iCMA statistics

These statistics are designed for use with summative iCMAs where students have just one attempt and complete that attempt.

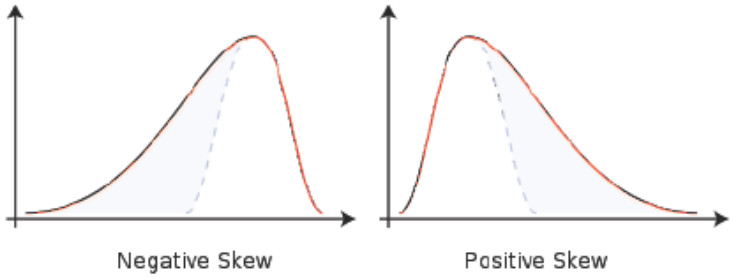
Test statistics

Average grade: For discriminating, deferred feedback, tests aim for between 50% and 75%. Values outside these limits need thinking about. Interactive tests with multiple tries invariably lead to higher averages.

Median grade: Half the students score less than this figure.

Standard deviation: A measure of the spread of scores about the mean. Aim for values between 12% and 18%. A smaller value suggests that scores are too bunched up.

Skewness: A measure of the asymmetry of the distribution of scores. Zero implies a perfectly symmetrical distribution, positive values a 'tail' to the right and negative values a 'tail' to the left.



Aim for a value of -1.0. If it is too negative, it may indicate lack of discrimination between students who do better than average. Similarly, a large positive value (greater than 1.0) may indicate a lack of discrimination near the pass fail border.

Kurtosis: Kurtosis is a measure of the flatness of the distribution.

A normal, bell shaped, distribution has a kurtosis of zero. The greater the kurtosis, the more peaked is the distribution, without much of a tail on either side.

Aim for a value in the range 0-1. A value greater than 1 may indicate that the test is not discriminating very well between very good or very bad students and those who are average.

Coefficient of internal consistency (CIC): It is impossible to get internal consistency much above 90%. Anything above 75% is satisfactory. If the value is below 64%, the test as a whole is unsatisfactory and remedial measures should be considered.

A low value indicates *either* that some of the questions are not very good at discriminating between students of different ability and hence that the differences between total scores owe a good deal to chance *or* that some of the questions are testing a different quality from the rest and that these two qualities do not correlate well – i.e. the test as a whole is inhomogeneous.

Error ratio (ER): This is related to CIC according to the following table: it estimates the percentage of the standard deviation which is due to chance effects rather than to genuine differences of ability between students. Values of ER in excess of 50% cannot be regarded as satisfactory: they imply that less than half the standard deviation is due to differences in ability and the rest to chance effects.

CIC	100	99	96	91	84	75	64	51
ER	0	10	20	30	40	50	60	70

Standard error (SE): This is SD x ER/100. It estimates how much of the SD is due to chance effects and is a measure of the uncertainty in any given student's score. If the same student took an equivalent iCMA, his or her score could be expected to lie within ±SE of the previous score. The smaller the value of SE the better the iCMA, but it is difficult to get it below 5% or 6%. A value of 8% corresponds to half a grade difference on the University Scale – if the SE exceeds this, it is likely that a substantial proportion of the students will be wrongly graded in the sense that the grades awarded do not accurately indicate their true abilities.

Question statistics

Facility index (F): The mean score of students on the item.

F	Interpretation	35-64	About right for the average student.
5 or less	Extremely difficult or something wrong with the question.	66-80	Fairly easy.
6-10	Very difficult.	81-89	Easy.
11-20	Difficult.	90-94	Very easy.
20-34	Moderately difficult.	95-100	Extremely easy.

Standard deviation (SD): A measure of the spread of scores about the mean and hence the extent to which the question might discriminate. If F is very high or very low it is impossible for the spread to be large. Note however that a good SD does not automatically ensure good discrimination. A value of SD less than about a third of the question maximum (i.e. 33%) in the table is not generally satisfactory.

Random guess score (RGS): This is the mean score students would be expected to get for a random guess at the question. Random guess scores are only available for questions that use some form of multiple choice. All random guess scores are for deferred feedback only and assume the simplest situation e.g. for multiple response questions students will be told how many answers are correct.

Values above 40% are unsatisfactory – and show that True/False questions must be used sparsely in summative iCMAs.

Intended weight: The question weight expressed as a percentage of the overall iCMA score.

Effective weight: An estimate of the weight the question actually has in contributing to the overall spread of scores. The effective weights should add to 100% - but read on.

The intended weight and effective weight are intended to be compared. If the effective weight is greater than the intended weight it shows the question has a greater share in the spread of scores than may have been intended. If it is less than the intended weight it shows that it is not having as much effect in spreading out the scores as was intended.

The calculation of the effective weight relies on taking the square root of the covariance of the question scores with overall performance. If a question's scores vary in the opposite way to the overall score, this would indicate that this is a very odd question which is testing something different from the rest. And the computer cannot calculate the effective weights of such questions resulting in warning message boxes being displayed.

Discrimination index: This is the correlation between the weighted scores on the question and those on the rest of the test. It indicates how effective the question is at sorting out able students from those who are less able. The results should be interpreted as follows

50 and above	Very good discrimination
30 – 50	Adequate discrimination
20 - 29	Weak discrimination
0 - 19	Very weak discrimination
-ve	Question probably invalid

Discrimination efficiency: This statistic attempts to estimate how good the discrimination index is relative to the difficulty of the question.

An item which is very easy or very difficult cannot discriminate between students of different ability, because most of them get the same score on that question. Maximum discrimination requires a facility index in the range 30% - 70% (although such a value is no guarantee of a high discrimination index).

The discrimination efficiency will very rarely approach 100%, but values in excess of 50% should be achievable. Lower values indicate that the question is not nearly as effective at discriminating between students of different ability as it might be and therefore is not a particularly good question.