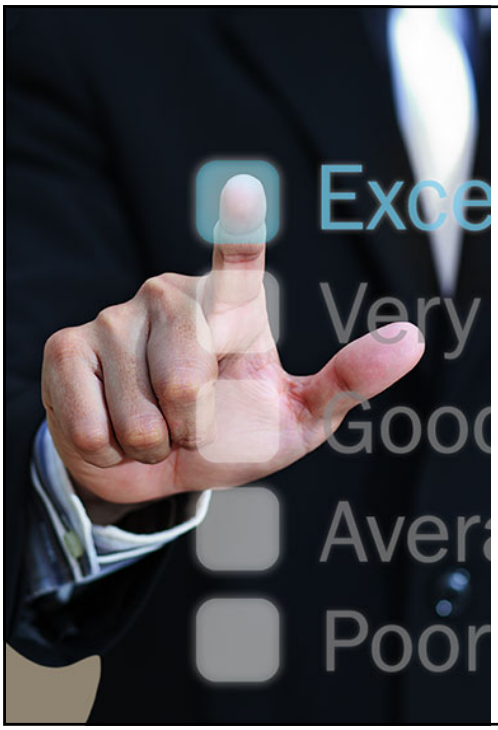


Field testing

Henrik Hein Lauridsen

Associate professor

Research Unit for Clinical Biomechanics

University of Southern Denmark

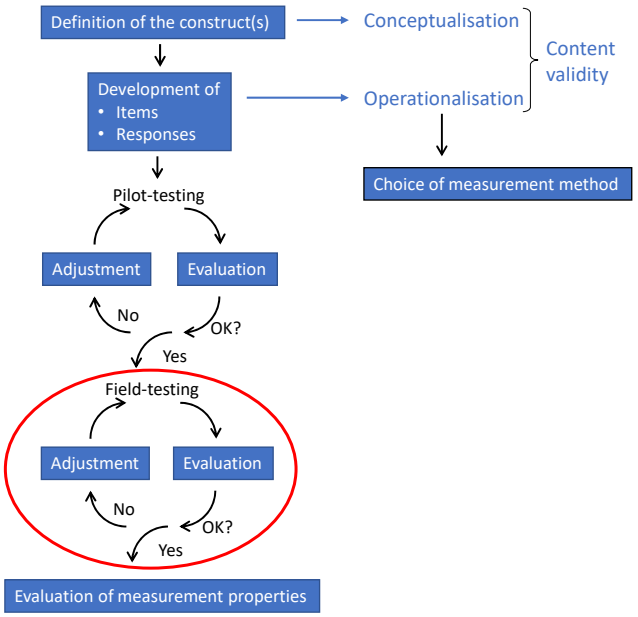
1

Programme for day 2

Time	Contents	Pages	Teacher
9.00 – 10.00	Field testing	MiM: 65-72 MiM: 80-84	HHL
10.00 – 11.00	Exercise: Field testing		HHL
11.00 – 12.30	Reproducibility (advanced) • Internal consistency • Reliability • Measurement error	MiM: 96-123	HHL
12.30 – 13.00	*Lunch*		
13.00 – 14.00	Exercise: Reproducibility		HHL
14.00 – 14.30	Responsiveness & interpretation	MiM: 202-220 MiM: 245-261 Articles 3 x vodcasts	HHL
14.30 – 16.00	Exercise: Responsiveness and interpretation		HHL

2

Questionnaire development



3

What is field testing?

A further refinement of our instrument after pilot testing

- Aim
  - Item reduction
  - Insights into data structure

N.B. Can also be carried out on existing questionnaires

4

Requirements

- Formative models
  - Item importance
    - Structured interviews & impact score
  - Item characteristics
    - Item scores, distribution, response curves
- Reflective models
  - Item characteristics
    - Item scores, distribution, response curves
  - Dimensionality
    - Factor analyses (EFA, CFA)
  - Reproducibility
    - Internal consistency
    - Reliability

de Vet et al., 2005; Raczek et al., 1998

5

Requirements

- Formative models
  - Item importance → N ≥ 100
    - Structured interviews & impact score
  - Item characteristics → N ≥ 100 (item scores & distribution); Response curves: Rasch ≥ 200, 2PL ≥ 1000
    - Item scores, distribution, response curves
- Reflective models
  - Item characteristics
    - Item scores, distribution, response curves
  - Dimensionality
    - Factor analyses (EFA, CFA)
  - Reproducibility
    - Internal consistency
    - Reliability

Target population ("rules of thumb")

de Vet et al., 2005; Raczek et al., 1998

6

Requirements

- **Formative models**
  - Item importance → Structured interviews & impact score
  - Item characteristics → Item scores, distribution, response curves

Target population ("rules of thumb")

- Item importance: N ≥ 100
- Item characteristics: N ≥ 100 (item scores & distribution); Response curves: Rasch ≥ 200, 2PL ≥ 1000

- **Reflective models**
  - Item characteristics → Item scores, distribution, response curves
  - Dimensionality → Factor analyses (EFA, CFA)
  - Reproducibility → Internal consistency, Reliability

- Item characteristics: See above
- Dimensionality: 4-10 subjects/variable, min. 100[1]
- Reproducibility: N ≥ 50

de Vet et al., 2005; Raczek et al., 1998

7

Item importance

- **Structured interviews**

- Asthma QoL Questionnaire (impairment with asthma)[1]
  - Symptoms, emotional problems, troublesome environmental stimuli, avoidance etc.

Sample questions of the AQLQ

IN GENERAL, HOW MUCH OF THE TIME DURING THE LAST 2 WEEKS DID YOU:

	All the time	Most of the time	A good bit of the time	Some of the time	A little of the time	Hardly any of the time	None of the time
7. Feel CONCERNED ABOUT HAVING ASTHMA?	1	2	3	4	5	6	●
8. Feel SHORT OF BREATHE as a result of your asthma?	1	2	3	4	5	6	●
9. Experience asthma symptoms as a RESULT OF BEING EXPOSED TO CIGARETTE SMOKE?	1	2	3	4	5	6	●
10. Experience WHEEZE in your chest?	1	2	3	4	5	6	●
11. Feel you had to AVOID A SITUATION OR ENVIRONMENT BECAUSE OF CIGARETTE SMOKE?	1	2	3	4	5	6	●

HOW MUCH DISCOMFORT OR DISTRESS HAVE YOU FELT DURING THE LAST 2 WEEKS?

	A very great deal	A great deal	Very Limited	Moderate amount	Some	Very little	None
12. How much discomfort or distress have you felt over the last 2 weeks as a result of COUGHING?	1	2	3	4	5	6	●

1. Juniper EF et al. Evaluation of impairment of health-related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax. 1992 Feb;47(2):76–83.

8

Item importance

- Structured interviews
  - Asthma QoL Questionnaire (impairment with asthma)[1]
    - Symptoms, emotional problems, troublesome environmental stimuli, avoidance etc.
  - Two extra questions:
    - Has 'item' been troublesome in the past year? (yes/no)
    - How important is the 'item'? (1 = not very important… 5 = extremely important)
  - Impact score

    Impact = proportion reporting the item as troublesome x mean importance (range 0 – 5)

    E.g. Shortness of breath: Troublesome = 92%, importance avg. = 3.60

    Impact = 0.92 x 3.60 = 3.31

1. Juniper EF et al. Evaluation of impairment of health-related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax. 1992 Feb;47(2):76–83.

Item characteristics

- Item scores, distribution and response curves

Oswestry Disability Index

- Questionnaire measuring 'pain-related function' in low back pain patients
- 10 items scored on an ordinal scale from 0-5
- Score range [0-50] where 0 = no disability and 50 = high disability

ODI item scores

Item score range [0 - 5]

Item	Content	Mean (N=233)	Median (N=233)
1	Pain intensity	2.60	3
2	Personal care (washing, dressing etc.)	1.02	1
3	Lifting	2.33	2
4	Walking	1.27	1
5	Sitting	1.99	2
6	Standing	1.92	2
7	Sleeping	1.33	1
8	Sex life (if applicable)	1.52	1
9	Social life	1.77	2
10	Travelling	1.64	1

Mean/median is low for all variables except 'pain intensity'

N.B. SD and IQR can be added

11

ODI item scores

Score range [0 - 5]

Item	Content	Mean (N=233)	Median (N=233)	Missing items (% of 233)
1	Pain intensity	2.60	3	0.0
2	Personal care (washing, dressing etc.)	1.02	1	1.4
3	Lifting	2.33	2	1.4
4	Walking	1.27	1	1.4
5	Sitting	1.99	2	1.4
6	Standing	1.92	2	0.7
7	Sleeping	1.33	1	2.7
8	Sex life (if applicable)	1.52	1	15.0
9	Social life	1.77	2	2.7
10	Travelling	1.64	1	0.7

What is acceptable?

Rule of thumb[1]:

< 3% = acceptable
> 15% = unacceptable

Decisions should balance missingness against the importance of the item

[1]Measurement in Medicine, de Vet et al. 2011

12

ODI item distribution

		Distribution (%) of response options					
Item	**Content**	**0**	**1**	**2**	**3**	**4**	**5**
1	Pain intensity	0	6.8	40.8	38.8	12.9	0.7
2	Personal care (washing, dressing etc)	27.9	47.6	19.1	2.7	0	1.4
3	Lifting	4.8	26.5	18.4	29.9	18.4	0.7
4	Walking	30.6	33.3	16.3	15.0	2.7	0.7
5	Sitting	9.5	18.4	44.2	18.4	6.8	1.4
6	Standing	10.2	30.6	25.2	23.1	10.2	0.7
7	Sleeping	12.9	52.4	21.1	8.8	2.0	0
8	Sex life (if applicable)	24.5	25.9	14.3	9.5	8.2	2.7
9	Social life	16.3	21.8	31.3	23.8	4.1	2.7
10	Travelling	9.5	40.8	30.6	13.6	3.4	1.4

→ Looks like a symmetrical distribution

→ Looks like very skewed distributions

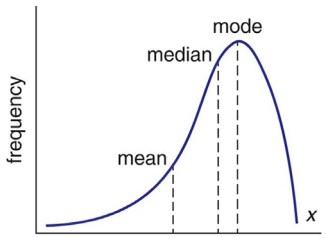
13

Skewness

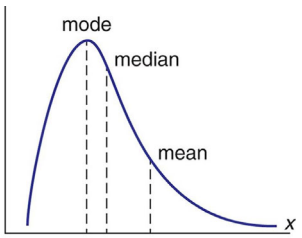
- A measure of the lack of symmetry of a distribution
  - 0 = symmetrical distribution
  - - coefficient → median>mean → left or negative skew
  - + coefficient → median<mean → right or positive skew

Usually!

Left/negative skew



Right/positive skew

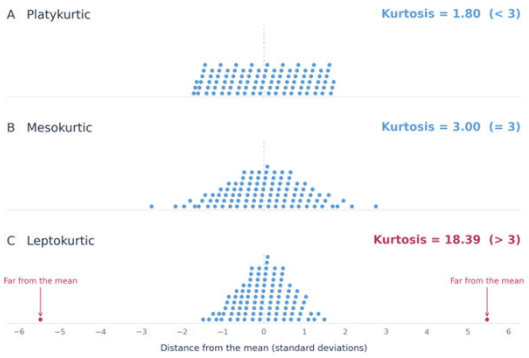


* Depends on package used

14

ODI item distribution

- Kurtosis
  - A measure of how extreme observations are relative to the overall spread of the data
  - Higher kurtosis indicates a stronger contribution from extreme observations
    - **Platykurtic:** below the normal-distribution reference, (Stata <3, R* <0)
    - **Mesokurtic:** equal to the normal-distribution reference, (Stata =3, R* =0)
    - **Leptokurtic:** above the normal-distribution reference, (Stata >3, R* >0)



* Depends on package used

15

ODI item distribution

- = left skewed
+ = right skewed

Stata:
3 = normal distribution
<3 = flat curve
>3 = peaked curve

		Distribution (%) over response options							
Item	Content	0	1	2	3	4	5	**Skewness**	**Kurtosis**
1	Pain intensity	0	6.8	40.8	38.8	12.9	0.7	0.2	2.7
2	Personal care (washing, dressing etc.)	27.9	47.6	19.1	2.7	0	1.4	1.3	6.7
3	Lifting	4.8	26.5	18.4	29.9	18.4	0.7	-0.1	1.9
4	Walking	30.6	33.3	16.3	15.0	2.7	0.7	0.7	2.7
5	Sitting	9.5	18.4	44.2	18.4	6.8	1.4	0.2	3.1
6	Standing	10.2	30.6	25.2	23.1	10.2	0.7	0.1	2.1
7	Sleeping	12.9	52.4	21.1	8.8	2.0	0	0.8	3.6
8	Sex life (if applicable)	24.5	25.9	14.3	9.5	8.2	2.7	0.8	2.6
9	Social life	16.3	21.8	31.3	23.8	4.1	2.7	-0.1	2.1
10	Travelling	9.5	40.8	30.6	13.6	3.4	1.4	0.1	3.6

→ ≈ Symmetrical, meso
→ ≈ right skewed, lepto
→ ≈ Symmetrical, platy

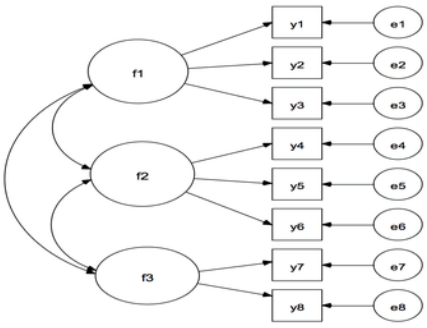
16

Requirements

- Formative models
  - Item importance
    - Structured interviews & impact score
  - Item characteristics
    - Item scores, distribution
- Reflective models
  - Item characteristics
    - Item scores, distribution
  - Dimensionality
    - Factor analyses (EFA, CFA)
  - Other
    - Internal consistency (reproducibility)
    - Item difficulty and targeting (IRT)
    - Floor/ceiling effect (interpretation)

Dimensionality

- Factor analyses (EFA, CFA)
- Eva

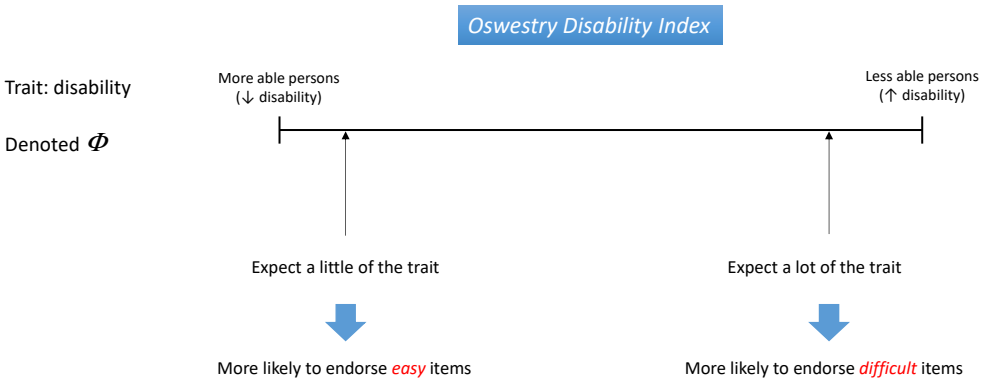


Item difficulty and targeting

- Established using Item Response Theory (IRT) → day 3
- Can be used to reduce number of items
- Uses
  - A persons trail level or person ability (Φ)
  - The difficulty of the item

19

Trait level or person ability (Φ)



20

Item difficulty

- *Item difficulty describes an item's location on the measured trait scale*
  - ***Dichotomous item:*** *The trait level at which the probability of endorsing the item is 50%*
  - ***Polytomous item****: Average of category thresholds. Each category threshold is the trait level at which two adjacent response categories are equally likely*

Oswestry Disability Index

An easy item

- An item that a patient with *a little* back pain will endorse →

Afsnit 1: Smerter
- ❑ Jeg har ingen smerter for øjeblikket
- ❑ Smerterne er meget svage for øjeblikket
- ❑ Smerterne er moderate for øjeblikket
- ❑ Smerterne er forholdsvis kraftige for øjeblikket
- ❑ Smerterne er meget kraftige for øjeblikket
- ❑ Smerterne er de værst tænkelige for øjeblikket

A difficult* item

- An item that a patient with *a lot of* back pain will endore →

Afsnit 4: Gå
- ❑ Jeg kan gå så langt jeg har lyst, selvom jeg har smerter
- ❑ Smerterne hindrer mig i at gå mere end 2 kilometer
- ❑ Smerterne hindrer mig i at gå mere end 1 kilometer
- ❑ Smerterne hindrer mig i at gå mere end 500 meter
- ❑ Jeg kan kun gå, når jeg bruger stok eller krykker
- ❑ Jeg ligger i sengen det meste af tiden og må kravle ud til toilettet

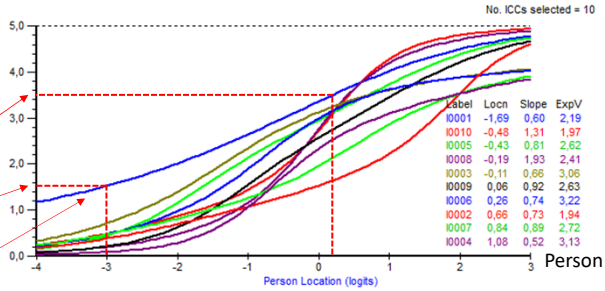
* A difficult item = a hard item

21

Item difficulty

- Item characteristic curves - ODI

Response categories



Person ability

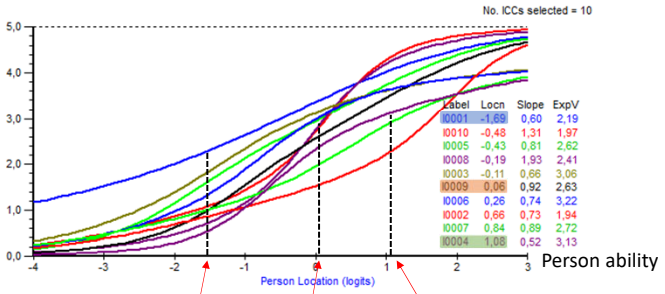
Afsnit 1: Smerter
- ❑ Jeg har ingen smerter for øjeblikket
- ❑ Smerterne er meget svage for øjeblikket
- ❑ Smerterne er moderate for øjeblikket
- ❑ Smerterne er forholdsvis kraftige for øjeblikket
- ❑ Smerterne er meget kraftige for øjeblikket
- ❑ Smerterne er de værst tænkelige for øjeblikket

22

Item difficulty

- Item characteristic curves - ODI

Response categories



Person ability

Item 1, Pain | Item 9, Social life | Item 4, Walking

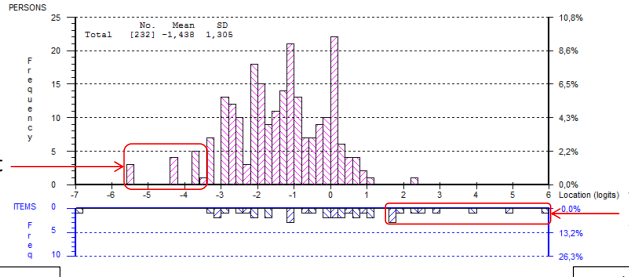
The easiest item

The most difficult item

The item closet to the person mean

Targeting

- Item difficulty has to correspond to the target populations' ability
- Poor targeting → low reliability → difficult to differentiate patients



Floor effect

Thresholds poorly targeted this sample

More able persons
Easier items

Less able persons
More difficult items

Interpretation:
Patients has less disability (mean: -1.44) compared to item difficulty → some items are rarely used as few respondents have limitations with them, and more items where respondents have little limitations (easier items) are needed.

Item response curves

Item threshold
The trait level at which two adjacent response categories have *equal* probabilities

We expect
Patients with a high amount of the measured trait should have a higher probability of endorsing say category 3 compared to category 2

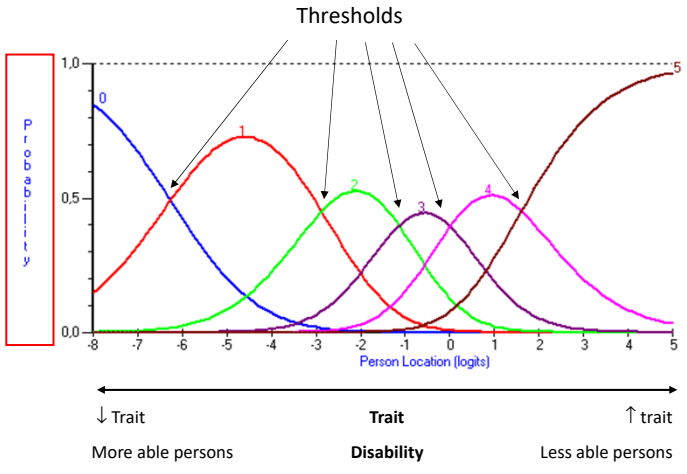
Oswestry Disability Index

Afsnit 1: Smerter

0 ❑ Jeg har ingen smerter for øjeblikket
1 ❑ Smerterne er meget svage for øjeblikket
2 ❑ Smerterne er moderate for øjeblikket
3 ❑ Smerterne er forholdsvis kraftige for øjeblikket
4 ❑ Smerterne er meget kraftige for øjeblikket
5 ❑ Smerterne er de værst tænkelige for øjeblikket

25

ODI item response curves



Oswestry Disability Index

Afsnit 1: Smerter

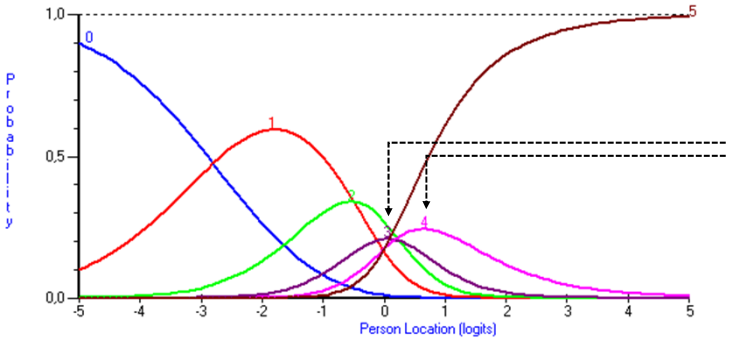
❑ Jeg har ingen smerter for øjeblikket
❑ Smerterne er meget svage for øjeblikket
❑ Smerterne er moderate for øjeblikket
❑ Smerterne er forholdsvis kraftige for øjeblikket
❑ Smerterne er meget kraftige for øjeblikket
❑ Smerterne er de værst tænkelige for øjeblikket

26

ODI item response curves

Afsnit 10: Rejse

0 ❑ Jeg kan rejse hvorhen jeg vil uden smerter
1 ❑ Jeg kan rejse hvorhen jeg vil, men det giver mig flere smerter
2 ❑ Smerterne er slemme, men jeg kan godt klare over 2 timers rejse
3 ❑ Smerterne begrænser mine rejser til mindre end 1 time
4 ❑ Smerterne begrænser mine rejser til korte, nødvendige rejser under 30 minutter
5 ❑ Smerterne hindrer mig i at rejse, undtagen for at få behandling



Disordered thresholds

Categories 3 and 4 are never the most probable response

27

Requirements

- Formative models
  - Item importance
    - Structured interviews & impact score
  - Item characteristics
    - Item scores, distribution
- Reflective models
  - Item characteristics
    - Item scores, distribution
  - Dimensionality
    - Factor analyses (EFA, CFA)
  - Other
    - Internal consistency (reproducibility)
    - Item difficulty and targeting (IRT)
    - Floor/ceiling effect (interpretation)

[1]de Vet HC et al. Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. Quality of Life Research. 2005;14(5):1203–18
* Item Characteristic Curve

28

Internal consistency

"Degree of inter-relatedness among items"

(COSMIN 2011)

- Used on multi-item questionnaires
- Assumes that all items measure the same concept
  - Item scorer are correlated
  - No items are irrelevant
  - Unidimensionality
- Cronbach's α

Alpha does not establish unidimensionality, distinguish reflective from formative models, or demonstrate validity

29

Alpha output

Include std (z-scores with mean = 0, variance = 1) if the response options are **not** on the same scale.

```
. alpha osw*, item label
```

Without std → alpha based on covariances
With std → alpha based on correlations

```
Test scale = mean(unstandardized items)

                        Item-test  Item-rest  Interitem
Item         |  Obs  Sign   corr.      corr.       cov.      alpha   Label
-------------+--------------------------------------------------------------------------
osw1         |  232    +    0.6571     0.5695    .5660488   0.8640   Pain intensity
osw2         |  232    +    0.6414     0.5772    .5944381   0.8658   Personal care
osw3         |  231    +    0.6898     0.5939    .5417706   0.8619   Lifting
osw4         |  230    +    0.6668     0.5643    .5469499   0.8640   Walking
osw5         |  230    +    0.6978     0.5978    .5364814   0.8621   Sitting
osw6         |  230    +    0.6716     0.5640    .5392544   0.8641   Standing
osw7         |  230    +    0.6164     0.5372    .5849837   0.8662   Sleeping
osw8         |  208    +    0.7386     0.6419    .5143936   0.8576   Sexlife (if applicable)
osw9         |  228    +    0.7080     0.6126    .5330442   0.8600   Social life
osw10        |  218    +    0.8066     0.7374    .5039011   0.8495   Travelling
-------------+--------------------------------------------------------------------------
Test scale   |                                   .5460109   0.8737   mean(unstandardized items)
-------------------------------------------------------------------------------------------
```

30

Alpha output

Pearson corr. calculated for pairs of scores where one item of each pair is an item score and the other item is the total test score.

Item correlation with sumscore of the **rest** of the items

Should be roughly the same for all items

Item-row entries describe covariance/correlation among the remaining items after deletion.

Covariance (non-standardized)/Correlation (standardized): Compare with the full-scale average. An increase after item removal may flag an item with relatively low average covariance/correlation with the other items.

```
. alpha osw*, item label

Test scale = mean(unstandardized items)

                          Item-test   Item-rest   Interitem
Item          |  Obs  Sign    corr.       corr.       cov.       alpha  Label
--------------+-------------------------------------------------------------------------
osw1          |  232    +    0.6571      0.5695    .5660488    0.8640  Pain intensity
osw2          |  232    +    0.6414      0.5772    .5944381    0.8658  Personal care
osw3          |  231    +    0.6898      0.5939    .5417706    0.8619  Lifting
osw4          |  230    +    0.6668      0.5643    .5469499    0.8640  Walking
osw5          |  230    +    0.6978      0.5978    .5364814    0.8621  Sitting
osw6          |  230    +    0.6716      0.5640    .5392544    0.8641  Standing
osw7          |  230    +    0.6164      0.5372    .5849837    0.8662  Sleeping
osw8          |  208    +    0.7386      0.6419    .5143936    0.8576  Sexlife (if applicable)
osw9          |  228    +    0.7080      0.6126    .5330442    0.8600  Social life
osw10         |  218    +    0.8066      0.7374    .5039011    0.8495  Travelling
--------------+-------------------------------------------------------------------------
Test scale    |                                   .5460109    0.8737  mean(unstandardized items)
----------------------------------------------------------------------------------------
```

Item-row entries describe Alpha among the remaining items after deletion.

If substantially higher than Chronbach's α → maybe misfitting

Average interitem covariance

Cronbach's α

31

Alpha output – item removal

```
. alpha osw*, item label

Test scale = mean(unstandardized items)

                          Item-test   Item-rest   Interitem
Item          |  Obs  Sign    corr.       corr.       cov.       alpha  Label
--------------+-------------------------------------------------------------------------
osw1          |  232    +    0.6571      0.5695    .5660488    0.8640  Pain intensity
osw2          |  232    +    0.6414      0.5772    .5944381    0.8658  Personal care
osw3          |  231    +    0.6898      0.5939    .5417706    0.8619  Lifting
osw4          |  230    +    0.6668      0.5643    .5469499    0.8640  Walking
osw5          |  230    +    0.6978      0.5978    .5364814    0.8621  Sitting
osw6          |  230    +    0.6716      0.5640    .5392544    0.8641  Standing
osw7          |  230    +    0.6164      0.5372    .5849837    0.8662  Sleeping
osw8          |  208    +    0.7386      0.6419    .5143936    0.8576  Sexlife (if applicable)
osw9          |  228    +    0.7080      0.6126    .5330442    0.8600  Social life
osw10         |  218    +    0.8066      0.7374    .5039011    0.8495  Travelling
--------------+-------------------------------------------------------------------------
Test scale    |                                   .5460109    0.8737  mean(unstandardized items)
----------------------------------------------------------------------------------------
```

Lets remove item 9

32

Alpha output

```
. alpha osw1-osw8 osw10, item label

Test scale = mean(unstandardized items)
```

	Interitem cov	alpha
Item 9:	.5330442	0.8600

Item	Obs	Sign	Item-test corr.	Item-rest corr.	Interitem cov.	alpha	Label
osw1	232	+	0.6744	0.5784	.550562	0.8466	Pain intensity
osw2	232	+	0.6382	0.5658	.5865907	0.8505	Personal care
osw3	231	+	0.7057	0.6008	.5233115	0.8439	Lifting
osw4	230	+	0.6716	0.5562	.5328375	0.8481	Walking
osw5	230	+	0.7044	0.5921	.5208688	0.8455	Sitting
osw6	230	+	0.6706	0.5476	.5266775	0.8492	Standing
osw7	230	+	0.6252	0.5368	.5737572	0.8501	Sleeping
osw8	208	+	0.7385	0.6281	.4986367	0.8409	Sexlife (if applicable)
osw10	218	+	0.8059	0.7263	.485573	0.8301	Travelling
Test scale					.5330442	0.8600	mean(unstandardized items)

33

Finalising the questionnaire

- Combine statistical findings with item relevance when selecting items
- Check content coverage: do the retained items adequately represent the intended construct?
- Revise or add items where problems or gaps remain
- Document changes and their rationale, including reasons for retaining or removing items
- Re-evaluate the revised questionnaire before using it in research or clinical practice

34

Finalising the questionnaire

Five takeaways for a considered final version

01 **Evidence + relevance**
Balance statistics with item relevance.

02 **Content coverage**
Keep the intended construct represented.

03 **Item revision**
Revise or add items where gaps remain.

04 **Decision record**
Record changes and the reasons for them.

05 **Re-evaluation**
Evaluate the revised questionnaire before use.

TAKEAWAY 01

Item selection

Combine statistical findings with item relevance when selecting items.

Show overall takeaway

35

Questions



36