## Measurement in healthcare research

**Martin Bland**

Prof. of Health Statistics
Dept. of Health Sciences
University of York

http://www-users.york.ac.uk/~mb55/msc/

## Measurement in healthcare research

“All science is measurement” -- Helmholtz

“All science is measurement but not all measurement is science.” -- Kelvin

Measurement: central role both in clinical care and in healthcare research.

Clinical care:

Diagnosis often depends on measurement.

Some diagnoses depend on a measurement being above some predetermined level (e.g. hypertension).

Monitor the progress of patients (e.g. serial lung function measurements in the management of asthma, or temperature measurements in fever).

Measurement: central role both in clinical care and in healthcare research.

Healthcare research

Most studies depend upon them.

Quality of data depends upon the measurement techniques.

Poor measurement techniques can introduce so much random variation into data that the research question cannot be answered.

Measurement: central role both in clinical care and in healthcare research.

Healthcare research

Much research is concerned directly with the development and testing of methods of measurement.

Other studies concern the interpretation of measurements, such as the evaluation of diagnostic tests.

**Wide definition of measurement:**

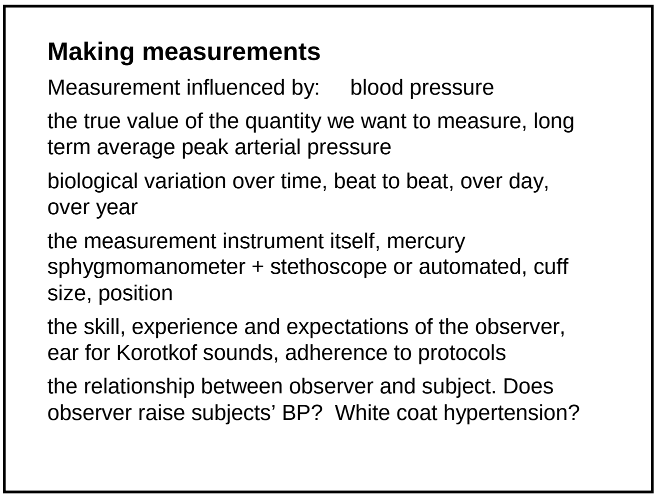
direct physical (height, weight, blood pressure),

questionnaire based scales (anxiety, depression),

subjective assessments (patient's condition as poor, fair, good or excellent),

presence or absence of a sign.

Same issues of repeatability, variation between different observers, etc., arise with all of them.



## Making measurements

Measurement influenced by:     blood pressure

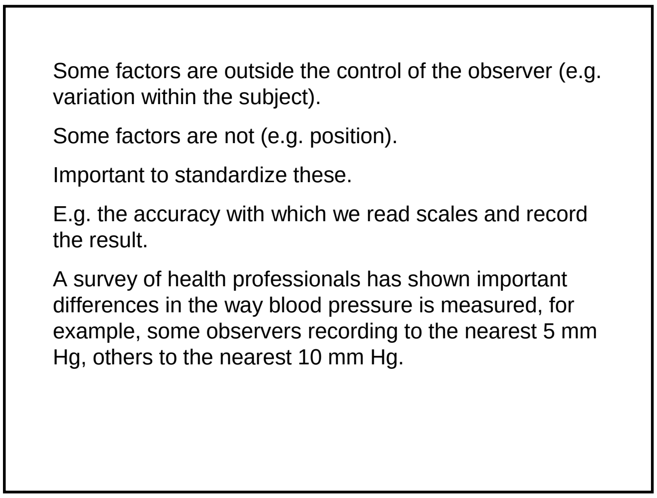
the true value of the quantity we want to measure, long term average peak arterial pressure

biological variation over time, beat to beat, over day, over year

the measurement instrument itself, mercury sphygmomanometer + stethoscope or automated, cuff size, position

the skill, experience and expectations of the observer, ear for Korotkof sounds, adherence to protocols

the relationship between observer and subject. Does observer raise subjects’ BP?  White coat hypertension?



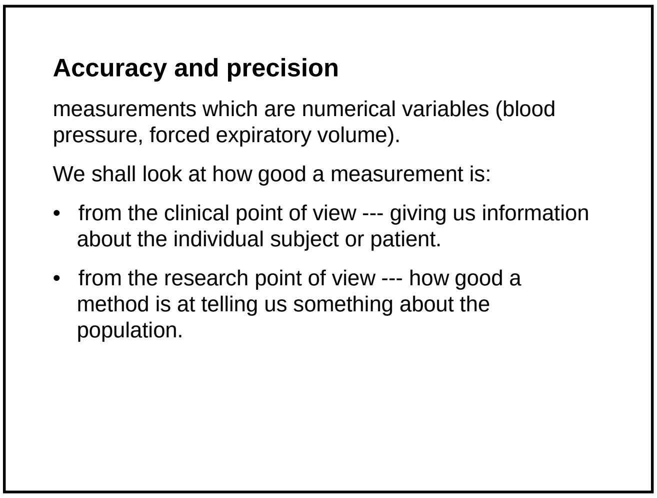
Some factors are outside the control of the observer (e.g. variation within the subject).

Some factors are not (e.g. position).

Important to standardize these.

E.g. the accuracy with which we read scales and record the result.

A survey of health professionals has shown important differences in the way blood pressure is measured, for example, some observers recording to the nearest 5 mm Hg, others to the nearest 10 mm Hg.

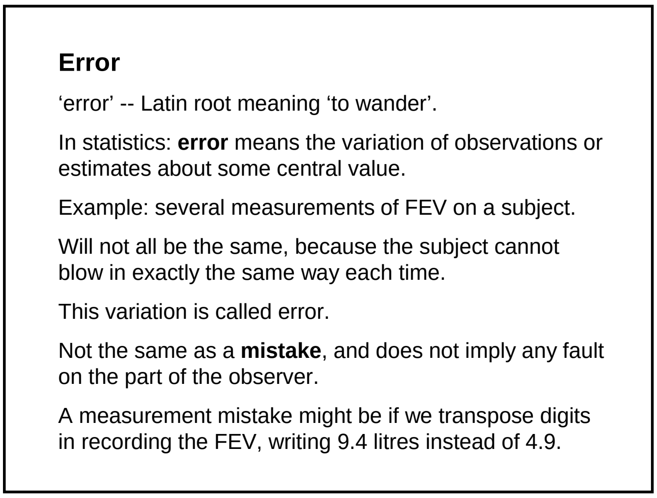


## Accuracy and precision

measurements which are numerical variables (blood pressure, forced expiratory volume).

We shall look at how good a measurement is:

- from the clinical point of view --- giving us information about the individual subject or patient.
- from the research point of view --- how good a method is at telling us something about the population.



## Error

‘error’ -- Latin root meaning ‘to wander’.

In statistics: **error** means the variation of observations or estimates about some central value.

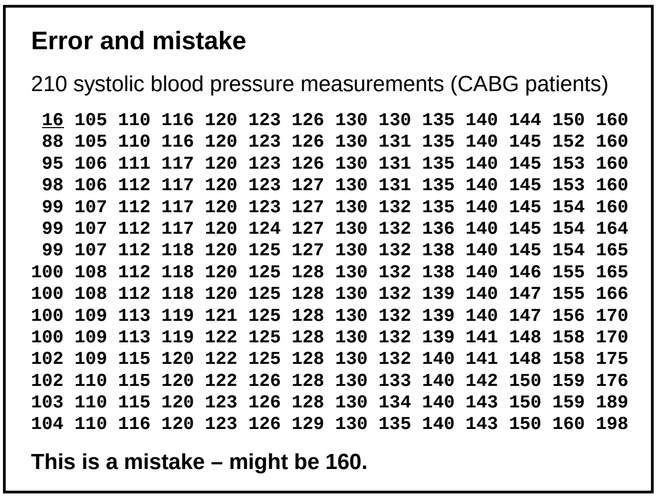
Example: several measurements of FEV on a subject.

Will not all be the same, because the subject cannot blow in exactly the same way each time.

This variation is called error.

Not the same as a **mistake**, and does not imply any fault on the part of the observer.

A measurement mistake might be if we transpose digits in recording the FEV, writing 9.4 litres instead of 4.9.

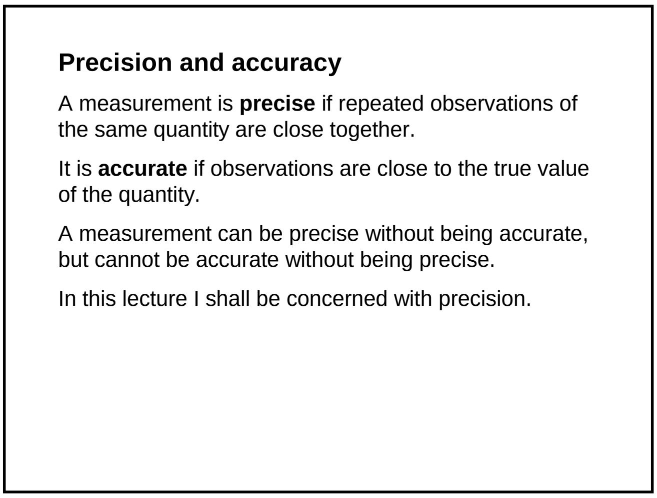


## Error and mistake

210 systolic blood pressure measurements (CABG patients)

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **16** | **105** | **110** | **116** | **120** | **123** | **126** | **130** | **130** | **135** | **140** | **144** | **150** | **160** |
| **88** | **105** | **110** | **116** | **120** | **123** | **126** | **130** | **131** | **135** | **140** | **145** | **152** | **160** |
| **95** | **106** | **111** | **117** | **120** | **123** | **126** | **130** | **131** | **135** | **140** | **145** | **153** | **160** |
| **98** | **106** | **112** | **117** | **120** | **123** | **127** | **130** | **131** | **135** | **140** | **145** | **153** | **160** |
| **99** | **107** | **112** | **117** | **120** | **123** | **127** | **130** | **132** | **135** | **140** | **145** | **154** | **160** |
| **99** | **107** | **112** | **117** | **120** | **124** | **127** | **130** | **132** | **136** | **140** | **145** | **154** | **164** |
| **99** | **107** | **112** | **118** | **120** | **125** | **127** | **130** | **132** | **138** | **140** | **145** | **154** | **165** |
| **100** | **108** | **112** | **118** | **120** | **125** | **128** | **130** | **132** | **138** | **140** | **146** | **155** | **165** |
| **100** | **108** | **112** | **118** | **120** | **125** | **128** | **130** | **132** | **139** | **140** | **147** | **155** | **166** |
| **100** | **109** | **113** | **119** | **121** | **125** | **128** | **130** | **132** | **139** | **140** | **147** | **156** | **170** |
| **100** | **109** | **113** | **119** | **122** | **125** | **128** | **130** | **132** | **139** | **141** | **148** | **158** | **170** |
| **102** | **109** | **115** | **120** | **122** | **125** | **128** | **130** | **132** | **140** | **141** | **148** | **158** | **175** |
| **102** | **110** | **115** | **120** | **122** | **126** | **128** | **130** | **133** | **140** | **142** | **150** | **159** | **176** |
| **103** | **110** | **115** | **120** | **123** | **126** | **128** | **130** | **134** | **140** | **143** | **150** | **159** | **189** |
| **104** | **110** | **116** | **120** | **123** | **126** | **129** | **130** | **135** | **140** | **143** | **150** | **160** | **198** |

**This is a mistake – might be 160.**



## Precision and accuracy

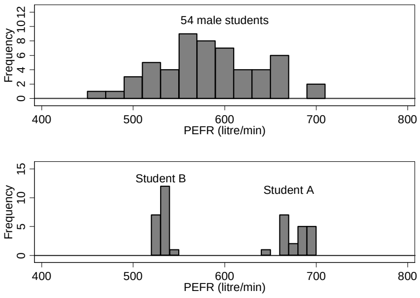
A measurement is **precise** if repeated observations of the same quantity are close together.

It is **accurate** if observations are close to the true value of the quantity.

A measurement can be precise without being accurate, but cannot be accurate without being precise.

In this lecture I shall be concerned with precision.

## Sources of variation



Two different kinds of variation:

- variation within individuals
- variation between individuals

Subject has a ‘true’ PEFR, which would be the mean of all possible measurements.

Difference between an individual measurement and the true value is its error.

Many factors could influence this error.

We would expect that a series of PEFR measurements made on a subject by different observers at different times spread over six months would vary more than a series over one morning by one observer.

We might be interested in different types of variability for different purposes.

Monitoring short term changes in blood pressure in a single patient requires one type of error, interpreting random blood pressure in a screening clinic another.

In the first case, we are detecting shifts in mean blood pressure over a short period of time.

In the second case, we are determining from one or two measurements whether the subject's mean blood pressure is above some cut-off point such as 90mm Hg diastolic.

### Repeatability and measurement error

Estimating the variation between repeated measurements for the same subject.

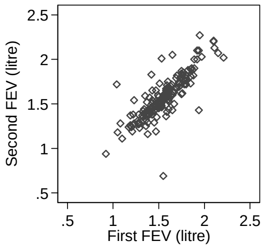
How far from the true value is a single measurement likely to be?

Simplest if we assume that the error is the same for everybody, irrespective of the value of the quantity being measured.

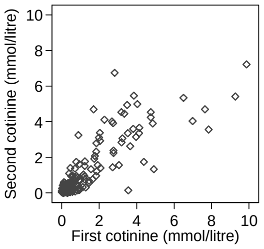
This will not always be the case, and the error may depend on the magnitude of the quantity, for example being proportional to it.

### Repeatability dependent on the magnitude of the variable

FEV and salivary cotinine in Scottish schoolchildren:



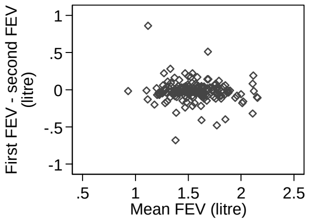
Independent



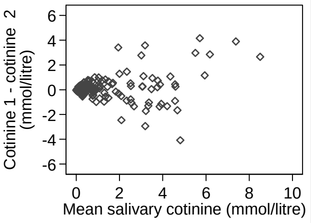
Dependent

### Repeatability dependent on the magnitude of the variable

Difference versus magnitude:



Independent



Dependent

## Repeatability and measurement error

Another term which is often used is **reliability**.

We usually specify the type of reliability, e.g.:

- **test-retest reliability**, correlation between observations by the same observer on different occasions,
- **inter-rater reliability**, the correlation between observations by different observers.

## How precisely should we record data?

Depends to some extent on the purpose for which the data are to be recorded.

Any data which are to be subjected to statistical analysis should be recorded as accurately as possible.

A study can only be as good as the data, and data are often very costly and time-consuming to collect.

The accuracy to which data are to be recorded and all other procedures to be used in measurement should be decided in advance and stated in the protocol, the written statement of how the study is to be carried out.

## Duplicate salivary cotinine measurements for a group of Scottish schoolchildren, ordered by magnitude

| 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ND | ND | 0.2 | ND | 0.3 | 0.1 | 0.4 | 0.3 | 0.9 | 0.3 | 1.0 | 0.2 | 1.9 | 2.8 | 3.9 | 5.5 |
| ND | ND | 0.2 | ND | 0.3 | 0.1 | 0.4 | 0.3 | 0.9 | 0.7 | 1.0 | 1.6 | 2.0 | 1.4 | 4.0 | 3.1 |
| ND | ND | 0.2 | 0.1 | 0.3 | 0.2 | 0.4 | 0.3 | 0.9 | 0.7 | 1.1 | 0.4 | 2.0 | 3.1 | 4.1 | 3.4 |
| ND | ND | 0.2 | 0.1 | 0.3 | 0.2 | 0.4 | 0.3 | 0.9 | 3.3 | 1.1 | 0.9 | 2.0 | 3.4 | 4.1 | 3.7 |
| ND | 0.1 | 0.2 | 0.1 | 0.3 | 0.3 | 0.4 | 0.3 | 0.6 | 0.8 | 1.1 | 1.0 | 2.1 | 2.9 | 4.1 | 5.0 |
| ND | 0.1 | 0.2 | 0.1 | 0.3 | 0.3 | 0.4 | 0.4 | 0.6 | 0.8 | 1.2 | 0.8 | 2.3 | 4.1 | 4.4 | 1.7 |
| ND | 0.1 | 0.2 | 0.1 | 0.3 | 0.3 | 0.4 | 0.4 | 0.6 | 1.0 | 1.2 | 0.9 | 2.7 | 1.4 | 4.7 | 4.5 |
| ND | 0.2 | 0.2 | 0.2 | 0.3 | 0.4 | 0.4 | 0.4 | 0.7 | 0.1 | 1.2 | 1.5 | 2.7 | 2.4 | 4.8 | 4.3 |
| ND | 0.2 | 0.2 | 0.2 | 0.3 | 0.4 | 0.4 | 1.1 | 0.7 | 0.2 | 1.2 | 1.8 | 2.7 | 4.0 | 4.9 | 1.4 |
| ND | 0.2 | 0.2 | 0.3 | 0.3 | 0.4 | 0.4 | 1.4 | 0.7 | 0.3 | 1.3 | 0.3 | 2.8 | 2.2 | 4.9 | 3.9 |
| ND | 0.2 | 0.2 | 0.3 | 0.3 | 0.4 | 0.5 | 0.1 | 0.7 | 0.3 | 1.4 | 0.7 | 2.8 | 3.9 | 6.5 | 5.4 |
| ND | 0.6 | 0.2 | 0.3 | 0.3 | 0.5 | 0.5 | 0.1 | 0.7 | 0.8 | 1.5 | 0.6 | 2.8 | 6.8 | 7.0 | 4.0 |
| 0.1 | ND | 0.2 | 0.5 | 0.3 | 0.6 | 0.5 | 0.3 | 0.7 | 0.9 | 1.6 | 0.8 | 3.1 | 1.6 | 7.6 | 4.7 |
| 0.1 | 0.1 | 0.2 | 0.6 | 0.4 | ND | 0.5 | 0.3 | 0.7 | 1.4 | 1.6 | 1.3 | 3.2 | 2.9 | 7.8 | 3.6 |
| 0.1 | 0.1 | 0.3 | ND | 0.4 | ND | 0.5 | 0.3 | 0.8 | 0.4 | 1.7 | 4.7 | 3.2 | 3.0 | 9.3 | 5.4 |
| 0.1 | 0.2 | 0.3 | ND | 0.4 | 0.1 | 0.5 | 0.4 | 0.8 | 0.5 | 1.8 | 0.9 | 3.2 | 4.5 | 9.9 | 7.2 |
| 0.1 | 0.2 | 0.3 | ND | 0.4 | 0.1 | 0.5 | 1.0 | 0.8 | 0.8 | 1.8 | 1.9 | 3.3 | 4.5 | | |
| 0.1 | 0.4 | 0.3 | ND | 0.4 | 0.1 | 0.6 | ND | 0.8 | 0.9 | 1.8 | 2.1 | 3.5 | 3.4 | | |
| 0.1 | 0.5 | 0.3 | ND | 0.4 | 0.1 | 0.6 | 0.3 | 0.8 | 1.8 | 1.8 | 2.3 | 3.5 | 4.9 | | |
| 0.2 | ND | 0.3 | ND | 0.4 | 0.2 | 0.6 | 0.5 | 0.9 | 0.2 | 1.9 | 1.2 | 3.6 | 0.2 | | |
| 0.3 | 0.1 | 0.4 | 0.2 | 0.6 | 0.6 | 0.9 | 0.2 | 1.9 | 1.5 | 3.7 | 2.6 | 3.8 | 3.6 | | |

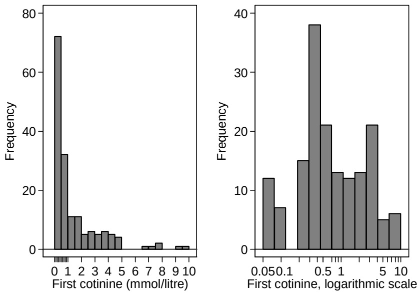
**How precisely should we record data?**

The observations 0.1 and 9.3, for example, are both recorded to one decimal place, but 0.1 has one significant figure and 9.3 has two.

The second observation is recorded more accurately.

This becomes very important when we come to analyse the data.

A bit of statistics: we may want to analyse the cotinine data on a logarithmic scale.



The greater inaccuracy of recording at the lower end of the scale is magnified by the transformation.

**How precisely should we record data?**

Accuracy of recording depends on the number of significant figures recorded, not the number of decimal places.

## Digit preference

Sometimes in measurement there is uncertainty in the last digit.

Observers will often have some values for this last digit which they record much more often than others.

Many observers are much more likely to record a terminal zero than a nine or a one, for example.

Observer training and awareness of the problem help to minimize digit preference.

If possible readings should be taken to sufficient significant figures for the last digit to be unimportant.

## Digit preference

210 systolic blood pressure measurements (CABG patients)

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 105 | 110 | 116 | 120 | 123 | 126 | 130 | 130 | 135 | 140 | 144 | 150 | 160 |
| 88 | 105 | 110 | 116 | 120 | 123 | 126 | 130 | 131 | 135 | 140 | 145 | 152 | 160 |
| 95 | 106 | 111 | 117 | 120 | 123 | 126 | 130 | 131 | 135 | 140 | 145 | 153 | 160 |
| 98 | 106 | 112 | 117 | 120 | 123 | 127 | 130 | 131 | 135 | 140 | 145 | 153 | 160 |
| 99 | 107 | 112 | 117 | 120 | 123 | 127 | 130 | 132 | 135 | 140 | 145 | 154 | 160 |
| 99 | 107 | 112 | 117 | 120 | 124 | 127 | 130 | 132 | 136 | 140 | 145 | 154 | 164 |
| 99 | 107 | 112 | 118 | 120 | 125 | 127 | 130 | 132 | 138 | 140 | 145 | 154 | 165 |
| 100 | 108 | 112 | 118 | 120 | 125 | 128 | 130 | 132 | 138 | 140 | 146 | 155 | 165 |
| 100 | 108 | 112 | 118 | 120 | 125 | 128 | 130 | 132 | 139 | 140 | 147 | 155 | 166 |
| 100 | 109 | 113 | 119 | 121 | 125 | 128 | 130 | 132 | 139 | 140 | 147 | 156 | 170 |
| 100 | 109 | 113 | 119 | 122 | 125 | 128 | 130 | 132 | 139 | 141 | 148 | 158 | 170 |
| 102 | 109 | 115 | 120 | 122 | 125 | 128 | 130 | 132 | 140 | 141 | 148 | 158 | 175 |
| 102 | 110 | 115 | 120 | 122 | 126 | 128 | 130 | 133 | 140 | 142 | 150 | 159 | 176 |
| 103 | 110 | 115 | 120 | 123 | 126 | 128 | 130 | 134 | 140 | 143 | 150 | 159 | 189 |
| 104 | 110 | 116 | 120 | 123 | 126 | 129 | 130 | 135 | 140 | 143 | 150 | 160 | 198 |

**Zeros: 62/210 = 30%, not 10%.**

## Non-numerical data.

We also find measurement error in non-numerical data.

Answers to the question: 'Have you ever smoked a cigarette?', by Derbyshire school children

| | | Interview | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Self-administered questionnaire | Yes | 61 | 2 | 63 |
| | No | 6 | 25 | 31 |
| Total | | 67 | 27 | 94 |

How closely do the children's answers agree?

## Non-numerical data.

We also find measurement error in non-numerical data.

**Answers to the question: Do you leak any urine/water when you don’t mean to? That means anything from a few drops to a flood during the day or night?’, by Leicestershire women**

| | | Self-administered questionnaire | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Interview | Yes | 21 | 3 | 24 |
| | No | 1 | 9 | 10 |
| Total | | 22 | 12 | 34 |

We cannot rely on answers to be invariably correct.

## Non-numerical data.

Two different methods:

**Anxiety for a group of osteoarthritis patients as recorded on the HADS scale and diagnosed at clinical interview by a psychiatrist**

| | | Anxiety diagnosed at clinical interview | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| HADS anxiety | Yes | 15 | 7 | 22 |
| score 8 or more | No | 2 | 30 | 32 |
| Total | | 17 | 37 | 54 |

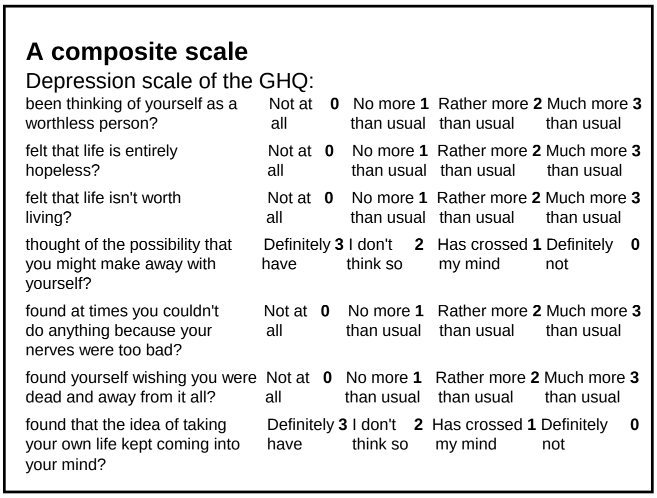
There is often uncertainty.

## Composite scales

Sometimes we combine a set of items together to make a composite scale.

The HADS scale (Hospital Anxiety and Depression Scale) is one of these.

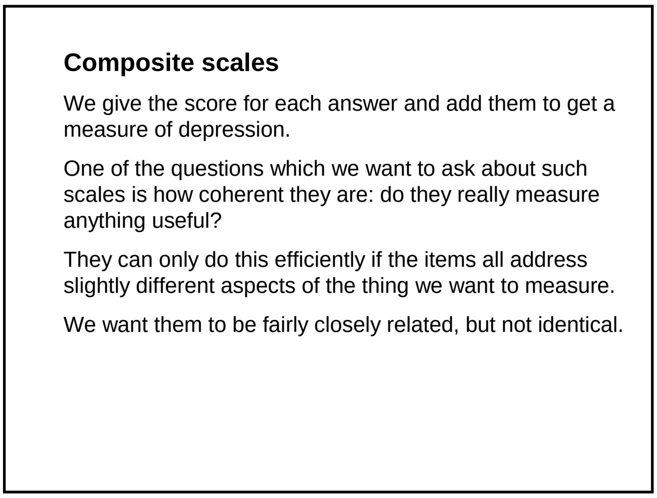
Another example, the depression scale of the GHQ (General Health Questionnaire).



## A composite scale

Depression scale of the GHQ:

| | | | | |
|---|---|---|---|---|
| been thinking of yourself as a worthless person? | Not at all **0** | No more than usual **1** | Rather more than usual **2** | Much more than usual **3** |
| felt that life is entirely hopeless? | Not at all **0** | No more than usual **1** | Rather more than usual **2** | Much more than usual **3** |
| felt that life isn't worth living? | Not at all **0** | No more than usual **1** | Rather more than usual **2** | Much more than usual **3** |
| thought of the possibility that you might make away with yourself? | Definitely have **3** | I don't think so **2** | Has crossed my mind **1** | Definitely not **0** |
| found at times you couldn't do anything because your nerves were too bad? | Not at all **0** | No more than usual **1** | Rather more than usual **2** | Much more than usual **3** |
| found yourself wishing you were dead and away from it all? | Not at all **0** | No more than usual **1** | Rather more than usual **2** | Much more than usual **3** |
| found that the idea of taking your own life kept coming into your mind? | Definitely have **3** | I don't think so **2** | Has crossed my mind **1** | Definitely not **0** |



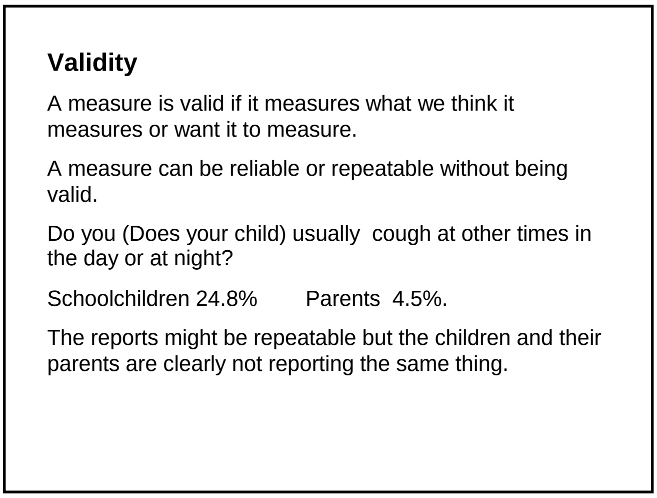
## Composite scales

We give the score for each answer and add them to get a measure of depression.

One of the questions which we want to ask about such scales is how coherent they are: do they really measure anything useful?

They can only do this efficiently if the items all address slightly different aspects of the thing we want to measure.

We want them to be fairly closely related, but not identical.



## Validity

A measure is valid if it measures what we think it measures or want it to measure.

A measure can be reliable or repeatable without being valid.

Do you (Does your child) usually cough at other times in the day or at night?

Schoolchildren 24.8% Parents 4.5%.

The reports might be repeatable but the children and their parents are clearly not reporting the same thing.

### Types of validity

❖ Criterion validity
Measurements are closely related to those given by some other, definitive technique, a 'gold standard'.

❖ Face validity
Instrument looks as though it should measure what we want to measure.

❖ Content validity
All the items appear relevant to the aim of the index, and all aspects of the thing we wish to measure are covered.

❖ Construct validity
Instrument is related to things to which we expect the concept we are trying measure to be related, and independent of those things of which the concept should be independent.