BUILDING TRUST IN ASSESSMENT DESIGN AND SCORE REPORTING

Linette P. Ross, PhD
Richard Feinberg, PhD

NERA Webinar: December 6, 2023

THE POWER OF TRUST IN ASSESSMENT

Linette P. Ross
12/6/23

NBME®

WHY THIS TALK?

Learning Objectives:

Help understand the power of trust in assessment design and score reporting

- What is trust/ trustworthiness?
- Why trust is important at the individual and organizational level?

Understand the relationship between high-quality assessments and the cornerstones of trustable assessments

Gain insights on how to build and foster trust in assessments through design, transparency, and fairness that instill confidence in stakeholders at all levels

What's trust got to do with it?

- Care
- Sincerity
- Reliability
- Competence

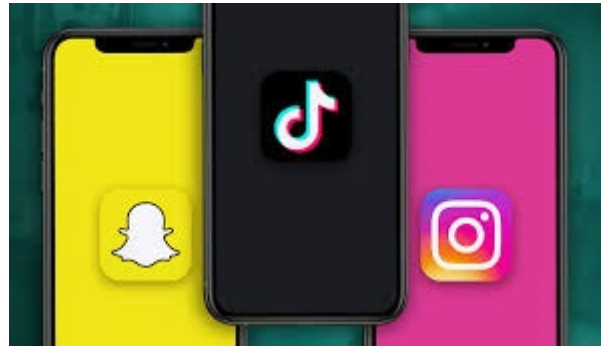
"Trust is fragile. Handle with care."

MOST TRUSTED BRANDS IN THE US IN 2023

Google

FedEx

YouTube



ups

Cash App

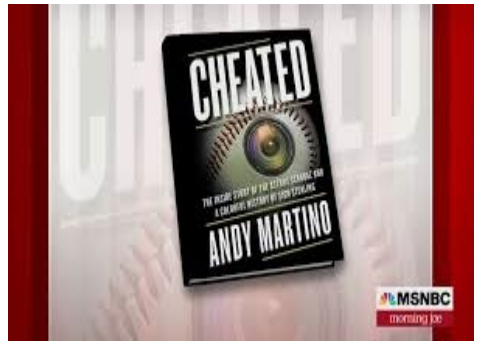
NETFLIX

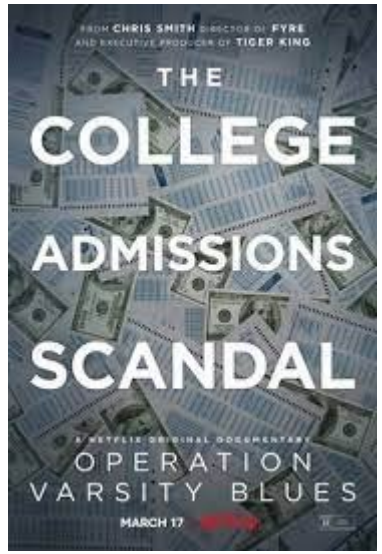
The Weather Channel

- The State of Consumer Trust: Morning Consult's Most Trusted Brands 2023 https://pro.morningconsult.com/analyst-reports/most-trusted-brands-2023
- Gen Z's Favorite Brands Report 2022. Morning Consult Pro, September 2022. https://morningconsult.com/gen-z-favorite-brands-2022/

TRUST FAILURES

What do we need to do to build trust again?





Consumers trust online reviews most, but new research finds a third of Amazon book, baby products, large appliances, computers & women clothing reviews are fake.

WHAT IS TRUST / TRUSTWORTHINESS?

Trust is a firm belief in the **reliability,** truth, **ability**, or strength of someone or something. (Oxford)

Trust is (Cambridge dictionary)

- **the belief that you can trust someone or something:**
- **to have confidence in something, or to believe in someone**

*When a person or organization is in a **position** **of** trust it comes with responsibilities, (especially to the public).*

Trustworthy / Trustworthiness: able to be relied on as honest or truthful.; able to be trusted (Cambridge dictionary)

Trustworthiness: Organizational Trust

Why is it important to trust organizations and the assessment results they provide?

Common Values: Do we share common values and beliefs?

Aligned Interests: Does the organization care about my welfare?

Benevolence: Does the organization care about my welfare?

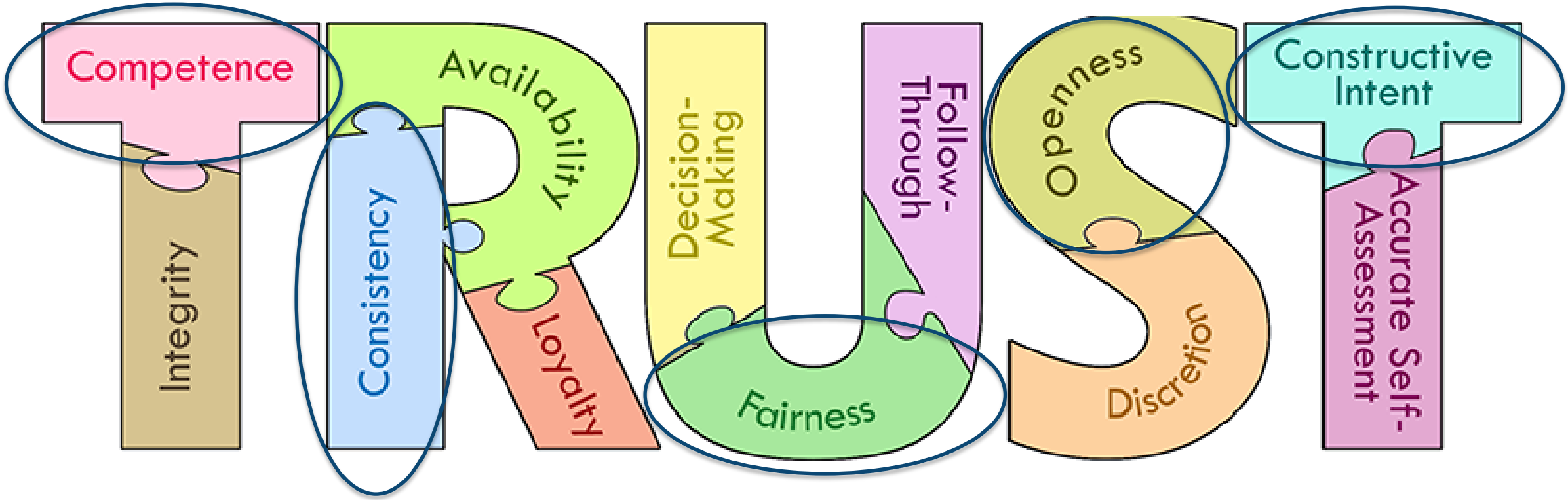
Competence: Is the organization capable of delivering on commitments?

Integrity: Does the organization abide by commonly accepted ethical standards (equity/fairness)?

Communication: Does the organization listen and engage in open and mutual dialogue?

Adapted from: **Robert F. Hurley**, Nicole Gillespie, Donald L. Ferrin, and Graham Dietz. Designing Trustworthy Organizations. *MIT Sloan Management Review*, 2013 June, 54 (4), 74-82.

12 DIMENSIONS OF TRUST



The keys to trust in assessment: Integrity, Credibility, Reliability, Fairness (Equity), Transparency, and Validity

Calvert, D. (2022). More than Honesty & Integrity! Know the 12 Dimensions of Trust. People First Productivity Solutions.

WHAT DO EDUCATORS WANT IN ASSESSMENTS?

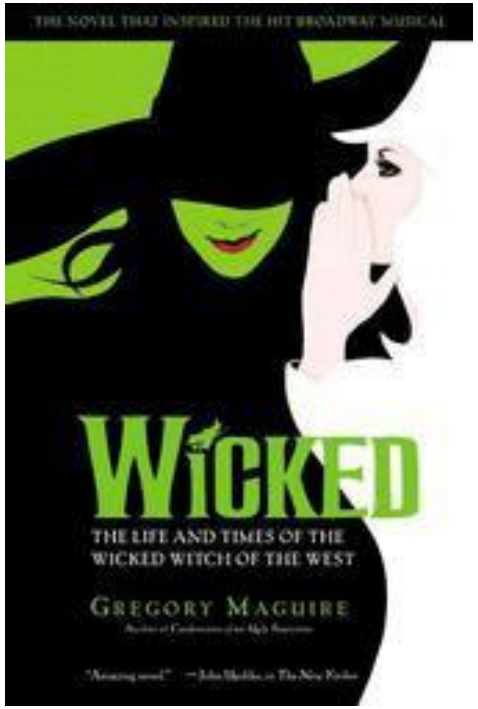
Top Factors When Evaluating Assessments

- Reliable (91)%
- Valid (91%)
- High-quality (90%)
- Alignment with state standards, district scope, and sequence (87%)

Trustable Assessment Results Matter!

Reliability and Validity are Keys to Trust!

MEDICAL EDUCATION'S WICKED PROBLEM: ACHIEVING EQUITY IN ASSESSMENT FOR MEDICAL LEARNERS



Intrinsic equity - selection and design of assessment tools

Contextual equity - fairness in the learning experiences and environment in which assessment occurs

Instrumental equity - uses of assessment data for learner advancement and selection and program evaluation

Consider these components when determining if the process and assessment outcomes support equity and fairness?

Lucey C.R., Hauer K.E., Boatright D., Fernandez A. Medical education's wicked problem: Achieving equity in assessment for medical learners. Academic Medicine 2020;95(12 suppl):S98–S108

CRITERIA FOR HIGH-QUALITY ASSESSMENT

- **Validity or Coherence** – assessment measures what it intends to measure
- **Reproducibility or Consistency** - assessment yields the same results (reliability)
- **Equivalence** – information is used similarly across settings
- **Feasibility** – practical to implement
- **Educational Effect** – methods motivate learners
- **Catalytic Effect** – effects of results on learners
- **Acceptability** – assessment tools are credible

Assessment outcomes are only useful and valid if users trust them for decision-making.

When designing assessments, reliability and validity are the keys to trust.

Is psychometric rigor enough ?

Norcini, J, Anderson, B, Bollela, V, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Medical Teacher, 2011, 33:206-214.

STANDARDS FOR TRUSTABLE ASSESSMENTS

*When a person or organization is in **a position of** trust it comes with responsibilities, especially to the public).*

STANDARDS for Educational and Psychological Testing. (2014). AERA, APA & NCME.

Validity
the assessment measures what it intends to measure

- Intended interpretations of test scores
- Forms of validity evidence
- Test construction
- Score reliability
- Accurate scoring

Reliability/Precision
the reproducibility and consistency of test scores

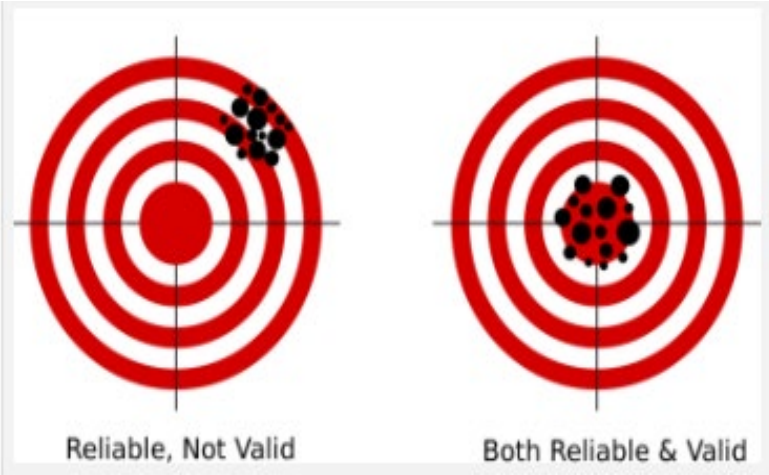
- Reliability coefficients
- Standard errors of measurement
- Decision consistency and accuracy

Fairness In Testing
fair and equitable treatment of all test takers

- Lack of measurement bias
- Access to the construct
- Minimize construct irrelevant components
- Valid interpretation of test scores

"Test and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population." (4.0)

Reliability and Validity: What's Trust Got To Do With It?



What is reliability?

the reproducibility and consistency of the data

What is validity?

the assessment measures what it intends to measure

the interpretation of the scores or assessment outcomes are for its intended use or purpose

Assessments must be **reliable** AND **valid** to be trusted.

Assessment outcomes are only useful and valid if users trust them for decision-making.

When designing assessments, reliability and validity are the keys to trust.

Reliability and validity matters!

INSTRUMENTAL EQUITY: BUILDING TRUST THROUGH TRANSPARENCY

Validity – what are some unintended outcomes and consequences of assessment results?

- Step 1 is a summative assessment that assures competency based on a criterion-referenced standard
- Designed to be pass/fail and its primary purpose was for medical licensure
- Numeric scores for the exam which was designed to ensure minimal competency were being used for residency selection

How do we build trust when the assessment tools are not used for its intended purpose?

USMLE Step 1 change to pass/fail

- In January 2022, USMLE changed from reporting a 3-digit numeric score to pass/fail
- Change reflects the intended purpose of Step 1: to assess minimal competency in medical knowledge in the basic sciences
- Addresses the inequity of using Step 1 for unintended purposes such as residency selection, while assuring minimal competency

How can we reduce bias, assure fairness, and build trust in assessments?

- Responsible test development and design
- Address misuse of scores for unintended purposes
- Hold users accountable for valid score interpretations and uses
- Apply psychometric rigor – fairness, reliability, and validity

INSTRUMENTAL EQUITY: BUILDING TRUST THROUGH TRANSPARENCY

One example: USMLE's decision to move to pass/fail for Step 1 to demonstrate minimal competency.

Some questions to ask:

- How are assessment results and data shared and used by stakeholders?
- How is the data used for learner advancement, progression, selection and evaluation?
- Do the assessment results over- or under-predict performance for some groups?
- When are population differences on assessments evidence of bias or inequity?

Other Ways to Achieve Instrumental Equity

- Advocate for structures and processes that support instrumental equity and equity in assessment outcomes.
- Strive to achieve the criteria listed for high quality assessment
- Hold institutions accountable for using assessment data for its intended purpose

FRAMEWORKS FOR BUILDING TRUST IN ASSESSMENTS

Apply criteria for high quality assessment **(Norcini, et.al, 2011)**

Effective summative assessment

- Criteria: validity, reliability (reproducibility) and equivalence are paramount
- Psychometric rigor will always be important to ensure trust in the decision-making process

Effective formative assessment

- Criteria: validity or coherence, feasibility, catalytic effect, and educational effect
- Provide useful and actionable feedback embedded in the process, on-going, timely & tailored

Desired outcome: Acceptability - are the assessment tools credible, acceptable and trustworthy?

Results and process indicators that build trust and indicate equity in assessment:

- Assessment procedures are fully aligned
- Assessment data is used for its intended purposes
- Programs routinely investigate issues of validity, fairness, and equity in their programs

Can I trust the process? Is the test developer trustworthy?

THANK YOU!

LINETTE P ROSS

LROSS@NBME.ORG

CONSIDERING TRUST IN SUBSCORE REPORTING

Rich Feinberg
NBME
12/6/2023

SCORE REPORTING IS A COMMUNICATION TOOL

Help users answer important questions about their performance:

- **Relative to other groups or standards**
- **Feedback for improvement**

Build trust to the extent that users can:

- **Understand the meaning and limitations of the information provided**
- **Take appropriate action**

SCORE REPORTING IS A COMMUNICATION TOOL

"Score reports are intended to provide stakeholders with the information they need, in a way that they understand, so that they may reasonably act on that information"

D. Zapata-Rivera (Ed.). (2019). *Score reporting research and applications (The NCME Applications of Educational Measurement and Assessment Book Series).* New York, NY: Routledge.

CollegeBoard SAT

SAT Score Report

Imagood Student
100 Main Street
Apt 2
Anytown, MA 00000-0000

Test Date: **March 13, 2021**
Registration Number: **0123456789**
Sex: **Female**
Date of Birth: **Feb. 12, 2004**
Test Center Number: **12345**
CB Student ID: **12345678**
High School Code: **123456**
High School Name: **John F. Kennedy High School**

Your Total Score

1010 | *400–1600*

50th Nationally Representative Sample Percentile

41st SAT User Percentile

Section Scores

490 | *200–800*
Your Evidence-Based Reading and Writing Score

44th Nationally Representative Sample Percentile
35th SAT User Percentile

You've met the benchmark!

520 | *200–800*
Your Math Score

57th Nationally Representative Sample Percentile
49th SAT User Percentile

You've scored below the benchmark.

Test Scores

22 | *10–40*
Reading

27 | *10–40*
Writing and Language

26.0 | *10–40*
Math

Cross-Test Scores | *10–40*

24
Analysis in History/Social Studies

23
Analysis in Science

Subscores | *1–15*

6 Command of Evidence	**9** Words in Context	**8** Expression of Ideas	**9** Standard English Conventions
8 Heart of Algebra	**9** Problem Solving and Data Analysis	**7** Passport to Advanced Math	

Get your full report online at sat.org/scorereport

Am I on Track for College?

Look for the green, yellow, or red symbols next to your section scores. They let you know if your scores are at or above the benchmark scores. Benchmarks show college readiness. If you see green, you're on track to be ready for college when you graduate.

If you score below the benchmark, you can use the feedback and tips in your online report to get back on track.

Benchmark scores:

Evidence-Based Reading and Writing: 480

Math: 530

How Do My Scores Compare?

A percentile shows how you scored, compared to other students. It's a number between 1 and 99 and represents the percentage of students whose scores are equal to or below yours.

For example, if your Math percentile is 57, that means 57% of test takers have Math scores equal to or below yours.

The Nationally Representative Sample Percentile compares your score to the scores of typical U.S. students.

SAT® User Percentile compares your score to the scores of students who typically take the test.

How Can I Improve?

To see which skills are your strongest and what you can do to boost your college readiness, go to your full report online and look for Skills Insight™.

What Are Score Ranges?

Test scores are single snapshots in time—if you took the SAT once a week for a month, your scores would vary.

That's why score ranges are better representations of your true ability. They show how much your score can change with repeated testing, even if your skill level remains the same.

Colleges know this, and they get score ranges along with scores so they can consider scores in context.

Your online score report shows your score ranges.

Boateng, Beatrice

SEPTEMBER 23, 2020

https://certs.duolingo.com/abc123

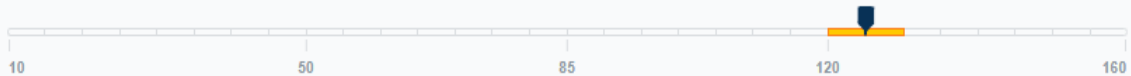
Overall

125

- Can understand a variety of demanding written and spoken language including some specialized language use situations.
- Can grasp implicit, figurative, pragmatic, and idiomatic language.
- Can use language flexibly and effectively for most social, academic, and professional purposes.

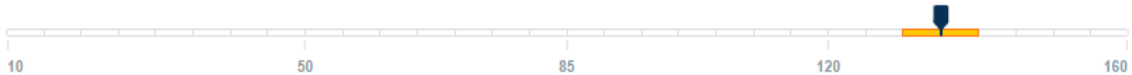
Subscores

Literacy Ability to read and write



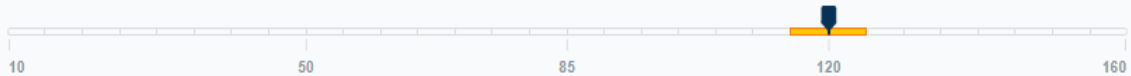
125

Comprehension Ability to listen and read



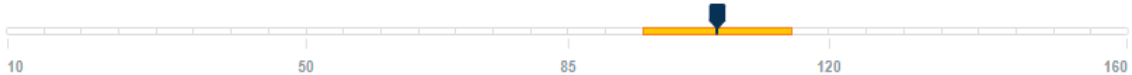
135

Conversation Ability to speak and listen



120

Production Ability to write and speak



105

Your score Your score range

What’s the problem with Subscores?

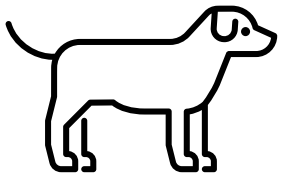
WHAT'S THE PROBLEM WITH SUBSCORES?

Often lack sufficient psychometric properties to be useful
(Folske, Gessaroli, & Swanson, 1999; Thissen & Wainer, 2000; Haberman, 2008; Sinharay, 2010; Feinberg & Wainer, 2014; Feinberg & Jurich, 2017)

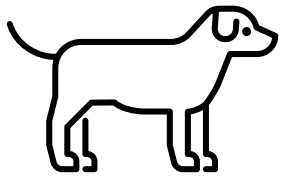
Poor Reliability:

Poor Validity:

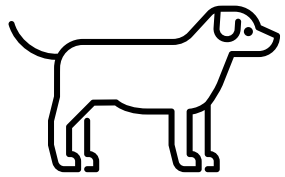
Dog

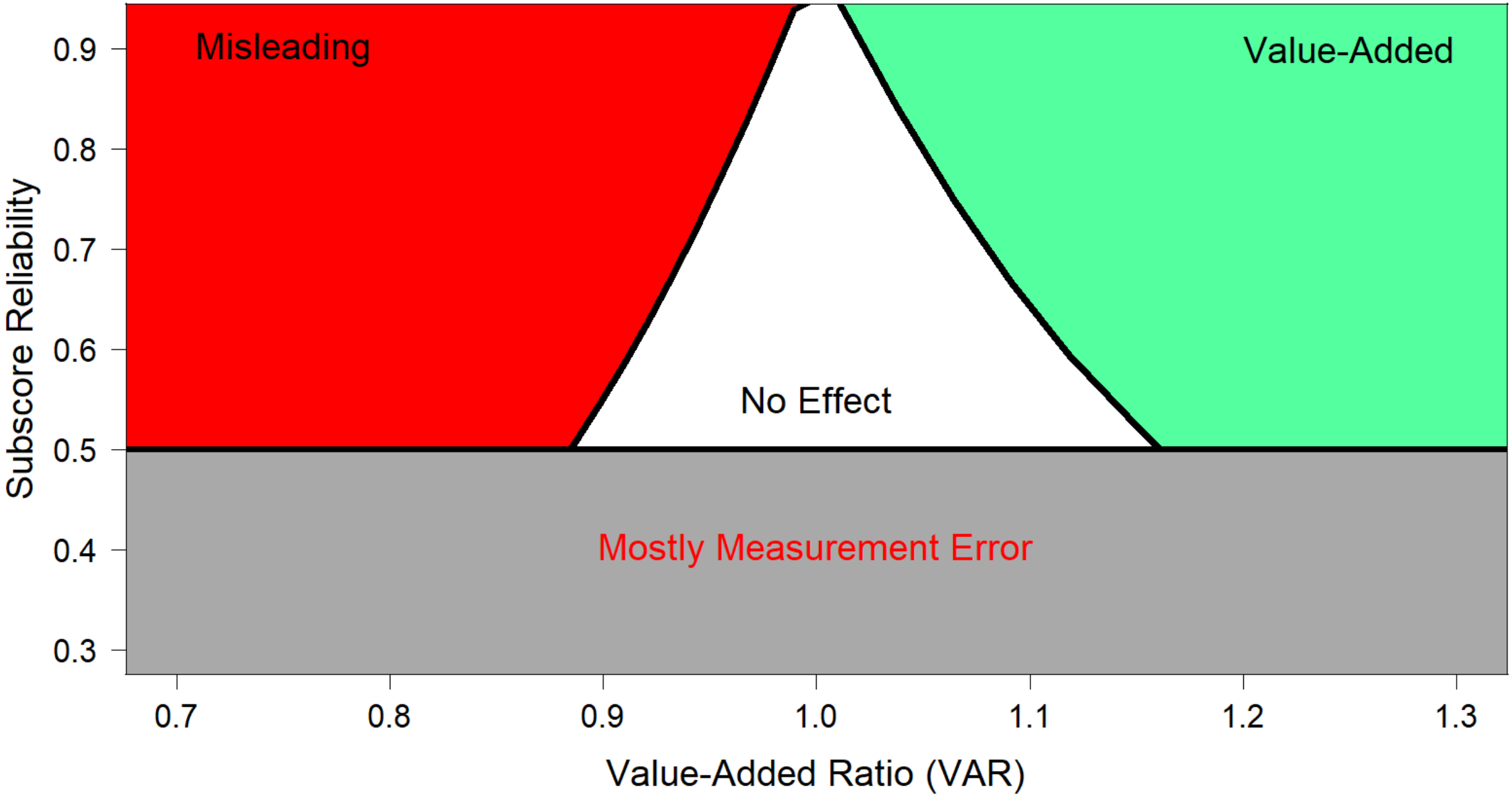


Canine

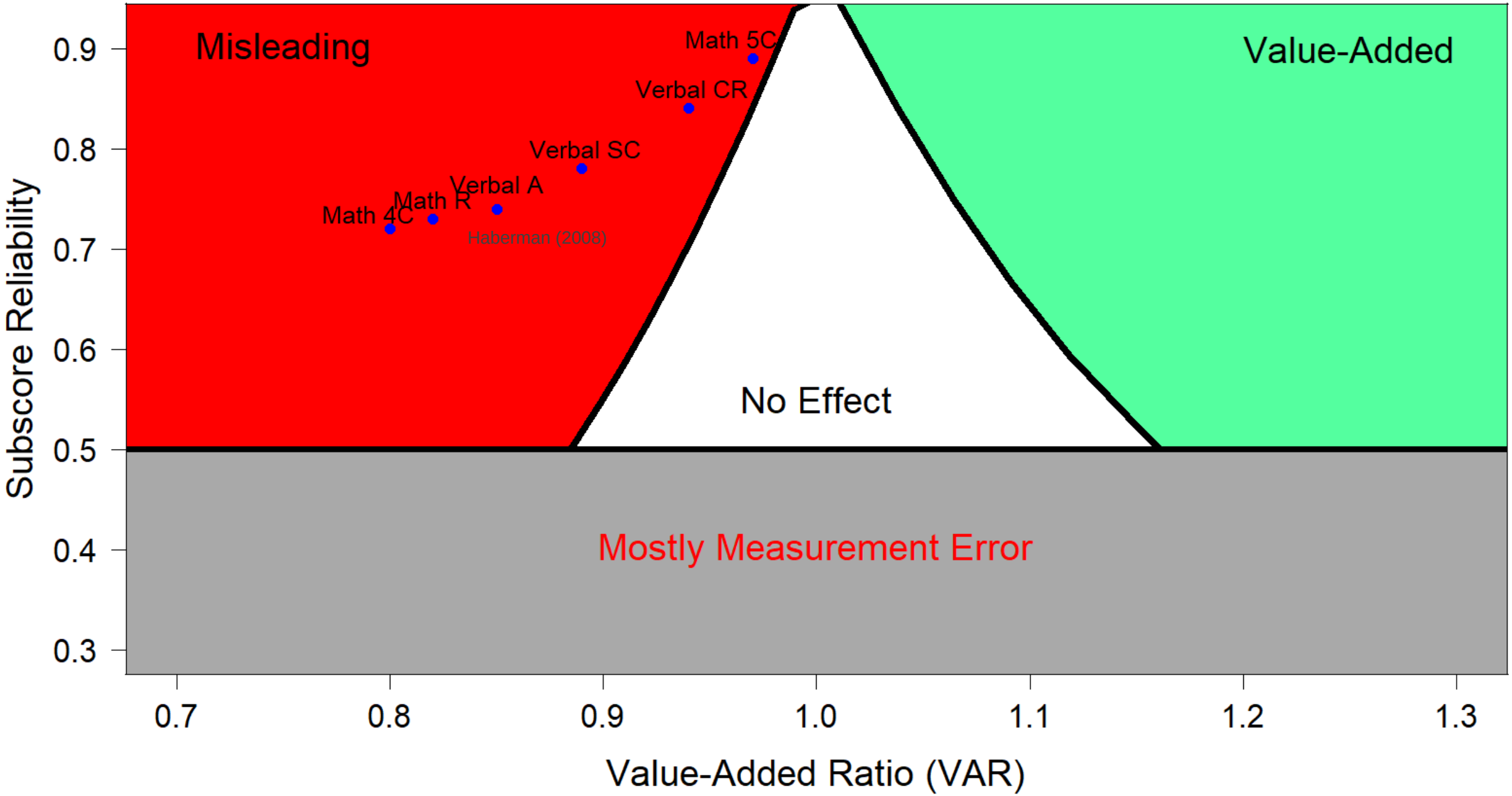


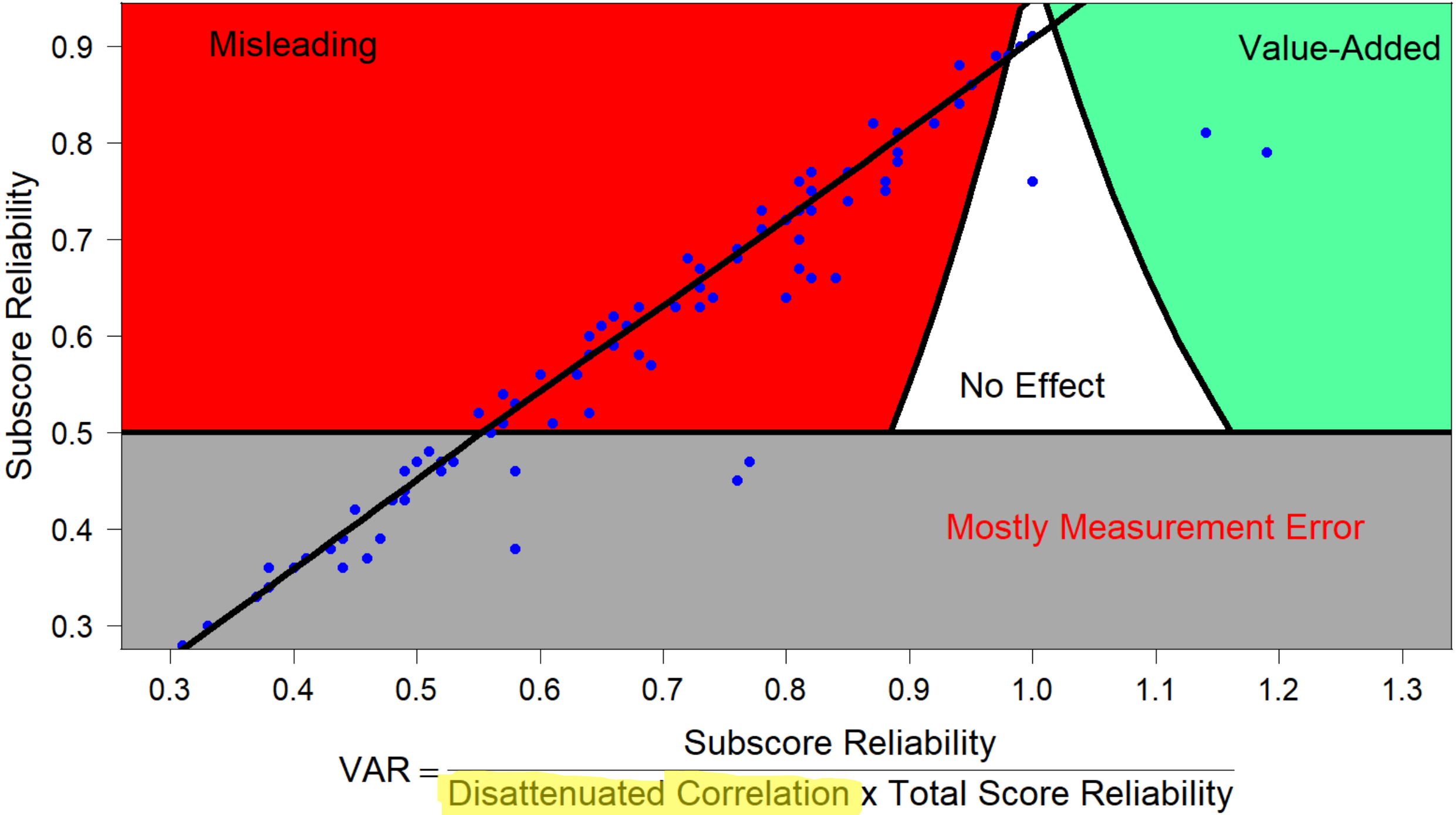
Pooch

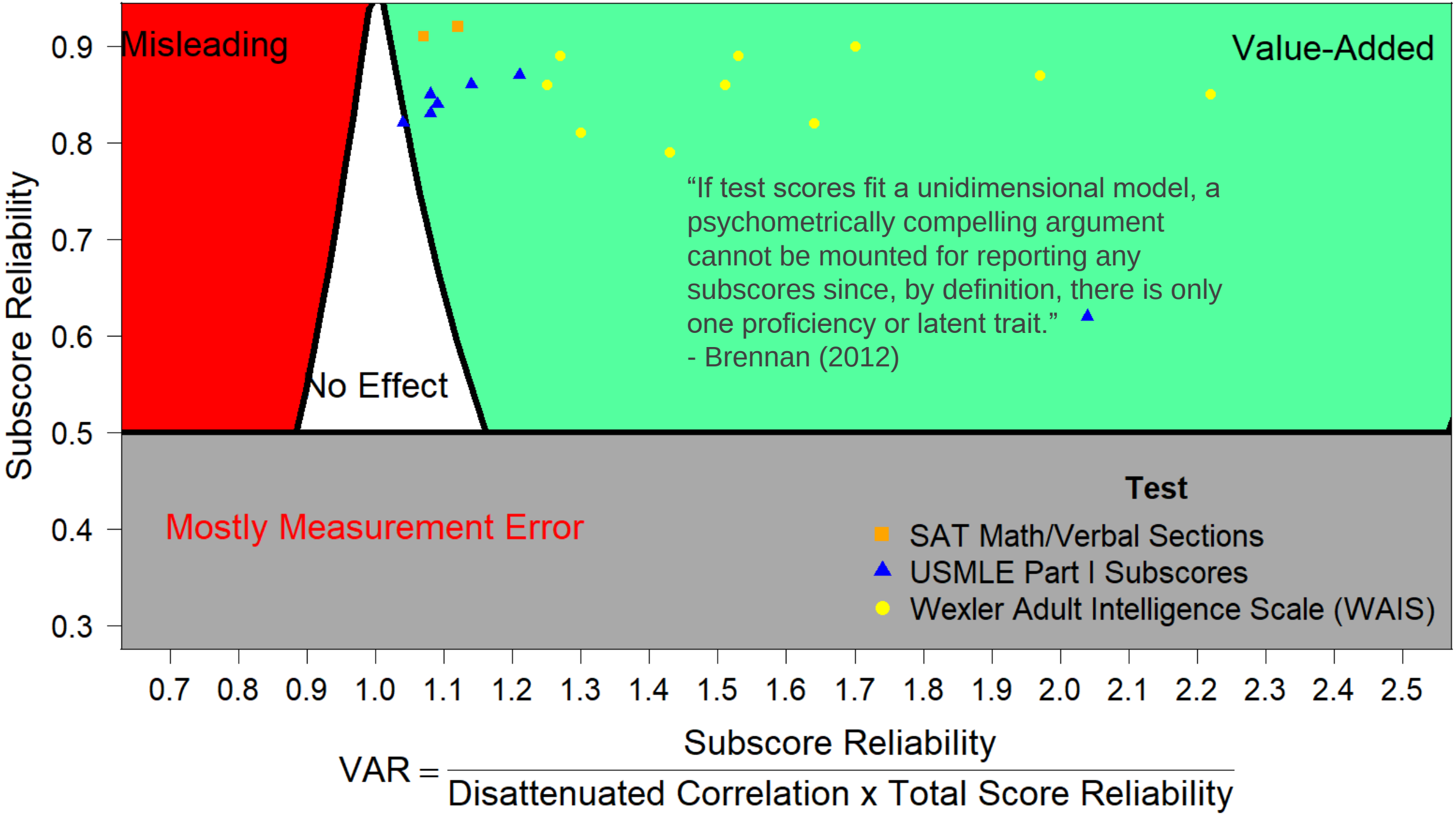


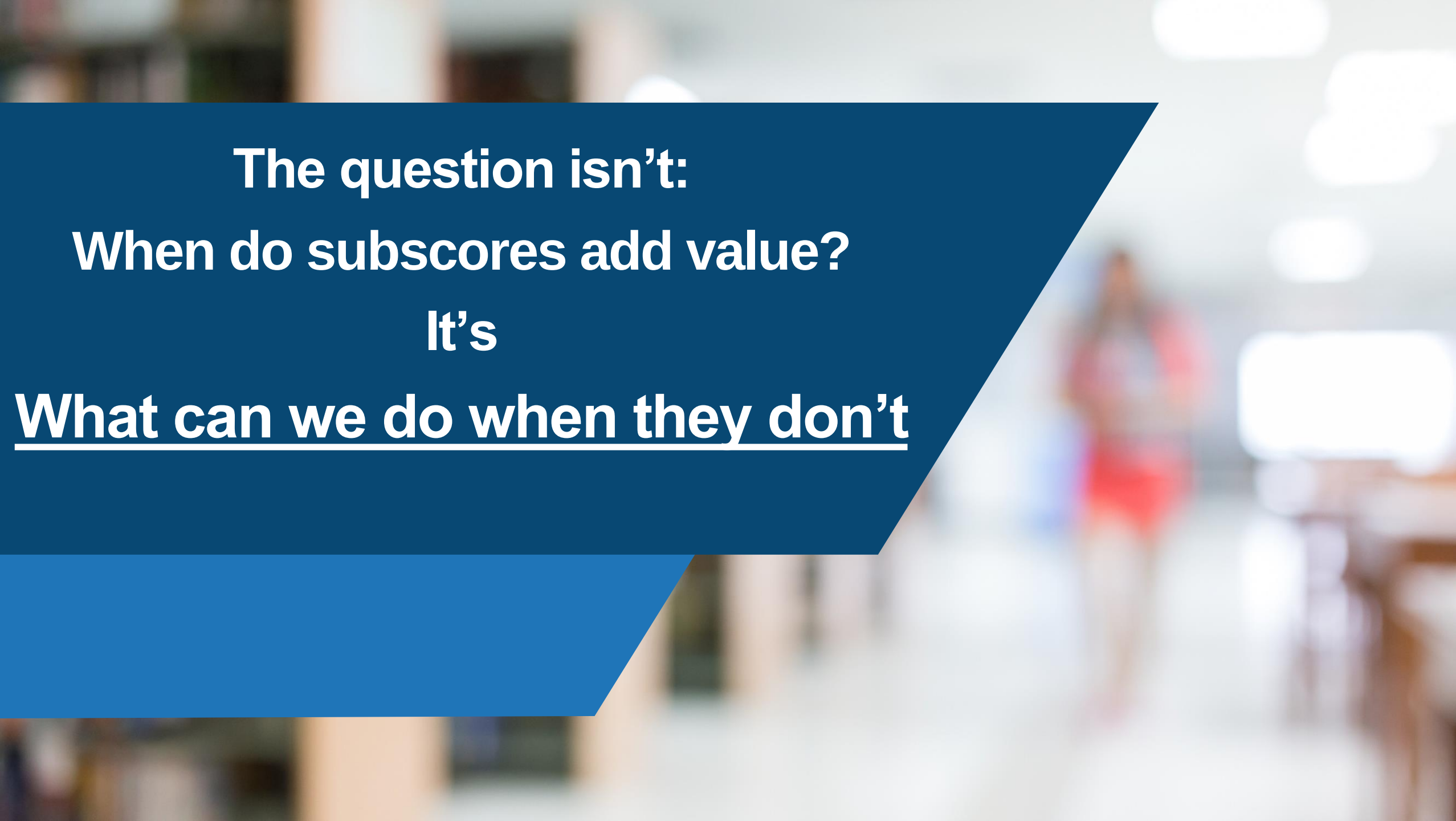


(Haberman, 2008; Feinberg & Jurich, 2017)


Misleading
Value-Added
No Effect
Mostly Measurement Error
Math 5C
Verbal CR
Verbal SC
Verbal A
Math R
Math 4C
Haberman (2008)
Subscore Reliability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.7
0.8
0.9
1.0
1.1
1.2
1.3
Value-Added Ratio (VAR)


Misleading
Value-Added
No Effect
Mostly Measurement Error
Subscore Reliability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
Subscore Reliability
VAR = Disattenuated Correlation x Total Score Reliability


Misleading
Value-Added
No Effect
Mostly Measurement Error
"If test scores fit a unidimensional model, a psychometrically compelling argument cannot be mounted for reporting any subscores since, by definition, there is only one proficiency or latent trait."
- Brennan (2012)
Test
SAT Math/Verbal Sections
USMLE Part I Subscores
Wexler Adult Intelligence Scale (WAIS)
Subscore Reliability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5
VAR = Subscore Reliability / Disattenuated Correlation x Total Score Reliability


The question isn't:
When do subscores add value?
It's
What can we do when they don't

OPTION 1: DON'T REPORT SUBSCORES

- ✓ **Avoids any misterpretation**
- − **Contractual obligations or risk confusing/angering stakeholders**
- **In 2014 the National Council of Bar Examiners (NCBE) eliminated the reporting of subscores on the Multistate Bar Exam. However, in response to negative stakeholder reaction MBE began providing some additional subscore information to failing candidates (*Pieper Bar Review*, 2017).**

OPTION 2: REDESIGN TEST BLUEPRINT

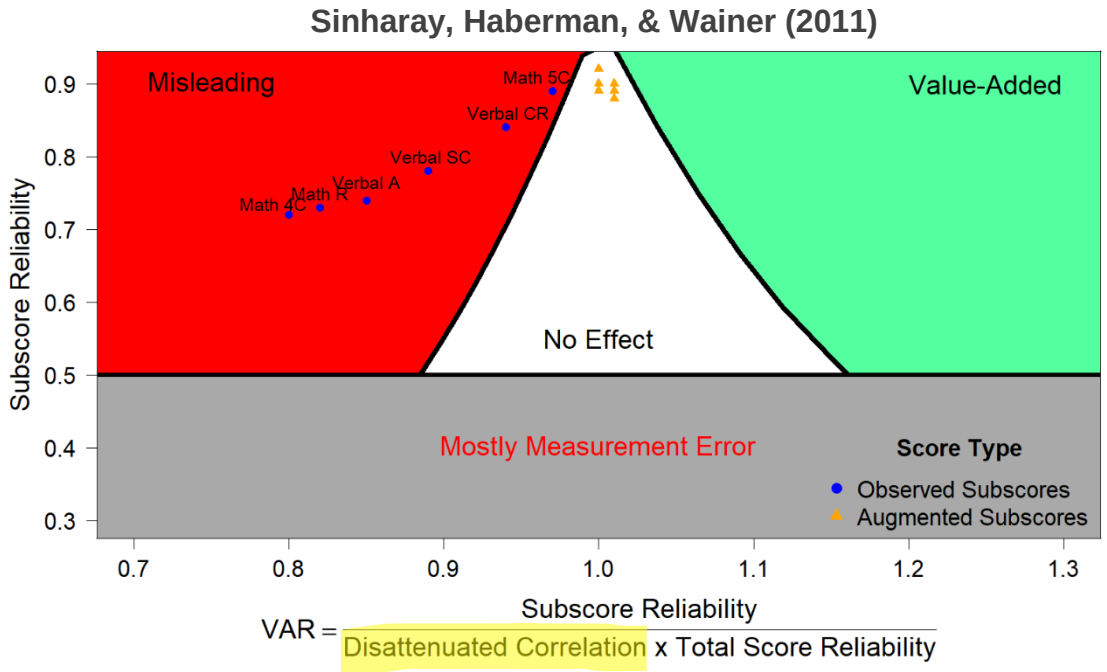
✓ **Use an Evidence Centered Design (ECD) approach where subscore inferences are planned in advance and factored into the test design process (e.g., content specification, item development)**

– **May be impractical**

- **A testing program may not have the resources to collect the necessary information (e.g., practice analysis)**
- **Not always straightforward to create good, targeted items**
- **Difficult to justify the expense when total score is fine**

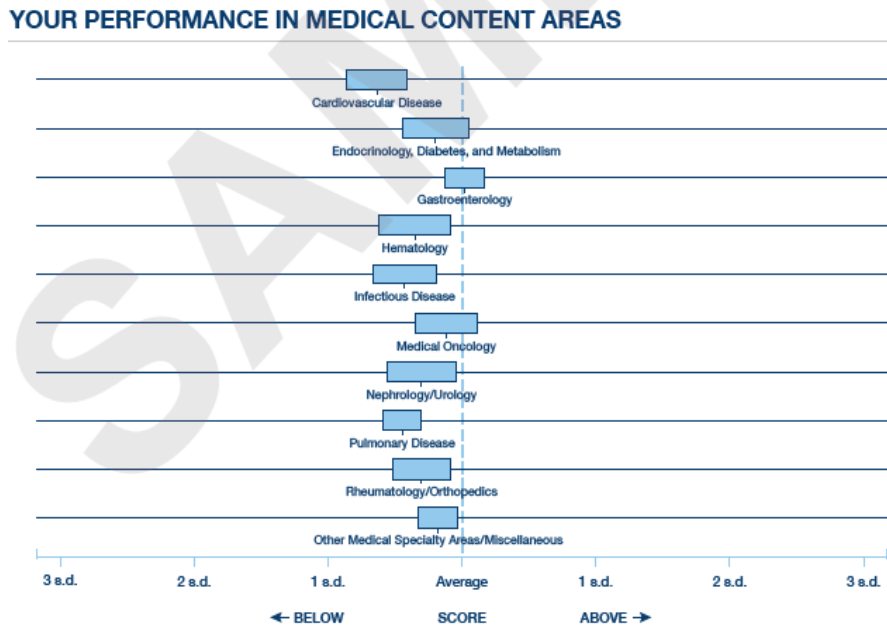
OPTION 3: IMPROVE EXISTING SUBSCORES

- ✓ **Add more subtest items or combining subtests of similar content areas to boost reliability**
- ✓ **Augment subscores to boost reliability**
- – **However, all these methods are unlikely to lead to value-added subscores**



OPTION 4: REPORT ACTUAL PERFORMANCE

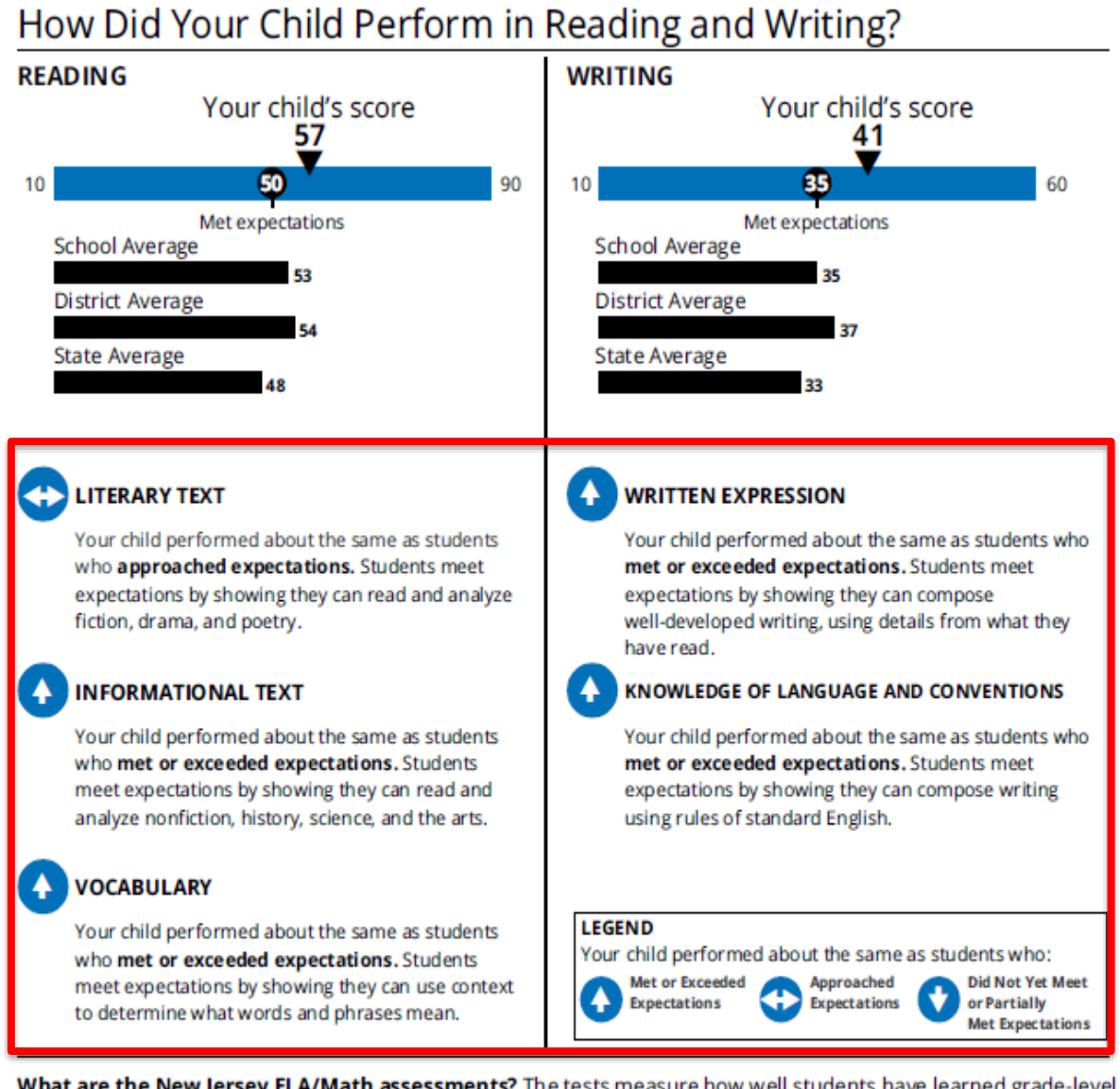
- ✓ **Data transparent and provide actual scores/profiles with SEM and interpretive language**
- – **However, research has suggested that SEM's or profile bands can be difficult to accurately interpret, even when detailed explanatory text is provided (Rick & Clauser, 2016).**



Content Area	Your Equated Percent Correct Score	SEM	Total Group Mean (SD)
Species			
Small Animal	32	3	70 (10)
Canine	32	5	71 (11)
Feline	32	5	70 (11)
Food Animal	34	5	66 (10)
Bovine	27	6	67 (11)
Porcine	47	10	63 (14)
Equine	31	6	68 (12)
Competency			
Clinical Practice	32	3	68 (9)
Communication	48	13	69 (16)
Preventive Medicine and Animal Welfare	35	11	65 (14)

OPTION 5: REPORT CATEGORICAL SUBSCORES

- **Loose information, power, and sensitivity (Royston, Altman, & Sauerbrei, 2006; Harrell 2008, Wainer, Gessaroli, & Verdi 2006)**
- ✓ **Discretization can help communicate results when less granularity is preferred for a broad audience (Gelman & Park, 2008)**
- ✓ **Research suggests categorical approach can be conservative to mitigate misinterpretation (Feinberg & von Davier, 2020; Feinberg, 2024)**



What *should* we do?

WHY DOES THIS MATTER?

Subscores are often included on summative (unidimensional) tests to support formatives inferences, helps to build trust

- **Identify individual relative strengths/weaknesses**
- **Recognize aggregate-level broader gaps in curriculum**
- **Inform plans for future study/preparation**

Can lead to the erosion of trust when this is not met

- **Future prep not in best interest of student**
- **Bad decisions across different levels of stakeholders**
- **Negative impression of testing program**
  - *"It takes 20 years to build a reputation and five minutes to ruin it. If you think about that, you'll do things differently."* – Warren Buffet

SCORE REPORTING IS A COMMUNICATION TOOL

Help users answer important questions about their performance:

- **Relative to other groups or standards**
- **Feedback for improvement**

Build trust to the extent that users can :

- **Understand the meaning and limitations of the information provided**
- **Take appropriate action**

BUILDING TRUST INTO SCORE REPORT DESIGN PROCESS

Promote trust in score reporting by working together with diverse stakeholder groups (e.g., surveys, focus groups, cognitive interviews)

- **Define the desired inferences that align with the test's purpose**
- **Being honest with assessment limitations (…ahem, subscores)**
- **Determine the type of score information and level of granularity that can best support the inferences and minimize misinterpretation**

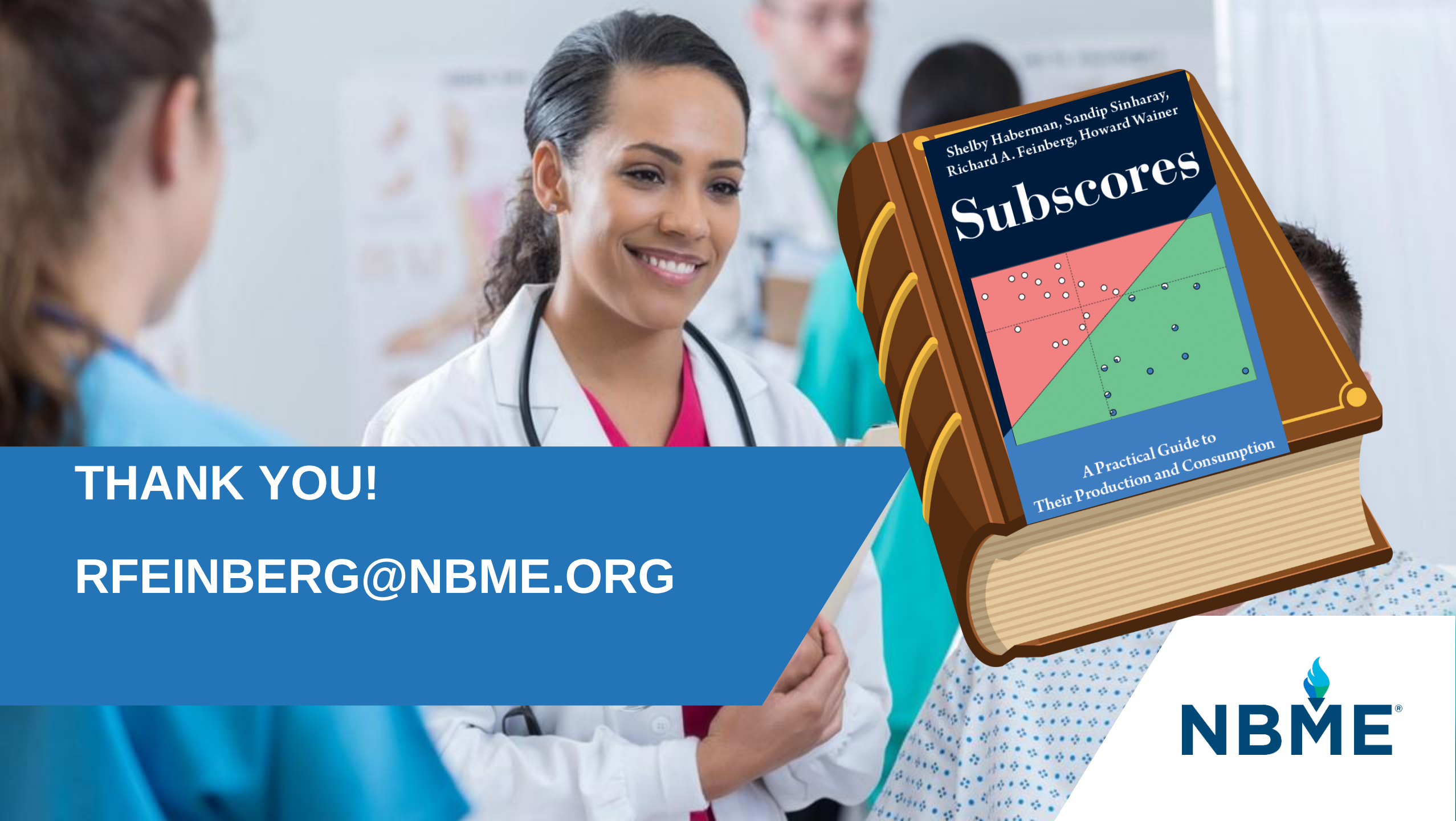
Build trust by appreciating the emotional interpretation of score results

- **Listening/seeking input**
- **Having a dialogue/sympathizing with potential outcomes**
- **Regular touch points to gather feedback and reassess design**

Demonstrate that you value trust by what you *do*

Whoever exercises mercy where strictness is required, will eventually be cruel where kindness is required

- Midrash Ecclesiates Rabbah 7.33



THANK YOU!

RFEINBERG@NBME.ORG

NBME®