A Closer Look at Diagnosis in Clinical Dental Practice: Part 1. Reliability, Validity, Specificity and Sensitivity of Diagnostic Procedures

· **Iain A. Pretty,** BDS(Hons), MSc, PhD ·
· **Gerardo Maupomé,** PhD ·

Abstract

Dentists are involved in diagnosing disease in every aspect of their clinical practice. A range of tests, systems, guides and equipment — which can be generally referred to as diagnostic procedures — are available to aid in diagnostic decision making. In this era of evidence-based dentistry, and given the increasing demand for diagnostic accuracy and properly targeted health care, it is important to assess the value of such diagnostic procedures. Doing so allows dentists to weight appropriately the information these procedures supply, to purchase new equipment if it proves more reliable than existing equipment or even to discard a commonly used procedure if it is shown to be unreliable. This article, the first in a 6-part series, defines several concepts used to express the usefulness of diagnostic procedures, including reliability and validity, and describes some of their operating characteristics (statistical measures of performance), in particular, specificity and sensitivity. Subsequent articles in the series will discuss the value of diagnostic procedures used in daily dental practice and will compare today's most innovative procedures with established methods.

MeSH Key Words: *decision support techniques; predictive value of tests; risk assessment/methods*

© J Can Dent Assoc 2004; 70(4):251–5
This article has been peer reviewed.

The need for cost-effective treatments is becoming increasingly important in resource-conscious health care systems. Yet appropriate treatment depends on accurate diagnosis, and earlier, more accurate diagnosis can lead to better patient outcomes as well as lower treatment costs. Advances in dentistry and the ability to reverse conditions once thought irreversible have also increased interest in early diagnosis based on quantifiable methods. In many cases, the mere assessment of the existing disease state is no longer appropriate, and procedures that allow some degree of longitudinal monitoring are becoming more popular. In conjunction with this move toward more timely diagnosis, the development of evidence-based clinical practice has led to an increased interest in determining the effectiveness of diagnostic procedures and re-evaluating their operating characteristics as a means of assessing performance. The results of these investigations are challenging many widely held beliefs about well-respected procedures in both medicine and dentistry.

Any dental clinician using a diagnostic procedure needs to understand how effective the procedure is, so that he or she can give appropriate weight to the result[1] in clinical decision making.[2] An objective assessment of a given diagnostic procedure would ascertain the reliability and validity of the diagnostic procedure, as well as its operating characteristics in terms of sensitivity and specificity. These features are not tests used for diagnostic or management purposes; rather, they are qualities of the diagnostic procedure itself. Lay persons and dentists alike use these terms, but often incorrectly. This article, the first in a series, defines these terms and provides some examples from dental practice to illustrate how the operating characteristics of diagnostic procedures are determined. In addition, a glossary, with concise definitions of terms, is provided (see **Appendix 1**, Glossary of epidemiology terms).

Reliability

Reliability is equivalent to repeatability or reproducibility.[1] A reliable diagnostic procedure is one that gives the same result, within accepted ranges, on repeat measurement

of the same variable. Reliability is linked to the precision of a procedure, that is, the degree of random variation that occurs during measurement of a constant value. A reliable procedure is one that is consistent, stable and dependable with minimal error. There are 2 major classes of error: systematic and random. Systematic error, or bias, leads to a one-sided deviation of the measured values from the actual values. The issue of bias as an operating characteristic of a diagnostic procedure constitutes a large field of research and is touched on only briefly in this paper. Random error, which may occur in either direction, has 3 main sources: the variation inherent among different observers; the variation related to the measurement tools, broadly referred to as their precision or accuracy; and the variation caused by changes occurring in the object being measured.

Because the largest source of variation is often ascribable to the individuals using the diagnostic procedure, 2 main aspects of reliability are usually assessed when determining the effectiveness of a procedure: intra-observer and inter-observer reliability.[3] Intra-observer reliability compares the results of a procedure performed by the same observer on several occasions with the same case materials.[4] Inter-observer reliability reflects the degree to which different observers classify individual cases in the same way.[5] For continuous data, reliability is often reported as a coefficient ranging from 0 to 1, which incorporates some measure of how scattered the individual values are (similar to a standard deviation or confidence interval). For most of these coefficients, values are generally designated as either "good" or "poor" to facilitate the interpretation of reliability, but in essence, most of the degrees between extreme values are arbitrary conventions.

Examples of situations where intra-observer and inter-observer reliability might arise in dentistry include diagnosing a dental condition on the basis of periodontal examinations, determining the need for orthodontic treatment and assessing teeth for restorative treatment.[6] Subjective procedures, such as the visual examination of dental radiographs, are often tested for reliability through repeated assessments by a number of observers.[7] Equipment reliability testing can involve in vitro work to determine the reliability of the equipment itself, followed by in vivo testing to determine if several operators arrive at the same diagnostic conclusion using the equipment, for example, apex locators.[8–11]

Dental diagnostic procedures can be divided into 3 major groups: those that provide results in terms of continuous values (e.g., orthodontic measurements or periodontal pocket measurements), those that provide dichotomous results (e.g., dental radiographic assessments as to whether or not caries extend into the dentin), and those that imply categories with discrete boundaries (e.g., the categorical data that denote different stages of cancer).

Table 1 **Decisions of 2 dentists, after examining 29 extracted teeth, to restore (Yes) or observe (No)[a]**

	Decision		
Tooth	Dentist A	Dentist B	Agreement
1	No	Yes	No
2	No	No	Yes
3	No	No	Yes
4	Yes	Yes	Yes
5	No	No	Yes
6	Yes	Yes	Yes
7	No	No	Yes
8	No	No	Yes
9	No	No	Yes
10	Yes	Yes	Yes
11	No	Yes	No
12	Yes	Yes	Yes
13	No	No	Yes
14	Yes	Yes	Yes
15	No	Yes	No
16	No	Yes	No
17	No	Yes	No
18	No	Yes	No
19	No	No	Yes
20	No	No	Yes
21	Yes	Yes	Yes
22	Yes	Yes	Yes
23	No	No	Yes
24	No	No	Yes
25	Yes	Yes	Yes
26	No	Yes	No
27	No	Yes	No
28	Yes	Yes	Yes
29	Yes	Yes	Yes

[a]*Total "yes" decisions: 10 for dentist A, 18 for dentist B; total number of cases with agreement: 21.*

For a continuous variable, reliability can be expressed as the standard deviation of the mean measurement, with a smaller standard deviation indicating greater reliability. Other methods are used for procedures with dichotomous results.[7] Agreement on dichotomous variables, either between a diagnostic procedure and the relevant "gold standard" (a standard widely accepted as the norm for a particular diagnosis) or between different procedures, is related to the similarity of the outcomes. For example, suppose 2 laboratory tests, X and Y, are used to determine whether disease Z is present in 29 patients. A total of 22 cases of agreement (75.9%) between the 2 tests is observed, with tests X and Y coinciding to indicate that disease Z is not present in 10 patients but does affect the other 12. Test X indicates that 19 patients have the disease,

Table 2 **A 2 × 2 contingency table of the data presented in Table 1, with percentages in parentheses**

		Dentist B		
		No	**Yes**	**Total**
Dentist A	**No**	11 (37.9)	8 (27.6)	19 (65.5)
	Yes	0 (0.0)	10 (34.5)	10 (34.5)
	Total	11 (37.9)	18 (62.1)	29 (100)

Table 3 **Kappa values and related estimates of strength of agreement[13]**

Kappa value	Strength of agreement
0.00–0.10	Poor
0.11–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

and test Y indicates that 12 patients have the disease. The probability of the 2 tests determining that a patient is affected is therefore 65.5% (19/29) and 41.4% (12/29), respectively.

A different perspective for a similar example would involve 2 dentists who are each given 29 extracted teeth and asked, in each case, whether they would restore the tooth. **Tables 1** and **2** illustrate the results from this example[12] of inter-observer reliability. A total of 21 cases of agreement (72.4%) are observed between the 2 dentists: they agree that 11 teeth do not require restoration and 10 do need operative intervention. Dentist A indicates that 10 teeth require filling, and dentist B identifies 18 as needing restoration. This second example does not relate purely to agreement on the result of the diagnostic procedure; rather, it concerns the application of data from the diagnostic procedure in making a clinical decision (whether or not to restore the tooth). Because many factors other than "pure" diagnostic data influence decision making, the model should not be misconstrued as a simple linear relationship. However, this example does illustrate the extent to which the 2 clinicians agree in their assessments. How can this level of agreement be further quantified?

A simple index would be the proportion of agreements between the 2 observers: 21/29 (i.e., there were 21 agreements out of 29 decisions) = 0.724, or 72.4% agreement. However, this measure ignores the agreement that would have occurred purely by chance. To correct for this chance agreement, Cohen's kappa statistic is used. While it is theoretically possible to achieve a negative value for kappa, the values normally fall between 0 (no agreement beyond chance alone) to 1 (perfect agreement). Landis and Koch[13] suggested a range of kappa values to express certain strengths of agreement, as shown in **Table 3**. These categories are purely arbitrary but are well accepted as reasonable benchmarks for determining agreement among observers.[12] In this example, where kappa = 0.49, it is possible to say that the 2 dentists had a moderate level of agreement regarding restorative decisions. This method can also be used to compare a specific observer with a gold standard. Kappa can thus be used to calculate agreement with a gold standard or to supply an estimate of the performance of an individual observer or method.

Validity

At a basic conceptual level, the validity of a diagnostic procedure is the extent to which it measures what it claims to measure, although more innovative conceptualizations of validity are much broader. In the past, the validity of a test was usually defined in terms of one or more of 3 specific types of validity: content, criterion and construct. In the context of diagnostic procedures, validity would thus be defined as a simple statistical association of test scores with some other objective measure of the criterion that the procedure was designed to quantify. It is now becoming more common to consider validation as an ongoing process, with validity being a property not only of a given procedure but also of its interpretation and the uses to which the findings are put. Inherent in this more recent view is the concept that much of the validity of a procedure resides in its consequences or effects on the individuals who undergo the procedure, as well as on programs, institutions and society.

Because the latter approach is too complicated for an introductory paper such as this one, this discussion of validity is limited to the operating characteristics of a diagnostic procedure, one of the preliminary steps in the validation strategy. Validity is thus determined in terms of the proportion of all procedure results that are correct (on the basis of comparison with the gold standard). Validity is often said to be synonymous with accuracy, but this is not necessarily the case. The accuracy of a measurement is the degree to which it is free from systematic error or bias.[1] Ideally, a diagnostic procedure should be both accurate and valid. Notably, a procedure can be accurate (i.e., no systematic error) without being valid, but it cannot be valid if it is inaccurate.

Sensitivity and Specificity

Sensitivity and specificity are 2 of the operating characteristics that indicate the accuracy of a diagnostic procedure, i.e., its ability to correctly identify those individuals with and those without the disease or condition of interest.

A typical diagnostic situation allows for 2 outcomes: either the person has or does not have the disease.[1] When the results of a procedure are compared with those of a gold

Table 4 **A 2 × 2 contingency table illustrating the outcomes of a comparison between a diagnostic or management procedure and a gold standard**

		Gold standard result		
		Positive	**Negative**	**Total**
Procedure result	**Positive**	True positive (TP)	False positive (FP)	TP + FP
	Negative	False negative (FN)	True negative (TN)	FN + TN
	Total	TP + FN	FP + TN	FN + TN + FP + TP

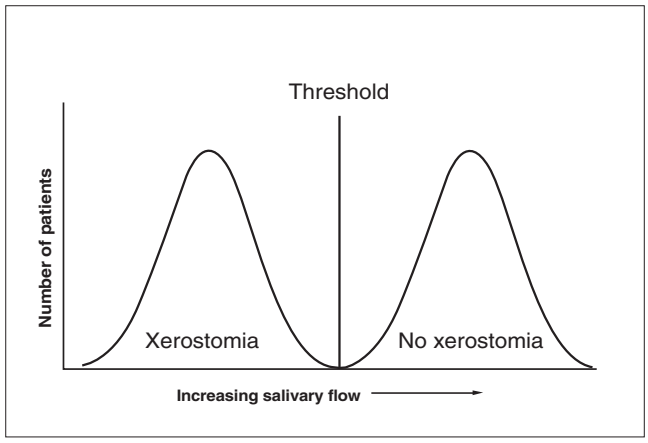

Figure 1: *Probability distributions of results for a hypothetical perfect diagnostic procedure. This procedure would correctly identify all those with and without the disease or condition, and therefore its specificity and sensitivity are both 100%. In reality, such a situation occurs only when the disease is so obvious, gross or advanced that a diagnostic procedure is not required.*

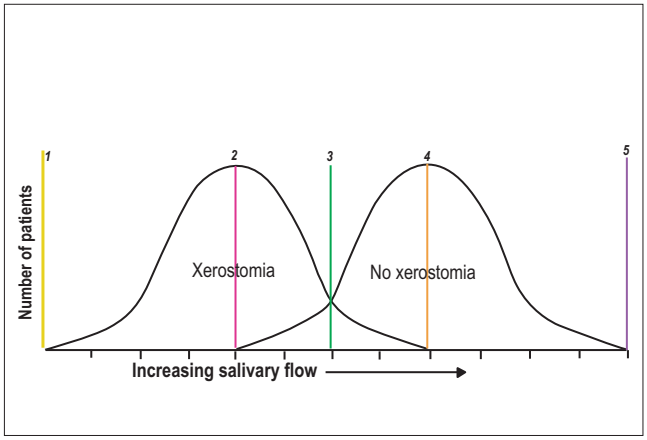

Figure 2: *A diagrammatic representation of a realistic diagnostic procedure. In this example, salivary flow rate is being used to diagnose xerostomia. Each numbered coloured line (1 through 5) represents a threshold value that could be used as a diagnostic cut-off. In clinical situations, the position of the threshold is usually dictated by the desire to limit either false positives or false negatives, depending on the implications of each of these outcomes.*

standard (either an established clinical procedure such as radiography for caries or a confirmatory test such as examination of histologic sections for caries), there are 4 possible outcomes:

- True positive (TP), whereby the procedure results indicate that the person has the disease, and this assessment is confirmed by the gold standard.
- False positive (FP), whereby the procedure results indicate that the person has the disease, but the gold standard indicates that the disease is absent.
- False negative (FN), whereby the procedure results indicate that the person does not have the disease, but the gold standard indicates that the disease is present.
- True negative (TN), whereby the procedure results indicate that the person does not have the disease, and this assessment is confirmed by the gold standard.

Table 4 illustrates these principles in a 2 × 2 contingency table.[14] Such tables are commonly used to present the results of this type of comparison.

The sensitivity of a procedure is its ability to correctly detect people who have the disease, expressed as the percentage of diseased people who are correctly diagnosed. A procedure with a sensitivity of 100% will identify *every* diseased individual; a procedure with very low sensitivity will be associated with numerous missed diagnoses. Typically, a procedure with high sensitivity yields very few false-negative results, and such procedures are used in situations where the consequences of a false-negative result are serious, for example, the screening of donated blood for HIV. Highly sensitive procedures are used for screening or ruling out disease; if the result of a highly sensitive procedure is negative, the disease can be ruled out with a high level of confidence.

The specificity of a diagnostic procedure is the percentage of disease-free individuals who are diagnosed correctly. A procedure that always yields a negative result for healthy individuals has a specificity of 100%. A procedure with high specificity produces few false-positive results, and such procedures are used in situations where the consequences of a false-positive diagnosis are serious, for example, when the diagnosis would lead to complex and painful surgery, would cause the patient to make irreversible life decisions (e.g., Alzheimer's disease) or could lead to labelling and stigmatization (e.g., schizophrenia).[1] These procedures are used for confirming the existence of a disease; if the result of a highly specific procedure is positive, the disease is almost certainly present.

An ideal test would be both highly specific and highly sensitive, but for many diagnostic procedures, these

characteristics are inversely related: an increase in one is often associated with a reduction in the other.[15] **Figure 1** represents the ideal situation, a diagnostic procedure with specificity and sensitivity of 100%. In this example, for a procedure that produces continuous variable data, stimulated saliva was collected as a means of determining xerostomia; the amount of saliva collected was then translated into a dichotomous decision as to whether the person did or did not have dry mouth, according to a threshold cut-off. The positioning of this cut-off point is crucial to the procedure's operating characteristics. In this case, there is no overlap between diseased and non-diseased subjects, and the threshold level for diagnosis is located between the 2 distributions; in other words, the subjects in each of the 2 populations are completely differentiated. If the procedure result for an individual subject is higher than the threshold, then it is considered positive; if lower, then it is considered negative.[16] Only rarely, however, is the distinction between 2 different states so unequivocal.

Figure 2 demonstrates a more realistic situation, where the patients' results overlap rather than form 2 entirely separate groups. In this example, salivary flow rate is again used to determine whether an individual is xerostomic or non-xerostomic. Clearly, the use of this measure to diagnose xerostomia requires the imposition of a cut-off or threshold point that will determine the sensitivity and specificity of the procedure. If the threshold represented by the pink line (labelled 2) is used, the procedure will be 100% specific and will correctly identify all patients without dry mouths. However, this choice of threshold will reduce the sensitivity and produce a large number of false-negative results, meaning that many patients affected by xerostomia will not be correctly diagnosed. If the threshold represented by the gold line (4) is used, the procedure will be 100% sensitive, correctly identifying all patients with xerostomia, but it will have low specificity. This choice of threshold will result in diagnosis of xerostomia in a large number of normal patients.

From these examples, it is clear that a procedure can be 100% sensitive and 100% specific only if there is no overlap between the normal and diseased populations, a rare circumstance. Moreover, when this does occur, the presence of disease is often so obvious that no diagnostic testing is required.[1]

The next article in this series will examine other operating characteristics of diagnostic procedures that can be used to help in ruling in or ruling out a specific disease. ◆

Dr. Pretty *is lecturer in prosthodontics, The University of Manchester, Manchester, UK.*

Dr. Maupomé *is investigator, Center for Health Research, Portland, Oregon; assistant adjunct professor, University of California at San Francisco, San Francisco, California; and clinical professor, University of British Columbia, Vancouver.*

Correspondence to: *Dr. Iain A. Pretty, Unit of Prosthodontics, Department of Restorative Dentistry, University Dental Hospital of Manchester, Higher Cambridge St., Manchester, M15 6FH, England. E-mail: iain.pretty@man.ac.uk.*

The authors have no declared financial interests.

References

1. Glazer AN. High-yield biostatistics. Baltimore: Williams & Wilkins; 1995.

2. White BA, Maupome G. Clinical decision-making for dental caries management. *J Dent Educ* 2001; 65(10):1121–5.

3. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.

4. Yerushalmy J. Reliability of chest radiography in the diagnosis of pulmonary lesions. *Am J Surg* 1955; 89:231–40.

5. Everitt BS. Statistical methods for medical investigators. London: Edward Arnold; 1989.

6. Eaton KA, Rimini FM, Zak E, Brookman DJ, Newman HN. The achievement and maintenance of inter-examiner consistency in the assessment of plaque and gingivitis during a multicentre study based in general dental practices. *J Clin Periodontol* 1997; 24(3):183–8.

7. ten Bosch JJ, Angmar-Mansson B. Characterization and validation of diagnostic methods. *Monogr Oral Sci* 2000; 17:174–89.

8. Pagavino G, Pace R, Baccetti T. A SEM study of in vivo accuracy of the Root ZX electronic apex locator. *J Endod* 1998; 24(6):438–41.

9. Ounsi HF, Haddad G. In vitro evaluation of the reliability of the Endex electronic apex locator. *J Endod* 1998; 24(2):120–1.

10. Kaufman AY, Fuss Z, Keila S, Waxenberg S. Reliability of different electronic apex locators to detect root perforations in vitro. *Int Endod J* 1997; 30(6):403–7.

11. Vajrabhaya L, Tepmongkol P. Accuracy of apex locator. *Endod Dent Traumatol* 1997; 13(4):180–2.

12. Dunn G, Everitt BS. Clinical biostatistics — an introduction to evidence-based medicine. London: Edward Arnold; 1995.

13. Landis JR, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33(1):159–74.

14. Brunette DM. Critical thinking: understanding and evaluating dental research. London: Quintessence Pub. Co.; 1996.

15. Smith AF. Diagnostic value of serum-creatine-kinase in coronary-care unit. *Lancet* 1967; 2(7508):178–82.

16. van Erkel AR, Pattynama PP. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *Eur J Radiol* 1998; 27(2):88–9.

Appendix 1 Glossary of epidemiology terms

Accuracy	The degree to which a measurement, or an estimate based on more than one measurement, represents the true value of the attribute being measured.
Area under the curve (AUC)	The cumulative response to an intervention, calculated by summing the area under the receiver operating characteristic (ROC) curve between each pair of consecutive observations.
Confidence interval (CI)	A statistical range of certainty with a specified probability (e.g., 95%) that a given parameter lies within the range.
Construct validity	The extent to which a given measurement corresponds to theoretical concepts (constructs) concerning the phenomenon under study.
Content validity	The extent to which a given measurement incorporates the domain of the phenomenon under study.
Continuous variable	A characteristic with an infinite number of possible values along a continuum.
Criterion validity	The extent to which a measurement correlates with some external criterion of the phenomenon under study.
Diagnostic test or procedure	A test or procedure conducted to identify a disease or condition.
Dichotomous variable	A characteristic with only 2 possible values.
False negative	A negative test result in a subject who actually possesses the attribute for which the test is conducted. Also, description of a diseased person as healthy on the basis of results of screening for the disease.
False positive	A positive test result in a subject who in fact does not possess the attribute for which the test is conducted. Also, description of a healthy person as diseased on the basis of results of screening for the disease.
Gold standard	A method, procedure or measurement that is widely accepted as being the best available for the phenomenon under study. Often used as a standard against which new methods are evaluated.
Inter-observer agreement (also known as **inter-observer reliability**)	The degree of agreement among different observers in classifying subjects or items into one of several groups.
Inter-observer variation	The degree of discrepancy among different observers in classifying subjects or items into one of several groups.
Intra-class correlation coefficient (ICC)	A statistical tool to assess consistency or conformity between 2 or more quantitative measurements.
Intra-observer	Agreement among 2 or more assessments by the same observer in classifying a subject or item into one of several groups.
Kappa coefficient	A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable, calculated as follows: $$k = \frac{P_o - P_e}{1 - P_e}$$ where P_o is the proportion of times the measurements agree and P_e is the proportion of times they can be expected to agree by chance alone. If the measurements agree more often than expected by chance, kappa is positive; if concordance is complete, kappa = 1; if concordance is the same as would be expected by chance, kappa = 0; if the measurements disagree more than expected by chance, kappa is negative.
Mean	The sum of all observations divided by the number of observations.
Negative predictive value	The probability that a person with a negative test result does not have the disease.
Positive predictive value	The probability that a person with a positive test result does have the disease.
Prevalence	The percentage of a population that is affected with a particular disease at a given time (point in time or interval of time).
Receiver operating characteristic (ROC) curve	A graphic means for assessing the ability of a screening test to discriminate between healthy and diseased individuals. The term "receiver operating characteristic" comes from psychometry, where the characteristic operating response of a receiver-individual to faint stimuli or nonstimuli has been recorded.
Reliability	The degree of stability exhibited when a measurement is repeated under identical conditions (in other words, To what degree can the results obtained by a measurement procedure be replicated?).
Resolution	Smallest change in the measured value that an instrument or test is able to detect.

Sensitivity	The proportion of diseased persons in a screened population who are identified as such by the screening test. Sensitivity is a measure of the probability of correctly diagnosing a case, or the probability that any given case will be identified by the test.
Specificity	The proportion of truly nondiseased persons in a screened population who are identified as such by the screening test. Specificity is a measure of the probability of correctly identifying a nondiseased person with the test.
Standard deviation	The average by which an observation departs from the mean.
Threshold	The point at which a physiological or psychological effect begins to be produced (e.g., the degree of stimulation of a nerve that just produces a response or the concentration of sugar in the blood at which sugar just begins to pass the barrier of the kidneys and enter the urine).
True negative	A negative test result in a subject who does not possess the attribute for which the test is conducted. Also, description of a nondiseased person as such on the basis of results of screening for the disease.
True positive	A positive test result in a subject who possesses the attribute for which the test is conducted. Also, description of a diseased person as such on the basis of results of screening for the disease.
***t*-test**	A statistical test to compare the mean values of 2 series of observations.
Validity	
Of a study	The degree to which the inferences drawn from a study, especially generalizations extending beyond the study sample, are warranted when the study methods, the representativeness of the study sample and the nature of the population from which the sample is drawn are taken into account.
Of a measurement	The degree to which a measurement measures what it purports to measure. Several varieties of measurement validity are distinguished, including construct validity, content validity, and criterion validity.
***z*-score**	The number of standard deviations by which a value lies below or above the mean; used to find the observation with a given rank from a normally distributed sample of a given size.