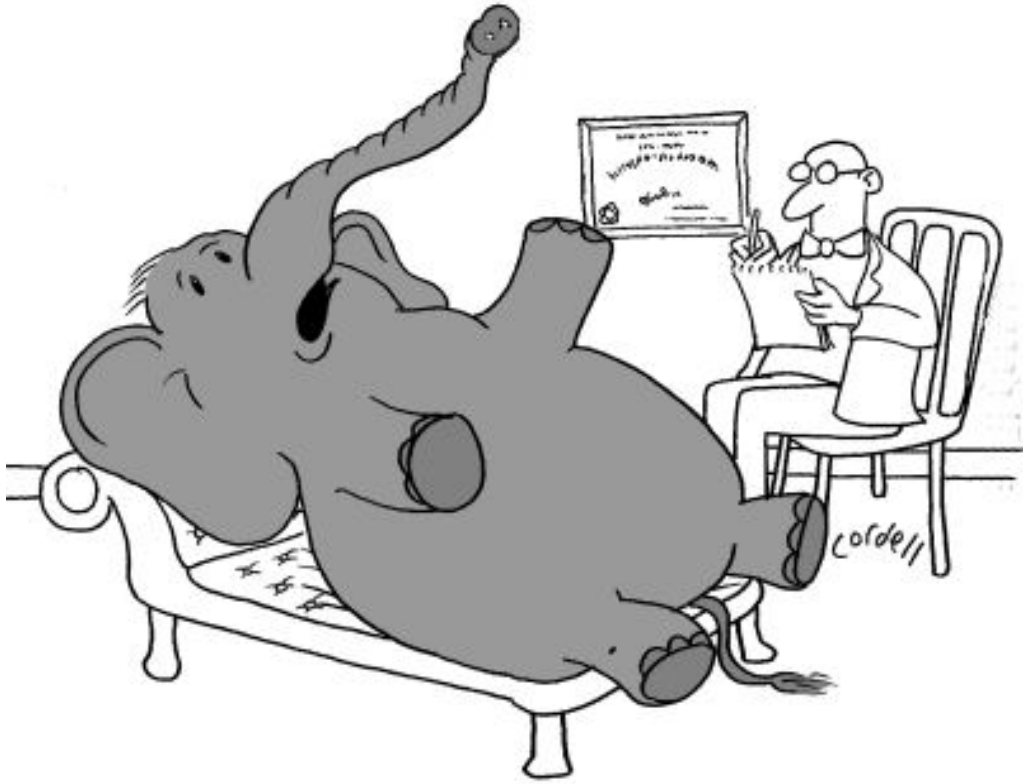
Design assessment tools fit for purpose and context

Rating Scales: Challenges

- Lack of faculty consensus on what they are assessing
  - Lack of clear understanding of program goals, objectives and outcomes
  - No standardized criteria for assessment processes or how to “translate” an assessment into a numeric code
- Need shared mental models (i.e. common framework) for assessment
- Too often the majority of variance resides in the rater

The Frame of Reference Problem

Several studies demonstrate that faculty heavily use self as the frame of reference in judging competence and entrustment. Assessment approaches assume faculty "self" is competent.



"Whenever I walk in a room, everyone ignores me."

Construct Aligned Scales


Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales

Jim Crossley,[1] Gavin Johnson,[2] Joe Booth[3] & Winnie Wade[3]

"Crossley and Jolly have suggested that effective assessment tools have construct alignment, which means that the tool reflects the expertise and priorities of the evaluator."

Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Medical Education* 2011; 45: 560–569

Entrustment Scales

- Per Rekman and colleagues, entrustability scales are a species of construct-aligned scales
- Entrustability scales are usually expressed by varying levels of supervision, oversight and/or actions of the attending

Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment. Acad Med. 2016 Feb;91(2):186-90.

PGME Entrustment Scales

Supervision level
1. Not allowed to practice (observer only)
2. Allowed to practice only under proactive, full (direct) supervision
3. Allowed to practice only under reactive/on-demand (indirect) supervision)
4. Allowed to practice unsupervised
5. Allowed to supervise others in practice

Chen C, et. l., The Case for Use of Entrustable Professional Activities in Undergraduate Medical Education. Acad Med. 2015; 90:431–436.

Co-activity
1. I had to do (observer)
2. I had to talk them through (direct supervision)
3. I had to be there from time to time (reactive/on-demand (indirect) supervision)
4. I needed to be there just in case (reactive/on-demand (indirect) supervision)
5. I did not need to be there (unsupervised practice)

Rekman J, et. Al. A New Instrument for Assessing Resident Competence in Surgical Clinic: The Ottawa Clinic Assessment Tool. J Surg Educ. 2016 Jul-Aug;73(4):575-82. doi:10.1016/j.jsurg.2016.02.003.

Entrustment Scales: Not a Panacea

Entrustment and Quality of Care

Condition	Beta Co-efficient[b]	95% CI	P-Value
Asthma	0.030	(0.014, 0.046)	**0.0004**[d]
Bronchiolitis	0.0004	(-0.0161, 0.0169)	0.96
Closed Head Injury	0.012	(-0.006, 0.031)	0.19
Conditions Combined[c]	0.014	(0.004, 0.023)	**0.006**[d]

[a]Parameter estimates are adjusted by post-graduate year, patient complexity, and patient acuity
[b]The change in RSQM composite score associated with a one unit change in entrustment
[c]Combined conditions are adjusted by diagnosis
[d]Significant at $p<0.01$

Schumacher D, et. al. Resident-Sensitive Quality Measures in the Pediatric Emergency Department: Exploring Relationships with Supervisor Entrustment and Patient Complexity and Acuity. Acad Med. 2019; in press.

Entrustment and Direct Observation

Table 3: Number and percent of incorrect and correct ratings by faculty participants.

Scripted Entrustment score*	2 (n=5 cases)	3 (n=3 cases)	4 (n=2 cases)	All (n=10 cases)
Total Ratings	384	231	231	768
Total incorrect ratings (%)	192 (50%)	106 (46%)	33 (22%)	331 (43%)
Ratings higher than scripted (%)	157 (41%)	66 (29%)	N/A	223 (29%)
Ratings lower than scripted (%)	35 (9%)	40 (17%)	33 (22%)	108 (14%)
Total Correct Ratings (%)	192 (50%)	125 (54%)	120 (78%)	437 (57%)

Bottom Line - Scales

- Numeric scales are nothing more than a synthesis "***code***" for the observations/questioning by the assessor
  - Numbers are convenient and can be analyzed quantitatively over time, however…
  - The "code" must be associated with a descriptive shared mental model of the competency being assessed
- Entrustment scales "easier" to use, but do not fix the frame of reference challenge

Narrative Evaluation

- Definition: A spoken or written account of connected events; a story
- Given that the “numbers” fail to discriminate between dimensions of competence, does “narrative” comments provide additional insight in resident performance?

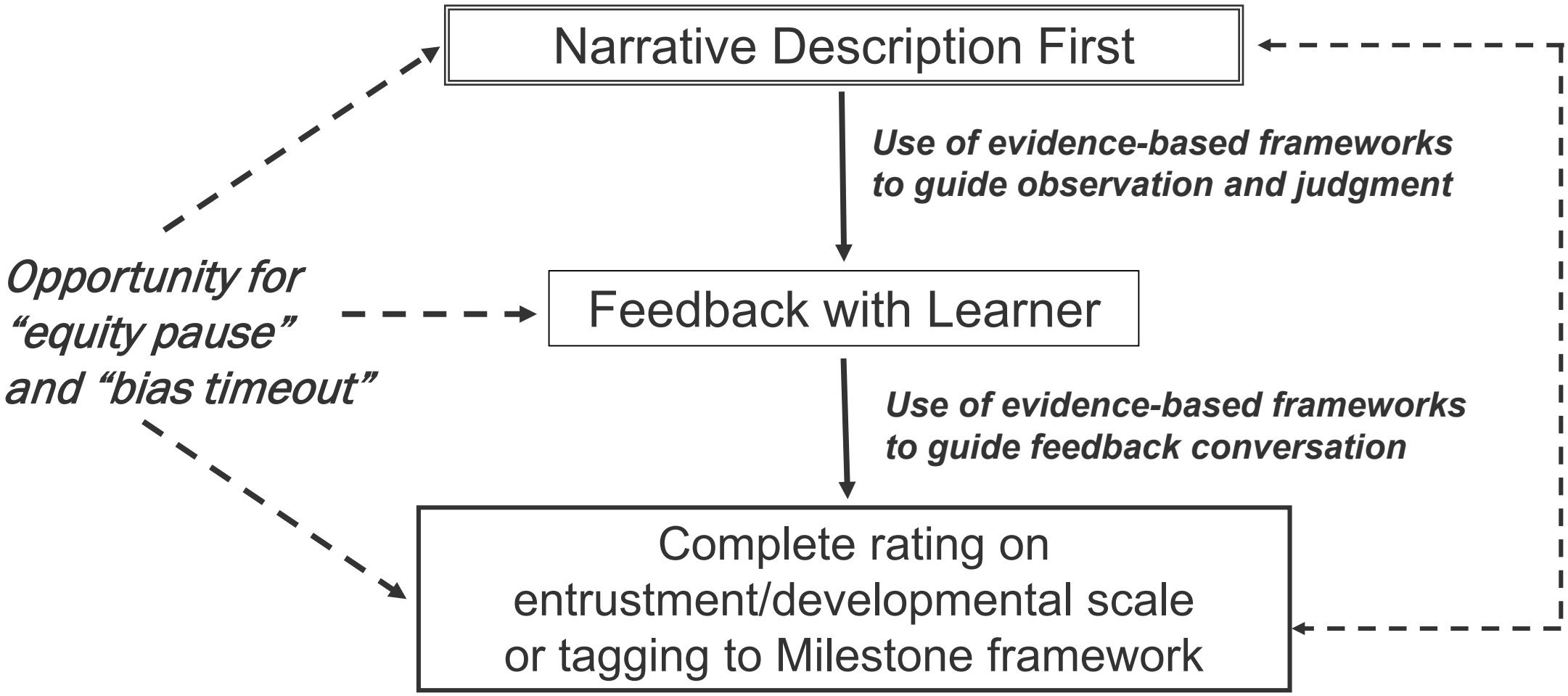
Narrative Assessment

The Hidden Value of Narrative Comments for Assessment: A Quantitative Reliability Analysis of Qualitative Data

Shiphra Ginsburg, MD, MEd, PhD, Cees P.M. van der Vleuten, PhD, and Kevin W. Eva, PhD

- *"Using written comments to discriminate between residents can be extremely reliable even after only several reports are collected. This suggests a way to identify residents early on who may require attention. These findings contribute evidence to support the validity argument for using qualitative data for assessment."*
- Reliabilities > 0.8 when 4 assessor's comments were included.

Re-thinking the Assessment Process



Faculty Development

- Methods of assessment are largely based on ***observations and questioning***
    - Faculty are the measurement instrument and they need training
    - Shared mental models essential
- Faculty development approaches need to use both “bolus” (e.g. workshop) and the “drip” (e.g. short aliquots of deliberate practice)

Questions and Discussion

eholmboe@intealth.org