COSMIN evaluation of the quality of the Neck Disability Index (NDI)

Eva Ørnbøl Statistician Functional Disorders AUH

The NDI questionnaire

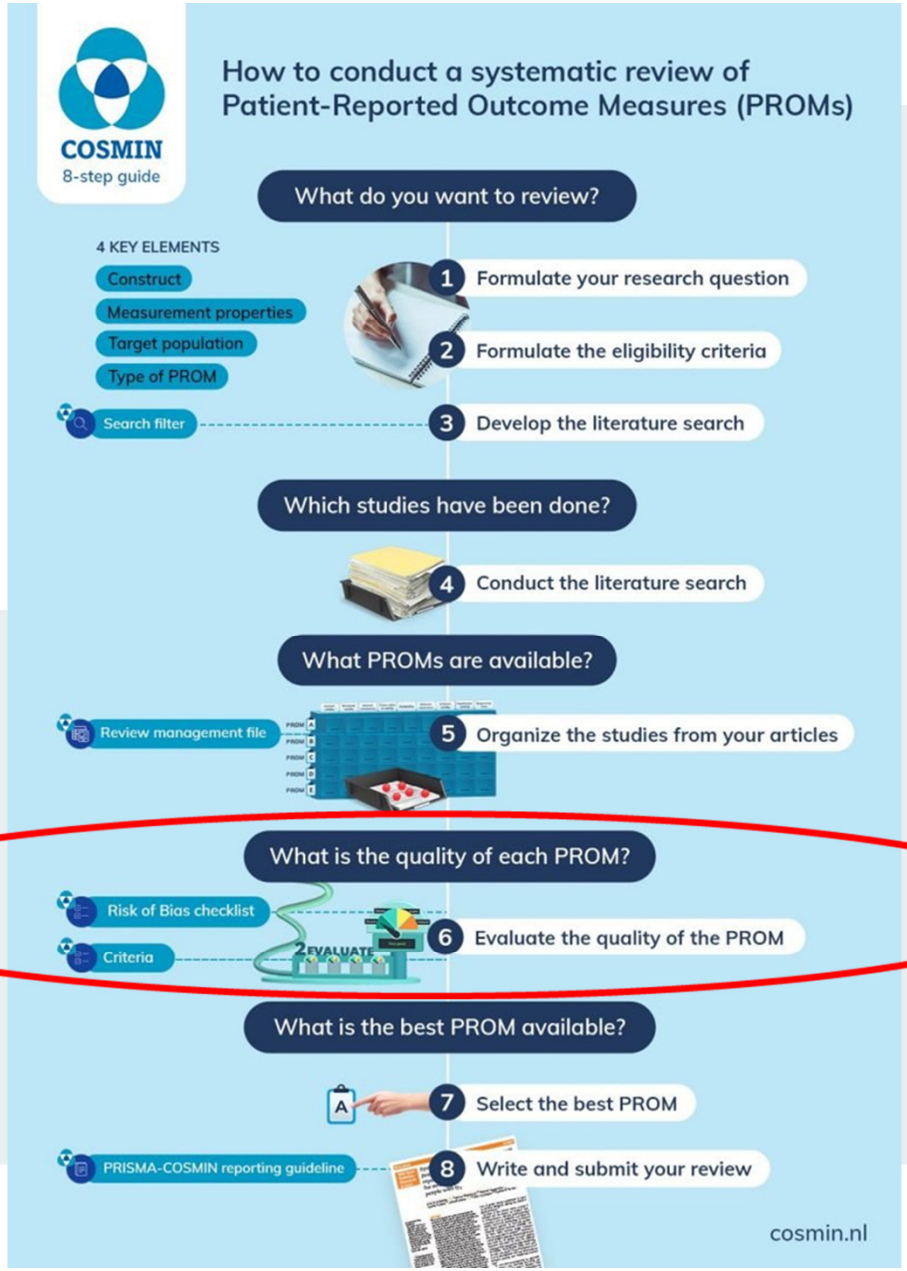
neck_disability_index.pdf

The focus is on the 10-item version – the only one that results in a score.

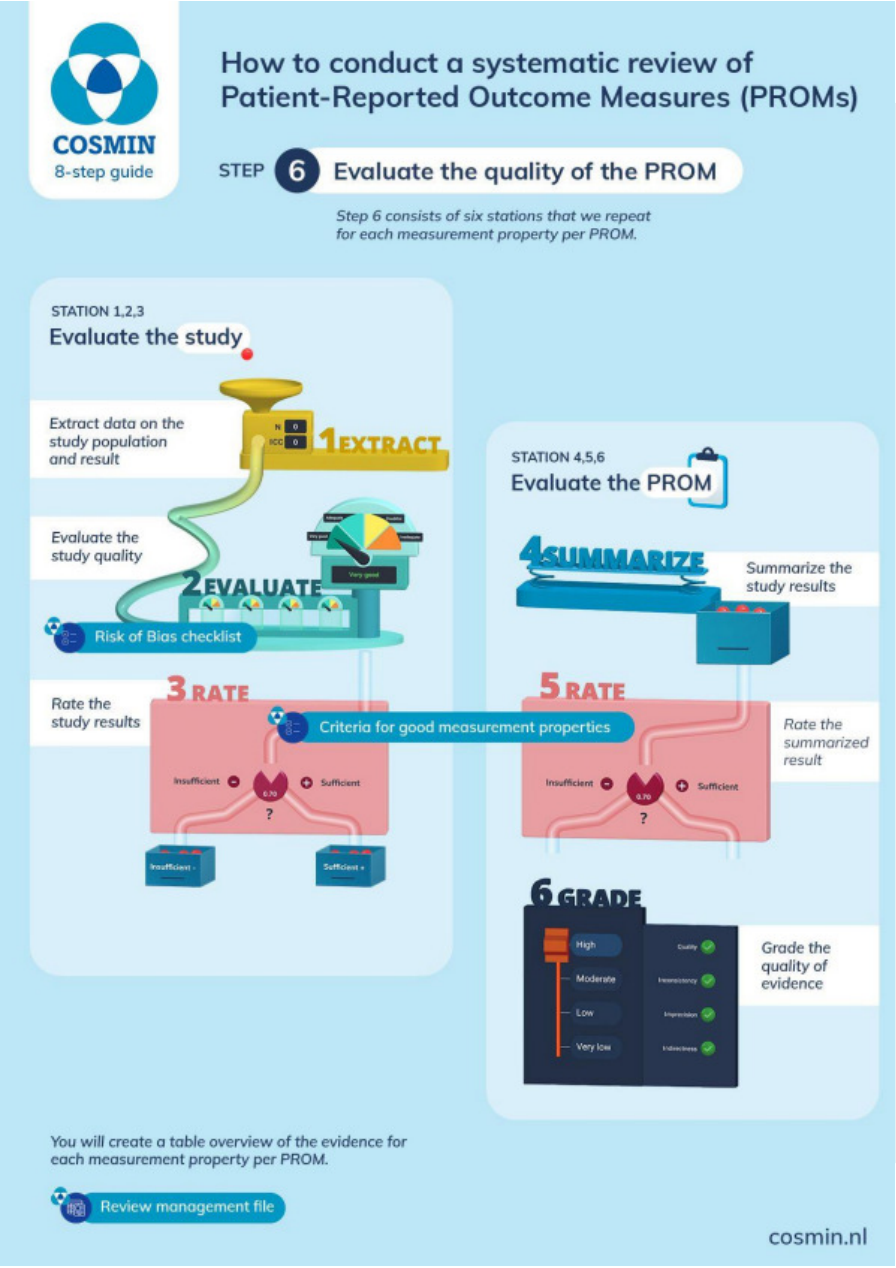
Where do we start?

- Visit the COSMIN homepage
  - https://www.cosmin.nl/
- Decide what you need help with
  - I am conducting a systematic review of outcome measurement instruments – the NDI
- Watch the short introduction:
  - https://www.youtube.com/watch?v=C3RFCilfCxU

We will not go further into the first 5 steps except to clarify that the items stems from a reflective model and the number of studies in the drawer labeled NDI 10-item and measurement property internal consistency is 6.

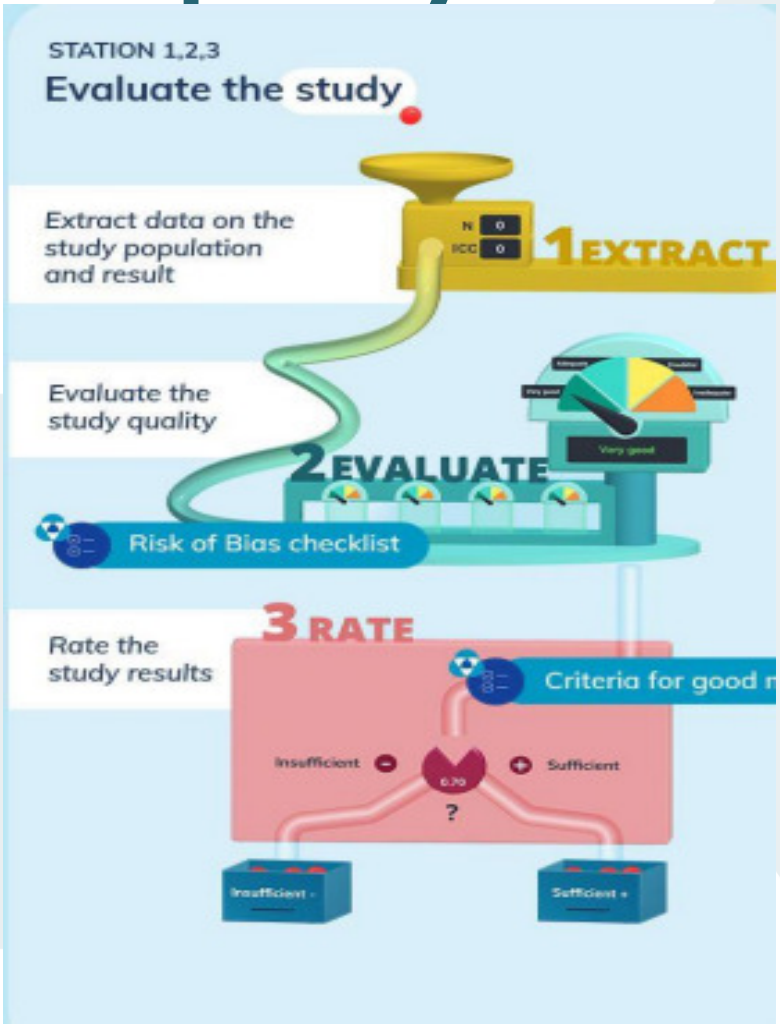


So, we will focus on step 6. This step consist of two parts.



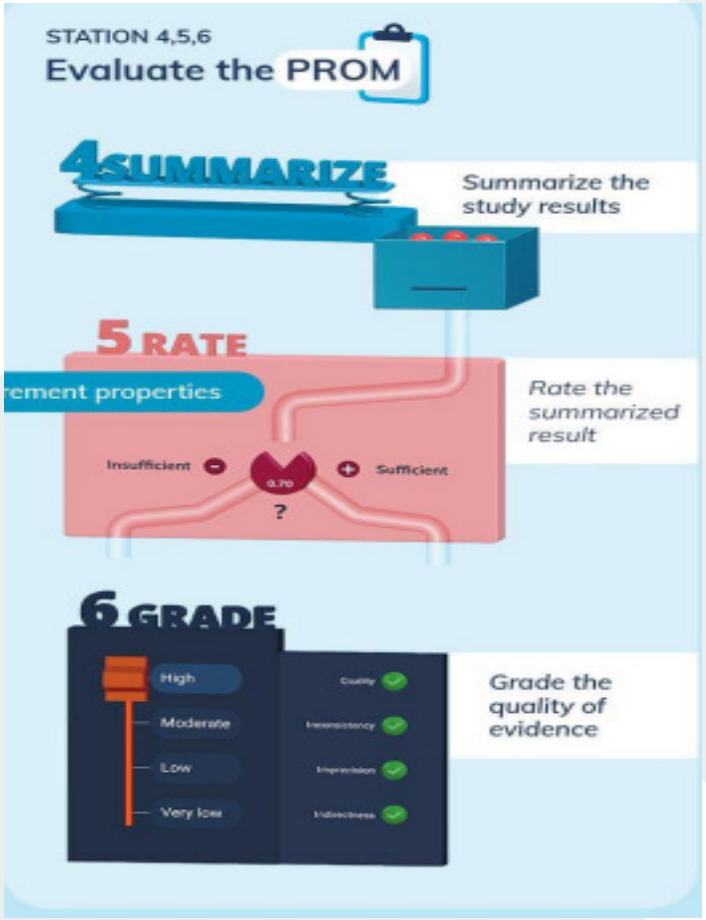
Evaluate the quality of the NDI

Part One: Here the focus is on the 6 individual studies.



Evaluate the quality of the NDI

Part Two:
Here the focus is on the NDI 10-item.



Part One - 1. extract

Exercise 1

Part One – 2. evaluate

The RoB checklist for Internal consistency

Box 4. Internal consistency					
Statistical methods	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
1 For continuous scores: Was Cronbach's alpha or omega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	NA
2 For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	NA
3 For IRT-based scores: Was standard error of the theta (SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	NA
Other					
4 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

Part One – 2. evaluate

The quality of each study on a measurement property should be assessed separately, using the corresponding COSMIN box. Each study is rated as very good, adequate, doubtful or inadequate quality. To determine the overall rating of the quality of each single study on a measurement property, the lowest rating of any standard in the box is taken (i.e. “the worst score counts” principle).

Part One - 2. evaluate

Exercise 2.

Part One – 3. rate

The criteria for Internal consistency

Internal consistency	+	At least low evidence for sufficient unidimensionality *AND* Cronbach's alpha ≥0.70
	?	Criteria for "at least low evidence for sufficient unidimensionality" not met *OR* Evidence for insufficient unidimensionality *OR* Not enough information reported
	-	At least low quality evidence for sufficient unidimensionality *AND* Cronbach's alpha <0.70

Part One – 3. rate

Exercise 3.

Part Two – 1. summarize

Decide whether the results of all available studies per measurement property are consistent.

- If consistent, the results of studies can be quantitatively pooled or qualitatively summarized.

Part Two – 1. summarize

Quantitatively pooling the results

Using meta-analysis to achieve pooled estimates - consult a statistician ☺

Qualitatively summarizing the results

- To rate the qualitatively summarized results as sufficient (or insufficient), in principle 75% of the results should met the criteria

Part Two – 1. summarize

Exercise 4.

Part Two – 2. rate

Rate the summarized results against the same criteria for good measurement properties as used for the individual studies.

Part Two – 2. rate

Exercise 5.

Part Two – 3. grade

The quality of the evidence refers to the confidence that the pooled or summarized result is trustworthy.

Table 5. Definitions of quality levels

Quality level	Definition
High	We are very confident that the true measurement property lies close to that of the estimate* of the measurement property
Moderate	We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different
Low	Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property
Very low	We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property

Part Two – 3. grade

The GRADE approach uses four areas to determine the quality of the evidence:

1. Risk of bias
2. Inconsistence
3. Imprecision (pooled/summarized sample size)
4. Indirectness (different target populations/contexts)

Part Two – 3. grade

Risk of bias area: downgrade based on the methodological quality of the individual studies.

Risk of bias	Downgrading for Risk of Bias
No	There are multiple studies of at least adequate quality, or there is one study of very good quality available
Serious	There are multiple studies of doubtful quality available, or there is only one study of adequate quality
Very serious	There are multiple studies of inadequate quality, or there is only one study of doubtful quality available
Extremely serious	There is only one study of inadequate quality available

Part Two – 3. grade

Inconsistency area: If your results are inconsistent, there are several ways, you can try to accommodate this. We will come back to this later.

Imprecision area: This refers to the total sample size included in all individual studies. Downgrade 1 if total sample size is below 100 and 2 if total sample size is below 50. Only this if sample size is not already a part of the RoB checklist.

Indirectness area: This is concerned with whether the results are directly applicable to the target population.

Part Two – 3. grade

The quality starts at the highest quality and is then downgraded if there are problems in the four areas.

Table 6. Modified GRADE approach for grading the quality of evidence

Quality of evidence	Lower if
High	Risk of bias -1 Serious -2 Very serious -3 Extremely serious Inconsistency -1 Serious -2 Very serious Imprecision -1 total n=50-100 -2 total n<50 Indirectness -1 Serious -2 Very serious
Moderate	
Low	
Very low	

n=sample size

Part Two – 3. grade

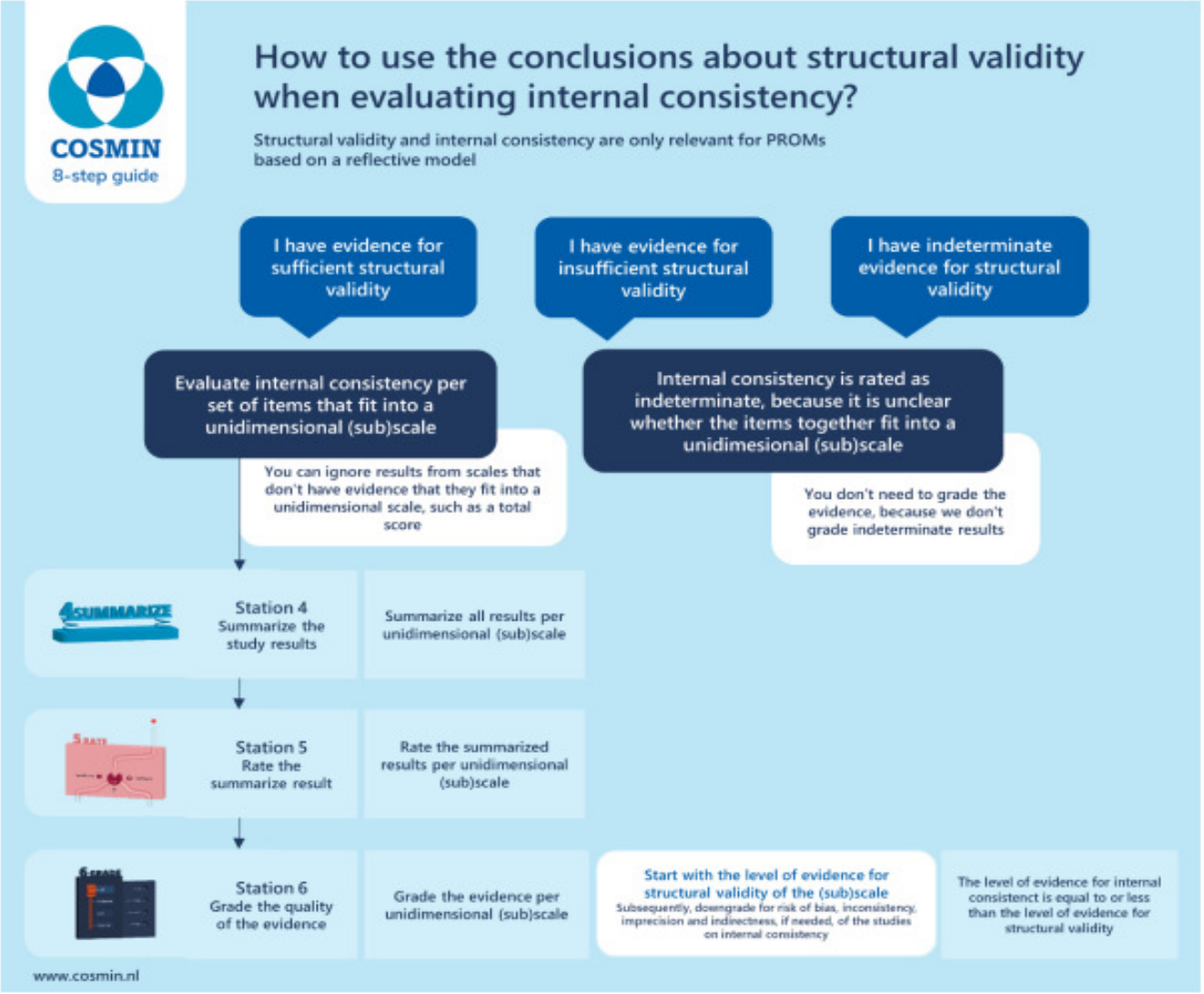
Exercise 6.

Part One – 3. rate

More on how to use conclusions on structural validity in the evaluation of internal consistency.

In the exercise you assumed that the criteria of at ”least low evidence of unidimensionality” was met. This could come from the overall evaluation of structural validity – the result from that specific drawer in the cabinet. Or it could come from the individual study itself.

The figure displays how to proceed with different ratings of the summary results of structural validity.



Part One: 1.-3.

Do we have ”at least low evidence of unidimentionalty” in the NDI 10-item?

Let have a quick look on the measurement property structural validity and go through Part One. The drawer we work with contains 5 studies. Let us extract data on study population, fill in the RoB checklist, rate the results against criteria for structural validity.

Box 3. Structural validity

		very good	adequate	doubtful	inadequate	NA
Statistical methods						
1	For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed	Only PCA was performed	No exploratory or confirmatory factor analysis performed	NA
2	For IRT/Rasch: does the chosen model fit to the research question?	Chosen model fits well to the research question	Assumable that the chosen model fits well to the research question	Doubtful if the chosen model fits well to the research question	Chosen model does not fit to the research question	NA
3	Was the sample size included in the analysis adequate?	FA: 7 times the number of items in the tested model and ≥100	FA: at least 5 times the number of items in the tested model and ≥100; OR at least 6 times number of items in the tested model but <100	FA: 5 times the number of items in the tested model but <100	FA: < 5 times the number of items in the tested model	
		Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models: 100-199 subjects	Rasch/1PL models: 50-99 subjects	Rasch/1PL models: < 50 subjects	
		2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokken scale analysis: 500-999 subjects	2PL parametric IRT models OR Mokken scale analysis: 250-499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250 subjects	
Other						
4	Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws (e.g. rotation method not described)	Other important methodological flaws (e.g. inappropriate rotation method)	

AARHUS UNIVERSITY

The RoB checklist for structural validity.

The criteria for Structural validity.

Structural validity	+	**CTT:** EFA/PCA: factor loadings of each item on its factor ≥0.30 *AND* Maximum 10% of the items have factor loadings ≥0.30 on multiple factors *AND* Explained variance ≥50% and structure is in line with the theory about the construct to be measured *OR* results on scree plot or Kaiser criterion (Eigenvalues >1) are in line with the theory about the construct to be measured CFA: CFI or TLI or comparable measure >0.95 *OR* RMSEA <0.06 *OR* SRMR <0.08 **IRT/Rasch:** No violation of unidimensionality: CFI or TLI or comparable measure >0.95 *OR* RMSEA <0.06 *OR* SRMR <0.08 *AND* No violation of local independence: residual correlations among the items after controlling for dominant factor <0.20 *OR* Q3s <0.37 *AND* No violation of monotonicity: adequate looking graphs *OR* item scalability >0.30 *AND* Adequate model fit: IRT: χ^2 >0.01 Rasch: infit and outfit mean squares ≥0.5 and ≤1.5 *OR* Z-standardized values >-2 and <2
	?	Not enough information reported
	-	Criteria for '+' not met

Part Two – inconsistent results

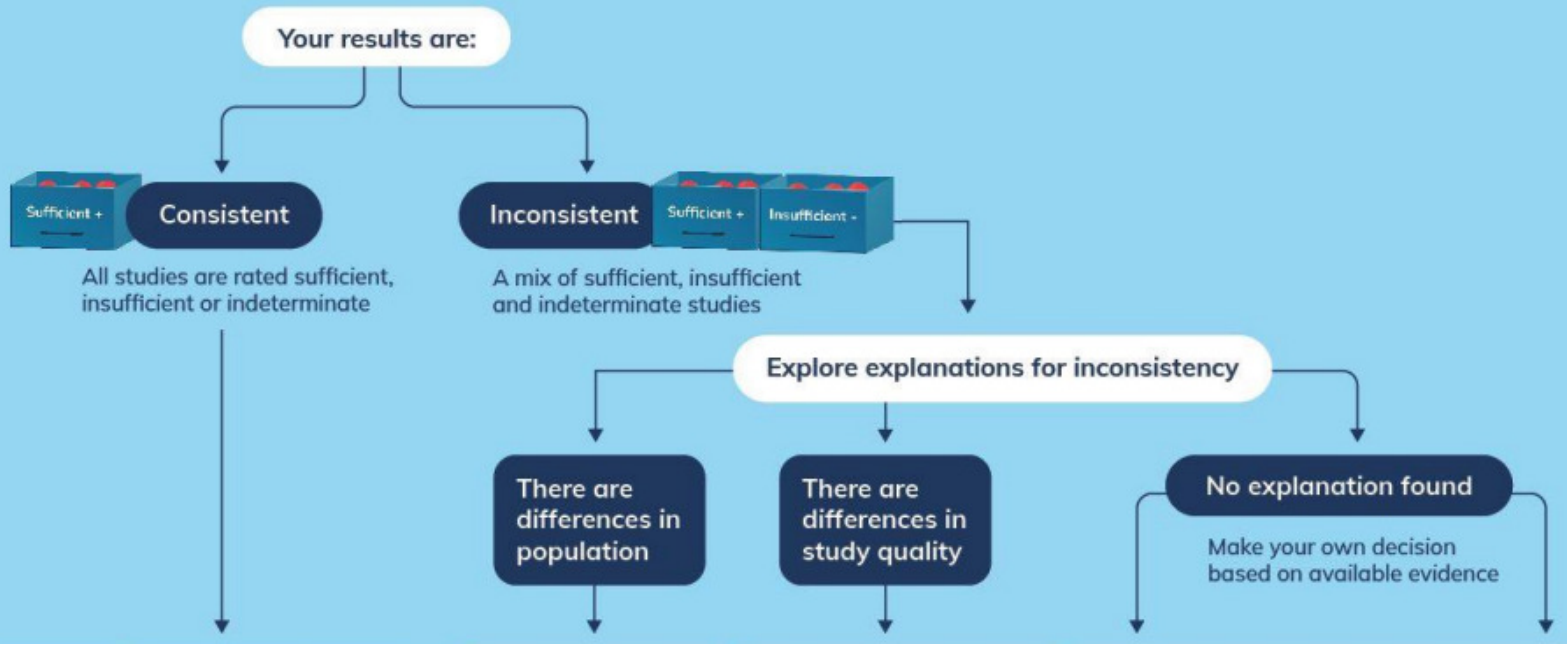
COSMIN
Decision tool

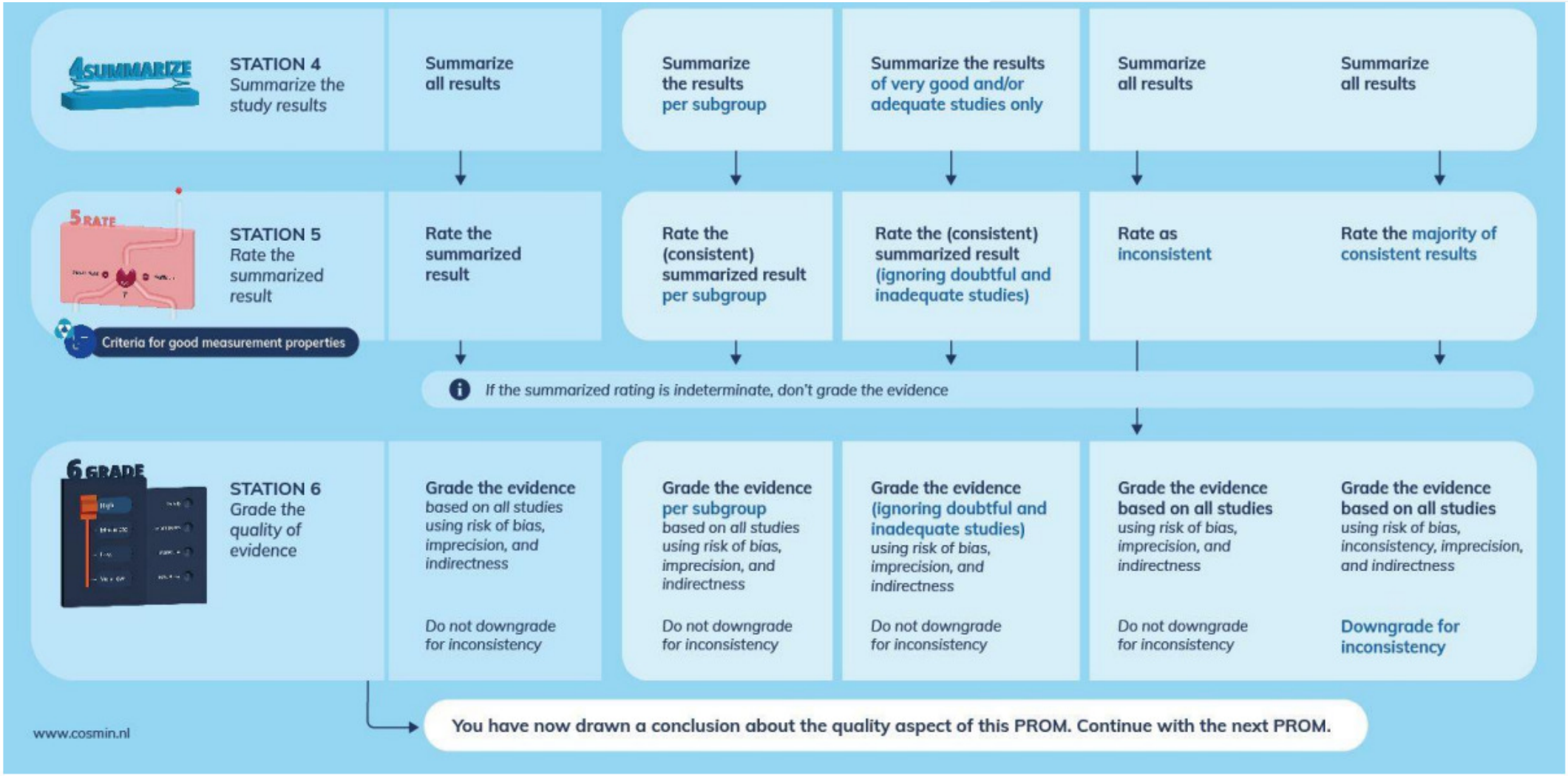
STEP 6 Conducting a systematic review of PROMs

How to deal with (in)consistent study results

while summarizing, rating and grading a measurement property of a PROM

In station 1, 2 and 3, you have extracted and rated all results of all available studies for a measurement property of a PROM. Follow this decision tool to help you draw a proper conclusion.




4 SUMMARIZE
STATION 4
Summarize the study results
Summarize all results
Summarize the results per subgroup
Summarize the results of very good and/or adequate studies only
Summarize all results
Summarize all results
5 RATE
STATION 5
Rate the summarized result
Rate the summarized result
Rate the (consistent) summarized result per subgroup
Rate the (consistent) summarized result (ignoring doubtful and inadequate studies)
Rate as inconsistent
Rate the majority of consistent results
Criteria for good measurement properties
If the summarized rating is indeterminate, don't grade the evidence
6 GRADE
STATION 6
Grade the quality of evidence
Grade the evidence based on all studies using risk of bias, imprecision, and indirectness
Do not downgrade for inconsistency
Grade the evidence per subgroup based on all studies using risk of bias, imprecision, and indirectness
Do not downgrade for inconsistency
Grade the evidence (ignoring doubtful and inadequate studies) using risk of bias, imprecision, and indirectness
Do not downgrade for inconsistency
Grade the evidence based on all studies using risk of bias, imprecision, and indirectness
Do not downgrade for inconsistency
Grade the evidence based on all studies using risk of bias, inconsistency, imprecision, and indirectness
Downgrade for inconsistency
www.cosmin.nl
You have now drawn a conclusion about the quality aspect of this PROM. Continue with the next PROM.

Back to internal consistency

What we have seen concerning the inconsistency of structural validity, is that how you choose to proceed, can influence the quality of internal consistency of NDI 10-item.

With the 5 studies available here I would go with, "internal consistency is indeterminate, as it is unclear whether the items fit into a unidimensional scale".