Two related theoretical notions are implicit in the concept of reliability.

First, the basic premises of classical test theory hold that actual measures (ie test scores) are “fallible” scores in the sense that they are not 100% reliable. Scores contain a degree of error because tests are not perfect measuring instruments. This “error” is random and unpredictable.

A “true” score is a term used to describe a “true” level of performance (ie an individual’s actual, or “real” ability). If a test existed which was perfectly reliable, the test score for a particular individual would be that person’s “true” score (ie their real ability). But, in practice, scores can be conceptualised as consisting of a “true” score, surrounded by a band of random error. This band of error decreases in size as the reliability of the test is improved.

Second, the notion of Generalisability Theory holds that the reliability of a test concerns the extent to which one can generalise from results obtained under one set of conditions to those which would be obtained under another.

Remember that a test can be defined as an objective and standardised sample of behaviour. The test reliability provides an index of the extent to which we can generalise from this “sample of behaviour” to other situations.

It is thus reasonable to assume that performance on a highly reliable test of basic arithmetic computation could be generalised to other situations, where similar kinds of arithmetic computation are required. However, if the test is less reliable, it is correspondingly less reliable to generalise to other situations.

Having accepted that all tests are fallible and result in a degree of error in test scores, we can nevertheless attempt to improve reliability by minimising the error inherent in tests.

On the following pages, we will examine ALL the theoretically possible sources of error that can occur in testing.

Please note, however, that in reality it is highly unlikely that ALL these possible sources of error will apply to the tests we are using.
SOURCES OF ERROR

Sources of random or systematic error include the following:

**Measurement error** concerns the quality and composition of the test questions (or items). Poor items, such as those that are ambiguous, misleading, or those overly dependent on specialised knowledge will reduce test reliability.

**Item sampling** concerns the process by which items are selected for inclusion in the final version of the test. Items measuring the discrete ability which the test purports to measure will improve the reliability of the test, whereas questions measuring subtly different abilities, or those contaminated by other factors, will represent a source of error.

**Administration of tests** should be standardised, as variations in any aspect of test administration could be a source of error. Variations in provision of background information, instructions, style and speed of oral instructions, feedback on practice questions, opportunities for candidates to ask questions etc all represent potential error if not standardised.

**Situational factors**, such as lighting, seating, workspace, noise level, distractions and interruptions, all represent sources of error if not adequately controlled.

**Scoring error** is particularly problematic for open ended tests (rather than multiple-choice) which require a degree of subjective evaluation on the part of the assessor. Reliability of scoring can be increased by two assessors independently scoring responses and subsequently resolving any disparities. However, multiple choice tests are also prone to a degree of scoring error, particularly if answer sheets and/or scoring keys are poorly designed. Some tests are machine scoreable. This tends to reduce scoring error but does not eliminate it altogether.

**Practice effects** are problematic in the application of widely used tests, because candidates applying to several organisations may be asked to take the same test on several occasions. This problem is explored in the article “Graduate Testing - The Problem of Practice Effects” available from the KCP Training College website.

**Guessing** the answers to ability test items introduces random error, and so test instructors should discourage candidates from wild guessing.