2011

Instruments and measures

Janet E. Sansoni
University of Wollongong, jans@uow.edu.au

Kate Senior

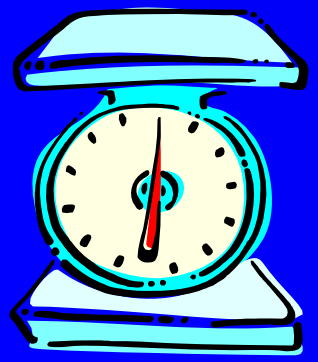
Publication Details
J. E. Sansoni & K. Senior "Instruments and measures", Menzies School of Health Research, Darwin, 8-11 March 2011, (2011)

Instruments and measures

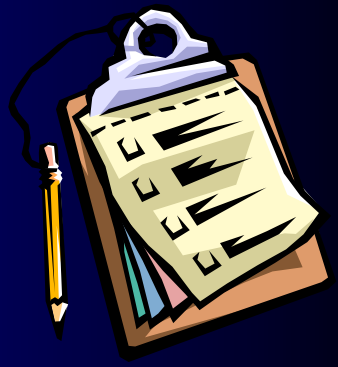
Abstract
Powerpoint presentation presented at Menzies School of Health Research, Darwin

Keywords
measures, instruments

Publication Details
J. E. Sansoni & K. Senior "Instruments and measures", Menzies School of Health Research, Darwin, 8-11 March 2011, (2011)



Session 2:
Instruments & Measures



Jan Sansoni (UOW) and Kate Senior (Menzies)

CHSD
Centre for Health Service Development

University of Wollongong

Instruments and Measures

1. Rank order the 5 most important things to you in your life at present (refer next slide)

2. SEIQOL exercise – do in pairs -1* participant, 1* administrator

SEIQOL

What are the things that you would rate as being the most important areas of your life?

1

2

3

4

5

On a scale of 1-10 (1 being worst possible, and 10 being best possible) how do you think you are doing in each of these life areas

Discussion

What are the most important dimensions?

Where did health rate? Did it rate as highly as you expected?

Is there a difference between QOL and HRQOL?

Might the importance of 'health' vary depending on age/ stage of life, gender, lifestyle and cultural factors?

What might be some issues in using SEIQOL as an outcome measure?

Where might it be useful?

CONSTRUCT OF HEALTH

Absence of disease, illness, injury?

or

‘A state of complete physical, mental and social well-being, and not merely the absence of disease or injury.’ (WHO, 1981).

The World Health Organization has recommended the development of measures of positive health…is this too broad?

Health and Well-being

Dimensions of Well-being

- health
- life satisfaction
- social well-being
- economic well-being
- environmental well-being
- spiritual or existential well-being
- other characteristics valued by humans

Dimensions of Health

- morbidity (disease or impairment)
- limitations to functional abilities (disability)
- role limitations because of health problems (handicap)
- bodily pain
- mental health (psychological distress & psychological well-being)
- vitality (energy/ fatigue)
- general perception of health

Discussion

- Do you think the dimensions in the construct of health outlined here are the same for all groups?
- Can you provide examples where the dimensions within the construct of health may differ across groups?

Difference in the meaning of health by social class - d'Houtaud and Field (1984)

Lower classes	Upper classes
Health as utility	Health as enjoyment
Health necessary to work	Health a value in itself
Fatalism	Control
Not being sick	Physical fitness, equilibrium and psychological well being
Focus on maintaining social relationships	Focus on the individual

The value of health

- Scottish women from a low socio-economic background (Blaxter and Paterson 1982).
- The importance of “coping”
- Stoicism in dealing with an inevitable part of life

“I suppose he would be really healthy because he has never been ill – had ulcer, cracked ribs, things like that, but never a cold or flu”

Perception of what is a normal state of health

- Health and disease are moulded by social context.
- Different groups of people will have different ideas about what a “normal state of health” entails.
- Influences people’s self assessed status
- ‘A person bought up in a community with a great many diseases and few medical facilities may be inclined to take certain symptoms as “normal” when they are clinically preventable’ (Sen, 2002, p. 860).

Healthism

- Crawford (1980) coined the term ‘healthism’ to mean the “preoccupation with personal health as a goal to be attained primarily through the modification of lifestyles… For the healthist, solution rests within the individual’s determination to resist culture, advertising, institutional and environmental constraints, disease agents or simply poor or lazy personal habits”

TYPOLOGY OF OUTCOME MEASURES

- **QUANTITY of LIFE**
- **PROCESS**
- **QUALITY of LIFE**
- **SATISFACTION**

- **Mortality, survival, avoidable premature mortality**
- **Practice Variations, ORPIs- readmission, complications etc**
- **Generic and specific measures: health status, HRQOL, QOL**
- **Client surveys, focus groups**

Health Related Quality of Life

Physical	Impairment	Disease/ Symptom	Single
Mental	Disability	Condition	Multiple Measures
Social	Handicap/ Capacitation	Generic Measures	Profiles/ Indexes

Types of Measures

Disease/ Symptom Specific: These are usually checklists of symptoms of a particular disease e.g. cancer. These may include symptom severity and impact items. Sometimes there will also be a single symptom measure such as sexual or cognitive functioning included in a battery (refer Rotterdam Checklist and Wexner Scale).

Condition Specific: Instead of a measure of depression you may have a broader measure that assesses mental health in general e.g. Beck Depression Inventory vs. HoNOS.

Functional Status Measures: ADL/ IADL

Blends: Where a quality of life or HRQoL measure is combined with a disease specific or condition specific measure (e.g Asthma QOL, FIQL). Some issues with these measures.

Types of Measures

Generic HRQoL/ Health Status Measures: SF-36, NHP

Generic QoL/ Well-Being Measures: COMQOL, WHOQOL

Generic Functional Status: FIM, Barthel

Health Utility Indexes: For economic evaluation, particularly cost utility analysis - AQOL, EQ5D, HUI –recent review by Hawthorne 2005.

Patient Satisfaction Measures: CQ8, CQ18, Picker Commonwealth, GUTTS

Outcome Measurement Suites: Stanford Q for CDM, COMS, DOMS

QOL/ HRQOL

- These terms are often used interchangeably but refer to quite different types of instruments
- Examine SF-36 V2
- Is this measuring quality of life, or is it measuring health related quality of life/health status?

Some Example Questionnaires

- Please fill in SF-36 V1 and SF-36 V2
- What changes have been made and why do you think that is?
- Were there any questions you found puzzling or difficult to answer?

Criteria for Selecting Measures

- **Reliability:** consistency of measurement; internal consistency and test-retest reliability

- **Validity:** does the instrument measure what it claims to measure? (content, construct, criterion…)

  Discriminant Validity: does the health status measure differentiate between the healthy public and the terminally ill

- **Responsiveness:** can the instrument detect change over time - if it is not sensitive to changes in a person's condition over time it is not much use as an outcomes measure

Criteria for Selection

- **Normative Data/ Clinical Data:** is information available for comparison purposes/ benchmarks?
- **Type of Instrument:** well-being measure, generic health status measure, health utility index, disease specific measure, symptom index, condition specific measure
- **Style of Instrument:** self-report inventory, clinical rating scale, goal attainment scale - issue of proxy reports

Criteria for Selection

- **Practical Utility:** respondent burden, costs, training
- **Freedom from Confounding Factors:** social desirability, inappropriate questions, literacy levels
- **Relevance and Suitability of Application:** does the instrument cover the dimensions of interest
- **Mode of Administration:** client fills in survey, structured interview, computer assisted telephone interview (norms can vary by method)
- **Culture, Gender, Age Appropriateness:** note there are a number of instruments specifically designed for children and adolescents. Some instruments need language modifications for Australia.

Some Statistical Issues

- When you are undertaking systematic reviews of an instrument or the literature relating to a particular research question you need to consider some statistical issues that are frequently reported in the literature
- What is the mean and what is the standard deviation and why do we use these measures of central tendency and dispersion
- What do the significance levels of $p<.05$ and $p<.01$ mean?
- Epidemiologists often use confidence intervals around the mean or confidence intervals relating to a proportion e.g. 68/363 people with asbestos exposure got the disease/lung cancer –why might we use these CIs?

Significance and Probability

- With our statistical test of the association between exposure (e.g. the intervention) and outcome provides we can estimate how likely this result is due to chance
- P<.05 –only 5 chances in 100 the result is due to chance alone
- P<.01 means?
- These are the commonly used significance levels for hypothesis testing
- A ‘convention’ and can be influenced by sample size – refer P 156 Webb et al. 2005

Type I and Type II error

- Random error means that whenever a hypothesis is tested, there is a finite possibility of either:
  - Rejecting the null hypothesis when it is true (Type I error)
  - Accepting the null hypothesis when it is false (Type II error)

Ref: Centre for PH 2002

Study result

Truth	Effect	No effect
Effect	✓	Type II error
No effect	Type I error	✓

Type 2 Errors

- In an analysis we may have found no effect when in fact there is one
- Was the study too small to detect the effect?
- Need to ensure the study is of sufficient size to detect the effect – that is has enough power to detect the effect
- Formulae for power calculations are available to estimate the minimum study size required to test a hypothesis with acceptable probabilities of type I and type II errors

Issues: Statistical and Clinical Significance

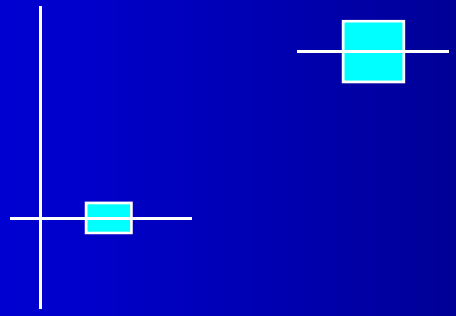
- A finding can be statistically significant but not clinically meaningful – we need both
- Using the drug finasteride (Hazard and Ward, 2000) found a stat. sig. improvement in symptom score from 2.5 to 2.8. However, for the patient to experience a subjective change in their quality of life it required a change of 3 points (Webb et al. 2005)
- This is often referred to as the minimum practically important difference or the minimum clinically importance difference and there are various ways of calculating this. Only a few papers will mention this but it is an important issue for instrument evaluation
- This also relates to the responsiveness of scales (capacity to detect change arising from the effects of an intervention)
- Usually a large change score with a smaller SD and a narrow confidence interval is more likely to be clinically meaningful as is a larger Odds Ratio or Relative Risk <2

Confidence Intervals

- Confidence intervals provide a useful measure of the amount of sampling error in a study. Due to sampling error our effect estimate may not be exactly right

- 95% CIs are often used –means if we repeat the study with different samples the 95% of the CIs would contain the true value

- Narrow confidence intervals (indicating good precision) are more informative than wide confidence intervals (indicating poor precision)



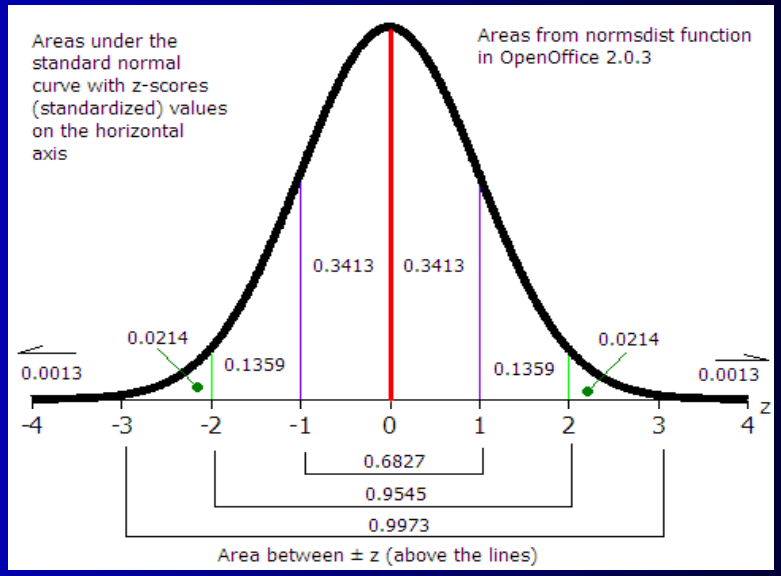
Confidence Intervals: Means

- In Psychology we use the Standard Deviation=SD = S^2 as an indication of variation or dispersion around the sample Mean= $\bar{X}$.

$$SD = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

- It is an estimate of how the sample scores vary from the sample mean. Show diagrams on board

- We also know that 95% of the population fall between 1.96 SD units on a normal distribution = z of -1.96 to + 1.96



Confidence Intervals: Means

- For CIs we take the Mean and SD of the sample and calculate the SE = Standard Error of the Mean
- We have a sample of 100 weight observations =N, the Mean is 68 kgs and the SD=10 kgs

- SE =SD / Square Root of N =10/10 = 1

$$95\%CI = (68 - 1.96\frac{10}{\sqrt{100}},\quad 68 + 1.96\frac{10}{\sqrt{100}})$$

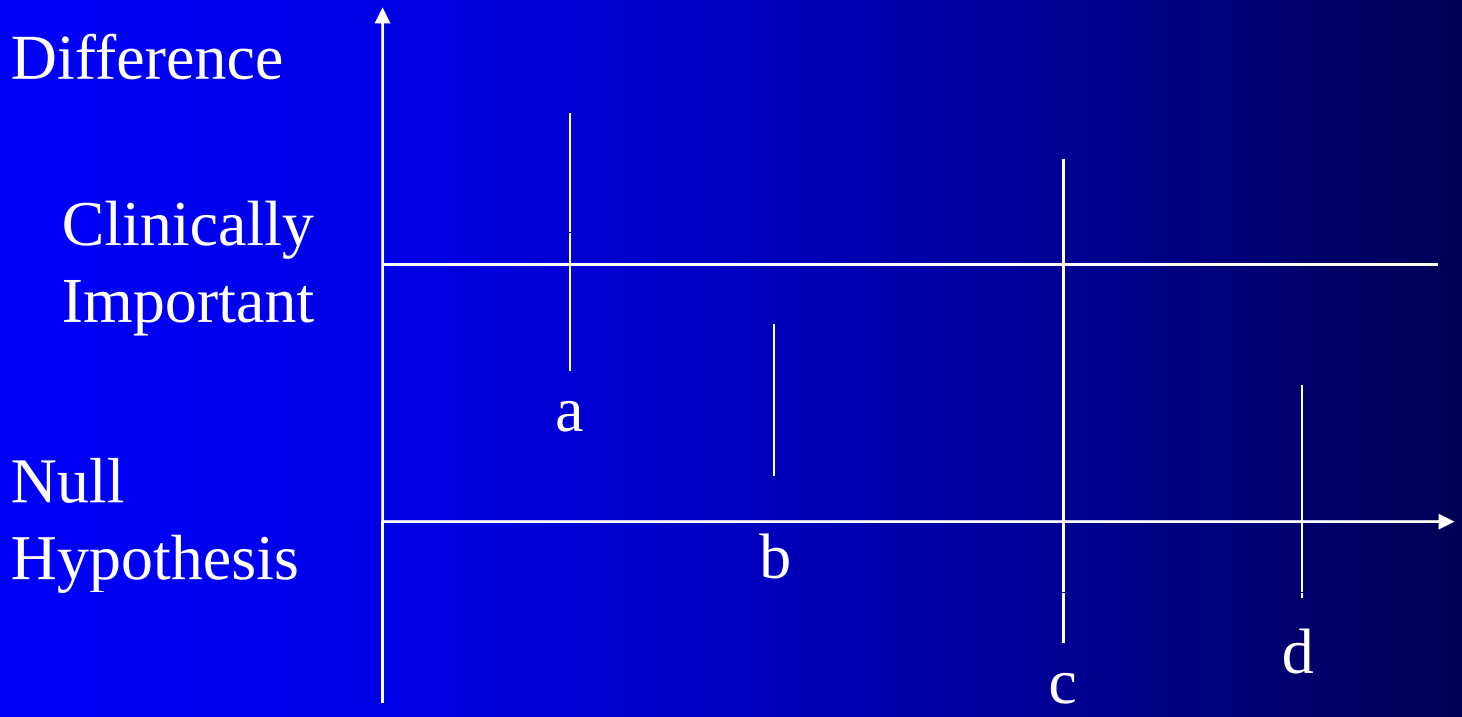
$$95\%CI = (\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

  - CI = Approximately 66 to 70 kgs
- The confidence interval of the mean gives the range of plausible values for the true population mean
- Similar methods are used to calculate CIs for proportions –use proportion vs. mean

www.chnri.org/.../WHO%20FETP%20India%20Presentations/CI%20for%20**mean%**20and%20**proportions**.ppt

Confidence Intervals

Difference for ‘a’ and ‘b’ are statistically significant, but only ‘a’ is clinically important. ‘c’ and ‘d’ not significant and …..



Reliability

- An instrument or measure is judged reliable when it consistently produces the same results. It refers to the consistency or stability of the measurement process across time, patients, or observers. The items in the scale should be tapping different aspects of the same attribute (e.g. be homogenous) and not different traits.

- An observed test score is made up of the true score plus measurement error. Measurement errors are random – a person's test score might not reflect their true score because they were sick, hungover, in noisy room etc

- Reliability estimates how much of the variability in test scores is due to measurement error and how much is due to variability in true scores

Reliability

- **Test-retest reliability:** For example, when a test is applied to same people at different times (usually @ 2 weeks apart) it produces the same results.
- **Inter-rater reliability:** The consistency between 2 independent raters observing the same set of participants.
- **Split half reliability:** Items are randomly divided into 2 subscales which are then correlated with each other. If the scale is internally consistent then the 2 halves should correlate highly. But there are many ways of splitting scales in half. Cronbach's alpha is a statistical way of deriving the average of all possible split half reliabilities for a scale (see below)

Reliability

- **Internal consistency:** **Assesses the degree to which each item correlates with others in the scale and with the total scale score (excluding this item). There should be a moderate correlation between items in a scale. If the correlations between 2 items are very high then one of these is probably redundant. If an item has a low item-total correlation (below .2) it is probably measuring something else and should be excluded.**

- **Cronbach's Alpha:** **Is used to test the internal consistency of scales. Generally a coefficient of .7 or greater is considered the minimum appropriate for a scale.**

- ***Note : A scale can be reliable but this does not make it valid. It is possible to have a highly reliable test which is meaningless. However, for a scale to be considered valid it must be reliable.***

Reliability : Glossary

- **In texts when reporting on reliability authors' generally refer to the statistical measures (or forms of correlation coefficient) they used to assess reliability on internal consistency. For example Pearson's *r*, intraclass correlation, and the kappa coefficient may be used to assess inter-rater reliability.**

- **Spearman's rho and Kendall's tau are often used when the data involve 2 sets of rankings rather than actual scores**

- **Cronbach's alpha is used to assess internal consistency**

- **It is useful to check a glossary concerning these terms such as the one found in McDowell's (2006) *Measuring Health***

Validity

- **Content Validity: Comprehensiveness e.g. in patient satisfaction (PS) measures – are all the dimensions of PS included and are all the items included relevant to patient satisfaction? (Includes face validity – on the face of it does the scale measure what it intends to measure?)**
- **Criterion Validity: should correlate highly with a gold standard measure of the same theme (e.g. compare new short version with accepted longer version of instrument) or hearing difficulties Q could be compared with results of audiometric testing. We could compare depression test results with the criterion of independent depression diagnoses made by a clinician who did not see the test results**

Sensitivity of a test would here refer to the % of people diagnosed with depression who are correctly classified as depressed by the test whereas specificity refers to the % diagnosed as without depression who are correctly classified by the test as not having depression.

Sensitivity and Specificity

Test Results	Person actually has condition (+)	Person does not have condition (-)	Totals
Positive (+)	True Positive (A)	False Positive (B)	A + B
Negative (-)	False Negative (C)	True Negative (D)	C + D
Totals	A + C	B + D	A + B + C + D

Sensitivity = A/A+C; Specificity = B/B + D

Construct Validity

- **Construct Validity concerns generating hypotheses about what measure should correlate with if it's a true measure of the construct. So for example a health status measure should correlate well with other measures of health (convergent validity) but should not correlate highly with things it is not related to such as intelligence (divergent validity).**

- **Discriminant validity refers to the ability of the scale to differentiate between relevant categories of respondents so a health scale, for example, should be able to differentiate between people who are sick or well.**

Responsiveness

- **This is the capacity for an instrument to detect change over time – for example a change in health status resulting from a health intervention. In this case health status would be assessed before treatment and after treatment. An instrument is not responsive when it cannot detect a change when one has occurred. Responsiveness is an important quality for instruments used to assess health outcomes**

- **May relate to the application: e.g. generic vs. disease specific Qs for coordinated care**

- **Effect size statistics are used as estimates for this purpose**

- **Clinical significance of change scores**

Generalizability

- Here we examine reliability and validity together as they are aspects of generalizability. We may want to know whether the results of an instrument used with a particular group can be generalized to other instruments or other groups

- A test can be both reliable and valid but the results may not be generalizable to other tests measuring the same construct nor to populations other than the one sampled

Generalizability

- **Example – one measures the levels of aggression of a random sample of primary school children in NT with the Aggro Scale.**

- **Could this be generalized this to all children in NT e.g. 5-18 years, or primary school children in NT?**

Why not?

Practicability

When selecting instruments you need one that is practical for the circumstances.

- **Is it too long? Respondent burden is a problem with long surveys. Also clinics don't have the time to administer lengthy surveys routinely – consider the setting**
- **Is it too short? It might have insufficient coverage**
- **Does it cost too much?**
- **Is it easy to administer and score? What are the training requirements?**

Confounding Factors

- Socially desirable responses – do the questions encourage socially desirable responses? (e.g. questions on sex and alcohol use)
- Readability – adult scales are generally set to a reading age of 12 years
- Are the questions appropriate…items with a lot of missing data provide clues
- Ambiguity
- Response categories mutually exclusive?
- Response sets –acquiescent and extreme response modes
- Forced choice items and reverse items are sometimes used – but note donkey vote issue that can sometimes occur when you reverse response categories from the previous item

Appropriateness Discussion: Kate

- Consider some cultural, age and gender issues.

Some Item and Scale Issues

- **Silly questions? Usually items selected are those that perform the best from a larger item pool…but if there are a lot of poorly written questions in the item pool then some poor items may remain (garbage in – garbage out)**
- **Issues of cultural relativity**
- **Inadequate response categories Yes/ No vs. levels of response**
- **Double barrel items**
- **Ceiling and floor effects.**

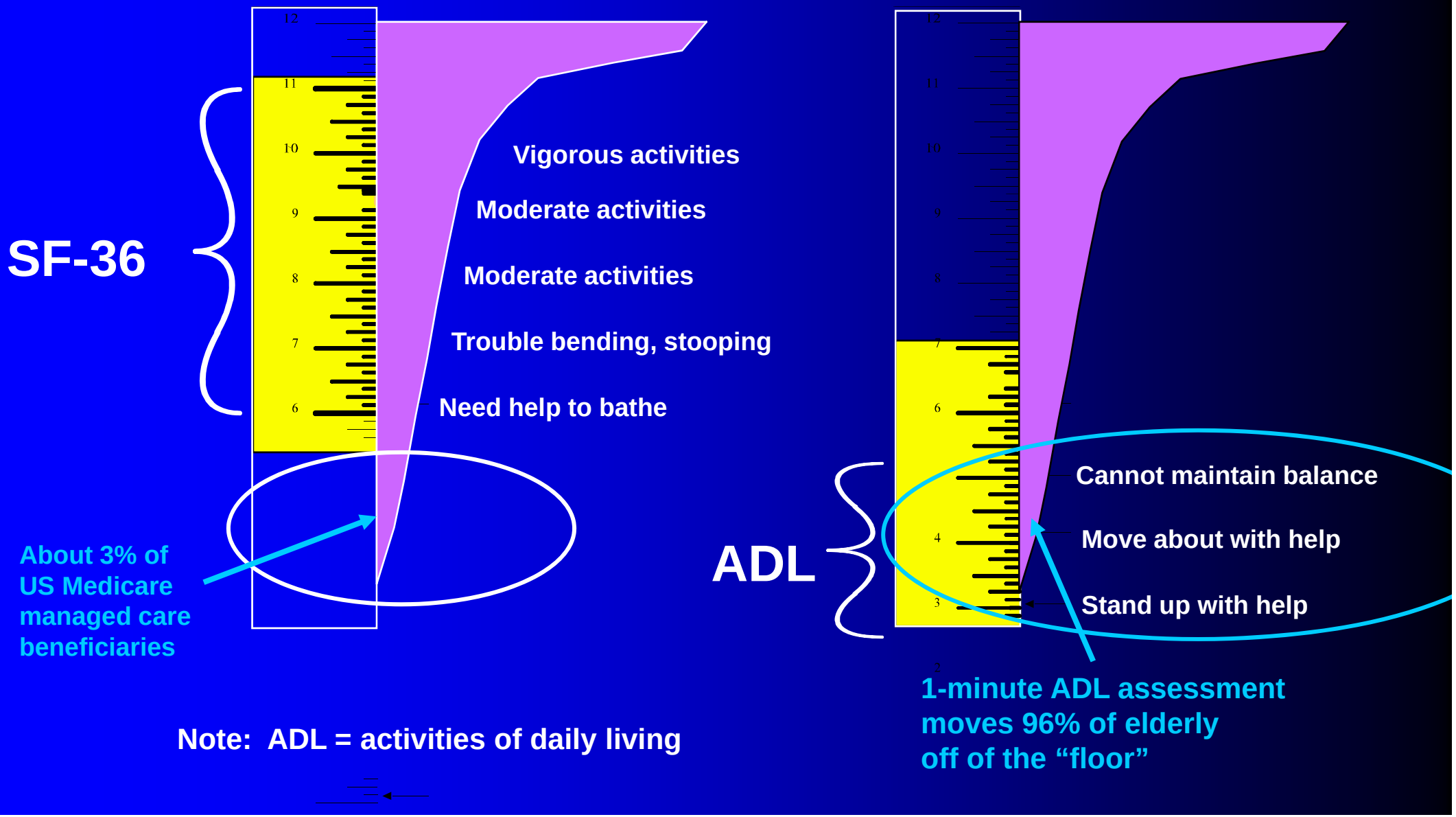
Ceiling and Floor Effects

- **Ceiling Effects – this is the % of people getting the highest possible score on an instrument/ scale e.g. if an exam is too easy and 50% of people get the maximum score your test does not differentiate adequately between these people. Many general health scales to not differentiate between people scoring at the top of the scale e.g. the healthy**

- **Floor Effects – this is the % of people getting the lowest possible score on the instrument or scale. Some general health scales do not differentiate between people who score at the bottom of floor of the scale e.g. elderly or chronically ill people**

- **Consider the distribution on the following slide**

Combining and Refining Measures (Ware)



Note: ADL = activities of daily living

Discussion: Some Instrument Issues

After examining the instruments provided consider the following:

Rating scales vs. self report – what might be some of the issues?

Is SF-36 V2 an improvement on V1 and if so why?

Response options and ceiling and floor effects.

Weighting and double counting issues.

Standardised instruments vs. DIY.

How might age, gender and cultural issues affect our design/ selection of instruments?

What instruments and items may be more prone to missing data?

When selecting instruments for an Outcomes Measurement Suite, how might we weight the criteria for selection?

Some Useful References

For instruments and measures

Bowling, A. (1995) *Measuring Disease*, Open University Press

Bowling, A (1997) *Measuring Health*, 2nd edit, Open University Press

Bowling, A. (2001).*Measuring Disease: A Review of Disease-specific Quality of Life Measurement Scales* (2nd ed.). Open University Press.

Bowling, A. (2005).*Measuring Health: A Review of Quality of Life Measurement Scales* (3rd ed.). Open University Press.

Child Health Nutrition Research Initiative (2011 download) *Confidence Interval for means and proportions.* Field Education Training Program, CHNRI, India

*www.chnri.org/.../WHO%20FETP%20India%20Presentations/CI%20for%20**mean**%20and%20**proportions**.ppt*

Dittmar, S.S. & Gresham G.E (1997) *Functional Assessment and Outcome Measures for the Rehabilitation Professional.* Aspen Publications

McDowell, I. & Newell,C. (1996) *Measuring Health*, 2nd edit, Oxford University Press

References (Cont.)

- McDowell, I. (2006) *Measuring Health*, 3rd edit. Oxford University Press
- Sansoni, J. et al (2008) *Final Report: Dementia Outcomes Measurement Suite Project*. Centre for Health Service Development, University of Wollongong, 2008.
- Streiner, D.L. & Norman, G.R. (2003) *Health Measurement Scales*, 3rd edit, Oxford University Press
- Thomas, S. et al (2006) *Continence Outcomes Measurement Suite Project*. Aust. Gov. Dept. Health & Ageing
- Webb P et al. (2005) *Essential Epidemiology: An Introduction for Students and Health Professionals*, Cambridge University Press, UK

and refer to the health outcomes reading list provided.

Materials

- **Instrument Kit, Seiqol materials.**