MMS Information Session

# From Research to Reality: Translating Reliability and Validity Findings into Actionable Insights

Jeff Geppert, Battelle

Matt Pickering, Battelle

# Welcome

The purpose of CMS's Measures Management System (MMS) Information Sessions are to:

- educate about quality measurement
- promote a standard approach to measure development and maintenance
- encourage public involvement throughout the Measure Lifecycle

Partnership for Quality Measurement
Powered by Battelle

# From Research to Reality: Translating Reliability and Validity Findings into Actionable Insights

Matthew Pickering, PharmD | Battelle
Jeff Geppert, EdM, JD | Battelle

October 23, 2024

# Meet the Presenters

## Matthew Pickering| E&M Technical Lead

- Oversees endorsement & maintenance (E&M) processes
- 10+ years quality experience

## Jeffrey Geppert | Sr. Research Leader

- Leads Measurement Science team for E&M
- 27+ years measurement science, healthcare and quality experience



CBE: consensus-based entity

# Session Objectives

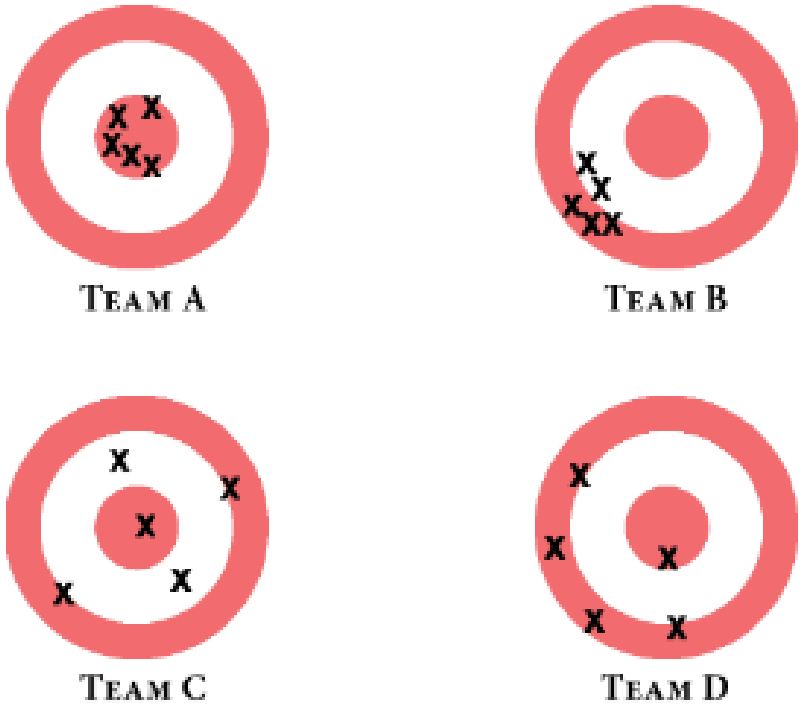
| | |
|---|---|
| **Purpose** | To discuss the *meaning* of entity-level reliability and validity claims and how to interpret some of the common approaches to substantiating such claims |
| **Agenda** | - Reliability vs. Validity<br>- Reliability<br>- Validity |

# Reliability vs. Validity

# Choice

- Choice-making is the agent that drives change and transformation
- The purpose of validity evidence is to establish a causal association between the person or entity response to the quality program and the measure focus
  - The person should respond to the quality program through *selection* of a better performing entity over a worse performing entity
  - The worse performing entity should respond to the quality program through *choice* to allocate resources to transform to a better performing entity
- "Should" causal claim:
  - Validity: There are known and effective ways of selection and choice that the person or entity *should* use . . .
  - Reliability: . . . and most of the variation in measure focus performance is attributable to variation in those ways (i.e., other possible explanations are "ruled-out")

# The Target Metaphor




- Which archery team should I select?
  - Team A – reliable and valid
  - Team B – reliable, but not valid
  - Team C – not reliable, but valid
  - Team D – not reliable, and not valid
- How should teams change-transform?
  - Team B - adjust bow sight, account for wind
  - Team C – build muscle memory, movements
  - Team D – standardize bow draw, stance, and release; progressive adjustment based on feedback

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Hachette UK.

# The Softball Hitting “Moneyball” Metaphor

**June**, **College**

Player A

**June**, **High School**

Player B


**March**, **College**

Player C

**March**, **High School**

Player D

Observe a batting average of 0.350

- Which softball player should I select?
  - Player A – reliable and valid
  - Player B – reliable, but not valid
  - Player C – not reliable, but valid
  - Player D – not reliable, and not valid
- How should players change-transform?
  - Player B – hitting in a variety of contexts
  - Player C – stance, swing mechanics, timing
  - Player D – focus on fundamentals of hitting, expand experiences to different types of pitchers, stadiums, weather

# Randomness

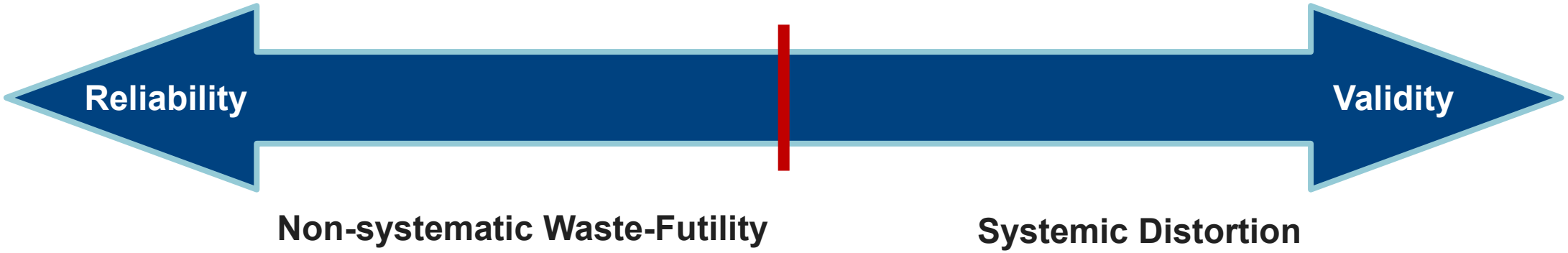
- Reliability is not the same as “randomness”
- “Randomness” in metrics of reliability (e.g., signal-to-noise) is a modeling assumption for the representation of model parameters, not a representation of reality
  - The only truly “random” phenomena found in nature are quantum mechanics and radioactive decay
- Moreover, associating reliability with “randomness” contributes to the perception of burden of quality measurement as somehow “unfair”
- Our well-meaning efforts to “account for” or reduce noise inadvertently led us to disregard essential signals, causing us to miss patterns, gaps, and perspectives that deserve our attention (Schmidt, 2020)

# Reliability

# What is Reliability?

"Validity is measuring the right thing; reliability is measuring the thing right"
– Thissen (2001)

Reliability is the degree to which a measure repeatedly and consistently produces the same result - ISO/IEC 25020 — Quality measurement framework (2019)



# Accountable-entity Reliability Calculation

- Reliability metric is generally calculated as:

$$\frac{\sigma^2_{Between}}{\sigma^2_{Within} + \sigma^2_{Between}}$$

- Reliability metric is high when $\sigma_{Between}$ is large relative to $\sigma_{Within}$
- Main methods of calculating reliability metric by entity
  - Binary (proportion): Signal-to-noise; beta-binomial (Adams; Rand 2009)
  - Complex (ratio, risk-adjusted): Intra-class correlation (ICC) with Spearman-Brown prophecy formula adjustment
  - Hierarchical ("borrowing strength"): mixed logistic regression, empirical bayes
    - Increases reliability by shrinking to a target, potentially decreases validity

# What is Good Reliability?

- Reliability is a feature of the entity(s), not the measure (CBE threshold 0.6)
- Reliability is *associated* with the likelihood of misclassification
  - Classifying the entity as worse when better and as better when worse
- Equity considerations if certain populations are more likely to receive care from low reliability entities
- A median reliability metric lacks meaningfulness and transparency
  - High reliability of some entities does not “offset” the low reliability of others
- Reliability and stability are related but not identical
  - The relation depends on whether the factors driving high reliability are systematic and/or persistent (e.g. a high performing surgeon leaves)

# Example: Reliability by Deciles

**Table 1.** Signal-to-Noise Estimates of CBE #4125 - Thirty-day Risk-Standardized Death Rate among Surgical Inpatients with Complications (Failure-to-Rescue)

| | Overall | Min | Decile 1 | Decile 2 | Decile 3 | Decile 4 | Decile 5 | Decile 6 | Decile 7 | Decile 8 | Decile 9 | Decile 10 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reliability** | 0.7039 (mean) | **0.2314** | **0.2571** | **0.3248** | **0.3879** | **0.4671** | **0.5379** | 0.6016 | 0.6697 | 0.7384 | 0.8106 | 0.8861 | 0.973 |
| **N of Entities** | 2055 | 21 | 205 | 206 | 206 | 206 | 205 | 205 | 205 | 206 | 205 | 206 | 1 |
| **N of Persons/ Encounters/ Episodes** | 1087624 | 525 | 15853 | 21776 | 29419 | 40024 | 53027 | 69384 | 92901 | 129893 | 199744 | 435603 | 8099 |

- Majority (51%) of entities have a reliability <0.6 (i.e., most of the variation in measure performance is not attributable to variation in known and effective ways)
- Developer/steward should consider mitigation for entities with low reliability estimates

# Mitigation of Low Reliability

- Low reliability does not mean the measure is not useful. It depends on the alternatives and the consequences.
- The closer the measure focus is on the continuum of “should not happen” the less consideration to reliability.
  - Reasoning “rules-out” other explanations for the association between the response to the quality program and the measure focus
- Efforts to mitigate the harm due to low reliability must be balanced against other considerations, such as reduced validity or quality program participation:
  - e.g., increasing the minimum sample size (participation) or the period of performance (validity)
- Actual harm depends on the decision context (features of quality program).
  - e.g., the threshold or benchmark, low volume payment adjustment, voluntary reporting

# Mitigation of Low Reliability, *continued 1*

- Start by understanding the features of the entities and persons at low-reliability and high-reliability entities

| Features | Low-reliability entities | High-reliability entities |
|---|---|---|
| Person-level features | | |
| Entity-level features | | |
| Average person-level features by entity | | |
| Geographic features | | |

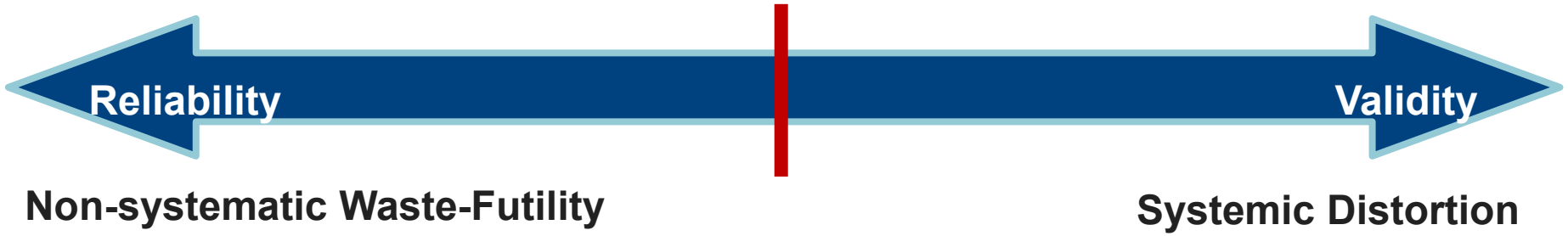
- Low reliability might be acceptable in certain contexts:
  - Measure focus with a direct causal association (e.g., complications, HAI)
  - No alternative for choice-making. A low reliability measure is usually better than no measure (e.g., Farmer's Almanac)
  - A related high reliability measure to inform the choice (structure, process, or outcome)
  - Trade-offs in harm favor low reliability over low validity or lower quality program participation

# Validity

# What is Validity?

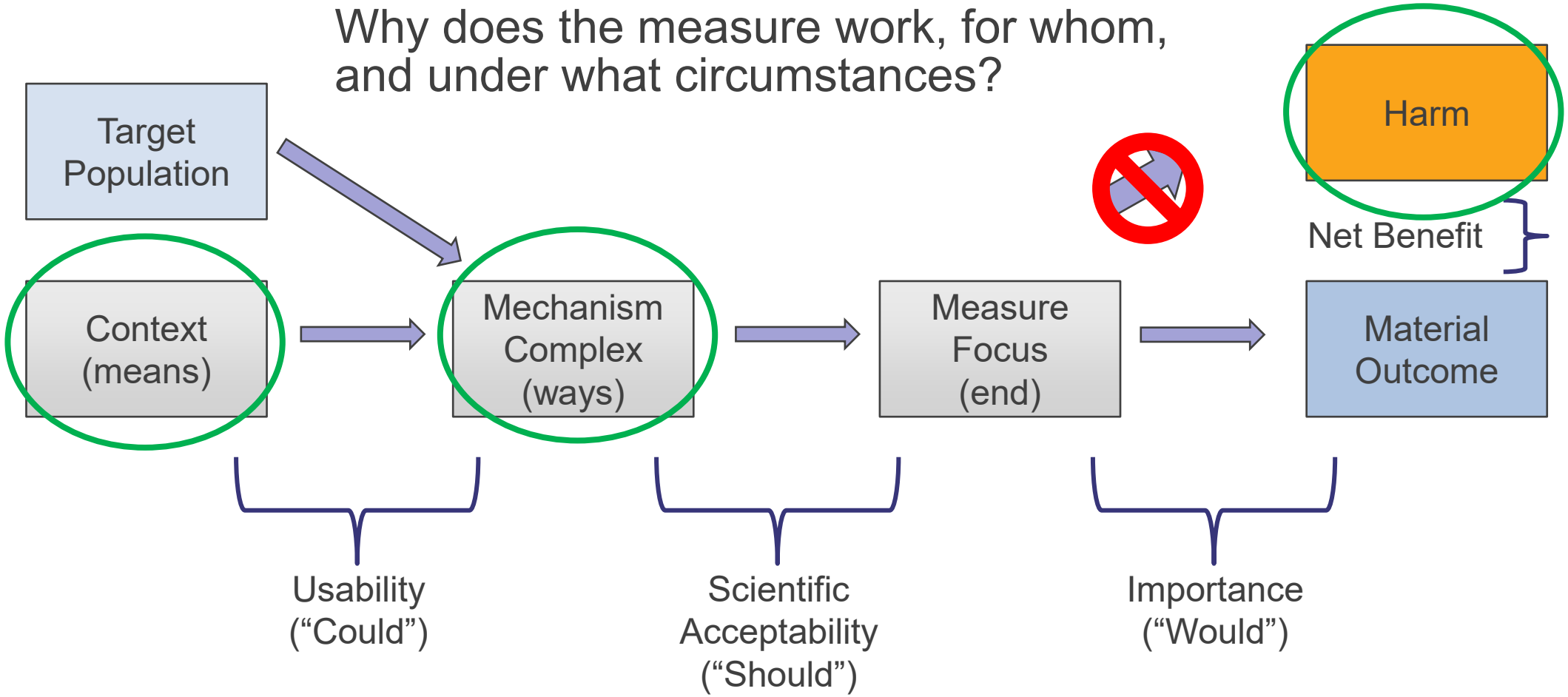
"Validity is measuring the right thing; reliability is measuring the thing right"
– Thissen (2001)

Validity is "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy of appropriate interpretations and actions on the basis of [the measure]"
– Messick (1989); Standards for Educational and Psychological Testing (2014)



# Validity and Purpose: What change/transformation are you trying to make?

Why does the measure work, for whom, and under what circumstances?



# Validity Claims: "Should"

Measure developers and/or measure stewards make certain explicit or *implicit* assertions or claims about the potential benefits and risks/harms associated with measure use (net benefit).

In general, there are three top-level claims related to measure properties necessary for a measure to yield positive net benefit to persons and entities:

| | |
|---|---|
| ***Would* claim:** | Person or entity *would* make decisions based on the measure because the measure focus is associated with a material outcome (end/importance). |
| ***Should* claim:** | There are known and effective ways of selection or choice that the person or entity *should* use (ways/scientific acceptability).<br>▪ Known: mechanism; effective: causal |
| ***Could* claim:** | Any barriers or facilitators to whether the person or entity *could* use those ways are known and addressed (means/usability). |

# Validity Claims: Misconceptions

- Absence of evidence of validity is not evidence of the absence of validity
  - Validity may not have been established, but that does not mean that lack of validity has been established
- Validity is not a (universal) property of the measure
  - The question is not “does the measure work” but “why does the measure work, for whom, and under what circumstances” (Pawson and Tilley, 1997)
- There are not different “types” of validity (e.g., concurrent, predictive, discriminant)
  - Rather there is one universal type (construct) and different forms of evidence-arguments
- Establishing a validity claims requires evidence and an argument
  - Evidence can be experience, expertise, empirical, reasoning, simulation, engineering, etc.
  - Arguments are logical inferences about why the evidence supports the claim
    - i.e., deduction, induction, inference to the best explanation (IBE)

# Validity Claims: Theory and Evidence

- Begin with a theory about why a measure works (known and effective way), and use that theory to guide empirical or other investigations
- Use a logic model or concept map to articulate the theory
- Use your TEP for input on the logic model for buy-in and face validity
- Identify the claims that need to be substantiated with evidence and argument

| Inputs (Resources-Means) | Activities (What the program does-Ways) | Outputs (Direct results of the activities) | Outcomes | Impact (Broad, systemic changes influenced by the program): |
|---|---|---|---|---|
| • Skilled healthcare professionals (nephrologists, surgeons, nurses).<br>• Training programs for AVF placement and maintenance.<br>• Medical equipment and facilities for surgery and follow-up care.<br>• Patient education materials.<br>• Funding for healthcare initiatives.<br>• Support from healthcare policy and administration.<br>• Access to patient data and healthcare records for monitoring. | • Early screening and identification of patients for AVF.<br>• Preoperative vascular mapping to assess suitability for AVF.<br>• Surgical creation of AVF.<br>• Postoperative monitoring and care for AVF maturation.<br>• Ongoing training and education for healthcare providers.<br>• Patient education and counseling about the benefits and care of AVF.<br>• Policy advocacy for supporting AVF use. | • Number of patients screened for AVF suitability.<br>• Number of AVFs surgically created.<br>• Number of healthcare providers trained in AVF-related procedures.<br>• Educational sessions conducted for patients.<br>• Policy changes or implementations supporting AVF use. | Short-term (Changes resulting from the outputs):<br>• Increased awareness among patients and healthcare providers about the benefits of AVF.<br>• Improved patient selection for AVF placement.<br>• Enhanced skills among healthcare providers for creating and maintaining AVFs.<br>• Improved patient readiness and compliance for AVF surgery.<br>• Policy and systemic changes facilitating increased AVF use.<br><br>Intermediate term (effects observed as the program matures)<br>• Increased rate of successful AVF placements.<br>• Reduced complications and failures in AVF post-surgery. | • Healthcare Policy and Funding<br>○ Program Influence: Advocacy and demonstrated success of the program can lead to changes in healthcare policies, prioritizing funding for AVF procedures and postoperative care.<br>○ Systemic Change: Shift in national or regional healthcare funding and policies to support early and efficient access to AVF for eligible patients.<br>• Standardization of Care Practices<br>○ Program Influence: Implementation of best practices for AVF creation and maintenance could set a benchmark for care quality.<br>○ Systemic Change: Adoption of these standards across healthcare systems, leading to a more uniform approach to hemodialysis vascular access.<br>• Training and Workforce Development<br>○ Program Influence: The focus on training and continuous education can highlight the need for specialized skills in nephrology and vascular surgery.<br>○ Systemic Change: Changes in medical education and professional development requirements, ensuring a well-trained workforce proficient in AVF management.<br>• Patient Education and Engagement<br>○ Program Influence: Comprehensive patient education initiatives can |

# Validity: Claim-Evidence-Argument

| Claim | Evidence | Argument |
|---|---|---|
| **The Way is associated with variation in the measure focus** | Variation in the measure focus across entities (observed or risk-adjusted) | • The Way is the best explanation of the variation across entities<br>• All other possible explanations have been ruled out |
| **The Way is associated with disparities in the measure focus** | Variation in the measure focus across sub-populations (observed or risk-adjusted) | • The Way is the best explanation of the variation across sub-populations<br>• All other possible explanations have been ruled out |
| **The Way is a confounder or common cause in the association between the measure focus and another known effect (of the Way)** | Entity-level co-variation (correlation) between the measure focus and a related process or outcome | • The Way is the best explanation of the co-variation between measures<br>• All other possible explanations have been ruled out |

# Validity: Claim-Evidence-Argument (continued)

| Claim | Evidence | Argument |
|---|---|---|
| **The Way is responsible for the measure focus** | Entity-level co-variation (correlation) between the Way and the measure focus | Generally, induction: the correlation holds in various populations, settings, and over time |
| **The Way is responsible for the measure focus** | Entity-level co-variation (correlation) between a structure measure (e.g., volume) that enables the Way and the measure focus | Generally, induction: the correlation holds in various populations, settings, and over time |

- Arguments are logical inferences: deduction, induction, or abduction (inference to best explanation)
- Inference to best explanation (IBE) is inferring causes from effects (most plausible)

| | |
|---|---|
| Association | There is an association between the person or entity response to the measure and the measure focus |
| Mechanism | There is an explicit articulation of the mechanisms (resources and response to those resources) responsible for the association |

# Validity: Claim-Evidence-Argument (continued)

**Table 3.1. Quality Levels of Evidence**

| Quality Level | Interpretation |
|---|---|
| High | Further research is highly unlikely to have a significant impact on our confidence in the claims |
| Moderate | Further research is moderately unlikely to have a significant impact on our confidence in the claims |
| Low | Further research is moderately likely to have a significant impact on our confidence in the claims |
| Very Low | Further research is highly likely to have a significant impact on our confidence in the claims |
| Unavailable | Further research is not possible |

**Table 3.3 Confidence Levels of Evidence**

| Quality Level | Interpretation | | |
|---|---|---|---|
| | **Independence** | **Consistency** | **Robust** |
| | Multiple studies using different methods demonstrate similar associations | Multiple studies using similar methods demonstrate similar associations | Multiple studies across different contexts demonstrate similar associations |
| High | Yes | Yes | Yes |
| More likely than not | Yes | Yes | No |
| | Yes | No | Yes |
| | Yes | No | No |
| | No | Yes | Yes |
| | No | Yes | No |
| | No | No | Yes |
| Low | No | No | No |

# Validity: Claim-Evidence-Argument (continued)

| Table 3.2. Status of a Claim | | | |
|---|---|---|---|
| **Status** | **Interpretation** | **Quality Level** | **Confidence Level** |
| Established | Community standards are met for adding the claim to the body of evidence (i.e., as evidence for other claims) | High | High |
| Provisional established | | Moderate | High |
| Arguably true | | Moderate | Claim more likely than not |
| Speculative | None of the other categories | | |
| Arguably false | | Moderate | Negation more likely than not |
| Provisionally ruled out | | Moderate | High |
| Ruled out | Community standards are met for adding the negation of the claim to the body of evidence | High | High |

| Document | Description |
|---|---|
| Endorsable | Importance, validity, and usability are all either established, provisionally established, or arguable true |
| Potentially endorsable | Neither endorsable nor unlikely endorsable |
| Unlikely endorsable | Importance, validity, and usability are all speculative or ruled out, provisionally ruled out, or arguable false |

# Validity and Inference to Best Explanation (IBE)

**Table 2. Other Possible Explanations to Rule Out**

| Possible Explanations | Description |
|---|---|
| Causation | A (quality program & response) is a cause of B (measure focus) |
| Reverse causation | B is a cause of A |
| Confounding | C is a common cause of both A and B (risk-adjustment) |
| Performance bias | A group identified and treated differently than not A group |
| Detection bias | B is measured differently in A group than in not A group |
| Chance | Random (reliability) |
| Fishing | Association between A and some B |
| Temporal trends | A and B change over time for independent reasons |
| Semantic relationships | A and B have overlapping meaning |
| Constitutive relationships | A is a component of B |
| Logical relationships | A and B are logically overlapping |
| Nomological (law) relationships | Association between A and B due to a natural law |
| Mathematical relationships | A = B + C |

# Validity and Risk-Adjustment

- The purpose of risk-adjustment is to “rule-out” that the variation in the measure focus (B) is due to other factors (C) and not due to response to the quality program (A)
  - C is a common cause of both A and B (risk-adjustment)
- Other factors may include (Conceptual Model demonstrates a good model)

| Factors | Factors |
|---|---|
| Pre-ecosystem (built environment) | Behavioral |
| Demographic | Access (selection) |
| Clinical | Post-ecosystem (e.g., community supports) |
| Functional | |

# Selected Resources

- **Pawson, R., Tilley, N. (1997). Realistic evaluation. India: SAGE Publications** – foundational text in evaluation, particularly the context-mechanism-outcome (CMO) ontology, especially on the need to go beyond whether a measure works to understanding why a measure works

- **Schmidt, R. (2020). The benefits of Statistical Noise. Behavioral Scientist** - The Benefits of Statistical Noise - By Ruth Schmidt - Behavioral Scientist

- **Thissen D, Wainer H. (2001). Test scoring. Lawrence Erlbaum Associates –** Reliability vs. validity.

# Questions & Answers

# Partnership for Quality Measurement

Powered by Battelle

# Announcements

## CMS Premieres Three Videos About Suite of MMS Tools

Visit **mmshub.cms.gov** to view a trio of animated videos illustrating the purpose and key functions of three MMS tools.

- CMS MMS Hub
- CMS Measures Inventory Tool (CMIT)
- CMS Measures Under Consideration (MUC) Entry/Review Information Tool (MERIT)

Watch these videos to better understand these valuable tools and how they can help streamline your quality measurement needs. Happy learning!


**Battelle**
MMSSupport@battelle.org

**CMS**

**Gequincia Polk**
gequincia.polk@cms.hhs.gov