This material represents the CASNR CIEQ Hand-Out prior to 2007 and the development of the Percentile Composite scores.

1. What is the purpose of the CIEQ?

Course evaluation or improvement should be based on a minimum of four sources of information. They include (1) self evaluation, (2) peer evaluations, (3) professional evaluation, and (4) student evaluation. Data from the CIEQ can provide the student evaluation portion of the process. CIEQ data do NOT provide a complete evaluation of course effectiveness. They are measures of the agreement between the instructor's expectations for a course and the students' expectations. As such, CIEQ data alone should never be the SOLE source of information leading to course evaluation or improvement.

2. How should the CIEQ be used?

The CIEQ provides data from students, most of whom do not have the credentials to judge the amount or quality of leaning going on in a course. However, they are very well qualified to judge their reaction to the instruction. These reactions can provide insight into learning successes and problems. Also, experience has shown that students often agree with more qualified raters.

3. When should I give the CIEQ?

CIEQ data can be collected any time during the last half of the course. The students should be given plenty of time to complete the questionnaire. Signatures or other identifying marks should never be required. Every effort must be made to insure that responses are anonymous.

4. How are the means and standard deviations calculated?

Each question allows four answers: strongly agree, agree, disagree, and strongly disagree. Some questions are stated in a positive manner and some in a negative manner. If the question is positive, these four answers are given the values 4, 3, 2, and 1 in the scoring process. If the question is stated negatively, these four answers are rated 1, 2, 3, and 4. The means are simply the averages of the answers. For a positive question example, if there were 55 students who answered "strongly agree," 32 students who answered "agree," 10 students who answered "disagree," and two students who answered "strongly disagree," then the mean will be given by:

(55 x 4 + 32 x 3 + 10 x 2 + 2 x 1)/99 = 3.41

where 99 is the total number of students. The standard deviations are calculated from the differences from this mean. The calculation is straightforward by standard statistical methods. It is a measure of consistency of student responses. If all students pick the same response, the standard deviation is 0.0. If one fourth of the class picks each of the four choices, the standard deviation will be maximum at approximately 1.12. The lower the standard deviation, the greater the agreement among the student responses. Note that this is a standard deviation, not a standard error.

5. How are the questions grouped to form the five general categories?

The 21 questions are grouped into five categories. The categories and the questions included are listed below:

General course attitude:	Questions 4, 9, 18, 24
Method of Instruction:	Questions 5, 10, 15, 21
Course Content:	Questions 7, 11, 16, 19
Interest and Attention:	Questions 8, 13, 20, 22
Instructor:	Questions 6, 12, 14, 17, 23
Total:	**All Questions**

Means for each of these areas are obtained by summing the numbers of each answer for all the included questions and performing the calculation of the mean and standard deviation as shown in Question 4. Note: These means are calculated directly form the student responses. They are the most accurate indicators of student data since no additional interpretation has been done. Further, these means and standard deviations serve as a basis to assess statistial significnce of data.

The "Instructor" mean is used in most summative evaluation systems as the best measure of teaching performance.

6. How are the deciles calculated and what do they mean?

The deciles are each calculated from a norm database. This database includes the means for a large number of courses which have been evaluated in the past. It can be subdivided by course or instructor characteristics such as course level, instructor rank, department, or class size producing database subsets. To generate the decile dividers for any given norm database, the means are listed in descending order. This list is divided into 10 groups each including 10% of the means, hence the name “deciles.” The group of highest means is called decile 9, the next lower group is called decile 8, and so forth. The lowest 10 % would be decile 0. This partitioning yields 11 numbers which indicate the borders of the deciles. A possible example would be:

Decile	9	8	7	6	5	4	3	2	1	0
Upper Limit	4.00	3.60	3.45	3.25	3.10	3.00	2.90	2.75	2.55	2.25
Lower Limit	3.60	3.45	3.25	3.10	3.00	2.90	2.75	2.55	2.25	1.00

An observed mean of 3.18 would fall in decile 6. Note that the means usually follow a normal distribution, so the boundaries in midrange will be much closer than on the extremes. A decile of 9 indicates that 90% of the means are less than yours.

The composition of any norm database remains constant until it is replaced with another database. A change in a norm database will not change the means and standard deviations, but may change the deciles.

The purpose of the deciles is to compare any of the six means for a course with those from other courses. If the characteristics of the norm database do not fit those for the course under study, the decile rankings mean nothing. Another problem with deciles is that there are no simple statistical procedures to determine the significance of the decile rankings.

7. How is the content of the norm database decided?

In CASNR, the Dean of the College determines the composition of the norm database. It is updated when resources are available to do so. The current norm database is...

8. How should the deciles be used by a teacher or evaluator?

If the norm database consists of similar courses to the one being evaluated, then the teacher can get an idea of how he or she compares with other teachers on the 6 areas of the report. However, the deciles are very sensitive in midrange, and may change based on very small and often insignificant changes in the mean. If the appropriateness of the norm database is unknown or doubtful, the decile rankings should be ignored. Trend analyses of decile data are very risky because of over sensitivity, lack of statistical support data, and possible changes in the norm database over the period of comparison.

9. What are the meanings of the headings on the summary report?

Subscale- The title of the group of questions included in the line.
P/Res - Percent of students responding to the questions.
Mean - The mean response on the questions as described above.
S. D. - Standard deviation.
REL - Reliability. The internal consistency of the evaluation document on these questions.

The remainder of the headings describe the titles of the norm databases used to calculate the decile rankings.

10. How does one interpret means for questions and categories?

Obviously, the means must occur between 1 (low) and 4 (high). A mean rating above 3.5 means that more than 25% of the class rated the item "strongly agree" (or "strongly disagree" if the question is negative). Overall, the following table relates the percentages of students (who "strongly agree" or "agree" in the case of positively stated questions or "strongly disagree" or "disagree" in the case of negatively stated questions) to the means.

Percent	100	95	90	85	80	75	70	65	60
Mean	3.5	3.4	3.3	3.2	3.1	3.0	2.9	2.8	2.7

Each teacher must decide for each course what satisfaction level is most appropriate. These values will vary slightly depending on the distribution of responses.

11. What can be derived from the student comments on the back of the form?
One must remember that the teacher rarely knows the context in which the comments were made. Therefore, single comments that are not supported by other comments or other data should probably be ignored. However, multiple comments on a subject that have a similar vein can be useful if observations or other data have supported them. The most common error in the use of student comments is to give too much weight to isolated and unsupported statements.

12. How does one validate changes in means or trends?
For each mean, the teacher needs to have some idea of statistical significance. There are at least three ways to document changes. First, one can put confidence limits on each mean. If these confidence limits do not include the previously observed mean then a change in behavior has been documented. Second, one can make a statistical comparison between the last two means with a t-test. If this difference is significant, then there has been a change. Note, in both cases the standard error must be computed from the standard deviation. Third, one can perform regression or correlation on the means for the last 4 or more years. If the regression or correlation coefficients are significant, then there is a valid trend. If none of these (or other) methods yield significant results, then one must conclude that no changes in student ratings have occurred irrespective of observed differences in means.

13. How should a teacher use CIEQ data for evaluation?
First prepare a table for each course showing means for the 6 categories for the last 5 years. Compute statistical data to document changes over the period or form year to year. Try to identify changes in the conduct of the course and student comments which support the documented trends. Utilize the data from the 21 questions where they are useful and needed. Where no trends can be documented, report mean values to prevent evaluators from using non-significant differences to produce unwarranted conclusions. Prepare a short narrative to explain the significance of your observations.

14. How does a teacher use CIEQ data for improvement?
If the mean for the total evaluation is above 3.4, it is doubtful that there is much in the evaluation that is useful for course improvement. Student acceptance of the course is very high and valid clues for improvement must come form other sources. Looking further is likely to lead at best to tinkering which will probably have little effect on course quality, or at worst, to undesirable changes.

If the total evaluation is below 3.5, look at means for the 5 subscales. If any of these are below 3.2, look at the means for the individual questions for that subscale. If question means are below 3.2, try to determine why students are answering the questions as they are. Utilize the standard deviations to assess agreement among students. In some cases, a low rating might be normal (Question 17, for example). Consider student comments which might help explain low mean ratings. Summarize your thoughts in writing so they can be used the next time the course is up for comprehensive review. These threshold values are only suggestions. Others might also be chosen.

It may be worthwhile to examine evaluation data further even if means are above cutoff values mentioned above. However, the teacher must be very careful to avoid "knee-jerk" changes in the course that are unlikely to provide significant course improvements. Major course changes should be made ONLY on the basis of a well-designed comprehensive course review.

What is the in-class procedure for evaluating courses?

The faculty member administers the evaluation with the choice of remaining or leaving during the time needed by the students for completing the form. A student of the class, eithr volunteer or designee, takes the completed forms (in a sealed envelope) to the departmental office. A designated staff member sends the forms to the scoring service. Following the filing of grades, the results are sent back to the department head who returns all raw data sheets and comments (along with the summary of the core questions) to the faculty member. The departments head retains a copy of the summary and forwaards one copy to the Dean's Office.

How do we interpret REL?
The reliability (REL) is based on an internal consistency calculation (alpha coefficient)which indicates the proportion of variation in the score due to the variation among people. Reliability figures over 0.90 can be indicative of insonsistent responses within a subscale. However, they can also represent artifacts of the calculation which are not representative of the consistency.

Reliability coefficients are affected by the size of the saample. They are also dependent on the vaariance component. If all responses are similar, the vaariance component will approach 0 and the reliability index will fall below 0.65.

The reliability figure should not be evaluated independent of all of the other information, particularly if it is below 0.65. Each item within a subscale should be evaluated for consistency for that subscale item.

How do I know if the results are adequate?
Refer to the top of the first page of CIEQ output. Check the "Sample Size". If the sample size or number of studentsresponding is less than one half of the course enrollment, results may be biased and should be interpreted with caution.
At the bottom of the first page is the section entitled "Subscale Results" that contains a column of figures labeled "Re[.". This column contains the obtained reliabilities for the six subscales of the CIEQ. Any subscale with a "Rel." below .65 should be interpreted with caution. Consult the Manual for further details.

Comparative Information
In all cases, comparative information is provided by decile rank (DEC). The decile rank describes the current course "MEAN" in relation to other courses that have administered the CIEQ. Decile ranks are always interepreted as follows:

1 - 3 Substantial improvement needed
4 - 7 Some improvement needed
8 – 10 No improvement needed

Differences between adjacent pairs of decile ranks within each interval (e.g. 1 vs. 4 vs. 5) are not considered to be significantly different.

First refer to the "Subscale Analysis" listing at the bottom of the output on page 1. Each subscale represents a different aspect of the course as indicated by its title. Decile ranks for the current course/instructor are listed for each subscale in comparison to six normative groups:

"IR", all instructors of the same faculty rank;
"CL", all course at the same grade level (e.g. freshmen, sophomore, etc.);
"D", all courses within the same department;
"C", all courses within the same college;
"UA", all courses at the University of Arizona;
"N", all courses that have used the CIEQ in the United States.

On the following two pages under the heading "Individual Item Results" are listed each of the 21 individual items of the CIEQ along with the proportion (%), frequency (#), mean, and standard deviation (S.D.) of responses to each individual item of the CIEQ. Also listed is the text of each item and the most favorable response or "BEST" answer, for each item.   All means have been scaled such that 4.00 is the most favorable response and 1.00 is the least favorable response regardless of the initial wording of the item.. To the far right of each individual item are listed decile ranks that compare each item mean to the item means obtained in all courses within the same college.

In interpreting results, refer first to the decile ranks for subscales. Low deciles for a subscale identify potential problem areas. Individual items can then be examined for more specific information. The subscales are composed of the following individual items:

"Attitude"	Items 1, 6, 15, 21;	"Interest"	Items 5, 10, 17, 19;
"Method"	Items 2, 7, 12, 18;	"Instructor	Items 3, 9, 11, 14, 20;
"Content"	Items 4, 8, 13, 16;	"Total"	Items 1 - 21.

Descriptive Information
Refer to the top of the first page of CIEQ output. Following the initial titles, information is listed on the composition of the responding sample under the heading "Class Description Results." Both the proportion and frequency of responses is listed for each alternative of the following items: "Class Information", "Gender", "Course Option", "Pass-Fail Option", "Major-Minor", and "Expected Grade".

The next portion of the output lists the proportion (%), frequency (#), "Mean", and standard deviation (S.D.) of responses to three global ratings: "Content Rating", "Instructor Rating", and the "Course Rating". A mean value of 6.00 is the most favorable rating. These three items have NOT been validated and should, therefore, be used only for the purpose of feedback to the instructor.

Optional Item Information
If Optional Items Section I was used, there will be two pages of output entitled "Optional Items Analysis" for items 22-42. Each of these items will have the same information that was presented for items 1-21 except for the decile information. NOTE: the best response is always AS for Optional Items Section 1. If the best response for an item should be DS, then subtract the MEAN of that item from 5.00 to obtain the correct MEAN value. The standard deviation (S.D.) is not affected by the positive or negative direction of the item statement.

If Optional Items Section 11 was used, there will be two pages of output for items 43-63. Each of these items will have the same information that was presented for items 22-42 except that the best response is always A. If the best response for an item should be E, then subtract the MEAN of that item from 6.00 to obtain the correct MEAN value.

What are some references I can read on the CIEQ?

Aleamoni, Lawrence M. “Issues in Linking Instructional-Improvement Research to Faculty Development in Higher Education”. Journal of Personnel Evaluation in Education, 11: 311–37, 1997.
Aleamoni, Lawrence M. “Evaluating Instructional Effectiveness Can Be a Rewarding Experience”. Plant Disease, April 1987.
Aleamoni, Lawrence M. “Student Rating Myths Versus Research Facts”. Journal of Personnel Evaluation in Education, 1: 111-119, 1987.
Aleamoni, Lawrence M. “Development and Factorial Validation of the Arizona Course/Instructor Evaluation Questionnaire”. Educational and Psychological Measurement, 1978, 38.