Getting More Meaning from the Results of CAA
Discussion forum introduction

John Kleeman MA MBCS C.Eng
Managing Director, Question Mark Computing Ltd
5th Floor, Hill House, Highgate Hill London N19 5NA
Tel: (020) 7263 7575
Fax: (020) 7263 7555
Email: john.Kleeman@qmark.co.uk
Web: www.qmark.com

Abstract:
This paper aims to stimulate discussion at a focus group on the topic of getting more meaning from the results of CAA. Some general thoughts are given, and then there is a specific example of Question Mark Perception's new Group Comparison report as illustration.

Introduction
Much of the reporting and analysis of CAA tends to follow the model for paper tests. The various reports that people used to make for paper tests are duplicated for computer tests. For example, analysis is made of the quality of questions and choices and of the reliability and validity of the test.

This analysis is important, but when a test is delivered on paper, the only analysis that you are likely to do is that which can be done easily without further calculation or data entry. In the more modern environment where CAA is commonplace, the full results from the assessment are in the computer automatically, without any work needed. This gives us the ability to think of new kinds of analysis, that would never have been considered when most tests were on paper, but which is worth doing now they are on computer.

Although sometimes overlooked, it's a truism that the primary purpose of CAA is the information and results it provides. The more meaning we can get out of the results, the more meaning and value the whole assessment process gives us, and the more it can be used to improve teaching and learning.

This paper only touches on the surface of what might be possible, but hopefully will trigger some thoughts and ideas for the future.

Analysis only possible with CAA

When conducting CAA, information can often be gathered that it simply isn't possible to know when conducting paper tests. For example the following information might be available:

- Time taken to complete the test or in some systems the time taken on each question or section of the test.
- Time of day the test was taken and perhaps the place or PC it was taken from.
- In some advanced question types, there may be information about the way a participant answers, or the order of selecting choices that can be significant.
- If questions are selected at random for the computer test, then the order of questions or which combinations of questions are included could have some meaning for the results.
- And if choices within a question are shuffled in order, this information could have significance.

The above information may or may not be significant, and perhaps in a well designed test, it shouldn't be significant. But some interesting questions one could ask might be:

- Is there a correlation between the time spent on a test and the score achieved at it?
- Are the results of one question influenced in a statistically significant way if a particular second question is included in the random selection for a test? This could pinpoint if the wording for one question included some information about another question's answer by mistake.
- Are there some PCs which people do consistently badly on? This could pinpoint some problem in the hardware or environment (e.g. screen glare).

Another related issue is that with results stored in a computer database, it becomes possible to change the data that is recorded. For example, it's possible to re-score an exam if one question is found to be ambiguous after the event.

Analysis that becomes much easier with CAA

When questions are answered on paper, a lot of work is required to input the results onto the computer for analysis, and the less that is input the better. This means that only the conventional analysis is likely to be performed. But when the answers, scores and other data are already on the computer, as is the case with CAA, the only work to perform further analysis can be to get some software to do some work for you. This means:

- It can be very straightforward to compare current data with historical data, for example to analyse the differences between this class and prior classes.
- It can be possible to correlate data from assessment with other computerised data in the institution, for example to do an analysis by age of student or other courses taken.
- It's possible to browse the data online, to find particular information quickly or to present and see information in different ways.
- It's possible to do speculative analysis, where you try something that may or may not be useful, but it costs nothing to have a go at.
- It's possible to consider data mining techniques, which use a computer to trawl through a large amount of data, trying to find new or useful patterns and trends.
- And of course, it's easy to make graphs to display things visually, to make numbers come more to life.

Question Mark Perception Group Comparison report

An example of a useful report that can give extra meaning to existing data is the Group Comparison report, recently introduced in version 2.2 of Question Mark's Perception product.

To use this report:

1. You define one profile of people (which can be an individual, a whole class, or some set of filters defining people in more complexity)

2. You define a second set of people in a similar way.

3. You define what you want to compare, typically the performance of the people on the same test or different tests that cover the same topics.

4. You then get a report illustrating the difference or gap between the two profiles of people. The gap can be presented in simple numbers, with statistical measures and/or graphically.

One of the most valuable ways to use the report is to use it to compare differences before and after a course (to perform level 2 analysis under the Kirkpatrick model).

This works by you delivering a test to students at the start of the course as a pre-test. Then you deliver a test at the end of the course (a post-test). The tests can be identical, or they can select at random from the same question bank, or they can be different, but designed to cover the same topic areas. The difference between the two test results reflects the effect of the course. If scores before and after the course are the same, then the course has not taught much (or the tests

have not been designed correctly). You can judge the effect of the course by looking at the size of the difference in scores, both for the test as a whole and for each topic.

The figure below shows an illustrative report for such an analysis. The course has improved scores overall and for each topic except Topic B, where for some reason, performance has actually decreased. Such a report should cause some urgent examination of how Topic B is taught on the course, or how it is being assessed.

	Before course	After course	Difference	
	Mean score	Mean score	Improvement	Improvement as a graph
Session score	50%	75%	+25%	
Topic A	40%	70%	+30%	
Topic B	50%	40%	-10%	
Topic C	50%	90%	+40%	
Topic D	60%	100%	+40%	

As well as comparing the same class before and after the course, you could also use the same style of report to compare the results after the course between different groups of people. For example, how does the class this year compare to the class last year? How do the results for females and males differ in the different topics? If your course has optional alternative courses, do people who do one of these alternatives do better on your course than people who do not? And so on.

You can also use the report to compare one individual with his/her peers and see where he/she is improved or needs improvement compared to the group. You might for example, get a report like the following stylised example:

	Individual	Group mean	Group SD	Difference vs mean	Divergence in SD
Test score	73%	73%	12%	0%	0
Topic A	64%	69%	12%	-5%	-0.42
Topic B	79%	90%	5%	-10%	-2.2
Topic C	76%	60%	18%	+16%	+0.89

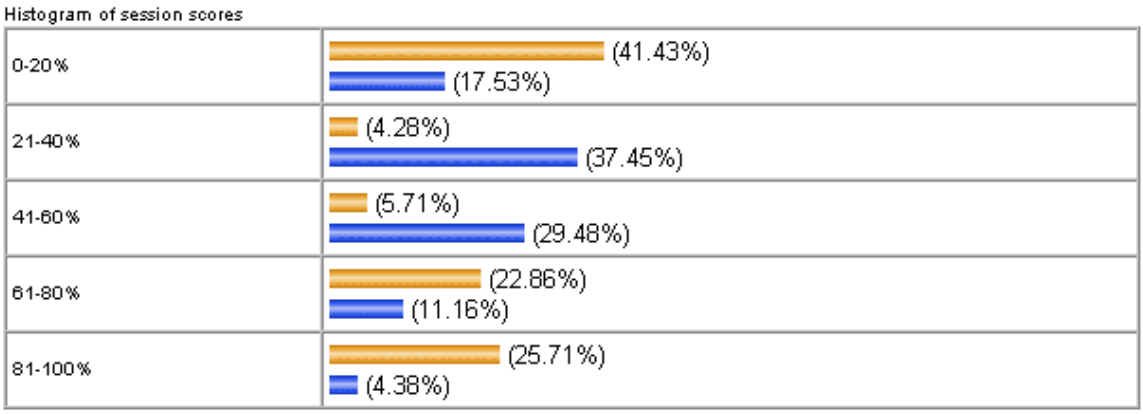
At first sight, the individual has got the same score as the group mean, and so is likely to be around average.

On second sight, the individual is a bit stronger in topic C and a bit weaker in topics A and B.

But a look at the far right column gives another picture. The scores for topics A and C are within one standard deviation of the mean, but the score for topic B is two standard deviations from the mean, implying that although the individual's overall score is average, his score on topic B is significantly below average, possibly in the bottom 2% of the class.

The Group Comparison report can also compute t values, which can tell you how significant differences between groups are. And it can present graphical histograms that compare distributions of scores in the two different profiles of people.

In the figure below, comparing the scores for two groups of people, both groups have around the same mean score of 40% or so. The second group has results spread around the mean and tailing off, whereas the first group has many good results, many poor results and few in between. This points out an important difference between the groups, which is likely to be meaningful.



Conclusion

The concluding message is to encourage us all to not just mimic the classical analysis of results with CAA, but to think about what sort of information might be possible and meaningful with CAA, and to go out and get it.