The validity of measurement methods

Validity

In this lecture I shall discuss some of the statistical procedures used in the validation of measurement techniques. There is no universally accepted definition of validity, but 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. Validation is very much an *ad hoc* process. For example, if we want to validate a new sphygmomanometer we can compare readings directly with those made with a random zero instrument, which we regard as a valid method. If there is good agreement between the two instruments, we can conclude that the new instrument is valid. However, if we want to validate 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.

Because of the great variety of measures for which we seek to investigate the validity, there are many terms used. Some of these do not have consistent interpretations and may overlap. They include concurrent validity, construct validity, content validity, convergent validity, criterion validity, discriminant validity, divergent validity, face validity, and 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. We can either compare our new method to an existing gold standard measurement method, or create an artificial ‘subject’ of known value, such as a radiological phantom. In this area researchers seldom need to use any other approaches to validation. The statistical methods of sensitivity and specificity for a categorical standard and limits of agreement for a continuous standard can be used, together with the usual statistical methods for comparisons of groups and relationships between continuous variables, such as t tests and regression.

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. However, we can compare questionnaire scores, for example, with clinical assessments, or new questionnaire scales with established ones. For example, in a study of the use of the Hospital Anxiety and Depression Questionnaire (HADS) in patients with osteoarthritis, patients were also given a clinical interview. This produced a psychiatric diagnosis of anxiety or depression, with the results shown in Figure 1. Patients with a positive clinical diagnosis tended to higher HADS scores. The HADS anxiety score was a better predictor than the depression score. We can quantify this using sensitivity, specificity, ROC curves, etc.

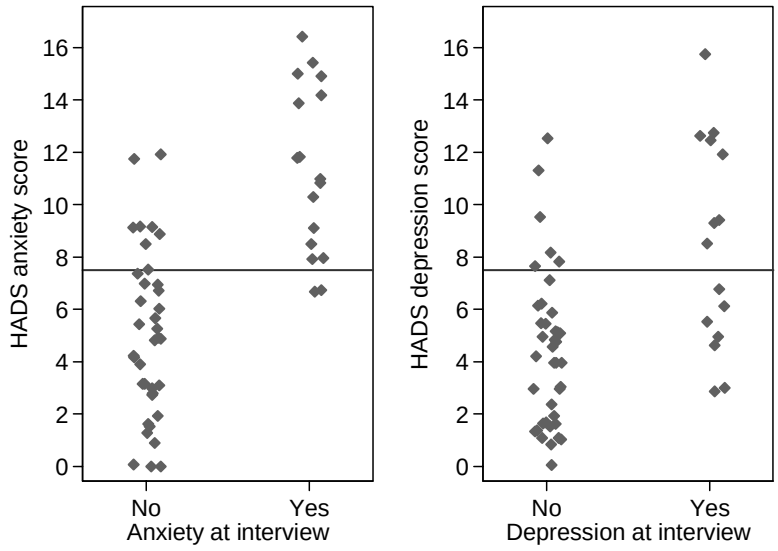

Figure 1. HADS measures of anxiety and depression and clinically diagnosed anxiety and depression, to investigate criterion validity.

New scales may be checked for a relationship with existing scales. For example, Pinar (2004) studied the Turkish version of the Multidimensional Quality of Life Scale - Cancer Version 2 (MQOLS-CA2) in 72 people with cancer. Pinar reported that '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.' 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, described below. These terms are not clear-cut.

We must be certain that we have a gold standard, that is that our criterion of validity really is itself valid. Sometimes this may not be so. Hamman *et al.* (1975) investigated the validity of parental reports of a history of respiratory disease (asthma, pneumonia and bronchitis) in their children. This included a survey of the child's General Practice records to see whether a diagnosis had been made. They found that many children whose parents reported asthma did not have this in the medical record. They did not conclude that the questionnaire instrument was wrong, but that the GP record, the criterion, was inadequate.

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

Face validity and content validity are terms which derive from the psychological literature and mainly relate to questionnaire instruments. Face validity means that the instrument looks as though it should measure what we want to measure. A question like 'Do you usually cough first thing in the morning?' has face validity as an indicator of respiratory disease, for 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.

Table 1. The OECD Long-Term Disability Questionnaire (abbreviated version)

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.)

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, for example dip sticks for measuring urine glucose. To the patient, they might as well be magic. However, this does not matter as they have criterion validity.

Sometimes we do not want instruments to have face validity, because we do not want the subjects to know what we are doing and so be able to conceal things from us. A good example (for which I thank Jeremy Miles) is assessing underlying attitudes to ethnicity.

Content validity is applied to scales made up of several items, which together form a composite index. It has two meanings. One is that the instrument appears valid to an expert, the other is that it covers all the required aspects of the concept being measured. For example, Table 1 shows such a scale, OECD long-term disability questionnaire (McWhinnie 1981, cited by McDowell and Newell 1987). Disability is measured both by answers to the individual questions and by the number of positive answers. The scale will have content validity if all the items appear relevant to the aim of the index, and if 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. The questions are all relevant to this, so the first requirement for content validity is met. However, McDowell and Newell (1987) note that 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.

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 content validity index. There is also a content validity coefficient. These methods seem to be little used and we will not pursue them here.

One other rather specialised problem which relates to the content validity of composite scales, where several variables are used to make up a single scale, is internal consistency. How well do the items form a coherent scale? We shall consider this separately in the lecture on formation of composite scales.

Table 1. 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')

Table 2. 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')

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, however. For example, Stallard and Rayner (2005) reported 'Face validity of the questionnaire items as assessed by a group of CBT experts (n = 16) was good.'

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. The way in which the construct validity of a given measurement technique is assessed depends so much on the particular circumstances that no general rules can be given. We shall 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.

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.

Morning cough in the child was reported by 3.7% of children and 2.4% of parents. Table 1 shows the relationship between morning cough reported by parents and by children. The two reports were significantly associated. If the child reported a morning cough, the parent was more likely to report the child to have the symptom than was the parent of a child who did not report the symptom. However, when the child reported a morning cough, only 14% of parents confirmed this, so the agreement was not close.

Table 3. 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

	Reported by parents					
	morning cough	day or night cough	morning phlegm	day or night phlegm	breath-lessness	more breathless than others
Reported by child:						
morning cough	0.15	0.09	0.06	0.04	0.10	0.09
day or night cough	0.08	0.12	0.06	0.07	0.09	0.08
breathlessness	0.09	0.09	0.05	0.05	0.18	0.18
Reported by parent:						
morning cough	1.00	0.46	0.44	0.29	0.26	0.27
day or night cough	0.46	1.00	0.27	0.35	0.29	0.28
morning phlegm	0.44	0.27	1.00	0.53	0.17	0.16
day or night phlegm	0.29	0.35	0.53	1.00	0.22	0.23
breathlessness	0.26	0.29	0.17	0.22	1.00	0.79
more breathless than others	0.27	0.28	0.16	0.23	0.79	1.00

Table 2 shows the relationship between cough at other times during the day or at night as reported by parents and children. Again, the symptoms are significantly associated. The relationship exists, but it is not a close one. The differences between data from these two sources is clearly shown by the prevalence of the symptom. Children report a prevalence of 26% compared to 4% reported by parents, so clearly they are not reporting the same thing.

Further evidence as to the validity of respiratory symptom questions is obtained from the relationships between them. As all the questions are dichotomous, we 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. Of course, this does not have the properties of the correlation coefficient found for Normal data, but in the dichotomous test we can interpret it. Under the null hypothesis of no relationship, V^2/n follows a Chi-squared distribution with 1 degree of freedom. (This may be verified quite easily by simple algebra.)

In the MRC Derbyshire Smoking Study, the children were asked about morning cough, cough during the day or at night, and breathlessness. Their parents were asked about the child's morning cough, day or night cough, morning phlegm, day or night phlegm, breathlessness and whether the child was more breathless than other children. When there was not a clear report of a symptom,

Table 4. 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

this has been treated as a ‘no’, that is, all questions where neither ‘yes’ nor ‘no’ was reported have been treated as a negative report, a ‘no’.

Table 3 shows the coefficients *V* between each pair of reported symptoms. 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. 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. This must be taken as strong evidence for the validity of the questionnaire method. The closest association with the child’s report of cough during the day or night is the parent’s report of cough during the day or at night. 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. The level of association between symptoms reported by the parents is higher overall than that between symptoms reported by the child. As might be expected, fairly high associations are found between breathlessness and more breathless than others (this was constrained by the questionnaire), between phlegm first thing in the morning and phlegm at other times during the day or at night, and between morning cough and day or night cough. There is also fairly high association between morning cough and morning phlegm, and between day or night cough and day or night phlegm. None of these are unexpected and, indeed, the last two would be necessary for the data to be internally consistent.

Table 4 shows the mean PEFR for children reported by parents to have each of seven respiratory symptoms. 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. To eliminate the possible effects of social class and area of residence, and to reduce the background variability, multiple covariance analysis was carried out, adjusting PEFR for sex, father’s social class, area of residence, height and weight. The differences in PEFR between children with and without reported symptoms after adjustment were slightly smaller than those shown in Table 4, but in every case the mean PEFR is less in those with the symptom than in those without.

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 refers to the ability of the instrument to predict some other variable, usually in the future. For 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. They reported that 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.

Researchers who use the term 'predictive validity' distinguish between this and **concurrent validity**. This refers to relationships with variables measured at the same time as the instrument under investigation. For 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. 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. 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. They reported that observed mobility was significantly correlated with EAMQ summary encounter ($r = 0.66$) and avoidance ($r = –0.58$) scores. 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. They concluded that self-reported frequency of encounter and avoidance of specific environmental features appears to be a valid method for determining environmentally specific mobility disability.

Convergent validity and divergent validity are terms used to distinguish between two aspects of construct 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. For 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. They report that '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).' There does not seem to be much difference between convergent and divergent validity in this usage, but there is an alternative usage. For 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). They reported that

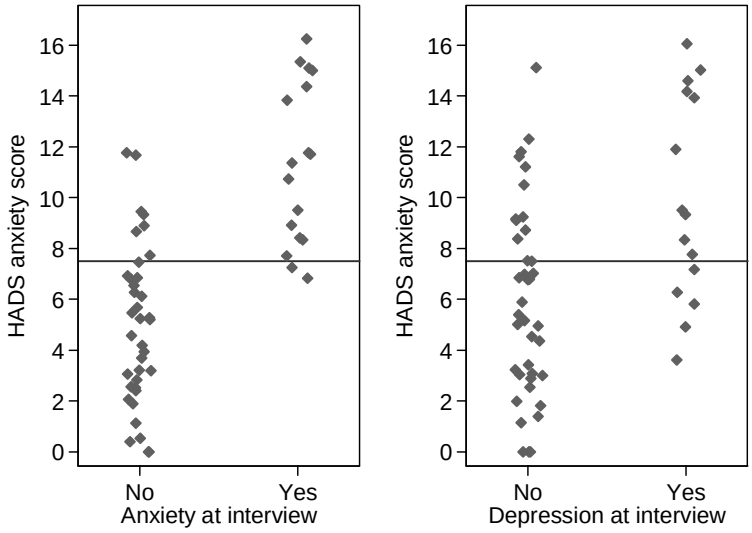

Figure 2. HADS anxiety subscale by anxiety and depression at clinical interview in a group of osteoarthritis patients.

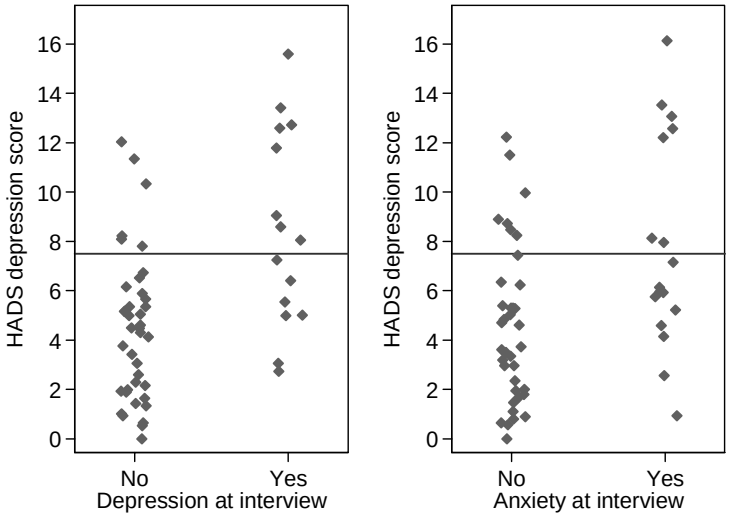

Figure 2. HADS depression subscale by anxiety and depression at clinical interview in a group of osteoarthritis patients.

Table 5. Anxiety and depression diagnosed at clinical interview in a group of patients with osteoarthritis

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

V = 0.47

‘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.

For another example, consider the HADS anxiety and depression subscales. For divergent validity, the HADS anxiety scale should be more closely related to clinical interview anxiety than to clinical interview depression. Figure 2 compares the two. The mean HADS anxiety score was greater by 6.3 points in patients with clinical anxiety ($P<0.0001$) and by 4.1 points in patients with clinical depression ($P=0.001$). There was a stronger relationship with clinical interview anxiety than clinical interview depression. Similarly, the HADS depression scale should be more closely related to clinical interview depression than to clinical interview anxiety. Figure 3 compares the two. The mean HADS depression score was greater by 4.1 points in patients with clinical depression ($P=0.0001$) and by 3.3 points in patients with clinical anxiety ($P=0.002$). There was a stronger relationship with clinical interview depression than clinical interview anxiety. We would not expect independence between anxiety and depression because clinical interview depression is related to clinical interview anxiety, as Table 5 shows. Similarly, HADS anxiety and HADS depression were positively correlated: $r = 0.55$, $P<0.0001$.

Discriminant validity is another term often regarded as interchangeable with divergent validity. For 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. Others give this term a completely different meaning: that the measure is able to discriminate between different groups of subjects. For example, Kleinman *et al.* (2005) reported the discriminant validity of the Gastrointestinal Symptom Rating Scale (GSRS) and Gastrointestinal Quality of Life Index (GIQLI). They reported that ‘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.’ These two interpretations of discriminant validity are quite incompatible.

In 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.

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. Clearly, no measurement technique can be valid if it is not repeatable. It can be repeatable without being valid, of course. 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.

As a result, 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.

References

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.

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

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.

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.

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

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.

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

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

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

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.

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.

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