NATIONAL QUALITY FORUM
Driving measurable health improvements together

http://www.qualityforum.org

Scientific Methods Panel Web Meeting

February 19, 2021

Funded by the Centers for Medicare & Medicaid Services under contract HHSM-500-2017-00060I –75FCMC19F0007.

Welcome

NQF Scientific Methods Panel Team

- Senior Leads
  - Sheri Winsper, RN, MSN, MSHA, Senior Vice President
  - Sai Ma, PhD, Managing Director/Senior Technical Expert
- Project Management
  - Mike DiVecchia, MBA, PMP, Senior Project Manager
  - Caitlin Flouton, MS, Senior Analyst
  - Hannah Ingber, MPH, Senior Analyst

Scientific Methods Panel Members

Panel Members	
David Nerenz, PhD, Co-chair	Paul Kurlansky, MD
Christie Teigland, PhD, Co-Chair	Zhenqiu Lin, PhD
J. Matt Austin, PhD	Jack Needleman, PhD
Bijan Borah, MSc, PhD	Eugene Nuccio, PhD
John Bott, MBA, MSSW	Sean O'Brien, PhD
Daniel Deutscher, PT, PhD	Jennifer Perloff, PhD
Lacy Fabian, PhD	Patrick Romano, MD, MPH
Marybeth Farquhar, PhD, MSN, RN	Sam Simon, PhD
Jeffrey Geppert, EdM, JD	Alex Sox-Harris, PhD, MS
Laurent Glance, MD	Ronald Walters, MD, MBA, MHA, MS
Joseph Hyder, MD	Terri Warholak, PhD, RPh, CPHQ, FAPhA
Sherrie Kaplan, PhD, MPH	Eric Weinhandl, PhD, MS
Joseph Kunisch, PhD, RN-BC, CPHQ	Susan White, PhD, RHIA, CHDA

Meeting Overview

NATIONAL QUALITY FORUM

Meeting Objectives

- To improve and clarify guidance for future measure development and evaluation cycles
- To clarify questions regarding measures in the Spring 2021 measure review cycle

NATIONAL QUALITY FORUM

Meeting Agenda

- Spring 2021 cycle updates
- Discussion of Evaluation Criteria & Terminology
  - Clarification of terminology
  - Reliability criteria and minimum acceptable threshold
  - Composite measure evaluation
- Opportunity for public comment
- Next steps
- 30-minute SMP closed door Q&A

Spring 2021 Cycle Updates

NATIONAL QUALITY FORUM

Spring 2021 Measures Overview

- 29 complex measures slated for SMP review
  - 18 maintenance and 11 new measures
- Breakdown by measure type
  - 11 outcome
  - 8 cost/resource use
  - 3 composite
  - 3 outcome: intermediate clinical outcome
  - 2 PRO-PM
  - 2 process

NATIONAL QUALITY FORUM

Spring 2021 Measures by Topic Area

Topic Area	Non-Complex Measures	Complex Measures	Total
All-Cause Admissions and Readmissions	0	6	6
Behavioral Health and Substance Use	1	0	1
Cancer	1	0	1
Cardiovascular	1	1	2
Cost and Efficiency	0	8	8
Geriatrics and Palliative Care	0	0	0
Neurology	1	1	2
Patient Experience and Function	0	2	2
Patient Safety	1	5	6
Perinatal and Women's Health	1	3	4
Prevention and Population Health	1	1	2
Primary Care and Chronic Illness	1	0	1
Renal	0	2	2
Surgery	0	0	0
Total	**8**	**29**	**37**

NATIONAL QUALITY FORUM

Important Dates

Review Cycle Step	Date
Measures distributed to subgroups	January 28, 2021
SMP members complete their reviews	January 28 – February 26, 2021
Preliminary reviews due back to NQF	February 26, 2021
SMP members identify additional measures to be pulled for discussion	Week of March 1, 2021
NQF distributes meeting materials to the SMP	Week of March 22, 2021
Spring 2021 measure evaluation meeting	March 30 – 31, 2021

Discussion of Evaluation Criteria

NATIONAL QUALITY FORUM

Topics for Discussion

- Action taken:
  - Adding ACO to the “settings” in testing forms
- Topics requiring further consideration
  - Clarification of terminology
  - Reliability criteria and minimum acceptable thresholds
  - Evaluation of composite measures

NATIONAL QUALITY FORUM

Clarification of Terminology in Testing

- **Data Element-level test**
  - For most measure types, suggest switching from "data element level" to "person or encounter level," as opposed to performance score-level (performance at the accountable entity level such as doctors, hospitals, plans).
    - » Examples: death at patient level for an adjusted mortality rate, or the pain score in a PRO-PM measure
  - *Data element*
    - » A data element includes information about the value set or the direct reference code for the eCQM, along with the QDM datatype and QDM attributes used by that data element.
    - » Measures are constructed from data elements (i.e., a numerator and denominator, in the case of ratio and proportion measures)
- **Performance Score-level test**
  - Could cause confusion as multi-item survey and PROMs generate a score at the patient/person level
  - Change to "accountable/reporting entity level"

NATIONAL QUALITY FORUM

Clarification of Terminology (con't)

- **Composite measure:**
  - Measures with two or more individual performance measure scores combined into one score for an accountable entity.
  - Measures with two or more individual component measures ***assessed separately for each patient*** and then aggregated into one score for an accountable entity, including all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient)
- **Not composite measures?**
  - **Summary score:** a summary score reflects many more measures that may address different issues. However, all the measures are about a single specific provider or service.[1] – *does each item/aspect have its own reliability and validity property?*
    - » Example: *Consumer Reports*
  - **Multi-item scale**: cannot be regarded as a composite because its components are not designed and assessed as individual measures of provider (accountable entity) performance.
    - » Examples: multiple questions about communication in a survey; multiple claims for one diagnosis

1. AHRQ: https://www.ahrq.gov/talkingquality/translate/scores/combine-measures.html

NATIONAL QUALITY FORUM

Reliability

- Objective: To provide more concrete guidance on evaluating reliability, according to level of testing and intended use
- **Suggestions**:

1. Landis & Koch that identifies a kappa statistic value of ≥ 0.4 as "moderate" - may be acceptable for data element reliability if the source data are based on written materials (e.g., survey, questionnaire, clinical notes)
2. Each reliability test should have its own rule-of-thumb guideline
   - Internal consistency (Cronbach's alpha): Bland & Altman
   - Inter-rater agreement (kappa): Landis & Koch
   - Test-retest reliability (simple correlation, ICC): Frost et al.
   - Signal-to-noise ratio (SNR): Adams et al.
   - Various split-half methods (e.g., ICC): Koo et al. 2016
   - Intra-unit reliability (IUR/PIUR): He et al. 2019

NATIONAL QUALITY FORUM

Kappa statistic

- A measure of “data element” agreement corrected for chance agreement
- However, kappa approach can generate surprisingly low and potentially “distracting” estimates when random allocation would lead to very high agreement, based on the marginal probabilities, and the two rating sources are asymmetric.

Rater 1

	NO	YES
NO	940	10
YES	20	30

- *In this example, Kappa coefficient is only around 0.6*

NATIONAL QUALITY FORUM

Two Different Kinds of Intra-Class Correlation Coefficients (ICCs)

- Comparing the variance of between-group (providers or accountable entities) random effects ("signal variance") with a variance estimate that includes within group, between-measurement effects (i.e., comparing a test period with a retest period, randomly splitting a data set into two or more data sets, comparing two or more observers or observations of the same phenomenon).
- Estimating how much of the TOTAL variation in performance at the patient level is explained by those provider-level random effects.

NATIONAL QUALITY FORUM

Reliability Criteria (continued)

3. Data element- and measure score-level analyses require different standards and thresholds
   - *Q: Under what circumstances is it necessary to assess reliability at both the accountable entity level and the person/encounter level?*
   - Note: It is explicitly NOT necessary for eCQMs based entirely on structured fields

NATIONAL QUALITY FORUM

Reliability Criteria (con't)

4. The standards and thresholds for reliability should not be the same for different intended uses of measures.
   - *Q: What are the different uses of measures (such as consumer choice/public reporting vs. financial incentive) that should require a higher threshold of reliability?*

Example:

The intended use of the proposed measure score is [Mark all that apply]:

- Administrative (e.g., identify outlier providers for more frequent review; internal quality improvement)
- Consumer choice/Public reporting (e.g., star ratings; Hospital Compare)
- Financial incentives (e.g., Value-based Purchasing programs; bonus payments/penalties)

NATIONAL QUALITY FORUM

Reliability (con't)

5. Require mis- or re-classification analysis or “stability of classification” (i.e., calculate rates of misclassification or reclassification) as tests of reliability for a given measure

- **Methods** – A nice feature of this approach is that it requires developers to describe specific plans for classification (like star rankings).
- **Tests** - Various methods and metrics for classification stability are possible (e.g., Staggs and Cramer 2016)
  - The number of units with percentile rank within five and within ten percentiles of their true percentile rank
  - Using Bayesian approaches to conjure true scores, you can estimate misclassification rates
  - Distribution of changes in score or percentile ranks
- **Threshold**: This might be especially important in cases where established measures of stability (split sample) are <0.70.

Table 5. Misclassification in a Two-Tiered Classification System, According to Specialty.*

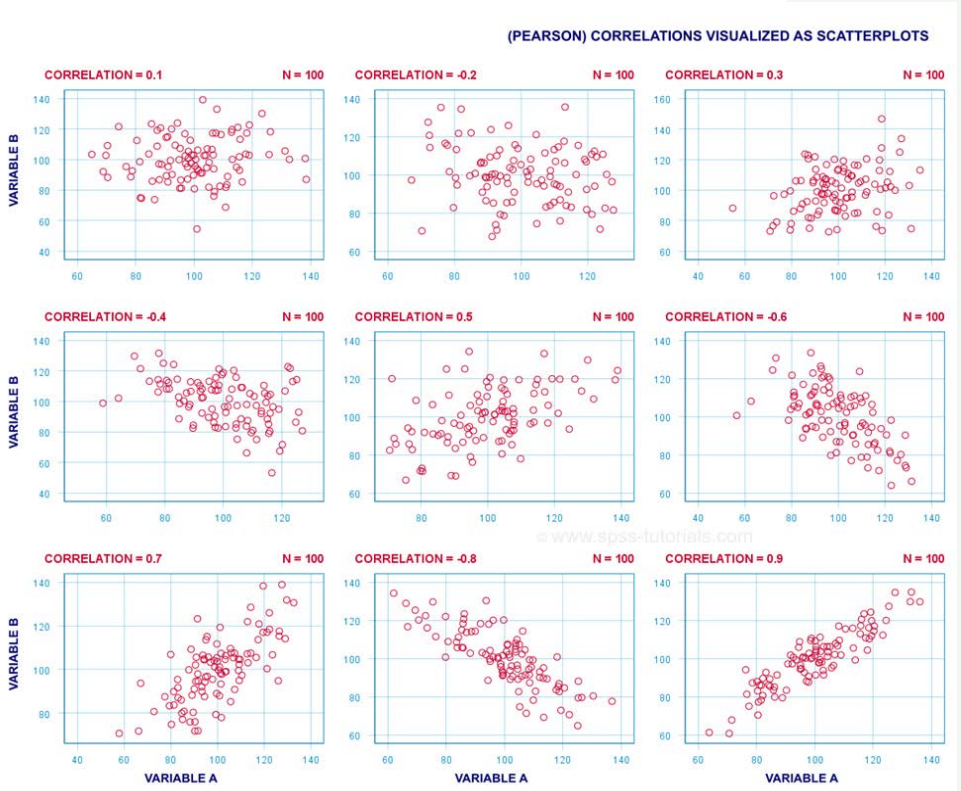
Specialty	No. of Physicians	Physicians Misclassified as Lower Cost†	Physicians Misclassified as Not Lower Cost‡	Overall Misclassification Rate
		percent		
Cardiology	708	40	13	20
Endocrinology	169	50	19	25
Family or general practice	1065	39	16	21
Gastroenterology	426	32	11	16
Internal medicine	2979	50	22	25
Obstetrics–gynecology	922	36	10	17
Orthopedic surgery	580	50	17	25
Otolaryngology	229	29	13	16
Pulmonary and critical care	362	58	21	28
Vascular surgery	72	67	22	36
Overall§	7512	43	18	22

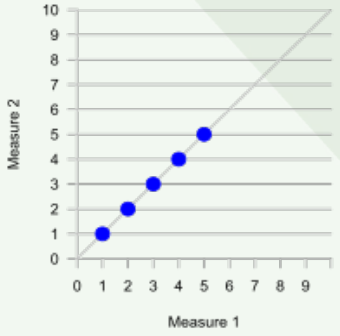
* In this two-tiered system, physicians whose cost profiles were in the lowest 25% of the distribution were labeled “lower cost.” The remaining 75% of physicians were labeled “not lower cost.”
† This is the proportion of physicians who were classified as lower cost but were not lower cost.
‡ This is the proportion of physicians who were classified as not lower cost but who were lower cost.
§ The numbers shown are for the 10 specialties listed in the table. When percentages were calculated across 26 of the 28 specialties included in the study, 43% of physicians were misclassified as lower cost, and 17% were misclassified as not lower cost; overall, 22% of the physicians were misclassified. In two specialties, the reliability was 0, making misclassification impossible to calculate.

Reference: Adams et al. 2010

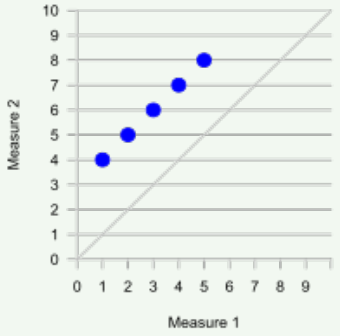
Reliability Criteria (con't)

6. Graphics as a supplement to numerical results

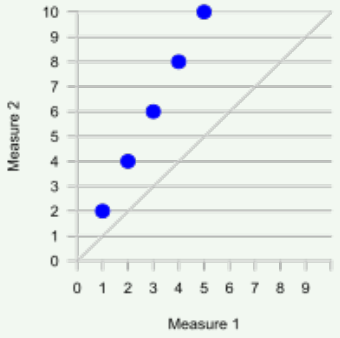




y = x
R^2 = 1
ICC(1) = 1
ICC(2,1),agreement = 1
ICC(2,1),consistency = 1
ICC(3,1),agreement = 1
ICC(3,1),consistency = 1



y = x+3
R^2 = 1
ICC(1) = 0.053
ICC(2,1),agreement = 0.357
ICC(2,1),consistency = 1
ICC(3,1),agreement = 0.357
ICC(3,1),consistency = 1



y = 2x
R^2 = 1
ICC(1) = 0.343
ICC(2,1),agreement = 0.476
ICC(2,1),consistency = 0.8
ICC(3,1),agreement = 0.476
ICC(3,1),consistency = 0.8

Reliability: Proposed Sample Table

Test/Use	Data element or measure score level?	What are you testing?	Unacceptable	Adequate	High
Cronbach's Alpha for multi-item survey	Data element (person or encounter)	How internally consistent are items in a multi-item scale?	< 0.7	0.7 - 0.9	>0.9
Pearson correlation (test-retest analysis of measure score)	Measure score (accountable/reporting entity)		< 0.6	0.6 - 0.8	>0.8
ICC [1 or 2, Agreement]	Measure scores (accountable/reporting entity)	How stable are entity-level scores when measures on split samples or closely spaced time points			
ICC [1 or 2, Consistency]	Measure scores (accountable/reporting entity)	How stable are entity-level ranks when measures on split samples or closely spaced time points			
SNR	Measure score				
IUR					
Kappa	Data element				
Classification stability	Measure score				

NATIONAL QUALITY FORUM

Articles Offering Rule-of-Thumb Guidance for Reliability Statistics

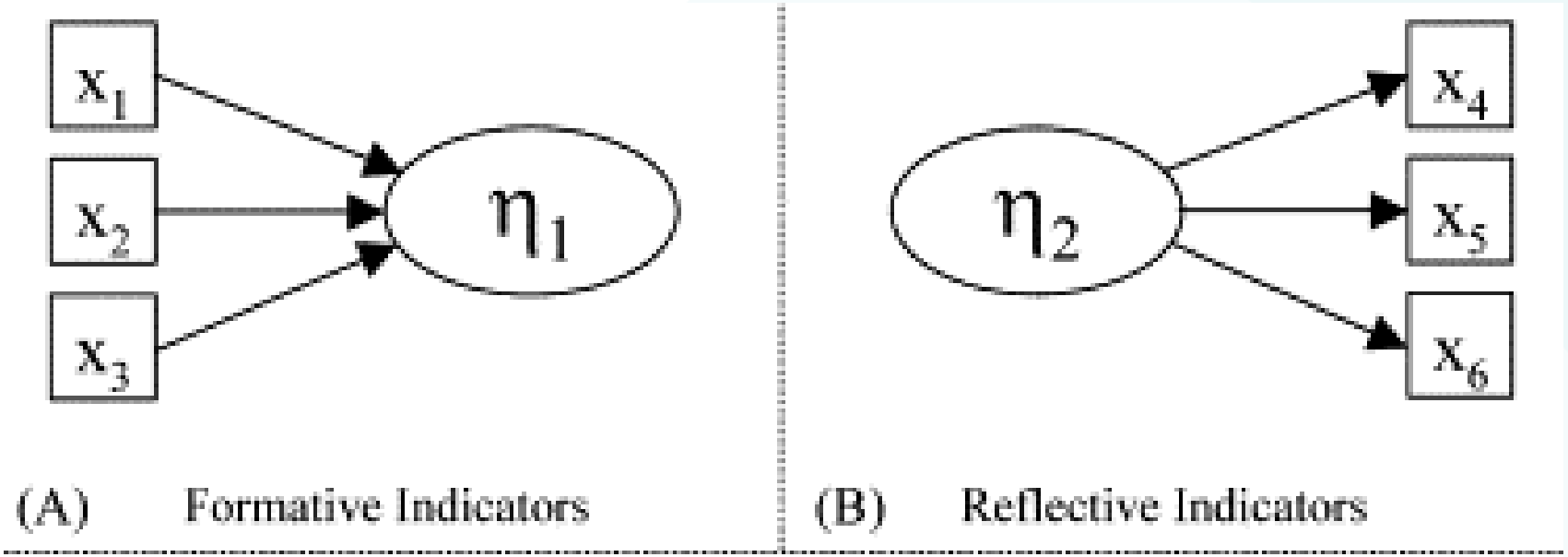
- Staggs, VS. & Gajewski, BJ. Bayesian and frequentist approaches to assessing reliability and precision of health-care provider quality measures. *Stat Methods Med Res* 2015; 26(3).
- Adams, JL, et al. Physician cost profiling–reliability and risk of misclassification. *N Engl J Med* 2010; 362: 1014–1024.
- McGraw, KO. & Wong, SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996; 1(1): 30-46
- Bland, JM. & Altman, DG. Statistics notes: Cronbach's alpha. *BJM* 1997; 314(7080): 572.
- Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-174.
- Koo, TK. & Li, MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15(2): 155-163.
- He, K. et al. Inter-Unit reliability for quality measure testing. *J Hosp Adm* 2019; 8(2): 1–6.

NATIONAL QUALITY FORUM

Composite Measure Evaluation

- The measurement model assumed by the developer (formative or reflective) will be key to assessing whether the analytic strategy for testing reliability and validity is appropriate for the unit of comparison proposed.
  - Depending on the scoring method (simple algebraic sums, all-or-none, etc.), appropriate methods could include intraclass correlation coefficients, structural equation modeling, generalized estimating equations, etc.
- Currently CAHPS surveys are regarded as multi-item measures, and therefore they are not composite measures. Therefore, one should not use individual items or subscales as performance measures.

Conceptual Frameworks:
Reflective vs. Formative Models



Conceptual Frameworks: *Causal vs. Effect Indicators*

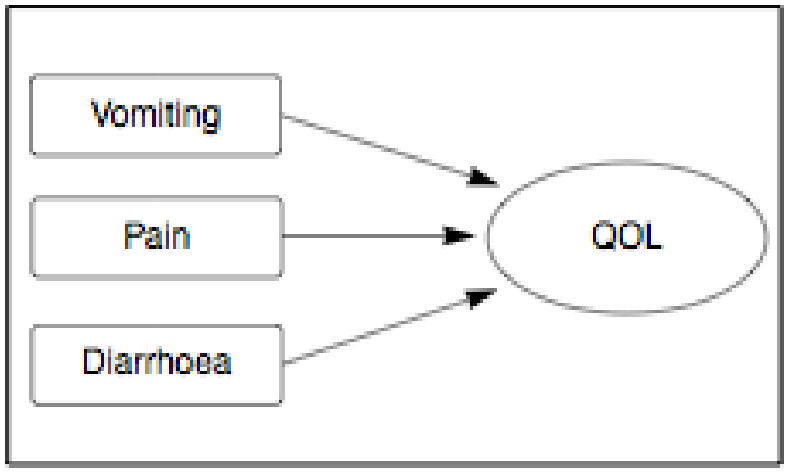


Figure 1: Causal indicators

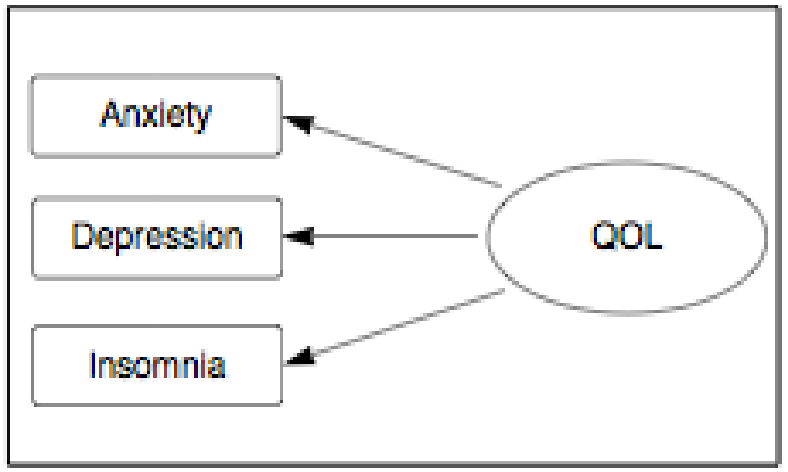


Figure 2: Effect indicators

NATIONAL QUALITY FORUM

Examples of Reflective and Formative Models

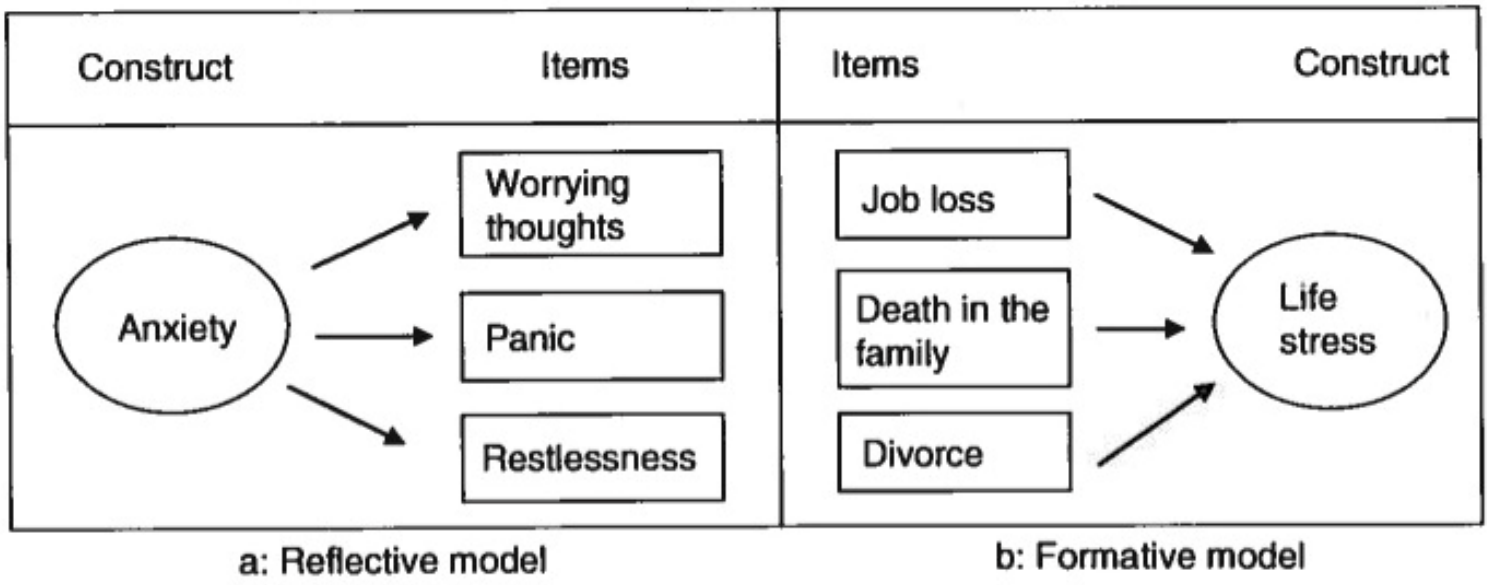


Figure 2.2 Graphical representation of a reflective model (a) and formative model (b).

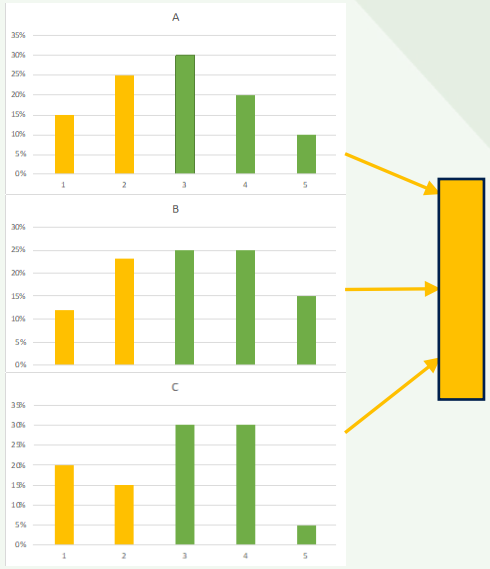
NATIONAL QUALITY FORUM

Reflective vs. Formative Models Described

Features	Formative	Reflective
Definition	Multiple semi-related measures (multi-dimensional)	All measures reflect same underlying construct (uni-dimensional)
Item characteristics	Uncorrelated	Correlated
Internal consistency test	Not required	Required
Causal relationship	Item changes cause construct change	Construct changes cause item changes
Descriptive equation	$\{Item\}=\beta_1\{Construct\}_1+\varepsilon$	$\{Construct\}= \beta_1\{Item\}_1+ \beta_2\{Item\}_2 \ldots + \zeta$
Examples	Total Illness Burden Index	CESD

Survey Items Using Top-Box Scoring & Evaluation

- Single survey item:
  - Validity can be assessed at the person/encounter level
  - Reliability is hard to operationalize at the person/encounter level (unless test/retest of the same person)
- Multiple survey items:
  - If sum top box values at the patient level, then they need to report internal consistency reliability
  - If combine rates for each item separately at the entity level, then this measure should be regarded as a composite measure (because each of the component items is scored at the accountable entity level)



Opportunity for Public Comment

Next Steps

What's Next

- Meeting summary review by the SMP members; will be posted on NQF website
- Important upcoming dates:
  - February 26: Return measures to SMP project team
  - First week of March: SMP members identify additional measures to be pulled for discussion to SMP team
  - Week of March 22: Distribute meeting materials to SMP
- Next meeting: March 30-31, 2021

SMP Closed Door Q&A

New Evaluation Form on Surveymonkey

- *[Screenshare]*

SharePoint Tutorial

- *[Screenshare]*

NATIONAL QUALITY FORUM

Other questions?

THANK YOU.

NATIONAL QUALITY FORUM

http://www.qualityforum.org