Assessing the Content Validity of PROMs (A Systematic Strategy based on the COSMIN Methodology)

Favour Adebayo
Under the supervision of Dr. Kalu
York University
Lab Presentation-June 2026

Do medical surveys actually measure what they should?

Overview

1. Development of Methodology: Content Validity
2. The Three pillars of Content Validity
3. The Evaluation Process
4. Box 1: PROM Development
5. Box 2: Evaluating Content Validity
6. Box 3: 10 Criteria matrix for good content validity
7. Study Ratings
8. Overall Conclusions
9. Reporting a systematic review on content validity
10. Conclusion - Next Steps

What is COSMIN?

COSMIN: COnsensus-based **S**tandards for the selection of health **M**easurement **IN**struments.

- **Core Goal:** An initiative designed to evaluate the quality of health measurement tools.
- **Standards vs. Criteria:**
  - **Standards:** The exact methodological requirements and statistical methods used to check study quality.
  - **Criteria:** The benchmarks for what actually makes a measurement property "good" in a Patient-Reported Outcome Measure (PROM).
- **The COSMIN Toolkit:** Offers a taxonomy of 9 measurement properties, a risk of bias checklist, a systematic review database, and strict guidelines to help researchers select the highest quality instruments.

Development of Methodology: Content Validity

- **Content validity** is widely recognized as the most critical and challenging property to assess
- Created via an international Delphi study involving 158 global experts across 21 countries
- **The Evaluation Core:** Shifting the focus from simple checks to looking at how well items measure the construct
- **The 3 Core Factors:** Evaluated comprehensively across Relevance, Comprehensiveness, and Comprehensibility

For a PROM to be selected for a Core Outcome Set(COS), it must have high quality of content validity

The Three Pillars of Content Validity

Relevance

Are all included items relevant to the construct, target population, and context of use of interest?

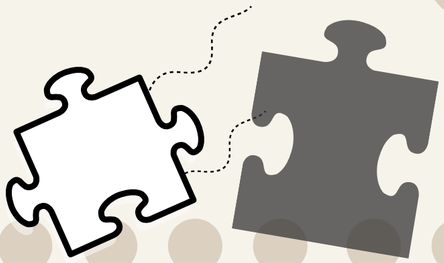
Comprehensiveness

Are there any key aspects or dimensions of the construct missing from the tool?

Comprehensibility

Are the PROM instructions, items, and response options easily understood by the target population exactly as intended?

"Researchers do not validate an instrument, they validate the application of it"
(Terwee et al., 2018, p. 8)



The Evaluation Process: From Review to Consensus

STEP 1: Evaluate PROM Development

How are the items originally generated and tested?

Step 2: Evaluate Content Validity Studies

What do patients and professionals say about relevance, comprehensiveness, and comprehensibility?
Comprehensibility does not apply to professionals

Step 3: Determine overall Content Validity

Combine PROM development quality and content validity study results into an overall rating and grade of evidence

Box 1: PROM Development

Box 1: Asks how well the PROM was built.

Part 1a: Concept Elicitation

· Construct, target population & context of use clearly described?
· Sample represents the target population?
* Appropriate qualitative methods (e.g. focus groups, interviews)?
· Skilled moderators, proper analysis, saturation reached?

Part 1b: Cognitive Interview/Pilot Testing

. Was a pilot test conducted at all? (If not -> automatically inadequate)
· Appropriate sample & qualitative methods used?
· Skilled interviewers, proper analysis?
· Were problems found used to adapt the PROM?

Rating system: Each standard -> Very Good / Adequate / Doubtful / Inadequate. "Worst score counts": the lowest rating across all standards becomes the overall PROM development rating.

Box 2: Evaluating Content Validity Studies

	Patients	Professionals
Relevance	Are questions relevant to their lived, daily experiences with the condition?	Are questions relevant in clinical context and useful for diagnosis?
Comprehensiveness	Is anything important missed from their lived experiences	Does the instrument cover the full theoretical scope of the context?
Comprehensibility	Target population required to validate if wording of questions are understood as intended.	N/A

Box 3a: 10 Criteria Matrix for good content validity

Relevance	Comprehensiveness	Comprehensibility
1. Construct of Interest 2. Target Population 3. Context of Use 4. Response options 5. Recall period	6. Are no key concepts missing?	7. Instructions understood 8. Items/response options understood 9. Items appropriately worded 10. Options match the question

Rating Key

✚ Sufficient

-- Insufficient

± Inconsistent

Box 3b and 3c: Study ratings to overall conclusion

Box 3b: Overall ratings

Results from the development study, content validity studies and the reviewer's rating are combined into Overall Ratings

If inconsistent: look for explanations (population, study quality, era) and consider the subgroups

Box 3c: Grading the Quality of Evidence

Each overall rating gets a grade using the modified grade approach

High | Moderate | Low | Very Low

Based on three factors:

- **Risk of bias** → quality of underlying studies
- **Inconsistency** → do studies agree with this?
- **Indirectness** → different population from target

Reporting a Systematic review on the content validity

Report the Overall Rating and the quality of evidence together

The systematic review should report:

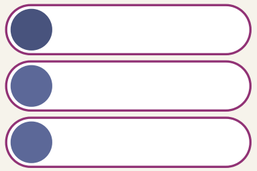
- Characteristics of the PROM (construct, population, context of use, number of items and recall period)
- Quality of PROM development (Box 1 ratings)
- Study characteristics of available content validity studies
- Quality of available content validity studies (Box 2 ratings)
- Overall ratings(relevance, comprehensiveness, comprehensibility, content validity)

Application of COSMIN to the Afrocentric CANE Adaptation

The CANE rates 24 domains of need (e.g. accommodation, food, memory, self care, e.t.c) as no need, met need, or unmet need, drawing on up to four informants (user, carer, staff, and rater). Once the items are added, COSMIN gives us a structured way to check whether the adaptation truly works:

Relevance

Does the new and reworded items across existing domains and any new cultural/spiritual domains genuinely reflect the needs and priorities of Black older adults.



Comprehensiveness

Beyond the original 24 domains, have we addressed the gap (personal, cultural, spiritual & intergenerational needs)that the original CANE items misses entirely?

Comprehensibility

Are the no need / met need / unmet need ratings, prompts, and new item wording clearly understood by participants during interviews and pilot testing?

Conclusion - Next Steps

- COSMIN gives us a structured, evidence-based way to evaluate whether our CANE adaptation actually achieves what it sets out to do.

- Concept elicitation with Black older adults → cognitive interviews/pilot testing on new and reworded items → documenting everything against Box 1/Box 2 standards

THANK YOU

For your attention

References

- Terwee, C. B., et al. (2018). COSMIN methodology for assessing the content validity of PROMs - User manual, version 1.0. Amsterdam, The Netherlands: VU University Medical Center.
- https://www.canva.com/s/templates?query=research