University of York Department of Health Sciences

Measurement in Health and Disease

The Validity of Measurement Methods

Martin Bland

http//:martinbland.co.uk/

Validity

No universally accepted definition of validity.

We shall regard a measurement technique as valid if it measures what we want it to measure.

Because of the great variety of measurement techniques, there is no strategy of validation which can be used in all cases.

Validity

Example: a new sphygmomanometer

We can compare readings directly with those made with an existing instrument, which we regard as a valid method.

Example: a set of respiratory symptom questions used in the study of possible effects on children of air pollution or passive smoking

We cannot compare the answers to these questions with any objective measurement of respiratory distress. We must rely on more indirect methods to assess validity, such as the relationship between answers to similar questions asked to children and their parents, or between answers and measured lung function.

Validity

Example: a scale for anxiety or depression or self efficacy. These do not exist in an objective sense, they are artificial constructs.

We ask whether they behave in the way a measurement of this construct should behave.

Validity

Many terms are used.

These do not have consistent interpretations and may overlap.

- concurrent validity,
- construct validity,
- content validity,
- convergent validity,
- criterion validity,
- discriminant validity,
- divergent validity,
- face validity,
- predictive validity.

Criterion validity

A measurement technique has criterion validity if its results are closely related to those given by some other, definitive technique, a 'gold standard'.

Most validation of physical measurements is criterion validation.

- compare new method to an existing gold standard measurement method,
- create an artificial 'subject' of known value, such as a radiological phantom.

Sensitivity and specificity, limits of agreement, usual statistical methods for comparisons of groups and relationships between continuous variables, such as t tests and regression.

Criterion validity

In the validation of non-physical measurements we cannot use agreement as a measure of criterion validity, because there is no objective reality for which we can set a criterion.

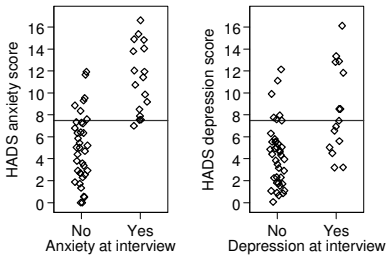
We can compare questionnaire scores:

- with clinical assessments,
- with established questionnaire scales.

Criterion validity

Example:Hospital Anxiety and Depression Questionnaire (HADS) in patients with osteoarthritis.

Patients were also given a clinical interview, which produced a psychiatric diagnosis of anxiety or depression:



We can quantify this using sensitivity, specificity, ROC curves, etc.

Criterion validity

New scales may be checked for a relationship with existing scales.

Example: Pinar (2004) studied the Turkish version of the Multidimensional Quality of Life Scale - Cancer Version 2 (MQOLS-CA2) in 72 people with cancer.

'The correlation between the global scores of the MQOLS-CA2 and Medical Outcomes Study 36-Item Short Form Health Survey was significant (r = 0.78, P = .0001), supporting the criterion validity of the MQOLS-CA2.'

Pinar R. (2004) Reliability and validity of the Turkish version of multidimensional quality of life scale - Cancer version 2 in patients with cancer. *Cancer Nursing* **27**, 252-257.

Criterion validity

New scales may be checked for a relationship with existing scales.

There are many studies which report a highish correlation with another questionnaire as an indicator of criterion validity.

However, other studies report very similar data as indicating construct validity.

These terms are not clear-cut.

Criterion validity

Do we have a gold standard? Is our criterion of validity itself really valid.

Hamman *et al.* (1975) investigated the validity of parental reports of a history of respiratory disease (asthma, pneumonia and bronchitis) in their children.

Survey of the child's General Practice records to see whether a diagnosis had been made.

Many children whose parents reported asthma did not have this in the medical record.

Did not conclude that the questionnaire instrument was wrong, but that the GP record, the criterion, was inadequate.

Hamman R.F., Halil T., and Holland W.W. (1975) Asthma in school-children. *Brit. J. Prev. Soc. Med.* **29**, 228-238.

Criterion validity

When a gold standard exists, validation is a straightforward process.

For many subjective measurement instruments there is no gold standard.

If we want to measure pain, for example, there is no objective standard. We must rely on what patients tell us.

Under these circumstances, criterion validity cannot be achieved and we must use more indirect methods.

Face validity and content validity

Derive from the psychological literature and mainly relate to questionnaire instruments.

Face validity: the instrument looks as though it should measure what we want to measure.

Example: question 'Do you usually cough first thing in the morning?' has face validity as an indicator of respiratory disease.

Example: perinatal mortality has face validity as a principal measure of the health status of national populations in developing countries, where infectious diseases are the major health problem and mortality in early life is very high, but does not do so for developed countries, where the quality of life of the elderly may be a much more important concern.

Face validity and content validity

Face validity is often used to refer to the appearance of the instrument to members of the general population.

Many physical measurements do not have face validity in this sense.

Example: dip sticks for measuring urine glucose.

Face validity and content validity

Sometimes we do not want instruments to have face validity.

We do not want the subjects to know what we are doing and so be able to conceal things from us.

Example: assessing underlying attitudes to ethnicity.

Face validity and content validity

Content validity is applied to scales made up of several items, which together form a composite index.

Two meanings:

- that the instrument covers all the required aspects of the concept being measured,
- that the instrument appears valid to an expert.

Face validity and content validity

The instrument covers all the required aspects of the concept being measured.

Example: The OECD Long-Term Disability Questionnaire

1. Is your eyesight good enough to read ordinary newspaper print? (with glasses if usually worn).
2. Can you hear what is said in normal conversation with one other person? (with hearing aid if you usually wear one).
3. Can you speak without difficulty?
4. Can you carry an object of 5 kilos for 10 metres?
5. Can you walk more than 400 meters without resting?
6. Can you walk up and down one flight of stairs without resting?
7. Can you move between rooms?
8. Can you get in and out of bed?
9. Can you dress and undress?
10. Can you cut your own food? (such as meat, fruit, etc.)

McWhinnie, J.R. (1981) Disability assessment in population surveys: results of the OECD common development effort. *Rev Epidemiol Santé Publique* **29**, 417.

Face validity and content validity

The instrument covers all the required aspects of the concept being measured.

Example: The OECD Long-Term Disability Questionnaire

The scale will have content validity if:

- all the items appear relevant to the aim of the index,
- all aspects of the thing we wish to measure are covered.

The OECD scale was intended to measure disability in terms of the limitations in activities essential to daily living: communication, mobility and self-care.

The disruption of normal social activity was seen as the central theme.

Face validity and content validity

Example: The OECD Long-Term Disability Questionnaire

The questions are all relevant to disability, so the first requirement for content validity is met.

Although the scale is intended to measure the effects of disability on behaviour, the wording of the questions concerns respondents' capacity to do things, not what they actually do.

Also, it does not contain any items concerning work and social activities.

As a scale measuring physical disability, there is reasonable content validity, but not as a scale to measure a wider definition including social disability (McDowell and Newell 1987).

McDowell, I. and Newell, C. (1987) *Measuring Health: a guide to rating scales and questionnaires* New York, Oxford University Press.

Face validity and content validity

The instrument appears valid to an expert.

There are several statistical indices which have been suggested to measure content validity.

If we can get several experts to review the instrument and rate each item in it, we can calculate the proportion who rate the item relevant. This is the criterion validity index.

There is also a criterion validity coefficient.

These methods seem to be little used and we will not pursue them.

Face validity and content validity

As a simple rule, we can think of face validity as appearing valid to the subjects, content validity as appearing valid to an expert.

The terms are not consistently used.

Example: Stallard and Rayner (2005) reported

'Face validity of the questionnaire items as assessed by a group of CBT experts (n = 16) was good.'

Stallard P, Rayner H. (2005) The development and preliminary evaluation of a Schema Questionnaire for Children (SQC). *Behavioural and Cognitive Psychotherapy* **33**, 217-224.

Face validity and content validity

Content validity of composite scales, where several variables are used to make up a single scale>

Require internal consistency.

How well do the items form a coherent scale?

Consider this separately in the lecture on formation of composite scales.

Construct validity

A measurement technique has construct validity if it 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.

The term comes from the validation of scales measuring artificial constructs without any physical reality, such as depression.

The usual statistical methods for comparison of groups and strength of relationships are used.

Construct validity

The way in which the construct validity of a given measurement technique is assessed depends on the particular circumstances.

Difficult to give general rules.

Illustrate the general principles by an example, the construct validity of respiratory symptoms questions to children (Bland 1980).

This was examined using the relationship between reports of the same symptom obtained from the child and from the parent, relationships between reports of different symptoms, and the relationship of reported symptoms to measured lung function.

Bland JM. (1980) *Epidemiological studies of respiratory symptoms in schoolchildren.* Ph.D. thesis, University of London.

Construct validity

Example: validity of child respiratory symptom questions

Reports of the same symptom obtained from the child and from the parent.

We would not expect to find a high level of association between child's and parent's answers, as the two questions do not necessarily measure the same thing.

For example, if the question is 'usually cough first thing in the morning', the child and the parent may interpret 'usually' and 'cough' differently, and it is quite possible that the parent would not see or hear the child until it had got up, that is, not first thing in the morning.

Also, the repeatability of these questions is poor.

Construct validity

Morning cough reported by children and parents, Derbyshire Smoking Study

Parent's report	Child's report: Yes		No		Not known		Total	
	n	%	n	%	n	%	n	%
Yes	29	14	104	2	0	0	133	2
No	172	83	1097	96	8	100	5277	95
Not known	6	3	132	2	0	0	138	3
Total	207	100	5333	100	8	100	5548	100

χ^2 = 119.4, d.f. = 1, P<0.001 (omitting 'not knowns')

Morning cough reported by 3.7% of children and 2.4% of parents.

The two reports were significantly associated.

However, when the child reported a morning cough, only 14% of parents confirmed this, so the agreement was not close.

Construct validity

Day or night cough reported by children and parents, Derbyshire Smoking Study

Parent's report	Child's report: Yes		No		Not known		Total	
	n	%	n	%	n	%	n	%
Yes	120	9	130	3	1	7	251	5
No	1206	88	3915	94	14	93	5135	93
Not known	48	3	114	3	0	0	162	3
Total	1374	100	4159	100	15	100	5548	100

χ^2 = 76.6, d.f. = 1, P < 0.001 (omitting 'not knowns')

The symptoms are significantly associated.

The relationship exists, but it is not a close one.

Children report a prevalence of 26% compared to 4% reported by parents, so clearly they are not reporting the same thing.

Construct validity

Example: validity of child respiratory symptom questions

Relationships between respiratory symptom questions.

Can measure the strength of the association between each pair of symptoms using a simple association coefficient, V, the product moment correlation coefficient obtained by putting 1 for yes and 0 for no.

Under the null hypothesis of no relationship, $V^2 \times n$ follows a Chi-squared distribution with 1 degree of freedom.

Construct validity

Association coefficients (V) between respiratory symptoms, Derbyshire Smoking Study

	Reported by child		
	morning cough	day or night cough	breath-lessness
Reported by child:			
morning cough	1.00	0.20	0.15
day or night cough	0.20	1.00	0.17
breathlessness	0.15	0.17	1.00
Reported by parent:			
morning cough	0.15	0.08	0.09
day or night cough	0.09	0.12	0.09
morning phlegm	0.06	0.06	0.05
day or night phlegm	0.04	0.07	0.05
breathlessness	0.10	0.09	0.18
more breathless than others	0.09	0.08	0.18

Construct validity

Association coefficients (V) between respiratory symptoms, Derbyshire Smoking Study

Every pair of symptoms showed a positive association and all are significantly associated at the 5% level.

Among symptoms reported by the child the closest relationship was between morning and day or night cough.

Breathlessness was more closely related to day or night cough than to morning cough.

Construct validity

Association coefficients (*V*) between respiratory symptoms, Derbyshire Smoking Study

When the relationship between symptoms reported by child and by parent are considered, the greatest association is between breathlessness reported by child and breathlessness reported by parent.

The closest association with morning cough reported by the child is morning cough reported by the parent. This association is actually greater than that with the child's own report of breathlessness.

Thus, for each symptom reported by the child the corresponding symptom reported by the parent is more closely associated than any other report by the parent.

Good evidence for the validity of the questionnaire method.

Construct validity

Example: validity of child respiratory symptom questions

Mean and standard deviation of PEFR (1/min) by reported respiratory symptoms (Kent Respiratory Study)

Symptom	Symptom present			Symptom absent			
	n	$\bar{x}$	*s*	*n*	$\bar{x}$	*s*	
Morning cough	56	296.6	64.0	1697	313.1	55.1	P=0.03
Day or night cough	92	294.8	57.1	1643	313.6	55.2	P=0.001
Cough for three months	43	295.7	68.8	1692	313.0	55.0	P=0.8
Morning phlegm	25	306.2	73.1	1710	312.7	55.2	P=0.5
Day or night phlegm	27	298.0	53.9	1708	312.6	55.4	P=0.2
Phlegm for three months	18	309.6	69.4	1717	312.6	55.3	P=0.8
Chest wheezy	31	285.3	82.4	1704	313.1	54.7	P=0.005

Missing values: 33

Construct validity

Example: validity of child respiratory symptom questions

Relationships between respiratory symptom questions and measured lung function.

For each symptom the mean PEFR was smaller in children reported to have the symptom than in children reported not to have the symptom than in children reported not to have the symptom.

Predictive, concurrent, convergent, divergent, and discriminant validity

Many different terms are used to describe validity.

Predictive, concurrent, convergent, divergent, and discriminant validity are referred to by different authors as aspects of criterion validity and of construct validity.

Predictive validity

The ability of the instrument to predict some other variable, usually in the future.

Example: Bader *et al.* (2005) examined the predictive validity of a simple subjective method promoted to dentists for assessing their patients' caries risk.

Bader JD, Perrin NA, Maupome G, Rindal B, Rush WA. (2005) Validation of a simple approach to caries risk assessment. *Journal of Public Health Dentistry* **65**, 76-81.

Predictive validity

The ability of the instrument to predict some other variable, usually in the future.

Example: Bader *et al.* (2005) examined the predictive validity of a simple subjective method promoted to dentists for assessing their patients' caries risk.

Data from practices that have used guideline-assisted caries risk assessment (CRA) for several years were analyzed retrospectively to determine the receipt of caries-related treatment following a CRA.

Patients categorized as being at high caries risk were approximately four times as likely to receive any caries-related treatment as those categorized as being at low caries risk and that those categorized as at moderate risk were approximately twice as likely to receive any treatment.

Concurrent validity

This refers to relationships with variables measured at the same time as the instrument under investigation.

Example: Shumway-Cook *et al.* (2005) set out to examine the concurrent validity of a new self-report measure of mobility function by comparing it with observed mobility, self-reported activity of daily living (ADL) function, and performance-based measures of gait and balance.

Shumway-Cook A, Patla A, Stewart AL, Ferrucci L, Ciol MA, Guralnik JM (2005) Assessing environmentally determined mobility disability: Self-report versus observed community mobility. *Journal of the American Geriatrics Society* **53**, 700-704.

Concurrent validity

This refers to relationships with variables measured at the same time as the instrument under investigation.

Example: Shumway-Cook *et al.* (2005)

Fifty-four adults aged 70 and older, completed the Environmental Analysis of Mobility Questionnaire (EAMQ), reporting frequency of encounter and avoidance of 24 features of the physical environment, grouped into eight dimensions, on two occasions 1 week apart.

Subjects were observed and videotaped during six trips into the community; frequency of encounters with environmental features within the eight dimensions was recorded.

Concurrent validity

This refers to relationships with variables measured at the same time as the instrument under investigation.

Example: Shumway-Cook *et al.* (2005)

EAMQ encounter and avoidance scores were compared with observed environmental encounters, with disability in ADLs and instrumental ADLs (IADLs), and lower extremity functional measures including the Short Physical Performance Battery (SPPB) and the Berg Balance Test.

Observed mobility was significantly correlated with EAMQ summary encounter ($r = 0.66$) and avoidance ($r = –0.58$) scores.

Concurrent validity

This refers to relationships with variables measured at the same time as the instrument under investigation.

Example: Shumway-Cook *et al.* (2005)

Moderate correlations were present between the EAMQ (encounter or avoidance) and observed mobility in the distance, temporal, terrain, posture, load, and density dimensions but not in the attention and ambient dimensions.

EAMQ encounter/avoidance was significantly associated with ADL and IADL ability and performance on the SPPB and Berg Balance Test.

Conclusion: self-reported frequency of encounter and avoidance of specific environmental features appears to be a valid method for determining environmentally specific mobility disability.

Convergent and divergent validity

Convergent validity asks whether the measurement is related to variables to which it should be related if the instrument were valid.

Divergent validity asks whether the measurement is unrelated to variables to which it should be unrelated if the instrument were valid.

Convergent and divergent validity

Example: Chou *et al.* (2005) studied the Chinese version of the Geriatric Suicide Ideation Scale in a sample of 154 Hong Kong Chinese older adults.

'In terms of convergent validity, the GSIS-C correlated significantly and positively with depression (assessed by CES-D), loneliness (assessed by Revised UCLA Loneliness Scale), and hopelessness (assessed by Beck's Hopelessness Scale).

'The divergent validity of the GSIS-C was demonstrated by the negative but significant, association between the GSIS-C and two variables including self-rated health status and life satisfaction (assessed by Life Satisfaction Inventory-Version A).'

Chou KL, Jun LW, Chi I. (2005) Assessing Chinese older adults' suicidal ideation: Chinese version of the Geriatric Suicide Ideation Scale. *Aging & Mental Health* **9**, 167-171.

Convergent and divergent validity

There does not seem to be much difference between convergent and divergent validity in this usage.

There is an alternative usage.

Example, Hoffman *et al.* (2004) evaluated the NCCN distress management screening measure (DMSM) in a sample of 68 cancer patients. The DMSM was administered with the Brief Symptom Inventory (BSI) and the Brief Symptom Inventory-18 (BSI-18).

Hoffman BM, Zevon MA, D'Arrigo MC, Cecchini TB. (2004) Screening for distress in cancer patients: The NCCN rapid-screening measure. *Psycho-Oncology* **13**, 792-799.

Convergent and divergent validity

There does not seem to be much difference between convergent and divergent validity in this usage.

There is an alternative usage.

'Convergent validity was established by the moderate positive correlation between the DMSNI and the BSI and BSI-18 global severity indices ($r = 0.59$, $p < 0.001$ and $r = 0.61$ $p < 0.001$, respectively).

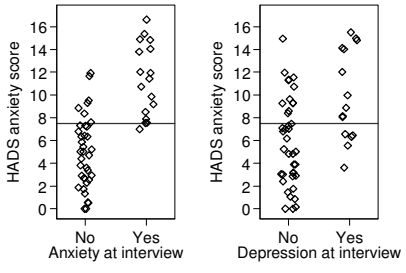
'Divergent validity was demonstrated by the lower correlations between the DMSM and the BSI subscales suggestive of psychopathology (e.g. paranoid ideation, obsessive-compulsive).'

Here divergent validity is taken as meaning a lack of relationship rather than a negative one.

Convergent and divergent validity

Example: divergent validity for HADS.

Is HADS anxiety scale more closely related to clinical interview anxiety than to clinical interview depression?



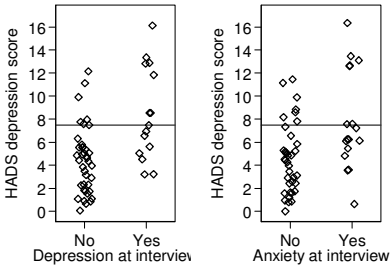
Clinical anxiety: 6.3 points, P<0.0001.

Clinical depression: 4.1 points, P=0.001.

Convergent and divergent validity

Example: divergent validity for HADS.

Is HADS depression scale more closely related to clinical interview depression than to clinical interview anxiety?



Clinical depression: 4.1 points, P=0.0001.

Clinical anxiety: 3.3 points, P=0.002.

Convergent and divergent validity

Example: divergent validity for HADS.

Would not expect independence between anxiety and depression because clinical interview depression is related to clinical interview anxiety:

Clinical anxiety	Clinical depression: No	Yes	Total
No	32	5	37
Yes	7	10	17
Total	39	15	54

Fisher's exact test, P = 0.001

Association *V* = 0.47

HADS anxiety and HADS depression: $r = 0.55$, P<0.0001

Discriminant validity

Sometimes used as interchangeable with divergent validity.

Example: Grover *et al.* (2005) 'developed and began construct validation of the Measure of Adolescent Heterosocial Competence (MAHC), a self-report instrument assessing the ability to negotiate effectively a range of challenging other-sex social interactions. . . Investigation of convergent and discriminant validity revealed that the MAHC was significantly related to measures of general social competence and anxiety in heterosexual situations and was not associated with a measure of socioeconomic status.'

Lack of association with socioeconomic status as evidence of validity is the same as divergent validity.

Grover RL, Naugle DW, Zeff KR. (2005) The measure of adolescent heterosocial competence: Development and initial validation. *Journal of Clinical Child and Adolescent Psychology* **34**, 282-291 JUN

Discriminant validity

A completely different meaning: that the measure is able to discriminate between different groups of subjects.

Example: Kleinman *et al.* (2005) reported the discriminant validity of the Gastrointestinal Symptom Rating Scale (GSRS) and Gastrointestinal Quality of Life Index (GIQLI).

'All GSRS subscales and the GIQLI total and four of the five subscale scores significantly differentiated between patients with/without GI complications ($P < 0.05$). . . . The GSRS and GIQLI differentiated between patients with/without GI side effects and by symptom severity better than did generic instruments, demonstrating excellent discriminant ability in this population.'

Kleinman L, Faull R, Walker R, Prasad GVR, Ambuehl P, Bahner U. (2005) Gastrointestinal-specific patient-reported outcome instruments differentiate between renal transplant patients with or without GI complications. *Transplantation Proceedings* **37**, 846-849.

Validity and repeatability

Repeatability is concerned with how precisely the technique measures what it measures, or how well the technique distinguishes between individuals.

Validity is concerned with how well it measures what we want it to measure.

No measurement technique can be valid if it is not repeatable.

It can be repeatable without being valid.

There may be a large bias, so that the measurements are always much higher than the true value, but they can still be same when measured again.

Validity and repeatability

Repeatability or reliability and validity are often studied together.

The appropriate methods to measure reliability are usually those using correlation or kappa statistics, as it is the properties of the measurement method with which we are concerned, rather than the interpretation of a single observation.

Validity and repeatability

Repeatability is concerned with how precisely the technique measures what it measures, or how well the technique distinguishes between individuals.

Validity is concerned with how well it measures what we want it to measure.

No measurement technique can be valid if it is not repeatable.

It can be repeatable without being valid.

There may be a large bias, so that the measurements are always much higher than the true value, but they can still be same when measured again.

Validity and repeatability

Repeatability or reliability and validity are often studied together.

The appropriate methods to measure reliability are usually those using correlation or kappa statistics, as it is the properties of the measurement method with which we are concerned, rather than the interpretation of a single observation.

Responsiveness to change

If the condition of the subject changes, does the measure change correspondingly?

Need alternative method for determining that a change has occurred.

- a different measure - objective criterion or subjective
- clinical trial, interventions produce change or not. (Has to be a real difference in the effects of the interventions.)

Responsiveness to change

Example: Hull Respiratory Questionnaire

41 subjects completed questionnaires before and after treatment.

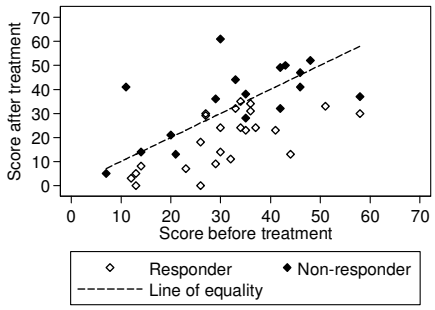
Asked whether they thought condition had improved.

Yes 24 (59%)
No 17 (41%)

Did the change in questionnaire score track the improvement?

Responsiveness to change

Did the change in questionnaire score track the improvement?



Responders have lower scores after treatment than before.

Non-responders have similar scores after treatment as before.

Responsiveness to change

Did the change in questionnaire score track the improvement?

	Mean questionnaire score	
	Pre-treatment	Post-treatment
Responders	31	19
Non-responders	33	36

Responsiveness to change

Did the change in questionnaire score track the improvement?

Regression of score post treatment on score pre-treatment and improvement:

```
      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  2,    38) =   26.18
       Model |    5579.426     2    2789.713           Prob > F      =  0.0000
    Residual |  4049.35449    38   106.56196           R-squared     =  0.5795
-------------+------------------------------           Adj R-squared =  0.5573
       Total |  9628.78049    40  240.719512           Root MSE      =  10.323

------------------------------------------------------------------------------
     total_2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     total_1 |    .669414    .130164     5.14   0.000     .4059108    .9329171
     respond |  -15.27374   3.283407    -4.65   0.000    -21.92065   -8.626825
       _cons |   13.77225   4.965197     2.77   0.009     3.720729    23.82376
------------------------------------------------------------------------------
```

Responders' mean score reduced by 15, P<0.0001, 95% CI = 9 to 22.

Responsiveness to change

Reliability of the change score.

How well can the measurement distinguish between those who change and those who do not.

Proportion of the variance of the differences from before to after which is due to real variation between people, without the measurement error.

For test-retest reliability, intra-class correlation coefficient:

$$\mathrm{ICC} = \frac{s_b^2}{s_b^2 + s_w^2}$$

where s_w^2 is the variance of repeated measurements on the same subject and s_b^2 is the variance of the true values between subjects, without error.

Responsiveness to change

Variance of a set of observed values for different subjects is the sum of these two variances: $s^2 = s_b^2 + s_w^2$.

Hence we could also write the ICC as

$$\mathrm{ICC} = \frac{s_b^2}{s_b^2 + s_w^2} = \frac{s^2 - s_w^2}{s^2}$$

Apply to the differences between the first and second measurement where a change should have occurred in at least some subjects, variance s^2.

Need a different estimate of the variance of differences on the same subject when we do not think that they are changing.

Responsiveness to change

Hull Respiratory Questionnaire:

Sample with two measurements before treatment.

Standard deviation of the differences = 8.229066.

Standard deviation of the pre- and post-treatment differences, where some people appear to change and others do not, is 13.37228.

Coefficient of reliability of change:

$$\frac{13.37228^2 - 8.229066^2}{13.37228^2} = 0.62$$

Interpret like any other ICC.

Responsiveness to change

Coefficient of reliability of change:

Interpret like any other ICC.

A relative index.

If we had bigger and more varied changes, it would increase.

It applies to the population and the size of change that we have sampled.

Responsiveness to change

Aspect of convergent validity.

We would expect a measure to change if the underlying quantity changes.

If we are measuring a quantity which can change and be changed, the measure should reflect this.

Summary

The validation process is an accumulation of evidence related to the particular measurement technique, using a variety of general and special statistical methods.

Because methods of validation have been developed in many different areas of application, terminology may be used inconsistently.

Because of the great variety of measurements which must be validated, we cannot lay down firm rules for doing it.