Sparks, R.E., L. Ganschow, J. Javorsky, and J. Pohlman (1992), 'Test comparisons amo students identified as high-risk, low-risk, and learning disabled in high-school foreign langua courses', *Modern Language Journal*, 76, 142-59

Van Patten, B. (1994), 'Evaluating the role of consciousness in SLA, terms, linguistic features, and research methodology', *AILA Review*, 11, 27-36.

Widdowson, H. G. (1989), 'Knowledge of language and ability for use', *Applied Linguistics*, 10, 2, 138-147.

Willing, K. (1987), *Learning Styles in Adult Migrant Education*, Adelaide, Adult Migrant Education Programme.

Yorio, C. (1989), 'Idiomaticity as an indicator of second language proficiency', in K.Hyltenstam, (ed.), *Bilingualism Across the Lifespan*, Cambridge, Cambridge University Press.

Five: Introduction to language testing and to C-Tests

Ulrich Raatz and Christine Klein-Braley†

1. What is a C-Test?

1.1 Construction

A C-Test is an integrative written test of general language proficiency based on the concept of reduced redundancy. A C-Test consists of five to six short authentic texts, each complete as a sense unit in itself. In these texts the first sentence is left standing. Then the " rule of two" is applied: beginning at word two in sentence two the second half of every second word is deleted. Numbers and proper names are usually left undamaged, but otherwise the deletion is entirely mechanical. The process is continued until the required number of blanks has been produced (in the canonical C-Test either twenty or twenty-five). Then the text is allowed to run on to its natural conclusion. The instructions for the examinees say something like "In this test parts of some of the words have been damaged. Please replace the missing parts." Texts are arranged in order of difficulty with the easiest text first.

1.2 Administration and scoring

In administering the test around five minutes is allowed for each text, so that a test with five parts would take twenty-five minutes to complete, one with six parts, thirty and so on. The tests are scored by giving each correct restoration one point. Spelling mistakes are counted as mistakes. In some few cases there may be alternative solutions, but this is rare. Here, a decision must be made whether to accept alternatives. This will probably depend on what the test is being used for: in a large-scale study accepting alternatives may complicate the scoring procedures considerably; in classroom use teachers will probably want to discuss the test results with their students, and here it would be appropriate to count genuine alternatives as correct. The final score is calculated by summing the scores for each individual text.

1.3 Interpretation

The result of a C-Test is a single number. Like any test score this number is only an estimate of the underlying trait (general language proficiency). Interpretation of this number can be norm-oriented (how well did the examinee perform as compared to

other examinees taking the same test?) or criterion-oriented (has a given level of proficiency been achieved?).

2. The theory behind C-Tests

2.1 Redundancy in language

C-Tests belong to the family of reduced redundancy tests, which also includes the Noise Test, various types of cloze tests, dictations, cloze elide tests. The concept of redundancy derives from Information Theory. A message which is redundant contains more information than is strictly essential for the understanding of the message. This ensures that if parts of the message are damaged or lost it can be restored from those parts which have been transmitted intact. Redundancy is a necessary feature of natural language since it is quite common for parts of a message to be distorted or missing - announcements over station loudspeakers or copies produced by defective photocopiers are obvious examples of damaged communication, as is cocktail party conversation. Redundancy is present in all levels of language from letters through words, sentences, paragraphs to texts. It is also found in the lexicon, the semantics and the pragmatics of a language. Redundancy is a feature of world knowledge: if two people share a common background - for instance in baseball or nuclear physics - the message can be conveyed much more elliptically than if expert knowledge is not shared. Cultural knowledge is another type of expert knowledge relevant for language learning.

2.2 Reduced redundancy as the basis for language tests

Tests using the concept of reduced redundancy start with the assumption that adult educated native speakers of a language can, in general, make use of the redundancy of their language to restore damaged messages through their knowledge of the rules, patterns and idiom of their own language and culture - their competence. From this it follows that a learner of the language – who by definition does not have a fully developed competence - will be less able to use redundancies to restore the message. We should stress that in this approach to testing it is not the redundancy of the text which is measured, but the examinee's ability to make use of the general redundancy of the language as a whole in order to restore the damaged text. This idea was put forward by Spolsky and his colleagues as a justification for the Noise Test (1968) and by Oller (1976) as a justification for cloze tests as foreign language tests. The Noise Test operates by adding successively larger amounts of noise (electronically produced hissing sounds) to individual sentences which have to be written down by the test subjects. Cloze tests, originally devised as readability

measures by Taylor (1953), operate by taking a written text and systematically removing every nth word, where n is a number between five and ten. Examinees are asked to restore the missing words. Scoring can be exact (only the deleted word is counted correct) or acceptable (any word which fits the text is marked right). In general, acceptable scoring has been used in foreign language testing.

Tests of reduced redundancy belong to Spolsky's (1981) psycholinguistic-sociolinguistic or post-modern tests, tests which were devised as a reaction to the non-language-like behaviour demanded by multiple-choice language tests. Psycholinguistic-sociolinguistic tests are in general integrative and based on a theory of language. They are intended to model authentic linguistic behaviour, but they are not necessarily intended to be communicative.

3. Cloze tests

3.1 Cloze tests as the major procedure in the reduced redundancy family

Cloze tests have been the most important representative of this family of foreign language tests. To construct a cloze test of the classical type a fairly long text is needed. Since it merely functions as a sample of the language the text should be authentic, but its content is not important, although it should not be biased in any way. After the initial run-in sentence every nth - very often every seventh - word is deleted. This is a simplification of a random sampling procedure for word deletion. It is recommended that a test should have at least fifty blanks in order to ensure adequate sampling. When the test is processed two sampling processes are operating simultaneously: the text (and its redundancy) is being sampled by the blanks; the examinee's competence is being sampled by restoring the missing elements. Cloze tests were repeatedly reported to have high reliabilities and to correlate highly with other language tests (validity). However, the majority of these studies were performed in the USA on highly heterogeneous groups of learners of English as a foreign language.

3.2 Criticism of cloze tests

In 1979 and 1983 Alderson and in 1981 Klein-Braley reported studies using more homogeneous samples of examinees and came to the conclusion that cloze tests, while theoretically sound, suffered from a number of major technical defects:

- In order to ensure a sufficient number of items, cloze tests have to be relatively long.
- Because of the deletion principle, cloze tests usually consist of only one longer text. This can result in text specificity and thus in test bias.

- The factors "text", "deletion rate" and "starting point" affect reliability and validity coefficients.
- If the exact method of scoring is used, then cloze tests are often too difficult for adult educated native speakers. If the acceptable method of scoring is used, a large subjective component enters the scoring, and the tests are much less reliable. Moreover scoring the tests consumes much more time.
- The difficulty of cloze tests depends on the proportion of structure and content words deleted.
- Many of the cloze tests reported in the literature are less reliable than originally assumed.

3.3 From cloze tests to C-Tests

In 1981 Raatz and Klein-Braley presented C-Tests as a technical improvement over cloze tests. They had set up a number of conditions which the new test format should fulfil:

- The new test should be much shorter, but at the same time it should have at least 100 items.
- The deletion rate and the starting point for deletions should be fixed.
- The words affected by the deletions should be a genuinely representative sample of the elements of the text.
- Examinees with special knowledge should not be favoured by specific texts, therefore the new test ought to consist of a number of different texts.
- Only exact scoring should be possible so as to ensure objectivity.
- Native speakers ought to be able to make virtually perfect scores on the test: 90% or higher. If native speakers cannot make scores higher than 90%, then the text should not be used for non-native speakers.
- The new test should be reliable, valid and easy to develop.

The first results presented in 1982 showed that all these criteria were fulfilled. Subsequent research (cf. Grotjahn, Klein-Braley, Raatz, this volume) has confirmed these early findings.

3.4 C-Tests in use

It is important to realise that there are a number of different types of language tests, and that no one test can be used for all the purposes for which tests are needed. C-Tests are proficiency tests and produce a single score, calculated from the sum of between four to six subscores, which represents the individual's global standing on the language proficiency continuum. C-Tests can be used for all learners of a language, including native speaker children. Since adult educated native speakers

are expected to make virtually perfect scores on the tests, it is obviously impossible to use the tests for adult speakers of a language unless adjustments are made, for instance by speeding the texts or by deliberately selecting extremely difficult texts (cf. Raatz, this volume). As was already stated, the score can be interpreted relative to the other scores in the group or absolutely in terms of a cut-off level as an indicator of progress towards the competence of a native speaker. This score can be used for rough and ready assessment of whole groups or individuals, for placement, and for selection or rejection. Since C-Tests are proficiency tests they cannot be used to provide information about areas of strength and weakness (diagnostic testing). There have recently been recommendations in Germany that teachers should use C-Tests as achievement tests for assessing class grades in the regular classroom tests which are given regularly at short intervals during the year. Klein-Braley and Grotjahn (1995) have strongly cautioned against such use since C-Tests do not reveal small increments in learning progress nor, since they use authentic texts, do they relate to the curriculum directly.

4. Quality criteria for C-Tests

If a test is to deliver accurate and interpretable results it has to conform to certain standards and fulfil quality control criteria. The most important quality criteria are standardisation, objectivity, reliability and validity. In this section we will discuss these criteria and examine how they can be applied to C-Tests. Empirical data relevant to this discussion are presented in the paper by Grotjahn, Klein-Braley and Raatz (this volume).

4.1 Standardisation

A test is standardised if its administration can be repeated at any time with the same conditions operating. This means that the test material, the administrative conditions and the scoring procedures must be identical for all subjects to be tested whenever they are tested. Without standardisation the test results are not comparable within or between sessions, and the other quality criteria cannot be attained.

Under normal circumstances C-Tests can be viewed as standardised. The texts are available in printed form, and the instructions for administering and scoring the test are unambiguous.

4.2 Objectivity

A test is objective if it cannot be influenced by the experimenter (test user) in its administration, scoring and interpretation. Standardisation is a precondition for

objectivity, as are exact instructions for test administration, overlays, keys, rating scales or check lists for scoring, instructions or examples for test interpretation. Naturally it is essential that the test user knows and adheres to the rules and instructions.

C-Tests can claim to be objective tests. The only problematic area is that of scoring the tests. There are two reasons for this. On the one hand use of an overlay is normally not possible since the scorer must read the test, blank for blank. However if the scorer is a native, or a very proficient non-native, and the original text is available for consultation, the probability of scoring errors is minimal. In a very few cases more than one solution is possible for a blank. During test construction it is possible to adjust the deletion system slightly in order to avoid such items, but if they do occur the scorer can be informed and both alternatives are marked correct.

4.3 Reliability

The reliability of a test is connected with the accuracy of measurement: a test which measures more accurately is more reliable because it has a smaller error of measurement. The reliability of a test is documented in a coefficient which can vary from 0 (the test consists entirely of measuring error) and 1 (the test is completely accurate and has no error of measurement). In order to evaluate this coefficient there are four methods available: the retest method, the parallel test method, the split-half method and the analysis of inner consistency. These methods are described briefly in the glossary (see appendix).

In evaluating C-Test reliability the most usual method has been consistency analysis, although in some cases retest coefficients have been determined. In the vast majority of cases C-Tests have been shown to have acceptable levels of reliability (cf. Grotjahn, Klein-Braley and Raatz, this volume).

In performing a consistency analysis on C-Tests it is not permissible to define the individual blanks in the test as items, since they are dependent on each other as a result of text structure and content. Consistency analysis assumes that all items entering into the equation are independent. For this reason, each C-Test text is classed as a super-item with a score for each subject between nought and twenty (depending on the number of blanks in the texts), and analysis is performed with the super-items. Thus there may be only four or five items in the reliability analysis. With such scaled items Cronbach's Alpha formula should be used instead of the Kuder-Richardson formulas.

4.4 Validity

A test is valid if it measures what it is supposed to measure, or if it does what it is supposed to do. The first approach is content-oriented and is based on the concept of content validity and in a broader sense on construct validity. The second approach is user-oriented, pragmatic, and based on the concepts of concurrent and predictive validity. These four approaches to validity are explained further in the glossary.

To a certain extent C-Tests can claim content validity since they consist of several randomly selected authentic texts (but see below, 6.1). The damaged words are a representative sample of all the words in the texts. Also the test-taking behaviour, namely reconstructions of missing portions, model real-life behaviour in the use of language. Naturally, the content validity of C-Tests is restricted to the areas of reading and writing. However if we assume that all language behaviour is related and thus integrative, then the validity of C-Tests can be extended to the use of the language generally (i.e. general language proficiency). In fact, studies have shown that performance on oral tests correlates highly with C-Test scores (cf. e.g. Raatz and Klein-Braley, 1983). In some cases, therefore, when the testing has to be swift and cheap, particularly, for instance, in placement testing, C-Tests may be the instrument of choice with oral testing restricted to those few candidates whose placement is in doubt.

To determine the construct validity of a test the test concept is embedded in a construct or nomological net. From this a variety of hypotheses are derived which are then upheld or rejected in validation investigations. If the results are in agreement with the assumptions of the construct, then construct validity can be said to have been demonstrated.

With language tests such hypotheses could address relationships between the test results and variables such as sex, age, social status, intelligence, school type etc, and could be investigated in a correlative study. Relations between language tests or other tests could use the approaches of factor analysis. The psycholinguistic processes underlying response behaviour could be investigated in think aloud experiments.

The construct validity of C-Tests has been investigated in a number of studies. The results are reported in the paper by Grotjahn, Klein-Braley and Raatz (this volume). In determining the concurrent and predictive validity of C-Tests the criteria have been school grades, the results of other language tests and test batteries, and self-evaluation procedures. In general the correlation coefficients which emerged have been extraordinarily high. This is not surprising when C-Tests are correlated with other language test batteries since we claim that the C-Test is an integrative test. What is surprising are the high correlations with teacher grades because in general

there is a consensus that teacher judgments are subjective and unreliable and that high correlations with objective and reliable test procedures can therefore not be expected. One explanation for this unexpected finding is that teacher judgments are also integrative measures so that the criticism should be applied rather to individual grading procedures than to the overall grades themselves.

C-Tests have extraordinarily high correlations with a variety of other procedures for measuring subjects' general proficiency in a language. Moreover, in factor analytic studies C-Tests frequently emerge as the tests with the highest loading on the general factor. Despite their extraordinarily good performance C-Tests have low face validity. Non-experts (teachers, students, parents) tend to view them as reading comprehension tests or even as a special form of intelligence tests. They find it difficult to accept C-Tests as integrative language tests. If C-Tests are to be used it may therefore be necessary to invest some time in public relations work first. We have found that if C-Tests are used together with other properly developed language tests, such as multiple-choice vocabulary tests, the acceptance level is higher. Since multiple measurement improves reliability, and since different test formats test different facets of language proficiency, it is both theoretically and practically desirable to use other tests in combination with the C-Test. However, as has already been stated, many factor analytic studies have shown that in a battery of different language tests the C-Test usually has the highest loading on the general factor. Thus, if only one test can be used because of time constraints, the C-Test often proves to be the most efficient test. The face validity of C-Tests improves as the test format becomes more familiar.

According to the theoretical assumptions C-Tests should decrease in difficulty as the learners' proficiency improves. Extensive investigations have shown that this is the case (Raatz, 1985; Klein-Braley, 1985). Therefore the texts must have some kind of 'inherent' difficulty levels based on linguo-statistic properties of the texts. Using this approach, regression equations both in English and in German have been developed which enable the prediction of the difficulty of a text for specific groups before the texts are used empirically (Klein-Braley, 1994).

In summary, we would conclude that C-Tests measure 'general language proficiency'. What is meant by this concept? In their 1982 model Bachman and Palmer (cf. Bachman, 1990) divide language competence into *operational competence* and *pragmatic competence*. These are then further subdivided into *grammatical competence* and textual competence on the one hand, and *illocutionary competence* and *sociolinguistic competence* on the other. Bachman (1990: 86) comments:

> The results of [Bachman and Palmer's] study suggest that the components of what they called grammatical and pragmatic competence are closely associated with each other, while the components they described as sociolinguistic competence are distinct.

The 'general language proficiency' tested by the C-Test and believed by us to be the bedrock of linguistic performance seems to be very similar to Bachman's operational competence - the superordinate category for lexical, morphological, syntactical, graphological knowledge on the sentence level, and for knowledge of cohesion and rhetorical organisation on the text level. We do not believe that the C-Test would be a suitable procedure for measuring pragmatic competence as a separate component and would suggest that if assessing this aspect of language proficiency is of supreme importance, then tests devised specifically for this purpose should be used. Nevertheless we would claim that adequate or superior performance in the area of sociolinguistic competence can only be achieved if the basic underlying organisational competence (= general language proficiency as assessed by the C-Test) is sufficient.

4.5 Scale attributes

A test should measure one trait unidimensionally on an interval scale. In order to achieve this aim tests are developed on the basis of a theory of testing. The best-known and most frequently used approach to test development is what is usually called classical test theory (Gulliksen, 1950). In this theoretical approach the central concepts are measurement error and reliability. This approach is extremely important when the test to be developed is a reliable norm-referenced test. However, it is not possible to determine within the framework of this theory whether the test genuinely measures on a unidimensional interval scale.

In order to do this a different theoretical approach is needed, namely item-response theories (IRT-theories). For 0-1 items extensions of the Rasch model (Rasch, 1960) can be used. The CLA-Model (Moosbrugger and Müller, 1982) is an extension of the Rasch approach for scaled items. This model has been used successfully for C-Tests.

5. How to construct a C-Test

In order to ensure that a C-Test performs satisfactorily as a test, measuring general language proficiency objectively, reliably and validly, certain rules based on classical test theory must be followed during test development. C-Tests do not come with an automatic seal of test quality, and must be investigated before they are used to make important decisions about people's lives.

The following steps are necessary:

1. Define the target population and the test format.
2. Choose suitable texts, more than necessary, using regression equation for predicting difficulty.
3. Bring the texts into C-format and combine them into one or more C-Tests.
4. Examine the tests in a group of educated adult native speakers.
5. Analyse each text: Is the mean difficulty 90% or more? Are there gaps in the text with more than one solution?
6. Decide either: text is satisfactory, OR text is satisfactory, but some words have to be changed; AND/OR in some damaged words letters have to be added; OR text is not satisfactory, i.e. too difficult.
7. Combine the good texts – possibly after correction – into one or more C-Tests in order of ascending estimated difficulty.
8. Examine the C-Test in a larger representative sample of the target population.
9. Perform item analysis, estimate test reliability and validity (if a criterion is available).
10. Perform CLA-analysis in order to investigate the dimensionality and the type of scale of the test.
11. Decide: C-Test is satisfactory, i.e. it is acceptably reliable and valid, unidimensional and interval scaled; OR C-Test is approximately satisfactory, but some texts or the order of texts have to be changed, or a text has to be excluded; OR C-Test is not satisfactory (in this case one has to start all over again).
12. Improve the test, construct the final form.
13. Perform additional studies for reliability and validity.
14. Administer the final form to a large representative sample of the target population.
15. Calculate the test norms on the basis of the distribution of the raw scores.

6. Special remarks

6.1 Selection of texts

The texts in a C-Test are a sample of all possible authentic texts of the language. Their function is therefore not to be interesting, but to be typical or representative. For this reason they should be as normal as possible. The following rules may help in choosing appropriate texts: choose written texts which are:

- complete in themselves
- appropriate in difficulty and content for target group
- with no specialised vocabulary and content,
- no literary texts, no verbal humour.

Possible sources of texts are non-fictional books, newspapers, magazines, brochures, information leaflets.

Some authentic texts, however, are too difficult for learners of a foreign language since they do not have the necessary vocabulary and grammatical rules at their disposal. In such cases it is permissible to work with quasi-authentic texts. Such texts can be specially written by native speakers of the language within the restricted repertoire of the learners. An alternative approach is to take material from equivalent text books for the same groups or to make minor alterations in authentic texts which would otherwise be suitable.

6.2 Item and test analysis

Item and test analysis is performed after the administration of the test to the target group (see step 8). Using classical test theory or an IRT-model statistical and other data such as item difficulty, item discrimination indices, test reliability and validity are calculated. These values are the basis for test improvement if necessary. The indices or statistical data shown in Table 1 are to be calculated or estimated. Only a C-Test which has been empirically trialled, investigated according to the demands of test theory, and improved if necessary can attain the criteria for a satisfactory test procedure. No student should be subjected to a test which does not conform to accepted test standards.

Table 1: Item and test analysis*

1	Raw score distribution, mean, standard deviation.	Aim: bell-shaped curve.
2	Difficulty of C-Test parts.	Aim: Difficulties between 80 and 20, mean 50, items to be arranged in ascending order of difficulty.
3	Discriminatory indices of C-Test parts.	Aim: equal, high.
4	Reliability: 4 possible methods, in most cases Cronbach's alpha.	Aim: .90 or more
5	Validity: concurrent validity: correlations with other language tests and school grades Construct validity: theory-driven empirical studies with variance and factor analyses, multiple regression etc.	Aim: high Aim: C-Tests conform to language and linguistic theories
6	CLA-analysis	Aim: unidimensionality, interval scale

* The conceps in Table 1 are explained in the glossary.

Bibliography

Alderson, J.C. (1979), 'The effect on the cloze test of changes in deletion frequency', *Journal of Research in Reading*, 2, 108-18.

Alderson, J.C. (1983), 'The cloze procedure and proficiency in English as a foreign language', in Oller, J.W. (ed.), *Issues in language testing research*, Rowley, Mass., Newbury House, 205-17.

Bachman, L. (1990), *Fundamental considerations in language testing*, Oxford, Oxford University Press.

Gulliksen, H. (1950), *Theory of mental tests*, New York: Wiley

Klein-Braley, C. (1981), 'Empirical investigations of cloze tests', unpublished PhD thesis, University of Duisburg.

Klein-Braley, C. (1985), 'Advance prediction of test difficulty', in C.Klein-Braley and U.Raatz (eds.), *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis*, 23-41.

Klein-Braley, C. (1994), *Language testing with the C-Test. A linguistical and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty*, Universität Duisburg, Habilitationsschrift.

Klein-Braley, C. and Grotjahn, R. (1995), 'Der C-Test: Eine eierlegende Wollmilchsau?', paper presented at the Congress of the Deutsche Gesellschaft für Fremdsprachenforschung, Halle.

Moosbrugger, H and Müller, H. (1982), 'A classical latent additive test model (CLA Model)', *German Journal of Psychology*, 6, 145-9.

Oller, J.W. Jr. (1976), 'Evidence for a general language proficiency factor: an expectancy grammar', *Die Neueren Sprachen*, 75, 165-74.

Raatz, U. (1985), 'C-Tests im muttersprachlichen Unterricht', in C.Klein-Braley and U.Raatz (eds.) *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis*, 66-71.

Raatz,U. and Klein-Braley, C. (1981), 'The C-Test - a modification of the cloze procedure', in T.Culhane, C. Klein-Braley and D.K.Stevenson (eds.), *Practice and problems in language testing*, Essex, University of Essex Occasional Papers, 113-48.

Raatz, U. and Klein-Braley, C. (1983), 'Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis', in R. Horn, K. Ingenkamp and R. Jäger (eds.), *Tests und Trends 1983, Jahrbuch der Pädagogischen Diagnostik*, Weinheim, Beltz, 107-38.

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*, Copenhagen, Pädagogisches Institut.

Spolsky, B. (1981), 'Some ethical questions about language testing', in C.Klein-Braley and D.K.Stevenson (eds.) *Practice and problems in language testing*, Frankfurt, Lang, 5-30.

Spolsky, B., Bengt, S.M., Sako, E.W. and Aterburn, C. (1968), 'Preliminary studies in the development of techniques for testing overall second language proficiency', in J.A.Upshur and J.Fata (eds.), *Problems in foreign language testing. Language Learning Special Issue 3*, 79-103.

Taylor, W.L. (1953), 'Cloze procedure: a new tool for measuring readability', *Journalism Quarterly*, 30, 415-33.

APPENDIX

Glossary of the most frequent concepts in statistics and test theory

Arithmetic mean $\bar{x}$: Average, main weight of a distribution of scores, representative of all scores.

CLA model: Probabilistic model (IRT-model) for investigating whether a test consisting of scaled items measures unidimensionally on an interval scale.

Classical test theory: Deterministic model for minimizing the error of measurement of a test, basis for constructing a norm-referenced test.

Concurrent validity: Conformity between test and criterion (i.e. other test, grades) measured at the same time, calculated by a correlation coefficient. A high coefficient says that test and criterion measure the same trait, but it does not say which trait.

Consistency analysis: Method for estimating the reliability of a test on item basis. The most important formulas are those of Cronbach and Kuder-Richardson.

Construct validity: Very important concept of validity. High construct validity means that the test fits the construct it is intended to measure. Construct validation is a never-ending sequence of theory-driven empirical research. Important methods are factor analysis and multi-trait-multi-method-matrix.

Content validity: Test behaviour and criterion behaviour are identical, the test items are a representative sample of the item domain of the criterion to be measured. No coefficient; expert rating.

Correlation coefficient r: Measurement of relationship of two variables in a sample. Ranges between r = -1 (maximum negative relationship) and r = +1 (maximum positive relationship). Important statistic for calculating the reliability and validity of a test.

Correlation diagram: Graphic representation of a relationship between two variables.

Difficulty index P: For 0-1-items: Percentage of persons in a sample who answered the item right; for scaled items: Mean score of an item related to the maximum score (in per cent).

Discrimination index: Correlation of the item score with the test score, validity of an item using the raw score of the test as a criterion. If the discrimination indices of each item are high, the test is homogeneous, and the reliability (Cronbach) is high. If there are only few items in the test, the indices must be corrected by the part-whole correction.

Face validity: A test looks as if it measures what it is intended to measure. Important for test buyers and test users. Tests which are not empirically valid may have high face validity; tests which are empirically valid may have low face validity.

Factor analysis: Statistical method for grouping tests or items. Gives the number of necessary dimensions and enables them to be named. The score of each test on a dimension is called its loading.

Frequency distribution: Graphical representation of the scores of a sample. In a coordinate system the X-axis represents the score, the Y-axis the frequencies.

Interval scale: Scale of scores with equal intervals. Higher order statistical procedures are permitted, such as calculation of the mean and standard deviation.

Item: Smallest unit of a test. Question to be answered, or stimulus to react to. Most important: multiple choice items , and completion items. For C-Tests, the single text is an item.

Item analysis: Statistical analysis of the single test item after trying out the test in a sample.

Item score: True-false item: 0 or 1 point. C-Testtext: point on a scale between 0 and 20 (depending on the number of gaps).

Multidimensional test: The items measure different traits in an uncontrolled manner. The raw score cannot be interpreted. Also: heterogeneous test.

Norm score: 1. shows the place of a person in a reference group (norm-referenced test). Two kinds: deviation norms (i.e. IQ) and percentile ranks.
2. show what percentage of the criterion the subject has mastered (criterion-referenced test).

Normal curve: Bell-shaped distribution of test scores, also GAUSS curve.

Objectivity: A test is objective if it is independent of the tester in administration, scoring and interpretation.

Parallel test method (or: alternate form reliability): Method of estimating reliability. Two equivalent forms of a test are given to the same sample. The correlation coefficient r is an estimate of reliability.

Part-whole correction: When calculating the discrimination index of an item the correlation coefficient between the item and the final score is artificially raised because the score on the item involved is also part of the test score. Therefore, especially when the test has only a few items, it is necessary either to use a special formula or to subtract the score achieved on the item from the raw score before calculating the correlation between item and raw score.

Predictive validity: Correlation of a test with a future criterion in a sample. Important for predicting criterion score, i.e. in connection with selection or classification.

Quality criteria of a test: Standards of a test. The most important are standardisation, objectivity, reliability and validity.

Rasch model: Probabilistic model (IRT-model) for investigating whether a test consisting of 0-1- items measures unidimensionally on a ratio scale.

Ratio scale: Scale of scores with equal intervals and real zero-point, where ratios are permitted to be interpreted. Statistical procedures of higher order are permitted, such as calculation of the mean and standard deviation.

Reliability: Degree to which a test score is influenced by random error. Measured by the reliability coefficient ranging between r = 0 (only error) and r = 1 (absolutely free of error). Methods for estimation of the reliability coefficient are the retest method, the parallel test method, the split-half method and the analysis of consistency.

Retest method: Method for estimating reliability. A test is given to a sample twice with an interval of some weeks. The correlation coefficient between the two results is the estimation. Precondition: The trait to be measured must be constant.

Split-half method: Method for estimating reliability. Correlation between the two halves of a homogeneous test given to a sample and corrected by special formulas (i.e. Spearman-Brown, Flanagan).

Standard deviation s: Measurement of variability of scores.

Standard error of measurement: Unsystematic error of test score, dependent on reliability. Is given in the norm tables as an interval for the true score.

Standardisation: Precondition for objectivity and reliability. A test is standardised if test material and test situation are fixed.

Test: Standardised arrangement of items given to a person.

Test battery: Combination of different equivalent unidimensional test parts, important if the trait is relatively heterogeneous. Raw scores are added to provide a global total score or a profile is interpreted.

Test profile: Combination of non-equivalent test parts measuring different traits or different aspects of a multidimensional criterion. No total score. Gives a test profile as result.

Test score: Sum of item scores. Also raw score.

True score: Error free result of a test, non-realisable ideal.

Unidimensional test: All items measure the same trait. Homogeneous test.

Validity: 1. A test is valid if it measures the trait which it is intended to measure (content validity, construct validity); 2. A test is valid if it does what it is intended to do (concurrent validity, predictive validity) Also empirical and institutional validity.

Variance: Square of standard deviation.