Dictionary of language testing

a

ability

Current capacity to perform an act. Language testing is concerned with a subset of cognitive or mental abilities, and therefore with **skills** underlying behaviour (for example, reading ability, speaking ability) as well as with potential ability to learn a language (**aptitude**).

Ability has a more general meaning than terms such as: **achievement**, **aptitude**, **proficiency**, **attainment**, while **competence** and **knowledge** are sometimes used as loose synonyms.

Ability is difficult both to define and to investigate, no doubt because, like all **constructs**, it cannot be observed directly. Much of language testing is concerned with establishing the **validity** of such constructs.

See also: **factor analysis**, **performance**, **true score**

Further reading: Bachman 1990; Carroll 1987

ability estimates

See **estimates**

accuracy

An assessment category commonly used in subjectively assessed speaking and writing tasks, particularly in relation to grammatical accuracy. Raters are required to give an estimate of each candidate's ability to produce well-formed grammatical structures; these estimates generally relate to a **proficiency scale** which may or may not include descriptions of typical

performance at each level.

Accuracy may also refer to word choice, or to surface features such as spelling and punctuation (in the assessment of writing) and pronunciation (in the assessment of speaking).

See also: **assessment criteria**

achievement

See **achievement test**

achievement test

Also **attainment test** (UK)

Achievement refers to the **mastery** of what has been learnt, what has been taught or what is in the syllabus, textbook, materials, etc.

An **achievement test** therefore is an instrument designed to measure what a person has learned within or up to a given time. It is based on a clear and public indication of the instruction that has been given. The content of achievement tests is a sample of what has been in the syllabus during the time under scrutiny and as such they have been called parasitic on the syllabus.

An achievement test may be distinguished both from a **proficiency test** and from an **aptitude test** by their uses. A set of grammar test items may be used as an achievement test if all the items have been part of the learners' syllabus; as a proficiency test if adequate performance on these items is required for some real world performance; and as an aptitude test if they provide the means of puzzling out the grammar of an unknown language so as to indicate language learning ability.

The view that an achievement test should measure success on ultimate course objectives rather than on course content is not widely held, largely because such an approach removes the achievement-proficiency distinction.

Because achievement tests are typically used at the end of a period of learning, a school year or a whole school or college career, their results are often used for decision making purposes, notably selection.

See also: **sampling**, **predictive validity**

Further reading: Anastasi 1988; Davies 1990

ACTFL

The **American Council on the Teaching of Foreign Languages** (ACTFL). Well known in language testing for the 'ACTFL Proficiency Guidelines' (1986), and the earlier provisional 'Proficiency Guidelines'

(1982). These provide **rating scales** with **descriptors** at six levels for the four **skills**, and are intended for assessing language students in schools and colleges. A descriptor is a representative sample of a particular range of ability, each one incorporating all lower level descriptors. The authors claim that the Guidelines are atheoretical (and therefore widely useable), proficiency-based and intended for global assessment. The ACTFL Guidelines were based on the **Interagency Language Roundtable (ILR)** Language Skill Level Descriptions. The best known ACTFL test is probably the **Oral Proficiency Interview** (OPI) test.

The ACTFL Guidelines have been much discussed in terms of their underlying holistic-universal view of proficiency, which takes account neither of separate **abilities** nor of differential progress across the skills. Their contribution has been mainly to the development of **direct** or real-life language **performance testing**. The oral proficiency tests using interview (OPI) and tape (**SOPI**) illustrate a real-life test and a semi-real-life one.

See also: **ILR**, **ASLPR**, **FSI**

Further reading: Bachman 1990; American Council on the Teaching of Foreign Languages (ACTFL) (1986)

adaptive testing

Also **tailored test**

A form of individually tailored testing, also known as sequential, branched, tailored, individualised, programmed, dynamic, response-contingent, deriving from the paper-and-pencil exercises of the programmed instruction movement of the 1950s and 1960s. These exercises were of two main kinds: linear and branching. They both provided a teaching unit in which successful progress through a series of frames indicated that satisfactory learning had taken place. In the linear programme success was judged by a prompt at the end of each frame; in the branching programme alternative routes were available for faster learners.

Adaptive tests have become fashionable through the use of variable-branching strategies based on **IRT** and administered by computer. Using a strategy which relates difficulty of **item** to **ability**, the student is repeatedly presented with the next most difficult item until there is no further **score** gain on the **trait** under test. This strategy is known as the maximum information adaptive testing strategy.

The principle underlying all types of adaptive testing is that items are chosen to reflect the level at which the candidate is estimated to be performing, thus generating a more efficient test. The entry level may be determined by a previously determined **proficiency** level, by performance on an initial item, **task** or pre-test, or by age (for example in deciding which of a

battery of L1 reading tests to administer). The accuracy of the final assessment may, therefore, be influenced by the accuracy of the initial judgement.

Adaptive tests may be administered in a variety of formats including individual interview (for example certain types of **oral proficiency interview**), pencil-and-paper format or computerised format. The most successful use of adaptive testing is in **computer adaptive testing**.

Further reading: Anastasi 1990; Bachman 1990

administration

Test administration involves the delivery of a set of test tasks to a group of **candidates** under specified conditions. Depending on the type of test, it may be administered by a language expert (a teacher or a trained **interlocutor**) or by a non-expert (often called an invigilator). Where the work is not specialised, training is rarely undertaken; however, for public tests administrators are frequently given detailed instructions regarding their duties and the resolution of problems. This information may cover the venue, lay-out of tables and chairs, registration of candidates, equipment needed (eg tape-recorders), aids allowed to candidates (eg dictionaries), sequencing of test components (including distribution and collection of papers and timing of tests), and instructions to be given to candidates. Expert test administrators, most commonly used in subjective assessment, are generally trained to ensure that the administration, including selection and delivery of **tasks**, is carried out correctly, thereby ensuring fairness to all candidates.

advantage

See **impact (2)**

affective reaction

The emotional reaction or engagement of a **test taker** to a test. Affective reactions are recognised as influencing the quality of the test **performance**, and as such will contribute to **measurement error**. While individual and group affective reactions to tests and test types have been recorded, the relationship between particular reactions and performance appears not to be stable (anxiety producing better performance in some cases and poorer performance in others, for example), and hence cannot be adjusted for.

See also: **motivation**, **test anxiety**

Further reading: Porter 1991

agreement (in CRM)

See **dependability**

alpha coefficient

See **Cronbach's alpha**

ALTE

See **Association of Language Testers in Europe**

alternate forms, alternative forms

Also **equivalent forms**

Although the terms alternate and **parallel forms** (versions) are used interchangeably, alternate form is a more general term.

Ideally, alternate forms of a test are two or more tests designed to the same specifications. Each form should correspond in terms of: number of test **items**, type of item, content of items, difficulty level of item, as well as in instructions, time allowed, etc. It is sometimes the case that actual alternate forms do not correspond to these demands and are simply two tests which have been designed to be parallel but lack the empirical evidence.

Alternate forms are useful in experimental and follow-up studies and in guarding against loss of test security because they avoid the need to administer the same test twice. They are also traditionally used in establishing test **reliability** based on the same principle as **test-retest** reliability. The argument is of course the same. If the test-retest reliability of Grammar test A can be measured by the size of the correlation between two administrations (as close in time as possible) of the same test to the same group of students, then it will also be measured by the correlation between the administration of Grammar test A1 and of Grammar test A2 to the same students. However, since equivalence is very hard to be sure about, it is usual now to assess reliability by the **split-half** procedure (in practice yet another permutation of the test-retest method) or by **internal consistency** reliability. An advantage of **IRT** methods of test analysis is that test equivalence can be guaranteed through the **anchoring** method.

Item banks from which items can be drawn to desired specifications (in eg **computer adaptive testing**) provide for the continuous construction of alternate forms.

See also: **validity**, **generalisability theory**

Further reading: Anastasi 1990; Cronbach 1964

analysis of covariance (ANCOVA)

See **ANCOVA**

American Council on the Teaching of Foreign Languages (ACTFL)

See **ACTFL**

analysis of variance

Also **ANOVA**

A statistical procedure which allows the comparison of the **means** and **standard deviations** of three or more groups in order to examine whether significant differences exist anywhere in the data. The procedure tests whether the observed values of the groups might all belong to the same **population**, regardless of group, or whether at least one set of observations seems to come from a different population. What the procedure does is to compare the variability of values within groups with the variability of values between groups. A significant result in ANOVA is achieved if the within-group variance is smaller than the between-group variance, the argument being that groups which are distinct are likely to represent different populations.

ANOVA is used in language testing in experimental situations, such as when tests are used to compare the effect of different treatments on language learning. For example the same test might be given to three groups, the first of which has been given intensive oral practice, the second a set of taped listening materials and the third reading-only practice. If the **variance** of 'between the groups' test scores is larger then the variance of 'within the groups' test scores, we would conclude that in our experiment there is an effect on language learning. To determine which treatment is producing the effect we need to perform a priori or post-hoc comparison of means using tests such as Tukey's or Sheffé's.

The research question asked in the above example is straightforward in that it asks only about the effect of different treatments on the language learning outcome. This use of ANOVA is known as a one-way ANOVA. More complex designs can be used. In a two-way ANOVA the **interaction** effect between treatment and the variable 'sex' of testees might be examined and in a three-way ANOVA the further variable 'first language' of testees might be added to the analysis. One possible result of the analysis could then be that the reading-only treatment group members do better but only when they are (a) female and (b) have a Romance language as their first language. On the other

hand, the taped-materials group members might do better when they have Japanese as their first language, but there is no interaction in this case with sex of testee. In other words, we might discover that the effect of the different treatments on test score depends on sex of testee or first language of testee, or both.

ANOVA is also fundamental to procedures used in **generalisability theory**.

See also: **multiple regression**, **sampling**, **hypothesis**

analytic scoring

A method of **subjective** scoring often used in the assessment of speaking and writing **skills**, where a separate score is awarded for each of a number of features of a **task**, as opposed to one global score. In the assessment of writing the functional trisection of content, organisation and structure is commonly represented in the assessment categories. In **speaking tests**, commonly used categories are **pronunciation** or **intelligibility**, **fluency**, **accuracy** and **appropriateness**.

Advantages claimed for the analytic method of scoring are that:

- raters are required to focus on each of the nominated aspects of performance individually, thus ensuring that they are all addressing the same features of the performance;
- it allows for more exact diagnostic reporting of literacy or oracy development, especially where skills may be developing at different rates (reflected in a marked profile);
- it leads to greater reliability as each candidate is awarded a number of scores.

A criticism commonly made of analytic scoring is that the focus on specified aspects of the performance may divert **raters**' attention from its overall effect. This problem may be at least partially overcome by requiring raters to give an overall impression score in addition to the analytic scores. A further problem with analytic scoring lies in the possibility of a **halo effect** distorting the score due to the number of judgements required. The main practical disadvantage of this method of scoring is that it is time consuming compared with **holistic scoring**.

An issue which has to be dealt with where analytic scoring is used is whether to, or how to, weight the different scores.

See also: **weighting**, **criteria**, **holistic scoring**, **multiple-trait scoring**

Further reading: Hamp-Lyons 1991b

anchoring

A technique used to **equate** two **test forms** but which avoids the need for a group of candidates to take both tests by simulating this statistically. Anchoring may be of two kinds: anchoring of **test items** and anchoring of **test takers**. In the first type, a subset of items (the anchor test) is administered to two groups of candidates in addition to one full form of the test (typically either an old form or a new form). Scores on the anchor test are then used to estimate the performance of the combined group on both forms of the test, and hence to compare the two tests. This procedure is used in normal administrations of operational tests as part of the process of development of new versions, obviating the need for special administrations where candidates are required to take two full tests. In the second type of anchoring, a small subset of candidates takes both forms of the test and their performance across the two is used to estimate the performance of the other candidates on both forms, and hence to compare the tests.

Anchoring is also used in the extension of an **item bank**: existing items with known properties are used as the anchors for the analysis of new items, again avoiding the necessity of **trialling** all the items on the same candidates.

Further reading: Petersen *et al.* 1993

ANCOVA

Analysis of covariance. A variant of **ANOVA** which allows for pre-existing differences in a variable which is not the focus of the research, but which might otherwise affect the results and lead to faulty conclusions to be controlled. For example, we may want to investigate the efficacy of three different test preparation methodologies by comparing the performance of three groups of students on a particular language test at the end of a course. Using ANCOVA, pre-existing **ability** differences amongst the groups can be controlled by taking into account a **pre-test** score on some common measure. In other words, there is no need to eliminate particularly high ability or low ability students in order to control the variable ability across these two groups: rather the effect of ability is controlled statistically.

ANOVA

See **analysis of variance**

answer key

See **key**

anxiety

See **test anxiety**

a posteriori test validation

Latin for 'from what comes after': procedures used to establish what a test actually measures after it has been developed. This process may include the use of statistical procedures, such as **multi-trait multi-method**, or the soliciting of expert opinion.

See also: **validation**, **a priori test validation**

Further reading: Weir 1988

applied linguistics

Two opposing views of applied linguistics as a discipline are prevalent: the first may be labelled the weak view, the second the strong view.

The view of applied linguistics as a weak discipline suggests that linguistic theories and procedures may be applied to other disciplines, eg the study of literary texts or writing a syntax for a computer program. An alternative weak view is that linguistics may be an important reference point in a number of (not necessarily connected) areas of language work. For example, speech pathology, communication engineering, speech technology, cognitive science, language in education, discourse analysis, interpreting and translating. This version of applied linguistics has been called 'linguistics applied' to distinguish it from the strong version, 'applied linguistics'.

The view of applied linguistics as a strong discipline normally refers to the institutionalised discipline of applied linguistics which concerns itself largely with language learning and teaching (and sometimes remediation, as in speech pathology). This version of applied linguistics attempts to be both problem based (or focused) and theoretically oriented. The theories (and methodologies) drawn on come not only from linguistics but also from other disciplines such as education, sociology, psychology, etc. To what extent this strong applied linguistics can achieve its own all-embracing theory has yet to be seen, although it is not clear that it need do so any more than other problem or professionally-oriented disciplines (eg engineering, medicine, social work, law).

Since language testing is focused on major 'problems' in language work, and is firmly committed to the measurement of language learning in context, it occupies a central position within a strong applied linguistics.

Further reading: Davies 1990; Bachman 1990

appropriacy

See **appropriateness**

appropriateness

Also **appropriacy**

An assessment **category** or **criterion** reflecting sociolinguistic competence: the relationship between language, the language users and the context of use. Such conventions of language use determine the extent to which the performance of a language **task**, or function, is viewed as appropriate. Assessments of appropriateness in relation to spoken or written language usually require **raters** to focus on the register or style of language used and the extent to which it is appropriate to the task.

a priori test validation

Latin for 'from what comes before': procedures to ensure that a test will actually measure what it is intended to measure before the test is developed. The process normally involves a **needs analysis** to provide a description of the test **domain** and may also involve review of **test content** during the course of test development.

See also: **a posteriori test validation**

Further reading: Weir 1988

aptitude

The extent to which an individual possesses specific language learning ability. Research is somewhat unclear on the existence of a general language aptitude **variable**; various aptitude tests have attempted to define and operationalise the **construct** in various ways.

See also: **aptitude test**, **Modern Language Aptitude Test**, **Language Aptitude Battery**

Further reading: Carroll 1981; Skehan 1989

aptitude test

An instrument to measure the extent to which an individual possesses specific language learning **ability**. Such tests are usually used for selection and diagnosis and for prediction of language learning success. Research is somewhat unclear on the existence of a general **aptitude variable** and the tests that exist normally claim to predict success only in terms of defined learning outcomes or distinct methodologies.