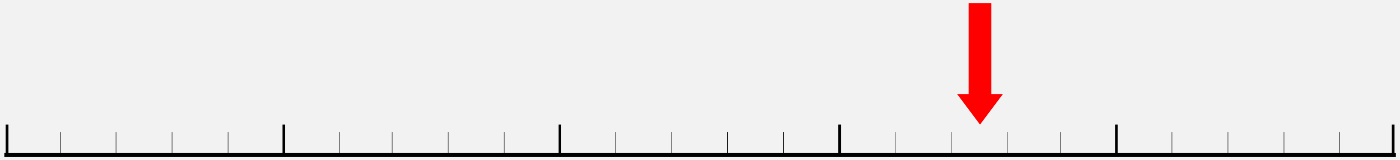
The dark arts – how to measure things we cannot see

An introduction to psychometric measurement

Andrew Mackinnon
Centre for Youth Mental Health
University of Melbourne

Fundamental concept

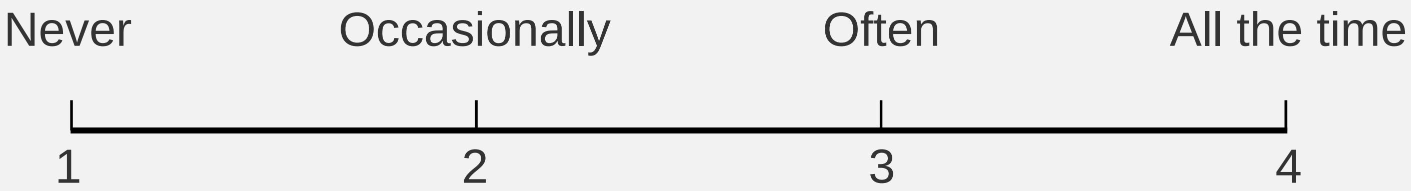


To locate individuals on a continuum/line representing a construct of interest

- **Scale should have appropriate mathematical properties (additivity) –**
  - **able to compare difference between individuals/groups**
  - **change in individuals/groups over time**
  - **calculate summary statistics, e.g., means**
- **Units will be arbitrary**
- **Distribution will be arbitrary**
- **Zero point arbitrary/undefined – ratio scales unlikely**

Scales/tests/inventories…

Getting up and going to school is a big hassle for me…

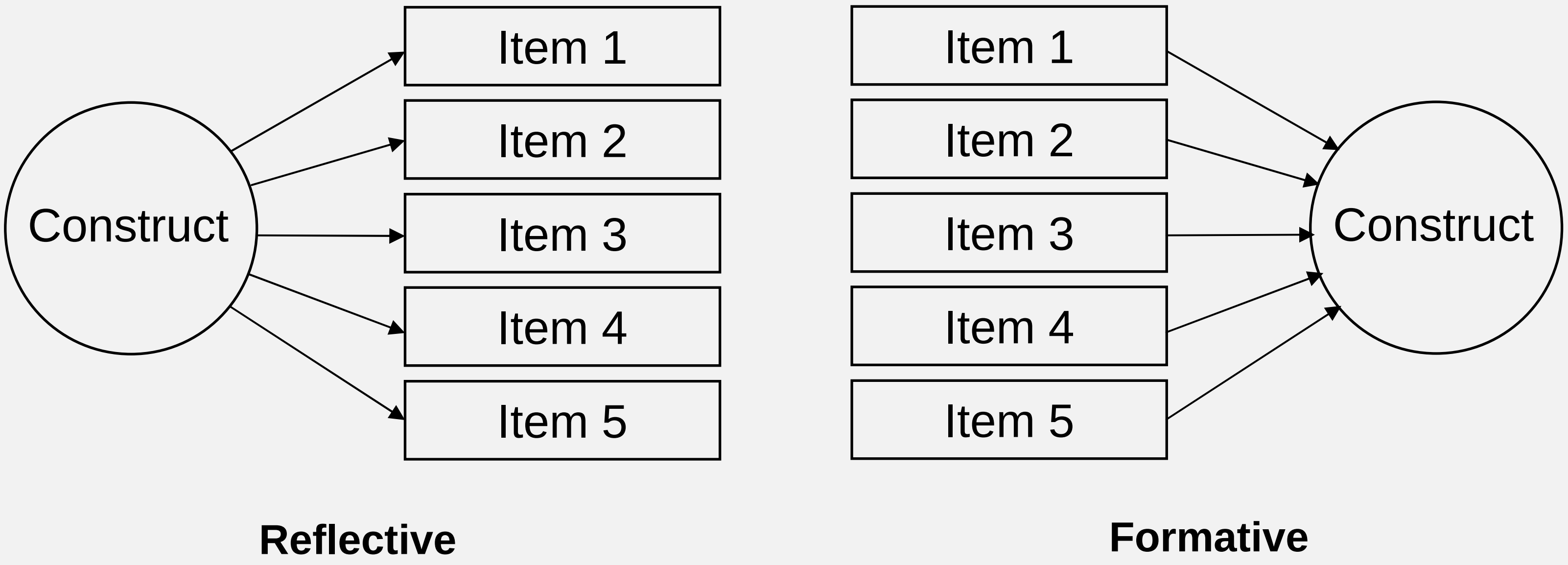


Never | Occasionally | Often | All the time

1 | 2 | 3 | 4

- **Multiple questions, statements or items**
- **Each has a response scale. These may be:**
  - **statements of frequency, severity, endorsement (agreement), self-appraisal**
  - **binary (yes/no)**
  - **multiple ordered categories (Likert scale)**
  - **have numbers associated with them**
  - **differ from item to item**
- **Ultimately, responses are combined – usually by (weighted) addition**
- **Measurements are not counts (even if they are)!**

Formative *vs* reflective scales

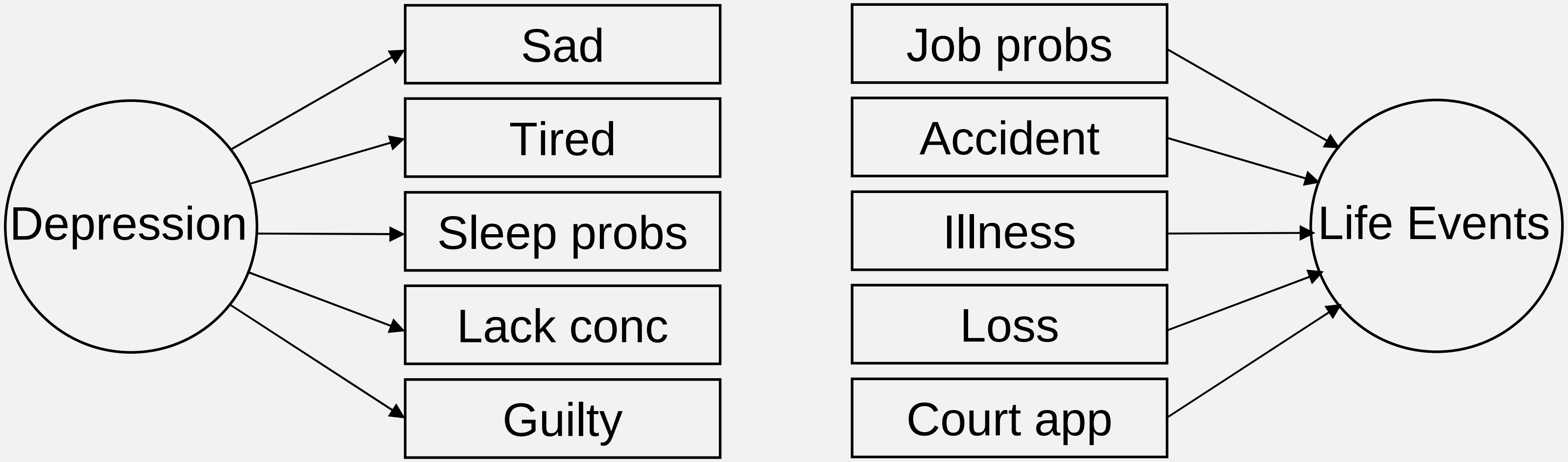


Reflective: The response to each item reflects a status of the respondent on the underlying construct or continuum

Formative: The construct is defined by the combination/cummulation of its indicators which may or may not be correlated

Different approaches are required to evaluate each type of scale

Formative *vs* reflective scales



Reflective

Formative

Reflective: The response to each item reflects the status of the respondent on the underlying construct or continuum

Formative: The construct is defined by the combination/cummulation of its indicators which may not be correlated

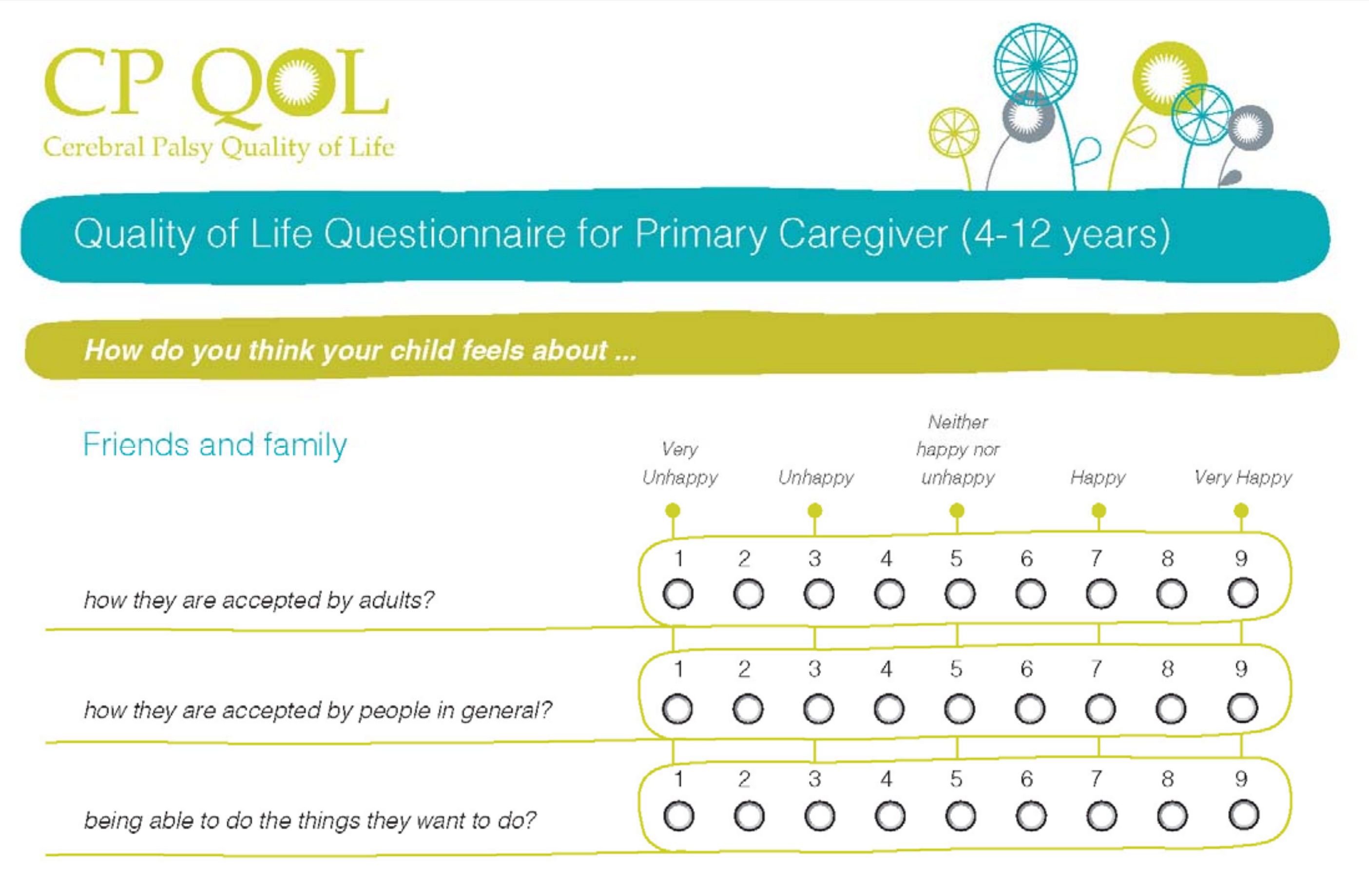
Different approaches are required to evaluate each type of scale

CP – QOL Child

Elizabeth Waters, Elise Davis
Dinah Reddihough, H. Kerr Graham, Roslyn Boyd
Sing Kai Lo, Rory Wolfe, Richard Stevenson, Kristie Bjornson
Eve Blair, Peter Hoare, Ulrike Ravens-Sieberer

- A condition-specific quality of life instrument for children with cerebral palsy
- Inquires about physical well-being, social well-being, emotional well-being, school, access to services, and acceptance by others.
- for children aged 4 to 12 years
- primary caregiver-proxy form – 66 items
- child self-report form (age 9–12y) contains 52 items
- 9-point rating scale: 1=very unhappy — 9=very happy
- yields 6 analytically derived subscales (plus family health)

CP – QOL Child



CP QOL

Cerebral Palsy Quality of Life

Quality of Life Questionnaire for Primary Caregiver (4-12 years)

How do you think your child feels about ...

Friends and family

	Very Unhappy		Unhappy		Neither happy nor unhappy		Happy		Very Happy
	1	2	3	4	5	6	7	8	9
how they are accepted by adults?	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
how they are accepted by people in general?	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
being able to do the things they want to do?	○	○	○	○	○	○	○	○	○

http://www.cpqol.org.au/

The big issues

- What to measure
- Who to measure
- What items to consider
- What are you measuring
- What items to retain
- How well are you measuring
- … and in whom and when

What to measure – the big picture

Central construct/motivation

- Specifically for children with CP
- Focus on well-being (rather than ill-being)
- Subjective – '*how do you feel about*' not 'how is'

Applications

- Change over time (including effect of interventions)

Alternatives

- generic health-related QOL e.g. KIDSCREEN
  - 'too' generic, inappropriate (e.g., Have you felt fit and well?)
- functioning – many available QOL measures are actually measures of functioning e.g., Pediatric QOL Questionnaire CP Module
  - items: difficulty moving one or both legs, difficulty using scissors, difficulty brushing teeth

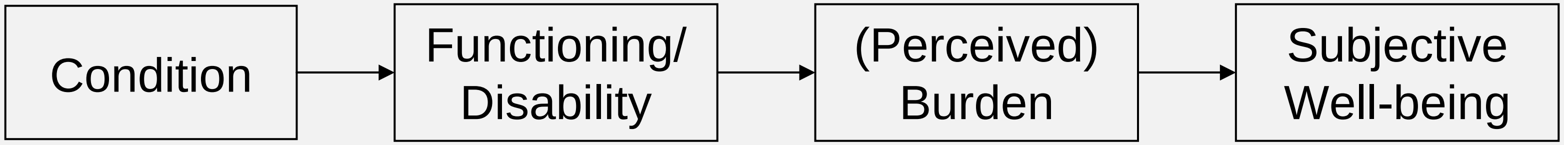
What to measure – a good idea?

Maybe

- Objective and/or self report measures of functioning are a potential incomplete picture of status or outcome

Maybe not

- Subjective well-being is quite resilient to positive and negative life events and circumstances
- Modest ‘sensitivity to change’



- Perhaps we should have measured a little closer to the ‘source’
- You need to comprehensively understand the construct you seek to measure

Who to measure

- Proxy reports are inevitable for young children and some people with disabilities
- Key Questions:
  - will proxies have enough knowledge of the child to make ratings?
  - do proxies have the background to make relative assessments?
- Teachers and professional carers may not know a child well
  - differential effects: some states or behaviour may be more apparent than other – absence of evidence
- Parents deal with only a few children. Can they place their child in the spectrum of possible response?
- Proxy report may become less complete as a child ages
  - adolescents and social or personal relations
- All proxies must infer internal states if these are inquired about
- Self and proxy reports may be complementary

Who to measure – development sample

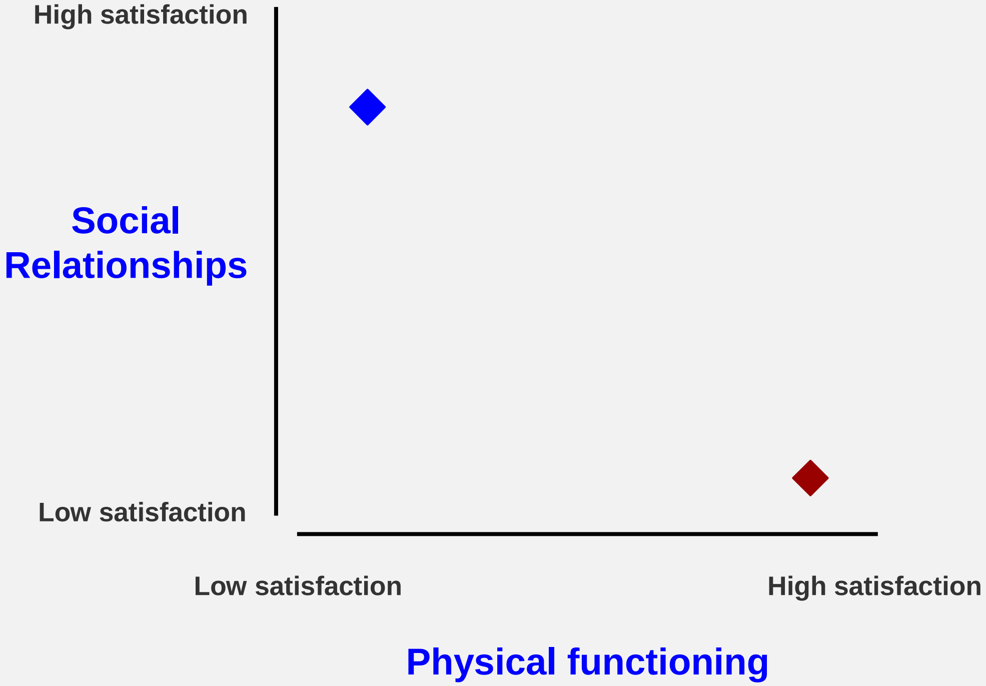
- Item content development may be based on small samples if they can give comprehensive insight into the target construct
- Development samples must include adequate numbers of respondents in regions of interest
  - in medical research, interest often lies towards the extremes of continua
- Initial analysis is possible with quite modest samples (100-200)
- Ultimately large (representative) samples are needed for good measurement work
  - >1000
  - response banking

What to measure – details

- The content of individual items must be decided
  - fiat or reference to standards and definition
  - focus groups and other qualitative methods
  - ‘theft’ and avoidance – comparison with current or like scales
- Most qualitative methods are likely to generate ‘positive’ instances of the attribute
- Explore to ‘edges’ of the construct to establish what it is not
- Consider location or severity in developing items
- Items which are inapplicable to particular groups make scaling very difficult
- Collect NA and “Don’t know” responses in pilot testing but not after

Dimensionality

- Good reflective scales must measure just one dimension (inventories may contain more than one scale)



If satisfaction with physical functioning and satisfaction with social relationships are combined ◆ and ◆ will have the same scores and be indistinguishable.

Dimensionality

- **Many types of analysis assume unidimensionality but do not test this**
- **Factor analysis is ~~the~~ my method of choice for investigating and establishing the dimensionality of a set of items**

$$\textit{Response} = \beta_0 + \beta_1 \textit{Factor}_1 + \beta_2 \textit{Factor}_2 \ldots + e$$

- **Factor_1, Factor_2 … are not observed and must be inferred**
- **Factor analysis can determine the number of dimensions underlying a set of items and the relationship of each item to each dimension**
- **For binary and polychotomous responses special methods of factor analysis are available and preferable**
- **Principal components analysis is pragmatically comparable to factor analysis – numerically simpler but often 'optimistic'**

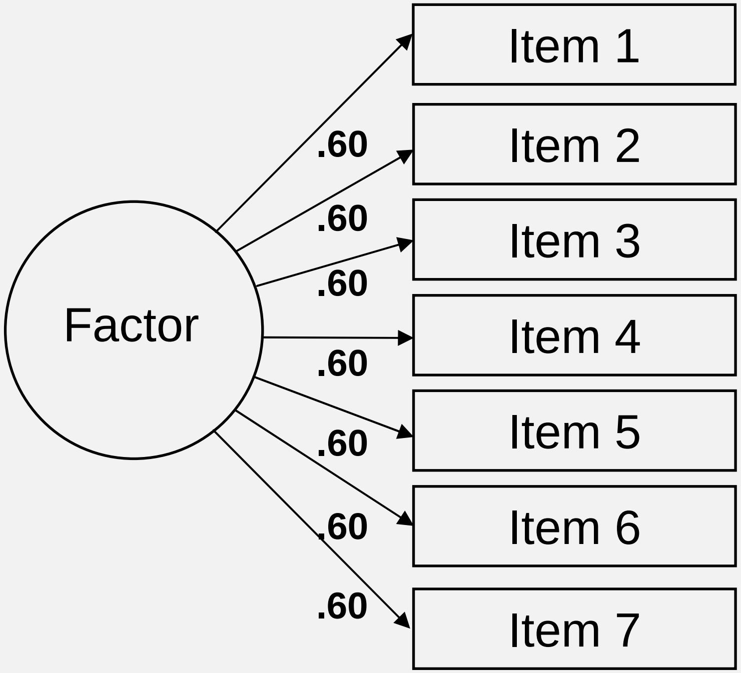
Factor analysis

Item

	1	2	3	4	5	6	7
1	1.00	0.36	0.36	0.36	0.36	0.36	0.36
2	0.36	1.00	0.36	0.36	0.36	0.36	0.36
3	0.36	0.36	1.00	0.36	0.36	0.36	0.36
4	0.36	0.36	0.36	1.00	0.36	0.36	0.36
5	0.36	0.36	0.36	0.36	1.00	0.36	0.36
6	0.36	0.36	0.36	0.36	0.36	1.00	0.36
7	0.36	0.36	0.36	0.36	0.36	0.36	1.00

Eigenvalues

1	2	3	4	5	6	7
3.61	0.64	0.64	0.64	0.64	0.64	0.64



Single factor with uniform loadings

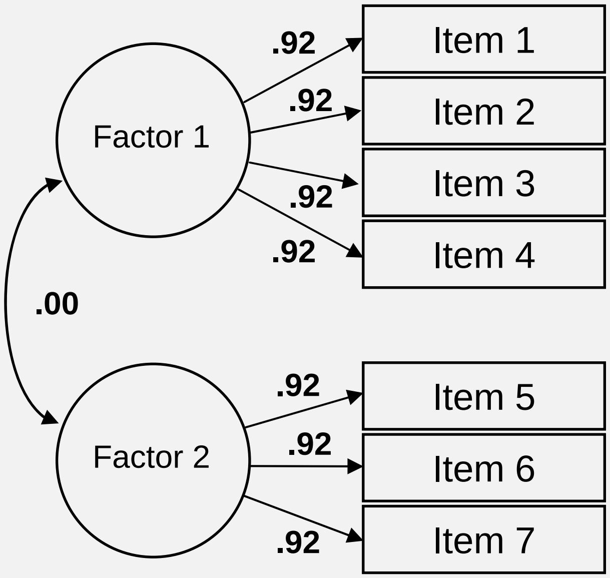
Factor analysis

Item

	1	2	3	4	5	6	7
1	1.00	0.85	0.85	0.85	0.00	0.00	0.00
2	0.85	1.00	0.85	0.85	0.00	0.00	0.00
3	0.85	0.85	1.00	0.85	0.00	0.00	0.00
4	0.85	0.85	0.85	1.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	1.00	0.85	0.85
6	0.00	0.00	0.00	0.00	0.85	1.00	0.85
7	0.00	0.00	0.00	0.00	0.85	0.85	1.00

Eigenvalues

1	2	3	4	5	6	7
3.55	2.70	0.15	0.15	0.15	0.15	0.15



Two uncorrelated factors

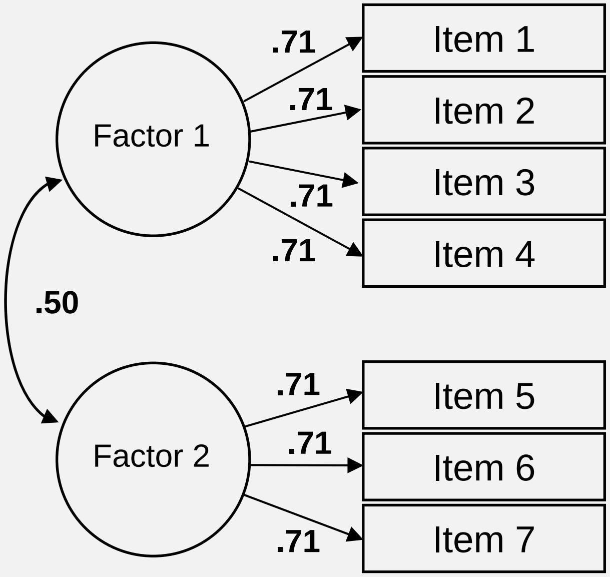
Factor analysis

Item

	1	2	3	4	5	6	7
1	1.00	0.50	0.50	0.50	0.25	0.25	0.25
2	0.50	1.00	0.50	0.50	0.25	0.25	0.25
3	0.50	0.50	1.00	0.50	0.25	0.25	0.25
4	0.50	0.50	0.50	1.00	0.25	0.25	0.25
5	0.25	0.25	0.25	0.25	1.00	0.50	0.50
6	0.25	0.25	0.25	0.25	0.50	1.00	0.50
7	0.25	0.25	0.25	0.25	0.50	0.50	1.00

Eigenvalues

1	2	3	4	5	6	7
3.15	1.35	0.50	0.50	0.50	0.50	0.50

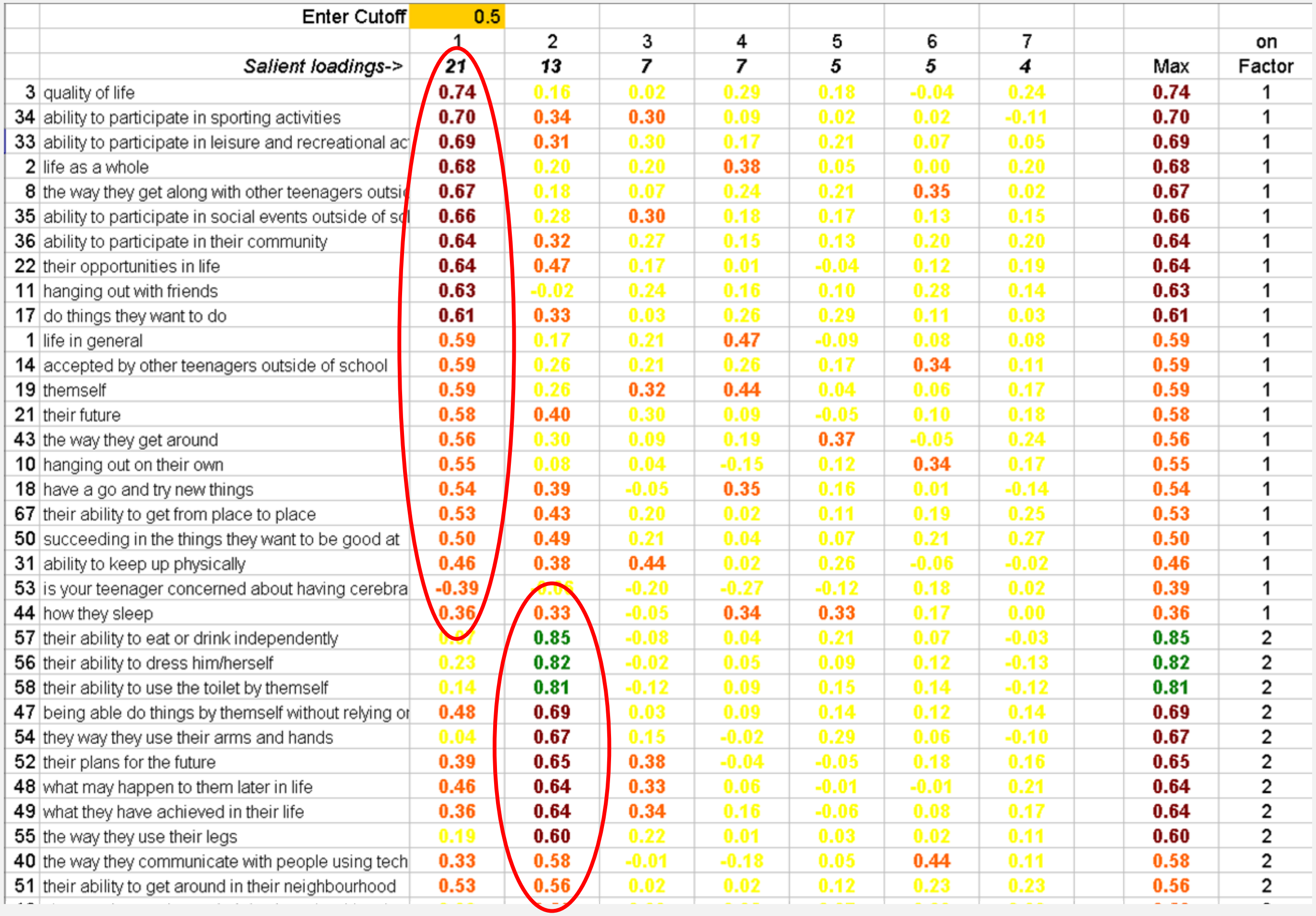


Two correlated factors

Dimensionality

- Items are a sample from a universe of items – more is (almost always) better
- Number of high loadings does not reflect the importance of a dimension
- Check items that don’t load anywhere – an item that doesn’t work or a single indicator of an important attribute?
- Routine FA gives no information about where an item is located on a dimension
- A considerable number of items are required to assess a dimension with any precision (bare minimum=3, 20 binary items desirable)
- Factors with a small number of loadings often reflect content specificity rather than a significant dimension of variation
- Factor analysis will not sort out causes, construct and consequences – substantive knowledge must ultimately inform item selection

Dimensionality in the CP - QOL



	Enter Cutoff	0.5									
		1	2	3	4	5	6	7			on
	Salient loadings->	*21*	*13*	*7*	*7*	*5*	*5*	*4*		Max	Factor
3	quality of life	0.74	0.16	0.02	0.29	0.18	-0.04	0.24		0.74	1
34	ability to participate in sporting activities	0.70	0.34	0.30	0.09	0.02	0.02	-0.11		0.70	1
33	ability to participate in leisure and recreational ac	0.69	0.31	0.30	0.17	0.21	0.07	0.05		0.69	1
2	life as a whole	0.68	0.20	0.20	0.38	0.05	0.00	0.20		0.68	1
8	the way they get along with other teenagers outsi	0.67	0.18	0.07	0.24	0.21	0.35	0.02		0.67	1
35	ability to participate in social events outside of sc	0.66	0.28	0.30	0.18	0.17	0.13	0.15		0.66	1
36	ability to participate in their community	0.64	0.32	0.27	0.15	0.13	0.20	0.20		0.64	1
22	their opportunities in life	0.64	0.47	0.17	0.01	-0.04	0.12	0.19		0.64	1
11	hanging out with friends	0.63	-0.02	0.24	0.16	0.10	0.28	0.14		0.63	1
17	do things they want to do	0.61	0.33	0.03	0.26	0.29	0.11	0.03		0.61	1
1	life in general	0.59	0.17	0.21	0.47	-0.09	0.08	0.08		0.59	1
14	accepted by other teenagers outside of school	0.59	0.26	0.21	0.26	0.17	0.34	0.11		0.59	1
19	themself	0.59	0.26	0.32	0.44	0.04	0.06	0.17		0.59	1
21	their future	0.58	0.40	0.30	0.09	-0.05	0.10	0.18		0.58	1
43	the way they get around	0.56	0.30	0.09	0.19	0.37	-0.05	0.24		0.56	1
10	hanging out on their own	0.55	0.08	0.04	-0.15	0.12	0.34	0.17		0.55	1
18	have a go and try new things	0.54	0.39	-0.05	0.35	0.16	0.01	-0.14		0.54	1
67	their ability to get from place to place	0.53	0.43	0.20	0.02	0.11	0.19	0.25		0.53	1
50	succeeding in the things they want to be good at	0.50	0.49	0.21	0.04	0.07	0.21	0.27		0.50	1
31	ability to keep up physically	0.46	0.38	0.44	0.02	0.26	-0.06	-0.02		0.46	1
53	is your teenager concerned about having cerebra	-0.39	-0.06	-0.20	-0.27	-0.12	0.18	0.02		0.39	1
44	how they sleep	0.36	0.33	-0.05	0.34	0.33	0.17	0.00		0.36	1
57	their ability to eat or drink independently	0.07	0.85	-0.08	0.04	0.21	0.07	-0.03		0.85	2
56	their ability to dress him/herself	0.23	0.82	-0.02	0.05	0.09	0.12	-0.13		0.82	2
58	their ability to use the toilet by themself	0.14	0.81	-0.12	0.09	0.15	0.14	-0.12		0.81	2
47	being able do things by themself without relying or	0.48	0.69	0.03	0.09	0.14	0.12	0.14		0.69	2
54	they way they use their arms and hands	0.04	0.67	0.15	-0.02	0.29	0.06	-0.10		0.67	2
52	their plans for the future	0.39	0.65	0.38	-0.04	-0.05	0.18	0.16		0.65	2
48	what may happen to them later in life	0.46	0.64	0.33	0.06	-0.01	-0.01	0.21		0.64	2
49	what they have achieved in their life	0.36	0.64	0.34	0.16	-0.06	0.08	0.17		0.64	2
55	the way they use their legs	0.19	0.60	0.22	0.01	0.03	0.02	0.11		0.60	2
40	the way they communicate with people using tech	0.33	0.58	-0.01	-0.18	0.05	0.44	0.11		0.58	2
51	their ability to get around in their neighbourhood	0.53	0.56	0.02	0.02	0.12	0.23	0.23		0.56	2

Genuinely exploratory factor analysis!

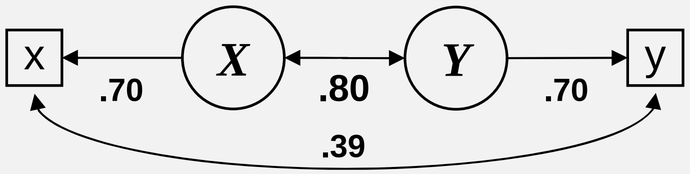
Dimensionality in the CP – QOL Child

Seven ‘proper’ interpretable factors (from 13 with eigenvalues > 1)

- Social well-being and acceptance (11 items)
  - the way they get along with other children at preschool or school?
  - how they are accepted by adults?
- Functioning (12 items)
  - their ability to play on their own?
  - the way they use their hands?
- Participation and physical health (11 items)
  - their ability to participate in recreational activities?
  - being able to do the things they want to do?
- Emotional well-being (6 items)
- Pain and impact of disability (8 items)
- Access to services (5 items – parent only)
- Family health (4 items – parent only)

Reliability – precision of measurement

- **The precision with which a test measures individuals is referred to as reliability**
- **Low reliability attenuates relationships with other variables**



- **Classical test theory model:**

  Observed Score (O) = True *Score* (T) + *e* **e ~ *D*(0, σ^2_E)**

- **Reliability:** $\rho_{TT'} = \rho^2_{OT} = \frac{\sigma^2_{OT}}{\sigma^2_O \sigma^2_T} = \frac{\sigma^2_T}{\sigma^2_O} = 1 - \frac{\sigma^2_E}{\sigma^2_O}$
- **But T is unobservable. Useable approaches**
  - **test-retest correlation**
  - **correlations between alternative measures**
  - **correlations between components**

Test-retest reliability

- Administer test a second time and compare results first and second measurements
- Statistics:
  - intraclass correlation
  - correlation and test comparing means (e.g. paired t-test)
- Issues:
  - learning / exposure effects
  - memory
  - individuals may have genuinely changed!

CP-QOL test-retest reliability

CP QOL- Child	Intraclass correlation
Social wellbeing and acceptance	0.87
Functioning	0.89
Participation & physical health	0.81
Emotional wellbeing	0.79
Access to services (Parent)	0.76
Pain and feelings about disability	0.78
Family health (Parent)	0.82

Different sources of information as ‘reliability’

- Where information is obtained from multiple sources it may be compared:
  - parent–child, parent–teacher, father–mother
- Differences may be ‘real’

CP QOL- Child	Parent-child correlation
Social wellbeing and acceptance	0.66
Functioning	0.77
Participation & physical health	0.65
Emotional wellbeing	0.74
Pain and feelings about disability	0.52

Internal consistency reliability

Rather than compare the whole test on two occasions, compare parts of it measured at the same time

- Split half – correlate two chosen or random subsets of items and adjust for the tests being half the length
- ‘Average’ over all possible divisions

Cronbach’s alpha $\alpha = \frac{k\overline{cov}}{\left(\overline{var} + (k-1)\overline{cov}\right)}$

The world’s most useless statistic?

Cronbach’s α is a function of number of items and inter-item correlation.

					Item			
		1	2	3	4	5	6	7
	1	1.00	0.36	0.36	0.36	0.36	0.36	0.36
	2	0.36	1.00	0.36	0.36	0.36	0.36	0.36
	3	0.36	0.36	1.00	0.36	0.36	0.36	0.36
Item	4	0.36	0.36	0.36	1.00	0.36	0.36	0.36
	5	0.36	0.36	0.36	0.36	1.00	0.36	0.36
	6	0.36	0.36	0.36	0.36	0.36	1.00	0.36
	7	0.36	0.36	0.36	0.36	0.36	0.36	1.00

Cronbach’s α = .8

The world’s most useless statistic?

					Item			
		1	2	3	4	5	6	7
	1	1.00	0.50	0.50	0.50	0.25	0.25	0.25
	2	0.50	1.00	0.50	0.50	0.25	0.25	0.25
	3	0.50	0.50	1.00	0.50	0.25	0.25	0.25
Item	4	0.50	0.50	0.50	1.00	0.25	0.25	0.25
	5	0.25	0.25	0.25	0.25	1.00	0.50	0.50
	6	0.25	0.25	0.25	0.25	0.50	1.00	0.50
	7	0.25	0.25	0.25	0.25	0.50	0.50	1.00

Cronbach’s α = .8

The world’s most useless statistic?

		Item						
		1	2	3	4	5	6	7
	1	1.00	0.85	0.85	0.85	0.00	0.00	0.00
	2	0.85	1.00	0.85	0.85	0.00	0.00	0.00
	3	0.85	0.85	1.00	0.85	0.00	0.00	0.00
Item	4	0.85	0.85	0.85	1.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	1.00	0.85	0.85
	6	0.00	0.00	0.00	0.00	0.85	1.00	0.85
	7	0.00	0.00	0.00	0.00	0.85	0.85	1.00

Cronbach’s α = .8

The truth about α

- In its own right α is an index of internal consistency/ homogeneity
- It establishes the *lower bound* of test reliability
- Highly reliable tests may have a low α
- alpha indexes average reliability over the range measured by the test – reliability at extremes will be lower
- Does not reflect critical measurement properties – dimensionality, severity, invariance
- High αs should be expected when items have been selected using factor analysis
- Not appropriate for formative tests
- There is no magic value for good reliability
- Should not drive item selection or test construction

NATIONAL HEART FOUNDATION · APPROVED ·

CP-QOL internal consistency/reliability

CP QOL- Child	Cronbach’s α (Parents)	Cronbach’s a (Parents)
Social wellbeing and acceptance	0.91	0.87
Functioning	0.90	0.87
Participation & physical health	0.92	0.90
Emotional wellbeing	0.85	0.85
Access to services (Parent)	0.80	—
Pain and feelings about disability	0.74	0.80
Family health (Parent)	0.77	—

A little bit about validity

Does the test/scale measure what it claims to?

- Difficult to determine when there is no objective standard
- Judgment about content
  - Face validity, Content validity, Criterion validity
  - Do the items tap the content they should?
    - expert/informed judgment
- Empirical assessment
  - Construct validity, Convergent validity
  - Do measurements agree with alternatives (e.g., clinical assessments)?
  - Are (theoretically) predicted patterns of association (and absence of association) observed?
  - Do groups differ in measurements in expected ways?
  - Can artefactual effects (e.g., social desirability, transient mental states) be ruled out as influences of measurements?

Statistical analysis can support validity but substantive knowledge of the construct must underpin its assessment.

Conclusion

- Despite being unobservable and lacking physical or objective definitions, it is possible to measure psychological constructs with surprising accuracy
- Good research in measurement must combine highly developed understanding of substantive aspects of the construct to be measured coupled with appropriate analysis and modelling
- Development of a new test should allow for multiple cycles of item development, analysis and refinement (unless you get lucky!)
- All psychometric tests are works in progress

References

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum: Hillsdale, N.J. [A noted text in this field – has a succinct but comprehensive chapter on classical test theory.]

Raykov, T. and Marcoulides, G.A. (2011) *Introduction to Psychometric Theory*. New York, NY: Routledge. [Recent and pretty rigorous text.]

Revelle, W. *eBook on Psychometric Theory* http://personality-project.org/r/book [Online text oriented towards using R for psychometric research.]