



# University of Colorado at Boulder

Office of the Associate Vice Chancellor for Undergraduate Education

---

306 Regent Administrative Center  
40 UCB  
Boulder, Colorado 80309-0040  
(303) 492-5538

28 February 2005

Phil DiStefano  
Interim Chancellor  
University of Colorado at Boulder

Dear Phil:

The Chancellor's FCQ Advisory Committee has now completed its work and hereby submit its recommendations regarding the FCQ instrument itself, and its recommendations for proper use of the results of this survey of student opinion. We submit them to you only as the product of this committee's work. We have not circulated this report for comment or advice.

Our report has eight parts:

- (1) Our annotated, question by question set of recommendations for a revised FCQ testing instrument which is shorter and contains, with one exception, only professionally recognized and generally validated questions,
- (2) Our specific, point by point response to the BFA recommendations about the FCQ to your office,
- (3) A list of thirty five nationally published recommendations regarding several aspects of the FCQ plus our recommendations about each one of them,
- (4) The results of our pilot study done last spring,
- (5) Our recommendations for interpretation and use of the statistical results by evaluators (personnel committees at the department, school or college and campus levels),
- (6) Our recommendations directly to faculty regarding administration of the FCQ itself and some guidance about interpretation,
- (7) A bibliography of key overview research literature citations, and

(8) the recommended FCQ form itself.

We hope you find these recommendations useful and that they will contribute to a better, more effective campus effort at improving the quality of our teaching, the quality of our evaluation of our teaching and the quality of our courses.

If and when these recommendations are accepted, I will consult with Planning, Budget, and Analysis on implementation details. If approval comes during spring 2005 we expect implementation in fall 2006. PBA cannot implement for fall 2005, and both committee and PBA agree that implementation should come in a fall for a full academic year.

The complete report (letter plus eight parts) is ready for Web posting but has not been posted yet. Please let me know if and when we should post. Should you decide to solicit comment on the report, having it posted would make this more convenient.

Please let us know if you need anything else from us.

For the FCQ Advisory Committee,

Michael Grant  
Committee Chair

CC: Committee Members: Susan Clarke, Political Science; Janet DeGrazia, Engineering; Susan Kent, Faculty Affairs; Basim Mahmood, UCSU; Lou McClelland, Institutional Analysis; Joe Neguse, UCSU; Christine Queja, UCSU; Christine Rohde, Honor Code Office; Lori Seward, Business; James Symons, Theater and Dance

The committee met January 2004 through January 2005. Some members participated in only spring 2004 or fall 2004.

Part 1:

### **The FCQ Instrument Recommended by the Chancellor's Advisory Committee**

Following extensive discussions and consultations with the published literature on the subject, the Advisory Committee (AC) has recommended a shortened instrument with a single narrative question. Every question, with the exception of question number nine, has been validated as appropriate and recommended in the student assessment literature. We have also recommended a change in rating scale from a letter format to a 1 (lowest) to 6 (highest) format for most questions. See the attached draft FCQ form for our recommended layout.

*1) Estimate the average number of hours per week you have spent on this course for all course-related work including attending classes, labs, recitations, readings, reviewing notes, writing papers, etc. {Scale: 0-4, 5-8, 9-12, 13-16, 17-20, 21+}*

This asks for a quantitative measure of time spent by students on all course-related work. The answer to this question is of great interest to many faculty and will provide student perceptions of their own workloads analogous to faculty self-reporting of their professional time investments. This survey question may help the university deal with the damaging reports from the Princeton Review about our students' study habits in a much more defensible and reliable manner. This question is structured differently from all the rest and has a unique response scale. We have placed it first on the instrument, based partly to highlight this difference based on our pilot study which tried different locations.

*2) Rate your personal interest in this material before you enrolled. {1 to 6 scale}*

The second question aims to partially assess student's motivation and prior interest in the course topics. One of the issues of substantial concern to faculty is that required courses may have a high proportion of students who 'don't want to be in this class' and that, consequently, FCQ ratings may be inappropriately influenced. We can more carefully address this question on our campus with the results of this question.

*3) Rate the instructor's effectiveness in encouraging interest in this subject. {1 to 6 scale}*

The AC was somewhat divided on the value and utility of this third question. Some members viewed stimulating and encouraging student interest in the subject at hand as a legitimate faculty responsibility while others argued that the faculty member should not be held accountable for creating interest in students who had no interest of their own. On balance, the former view was accepted and this question included.

4) *Rate the instructor's availability for course-related assistance such as e-mail, office hours, individual appointments and phone contacts, etc. {1 to 6 scale}*

This question asks for student opinions regarding availability and inclusive 'interactive connectedness' of the instructor to and from the students for, specifically, course-related assistance. This is a question where users of FCQ data should be especially cognizant of class size.

5) *Rate the intellectual challenge of this course {1 to 6 scale}*

This question asks for student opinions regarding the academic quality, depth and rigor of the course. It will be very interesting to compare the results of this question with those for question #1.

6) *Rate how much you have learned in this course. {1 to 6 scale}*

The sixth question, while related to the fifth, asks the student to self-evaluate how much they believe they have learned. Student's perceptions on this question can likely be usefully compared to grade distributions in the course and, of course, to questions one and five.

7) *Rate the course overall. {1 to 6 scale}*

The seventh question represents a 'global-style question' and – based on previous FCQ studies--correlates very highly with the eighth question although the numerical values tend to be slightly lower than on the eighth. The results of this question can be useful in curriculum planning as well as in faculty evaluation.

8) *Rate the instructor overall. {1 to 6 scale}*

The eighth question constitutes a key 'global-style question' and is the one most often singled out by evaluators, sometimes to the exclusion of the other items on the instrument. There is fairly substantial literature support for this practice, especially when results from multiple courses, multiple levels over multiple years are considered.

9) ***Rate this instructor's respect for and professional treatment of all students regardless of race, color, national origin, sex, age, disability, creed, religion, sexual orientation, or veteran status. {1 to 6 scale}***

The ninth question is the one question generated locally without support from the research literature on effective student assessment questions. It is also the one

with which the AC struggled mightily. Briefly, the spirited discussions centered on the disagreement about whether specific groups should be named in this question or not. One perspective argued that faculty should display respect for and professional treatment of all students, period. The counter-perspective argued that for very good reasons, certain groups should be specifically named in the question itself. As a compromise, the AC recommends that the first part of the sentence be made more prominent by being italicized and that the rest of the sentence specifically name the groups which have already been formally approved as part of the University non-discrimination policy by the Board of Regents. This is the one question which elicited some comments of disapproval from the respondents in the pilot trial.

We conclude our instrument with an open ended question inviting comments:

*Your comments, for the instructor ONLY. Please highlight or comment on any specific aspects or elements of this course, positive or negative, which you feel are important for the instructor to consider. Please note that you may also send separate comments directly to the school/college dean or the department chair. This form will be viewed only by your instructor.*

The last recommended question invites students to write to the instructor--and *only* to the instructor--about any part of the course they deem important. Because students have, in the past, so often written items clearly designed to be read by a faculty supervisor, we have included a note indicating that if they want to do that, they need to do it separately from the FCQ instrument itself.

A facsimile of the recommended form is in part 8 of this report.

Part 2:

**Chancellor's FCQ Advisory Committee's responses to BFA letter to the Chancellor, dated December 5, 2003.**

References in parentheses refer to the bibliography of this report.

**Formal Resolutions**

*(1) ... replace the current Faculty Course Questionnaire*

The Advisory Committee (AC) agrees with this resolution and has produced a substantially revised FCQ instrument. (10, 11)

*(2) ... to approve in principle the set of questions included in appendix A ...*

The AC recommends acceptance of several of the recommendations, but not all. The recommendations will be addressed individually below.

*(3) ... define policy of access to narrative questions in the FCQ ...*

The AC makes a specific and clear recommendation on this issue: the narrative portion of the FCQ should be addressed to the faculty member and only the faculty member should have automatic access to it. The faculty member may share those comments, if so desired. If Chairs or Deans or other campus offices need additional information beyond the statistical summary of numerical FCQ scores, it will be their responsibility to devise, administer and evaluate the means to gather that additional information. (10, 11)

The AC also recommends that campus academic officers encourage faculty to employ some means of asking for student feed-back early in each course (preferably by the 4<sup>th</sup> or 5<sup>th</sup> week of class) which should be aimed at assisting the course instructor in improving the quality of instruction. This material should go directly, and only, to the course instructor, unless he or she chooses to use it elsewhere. (10, 11, 13, 17)

*(4) ... outlines how summative information derived from the FCQ is used in the evaluation of faculty.*

The AC believes this is an extremely important issue and offers an extensive set of recommendations, mostly following the lead of nationally recognized scholars in this area. These are detailed below.

**BFA 'Findings'** (numbering follows the original BFA report)

*(1) The current form lacks credibility among the faculty.*

The issue of 'credibility' of student ratings instruments has generated a very large scholarly literature over the last four decades since they began to be used widely in higher education. There have been many studies with individually contradictory results and there are undoubtedly many faculty who take the view that student ratings can never be valid, useful or even appropriate. The AC has consistently adhered to the principle of basing our recommendations upon well-founded scholarship, taken primarily from the national literature.

That national literature argues, on balance, that well-crafted, well-validated questions on an FCQ type instrument are credible and provide valuable data to evaluators when used properly. The FCQ instrument the AC has generated contains only questions which have, indeed, been validated and are recommended by the scholars in this field. (4-7, 11, 12, 14, 18-22, 25-31, 34-36) except for question number nine as described elsewhere in this report.

*(2) The current form is susceptible to abuse.*

The AC understood this to refer faculty concern that students may use the opportunity to verbally abuse faculty. We agree that this type of abuse is not only possible, but occurs with distressing frequency. After extensive discussion of possible options designed to stop or reduce this abuse, the AC recommends that the narrative portion of the FCQ – where most abusive language occurs – be made available only to the faculty member and not to anyone else including department chairs, deans, promotion committees, etc., except at the option of the faculty member. The AC also recognizes that this only contains the abusive language; it will not stop it. The AC recommends that workshops for faculty provide experienced guidance and counsel, especially for new faculty, in how to deal with the stress such destructive language can cause. The committee felt that reducing the distribution of the comments would partially mitigate the embarrassment, anger and frustration for faculty receiving them. The AC committee notes that abusive comments to and about faculty cannot be stopped entirely for the FCQ is but one method of communication chosen by a few students. On balance, the narrative comments from students were deemed more valuable than the cost of abusive ones and so one narrative question has been recommended for the FCQ instrument. The AC also recognizes that a program of student education in the effective and appropriate use of FCQs by students could improve the value of the process.

*(3) Faculty who teach courses in subject matters that students might categorize as 'non-traditional' or 'out of the main stream' are especially susceptible to abusive comments.*

The AC does not make a recommendation for special procedures for individual courses dealing with certain topics because we did not find such recommendations in the scholarly literature. We have made a recommendation in the FCQ instrument that students be asked about their prior motivation for taking the course which may provide some useful insights in this area in the future. The FCQ results can also be useful in making within-discipline comparisons to other courses or other curriculum questions.

*(4) Faculty believe that the FCQ contributes to grade inflation.*

This is a very powerful and very widespread faculty perception, locally and nationally. Consequently, there have been several hundred scholarly papers published on this topic and, very likely, several thousand 'informal' local studies by concerned individual faculty (many on this campus). Further, the common faculty belief that it is 'you' not 'me' that has let grade inflation become a serious problem, highlights conflicting faculty perceptions on this issue. This might well be the issue of greatest concern to those who tend to view FCQs as inappropriate, mis-leading, not credible, or decidedly harmful. The relevant literature is large and, in general, indicates the phenomenon is smaller than most faculty believe. (2, 4, 5, 7, 11, 18, 22-24, 26-28, 36) On our campus, the rise in academic grades was shown to be 0.09 points over the last decade. The effects of the FCQ could only be a fraction of that total.

The AC endorses the recommendations of national scholars on this issue which are, in turn, based on a broad overview of the vast literature. In particular, we note that a moderately industrious individual can easily muster a few dozen carefully selected publications which support their own particular view, pro or con; we have eschewed this strategy, relying instead, on summary, overview papers. The general perception among faculty of a strong, positive association between course grades and FCQ ratings is not well supported by extensive statistical analyses. Cashin (10, 11) and others report that the statistical association between student grades and student ratings of the instructor generally accounts for 1 to 10% of the variance in student ratings. Cashin (10, 11) recommends that the issue of grading leniency be addressed by scholarly peers directly by reviewing course material, exams, essay samples, projects and other graded work and not be directly taken into consideration in the evaluation of FCQ data.

We have provided several specific recommendations: one set for individual faculty and a separate set for evaluators such as chairs, deans, personnel committees and VCAC.

### **Background and Findings of the CEDA workshop**

#### *1. At least 90% of course evaluations forms ... are not scientifically designed ...*

The AC agrees that our FCQ needed to be re-done and we have submitted our recommendations which reflect our understanding and reading of the scholarly literature. We have considered all the specifics itemized in this section of the BFA report and believe we have supported their recommendations except for two specifics. The issue of statistical reliability, at least as we read the literature, rests largely on the validity of the sampled population with respect to the purposes for which the survey data are collected. We have followed the strategy of a short, concise form with carefully focused questions aimed to elicit student opinion about appropriate elements of the faculty member's performance plus an open ended opportunity for the student to address any other issues they might choose to address. In order to maximize student participation, the AC recommends a short, non-redundant instrument. All the questions, save #9, are recommended and vetted by experts in this area of research.

The second point – switching the evaluation scales in direction within a single questionnaire – is specifically and strongly condemned by questionnaire experts. Although faculty would like for students to present very thoughtful, individually considered responses as described in the BFA report, the questionnaire experts point out that switching directionality between one question and the next almost guarantees that the results will be useless. Students will generally consider this opportunity to be solely designed to collect, not form, their opinions. They will have already decided their views over the course of the entire semester and simply want to record those views as quickly and efficiently as possible.

#### *(2) Course evaluations should have a clearly defined scale.*

The AC recommends a complete revision of the rating scale with a six point scale, numbered from one to six so the feared connection between course letter grades and the previous letter 'grade' FCQ rating system will be severed. Additionally, the pilot study indicates the new version may provide some improvement in the resolving power of the generated data, especially at the extremes. We have, for practical reasons, retained the labels 'lowest' for the rating of 1 and the label 'highest' for the rating of 6. Physical space limitations, as well as the difficulty and advisability of subjective characterizations for each step, prevent our supporting the BFA recommendation that every level be specifically defined. It is probably worthwhile to remember that this instrument only asks for student

opinion which, by its very nature, is subjective and, consequently, so to are the triggers which motivate a student to move from one rating level to the next higher or next lower one.

*(3) Course questionnaires should ask specific, narrow questions from which global assessments can be deduced.*

We believe we have recommended questions in the FCQ instrument which have been validated and recommended by national experts and other universities, with the single noted exception of question #9. We have addressed the evaluation part of the FCQ process in the recommendations to instructors and evaluators sections.

*(4) There are two documented ways in which course evaluation results are skewed: lower-level classes, regardless of class size, consistently produce lower ratings; and science and mathematics courses produce lower ratings than humanities and social science courses.*

The AC has endorsed the specific recommendations from the national literature that comparisons of FCQ ratings be made only within disciplines; this is, indeed, a measurable effect and should be accounted for by evaluators of the data, especially at the Vice Chancellor's Advisory Committee level. The simplest way to do this is to only use within-discipline comparisons. We also recommend that class level (upper vs lower division for undergraduates) be incorporated into comparisons wherever possible. This strategy also often ameliorates the small effects of class size which are often confounded with upper and lower division status. We also strongly endorse the recommendation that scores from multiple classes over multiple years be utilized by evaluators wherever possible. We address the topic of additional 'corrections' or biases (skew, in BFA's language) in accordance with the literature cited in the bibliography in our recommendations below.

### **Committee Deliberations**

*(1) Confidentiality of responses is not always observed ...*

This issue falls under administrative handling of the FCQ process. Our committee places high value on student anonymity so we recommend employing whatever devices are required to assure anonymity (recognizing that it will sometimes be used inappropriately by students).

*(2) Timing of the FCQs might contribute to hostile comments.*

One of the recommendations by Cashin (reference #10) is that the FCQs be administered specifically in the next to last week of the semester. We recommend

that the campus follow this suggestion assuming that the Office of Institutional Analysis can successfully switch to this new schedule.

*(3) Students are not fully aware of how FCQs are used in assessing teaching and allocating merit points.*

The AC certainly agrees with this and adds that some students who are fully aware choose not to participate; we see no ideal solution to this (i.e. how to 'impress upon them') the importance of their participation. In our recommendations to faculty, we do suggest some language faculty might use and we also emphasize how important it is to have a 'standardized administration' process. We also would like to encourage discussion of how the campus can better inform students of the importance of their opinions and encourage responsible, thoughtful participation.

*(4) On-line evaluations might prompt students to be more thoughtful in their comments.*

The campus has conducted pilot studies of this approach and has found no differences in the pattern of student responses online or by paper. The largest concern here, based on pilot results, is reduced response rates when an on-line process is used.

*(5) ... we should move away from the A-F scale ...*

The AC agrees and has recommended a basic 1 to 6 rating scale.

*(6) ... appropriate to propose an alternative form consistent with findings of the scholarly literature.*

The AC has wholeheartedly endorsed this strategy and the FCQ instrument we have recommended adheres strictly to this idea, with the single noted exception of question #9.

*(7) BFA question-of-the-month to poll faculty ...*

The AC committee members have carefully read all of the BFA report and took due cognizance of its contents; we have primarily emphasized the scholarly literature in making our recommendations.

Part 3:

**Committee Response to Recommendations from: IDEA Paper No. 22**

[http://www.idea.ksu.edu/papers/Idea\\_Paper\\_22.pdf](http://www.idea.ksu.edu/papers/Idea_Paper_22.pdf)

Center for FACULTY EVALUATION & DEVELOPMENT  
Division of Continuing Education Kansas State University

**Student Ratings of Teaching: Recommendations for Use**

William E. Cashin

Recommendation	Form	Collection	Reporting	Use	Other
1. Use <i>multiple sources</i> of data about a faculty member's teaching if you are serious about accurately evaluating or improving teaching.				Strongly recommend. Cashin suggests 20-50% of evaluation be FCQ based	Regent policy requires multiple means.
2. Do use student rating data as <i>one source of data</i> about effective teaching.	Revised	Every course	Every semester	Self improvement and Personnel decisions	Done
3. Discuss and decide upon the <i>purpose(s)</i> that the student rating data will be used for <i>before</i> any student rating form is chosen or any data are collected.				Stated on FCQ website: for instr improvement, admin decisions, & student course selection.	
4. To obtain <i>reliable</i> student rating data, collect data from <i>at least ten raters</i> if this is possible.		Default is no form printed for <2 students.	Number of raters always reported.	Averages, medians and dist shown in one report or another.	Recommend minimal use if fewer than 10 forms returned. Leave off web site if fewer than 10 in UG course, 5 in a grad course.
5. To obtain <i>representative</i> student rating data, collect data from <i>at least two-thirds of the class</i> .	Try to obtain from all students in every class	No sampling done, but Raters are self-selected and may not be representative. In-class averages responses from 67% of enrolled. Problem could be worse with online out-of-class collection	Response rate always shown. Remedies: publicity, reminders. Telecomm experimenting with failing grades for non-responders. Should we work to increase response rates?	Anecdotal reports indicate some reviewers (e.g. on promotion committees) may use response rate as a measure of teaching quality. Not a valid method.	

Recommendation	Form	Collection	Reporting	Use	Other
6. To <i>generalize</i> from student rating data to an instructor's overall teaching effectiveness, sample across <i>both courses and across time</i> .			Annual instructor reports and online reports include all courses, for several terms.	Will recommend users (personnel issues) use multiple courses and multiple years, always.	Faculty will receive cumulative reports with ratings distributions in future
7. For <i>improvement</i> , develop a student rating system that is <i>flexible</i> .	Optional questions pre-stored in data bank and available to each instr, to dept. etc.	<u>Issue:</u> Priorities in using optional slots. Current order: instructor, dept, college, campus.	Optional question results go to instructor plus anyone else who requested them.	Recommend Deans or Chairs who employ optional questions, make faculty fully aware of plans and specifics in advance.	
8. Provide <i>comparative data</i> , preferably for all the items.			Section reports compare to dept and college cluster (natural science, humanities, etc.) for same instructor groups (TTT, TA, other). Instructor annual reports compare to dept only. Drop campus-wide comparisons?	Recommend users in personnel decisions be thoroughly briefed on how to make legitimate comparisons and avoid misleading and inappropriate comparisons.	Note that FCQ office gets frequent requests for campus and college averages. Recommend that we do not provide because of potential for mis-use. Use clusters instead.
9. Discuss and decide what <i>controls for bias</i> will be included in your system.			Class level & size and subject are shown on all reports. Grades assigned by instructor are shown on instructor summaries. Recommend no centralized 'corrections'.	Provide specific briefings for users in personnel decisions on relationships, both real and perceived to be real by faculty.	.

Recommendation	Form	Collection	Reporting	Use	Other
10. Do <i>not</i> give undue weight to: the instructor's age, sex, teaching experience, personality, or research productivity; the student's age, sex, level (freshman, etc.), grade-point-average, or personality; or the class size or time of day when it was taught.		No controls for any. Optionals can be used to collect these data on students, but many jeopardize anonymity. <u>Recommend extreme caution</u> about questions that jeopardize anonymity. Faculty member responsible.		Thoroughly brief users on the usually negligible contributions of these extraneous variables.	Other topics here will include gender, age, student grades, interdisciplinary variance, faculty rank, instructor style, etc.
11. Take into consideration the <i>students' motivation level</i> when interpreting student rating data.	Prior interest assessed on the recommended new FCQ form.		Report within and between courses. Latter likely to be most informative.	Brief users on how to utilize and interpret.	
12. Decide how you will treat student ratings from <i>different course levels</i> , e.g., freshman, graduate, etc.			Course level is part of reporting: lower, upper and graduate.	Recommend users employ within course level comparisons wherever possible.	
13. Decide how you will treat student ratings from <i>different academic fields</i> .			See #8	Brief users about differences (e.g. VCAC).	
14. For <i>improvement</i> , develop a system that is <i>diagnostic</i> .	Base form bubble questions are not diagnostic but many of the optional questions can be.	Open ended question is diagnostic and aimed only to instructor as recipient.	Narrative, open ended question goes ONLY to instructor as its purpose is strictly formative. Deans/Chairs who want more should devise a custom process outside the FCQ process.		As a best practices technique, we should encourage faculty to employ teaching feedback questions early and often.

Recommendation	Form	Collection	Reporting	Use	Other
15. Develop a system that is <i>interpretable</i> . Report with text, not just numbers. Use faculty consultants to help with interpretation.		Recommend major effort in improving readability and interpretability of FCQ data.	The reporting changes will be done centrally by PBA.	At present FTEP is only formalized source of consultation. Recommend each department identify knowledgeable and helpful senior faculty member(s) to play this role.	
16. For evaluation, <i>use a few global or summary items or scores</i> .	Done ('course overall,' 'instr overall' & 'how much learned')	We have kept to a shorter, more global FCQ form in our recommendations.			
17. Use the short, evaluation form (or items) in every class every term.	Done	Done			
18. Use a <i>long, diagnostic form in only one course per term</i> – in the course that the instructor wishes to focus upon for improvement.	Instr/dept could use diagnostic optionals to assist.	Long, diagnostic forms would be instructor's choice and responsibility. Mid-term collection recommended.			
19. For improvement, use <i>items that require as little inference as possible</i> on the part of the student rater and as little interpretation as possible on the part of the instructor.	The FCQ form recommended has followed vetted and recommended forms and questions, only.				
20. For improvement, do <i>not use a single, standard set of items</i> for every class. <i>Provide a pool of items or some kind of weighting system</i> .	Optional questions available at instructor, chair, or dean request.				
21. Use a <i>5-point to 7-point scale</i> .	We have recommended a six point numerical scale.	Pilot results suggest some added discrimination ability.			

Recommendation	Form	Collection	Reporting	Use	Other
22. In the analysis of the results, report computations <i>only to the first decimal place</i> .			With change to 6-point scale, will definitely go to 1 decimal place.		
23. Do <i>not overinterpret the data</i> , allow for a margin of error.			STD and SEM on all section reports.	Will brief users on this danger.	
24. Use <i>frequency distributions</i> – what number or percent of the students rated the item “1” or “2,” etc.			On section report. Plan to upgrade to visual, graphical representation.		
25. For <i>improvement</i> , ask for <i>open-ended comments</i> as well as quantitative ratings.	Done with the single open ended question to be directed only to course instructor.				
26. Use the open-ended comments <i>only</i> for improvement.		Since we recommend open ended comments go only to instructor, they cannot be used for evaluation.	Chairs and deans, will need to design a custom process if they want to assess variation in student reactions beyond the numerical questions.		
27. For evaluation, <i>develop standardized procedures</i> covering all relevant aspects of your student rating system and <i>monitor that the procedures are followed</i> .		We post (on the web) and distribute administration instructions. Monitoring is weak.	Standardized procedures are important and we presently invest minimal resources in seeing this is done.	This is a weakness in our system; loosely monitored by dept. Chairs and student complaints.	
28. For evaluation, <i>administer the ratings about the second to the last week of the term</i> .		Administration week is standardized as last week of term before finals.		Recommend to faculty not to do on last day or immediately after an exam.	

Recommendation	Form	Collection	Reporting	Use	Other
29. Develop <i>standardized instructions</i> that include the purpose(s) for which the data will be used, and who will receive what information, and when.	Done				
30. Instruct the students <i>not to sign their ratings</i> .	Not explicitly stated, but implied, and very few do.				
31. The instructor may hand out the rating forms and read the standardized instructions, but the instructor should <i>leave the room until the students have completed the ratings and they are collected</i> .		Our procedure requires administrator (not instr) to distribute forms.			
32. The ratings should be collected by a <i>neutral party</i> and the data taken to a predetermined location – often to where they are to be scored – and they should <i>not</i> be available to the instructor until the grades are turned in.		Done			
33. Develop a <i>written explanation</i> of how the analyses of the student ratings are to be interpreted.				We will review and expand info on FCQ website, and provide recommendations and briefings to users of FCQ data.	

Recommendation	Form	Collection	Reporting	Use	Other
34. Appoint a faculty member to serve as <i>instructional consultant</i> to help faculty interpret their results and to improve their teaching.	See #15 above.			Desirable to improve our reporting on FCQ results for each department as a unit.	

Part 4:

## **Spring 2004 FCQ Pilot Form Results**

August, 2004

Perry Sailor and Lou McClelland, PBA

### **Background**

In May 2004 the Faculty Course Questionnaire Revision Committee proposed a revised FCQ instrument, based on (1) recommendations for such instruments in published, scholarly work, (2) question formulations well vetted and well used in higher education, (3) professional advice and recommendations for general structure and form, (4) vigorous discussion and argument among all faculty, student, and staff members of the FCQ committee, and (5) recommendations from a Boulder Faculty Assembly committee.

Two versions of the revised instrument were piloted in undergraduate spring 2004 courses of three volunteer instructors, in business, engineering, and political science. The versions differed only in order of the questions. The courses had 443, 47, and 28 enrolled students, respectively. Both versions were administered to random halves of students in all three courses in lieu of the usual FCQ instrument, in class at the usual FCQ administration time, with 375 total forms returned from the three courses. Both forms are shown at the end of this document as Displays 1 and 2.

Results for the three pilot courses have been returned to the instructors with the forms. Those sections are listed as "Piloted new FCQ form" on web postings and annual instructor summaries.

### **Summary of results**

The pilot was designed to collect information on

- Ratings on and reactions to the nine check-off items, particularly
  - Use of the new 1-6 (vs. A-F) rating scale – Did the scale
    - Pose problems for students? No.
    - Spread responses more than the A-F scale? Yes; the proportion of responses in the highest scale category is lower with the revised forms.
    - Produce higher, lower, or more or less the same means, expressed as percentage of maximum possible rating, as the current form? Means were more or less the same, with probable reduction in ceiling and floor effects of the current form.
  - Reactions to and missing data on a radical revision of the diversity item

- This single item produced half of the total missing responses, but even so was missing on only 3% of the forms.
- 7 students commented about the item, all negative. While this is only 2% of all respondents, it's far more comments than any other item received, and was unanimously negative. Most found the item inappropriate or irrelevant.
- Reactions to the absence of a "not applicable" option on any item
  - Missing responses were few. Aside from the diversity item (discussed below) only 11 of 3000 possible ratings, under 0.5%, were missing.
- Any effects of item revisions on rating patterns over items
  - None discernable. The only consistent pattern observed in these courses in prior terms was higher instructor than course ratings (the same pattern is found in most other courses as well). This pattern persisted in ratings from the pilot forms.
- Any effects on ratings, missing data, or student reactions of the position of the workload item, which was the first item on one version, the fourth item on the other
  - The workload question is long, wrapping to three lines. Student mismarkings indicate that any items which wrap to multiple lines should be formatted, numbered, and demarcated carefully so that the first line alone does not read as a complete question, and so that students are immediately aware that a single question is on multiple lines.
  - Putting the workload question, which uses a different scale than other items, in the middle clearly created problems for students. 13 mismarked or double-marked this item when in the #4 position, vs. only one mismark when in the #1 position.
  - Both the workload and the intellectual challenge items had statistically reliable differences in mean ratings on the two forms. Students estimated spending more hours per week on the course on the workload-middle form than the workload-first form; they also rated intellectual challenge - the item that followed workload - lower on the workload-middle form. The reasons for this are unknown, and the result should probably be treated with caution until replicated.
- Responses to and comments about the revised narrative item, which replaced four comment items on the current form
  - 60% did make some comment, a usual rate, and none commented about the change in questions.
- Response to a new instruction on the narrative item, "These comments will go ONLY to the instructor."
  - The instruction appears to have been ineffective -- almost 25% of comments appeared to be directed at a supervisor, about the same proportion as in three classes using the current form.

- One student commented that the form “needs a comment section *not* to be read by the instructor.”
- Any other student reactions recorded in a box inviting comments on the new form itself
  - Most of the 59% of students who made comments liked the new form, with positive comments outnumbering negative more than 3 to 1 (62% vs. 18%).

## Detailed results

### Comparison of two alternate versions of the new form

The new form came in two variants, with identical items and wording, and identical response scales. The only difference was in the ordering of items, in particular the position of the one item – concerning hours per week spent on course-related work, hereafter referred to as “workload” – that had a different response scale from the other eight items. While all other items were marked on a 1-6 scale, with the end points labeled “lowest” and “highest” and the intermediate points unlabeled, the workload item had each of the six points labeled with a range of hours, e.g., “0-4,” “6-8,” etc., up to “21+.”

One variant of the new form, buff in color, had the workload item, with its different scale, in the fourth position, while the other (white) had this item in the first position. From the point of view of the respondents, then, those with the buff form started out with one scale (the 1-6 scale labeled “lowest” to “highest”), switched to the “hours/week” scale for item 4, then back to the original scale for items 5-9. Respondents using the white form started out with the “hours/week” scale for item 1, then used the “lowest to highest” scale for the remaining eight items. We will refer to the form with the workload item first, and requiring users to make one change in scales, as the “workload-first” form, and the form that required users to change scales at item 4, then change back again, as the “workload-middle” form.

In all, 196 students were randomly given the workload-first form, 179 were given the workload-middle form. We compared the two forms on (1) mean responses to the items, (2) number of missing responses, and (3) number of obvious mismarkings in using the scales, i.e., circling the scale label rather than or in addition to properly blackening the bubble.

### Mean item responses

Only two items showed statistically significant mean response differences between the two forms: workload and intellectual challenge. Users of the workload-first form estimated lower hours/week than did users of the workload-middle form. The only other item with a statistically significant mean difference, intellectual challenge, immediately followed the workload item on the workload-

middle form. On that form, intellectual challenge was rated lower than on the workload-first form. To reiterate, workload was estimated lower when it appeared first than when it appeared in the middle, and intellectual challenge was rated lower when it immediately followed the workload estimate than when it didn't. (It followed instructor availability on the other form.)

Table 1. Mean item ratings (1-6 scale).

Item	Workload-middle	Workload-first	Effect Size
			(s.d. units)
Workload	<b>2.5</b>	<b>2.2</b>	<b>.25</b>
Interest before	2.9	3.0	-.07
Instructor encouraged interest	3.4	3.4	.00
Instructor availability	4.5	4.3	.13
Intellectual challenge	<b>4.4</b>	<b>4.7</b>	<b>-.25</b>
How much learned	3.7	3.8	-.07
Course overall	3.5	3.5	.00
Instructor overall	3.8	4.0	-.12
Instructor respected diversity	5.2	5.3	-.08
N	179	196	

**Bold: Difference statistically significant,  $p < .05$ .**

What does this mean, if anything? There's no way to know, one can only speculate. It's possible that answering three preceding items about the course somehow caused respondents to make a higher time estimate, and that making this time estimate somehow affected the rating of intellectual challenge. Or, it's possible that something about the scale changes, rather than item positions, caused the differences. It's also possible that the statistically significant differences in ratings are simply flukes, that there's no causal effect operating at all. In the absence of replication, it's probably best to treat these results with great caution, and not to consider them proof that the different ordering of items somehow caused the differences. It may be worth noting, however, that in one sense, we *do* have independent replication here -- the differences on each of these two items were in the same direction in each of the three classes who participated in the pilot study, although the differences within each class were not statistically significant (with one exception, intellectual challenge in the BCOR class). Differences were very small on the intellectual challenge item, but quite sizable on the workload item. *Something* appears to be going on that caused students to make higher workload estimates when the item is in the middle than when it is first. This might be worth further study if accuracy in estimating the absolute number of hours is desired.

### Missing responses

There were twice as many missing responses on the workload-first form as on the workload-middle form (15 to 7). However, the numbers were very small, were scattered across the items, and appear to have nothing to do with the scale differences. Half of the missing responses (11 of 22) were on the diversity item, which was in the same place – last -- on both forms. The workload item itself had only 1 missing response when it was in the middle and 3 when it was first, out of 375 total participants.

### Mismarkings

There was one type of mismarking that is worth noting, because it happened repeatedly: On the workload item, 14 respondents circled the scale label (“0-4,” “5-8”, etc.), and also bubbled in a response. In every case, the bubbled response was different from, and higher than, the circled scale label. Clearly, these 14 respondents were reading the item as two separate items. The item’s wording and format was:

“Estimate the average number of hours per week you have spent on this course for all course-related work including attending classes, labs, recitations, readings, reviewing notes, writing papers, etc.”

The line breaks were exactly as shown above, as follows:

Estimate the average number of hours <u>per week</u> you have spent on this course for	0-4	5-8	9-12	13-16	17-20	21+
1. all course-related work including attending classes, labs, recitations, readings, reviewing notes, writing papers, etc.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Note that the scale labels appeared opposite the end of the first line (the word “course”), while the response bubbles were set below the labels, opposite lines 2-3. What seems to have happened is that the 14 respondents marked one answer in response to the first line of the item, then a second answer in response to either lines 2-3 or to the total question. And since 13 of the 14 items mismarked in this way were on workload-middle forms, it seems clear that putting this item in the midst of the form and requiring the respondent to switch to a different item scale in midstream confused some respondents (about 4% in this study).

We recommend that (1) the workload-first form be adopted, and that formatting changes be made to make the intended use of the response scale clearer. These might include some combination of boxes, brackets, color, shading, different line breaks, or some other changes.

### Response distribution compared with the current FCQ form

Direct comparisons of means between the current and pilot forms are not meaningful, because the current form has five points on the evaluative scale, while the pilot form has six. But the distributions of responses can be compared. The following tables show response distributions for the pilot form and the

current form, for courses taught by the same instructor. For BCOR, the comparison is to the same course in spring '03; for PSCI, the best available comparison was two 5000-level courses in spring '03; for CHEN, the comparison is to two other sections of the same course in spring '04.

One issue that has long characterized FCQ ratings is a frequent “ceiling effect,” i.e., a tendency for the vast majority of ratings be at the top end of the scale. This makes it difficult to discriminate among degrees of excellence in ratings. It was hoped that using a 6-point scale instead of a 5-point scale would mitigate this effect, and it seems to have done so – the percentage of “top” ratings is lower on the pilot form than the current form in 5 of the 6 comparisons (2 items, overall course and instructor ratings, times 3 instructors). In most cases the differences are sizable.

In addition to the distribution, the table shows the percentage of maximum points earned (6 points for each 6, etc., on the new form; 5 points for each A, 4 for each B, etc., on the current form). If the ceiling effect often found on the current form is mitigated by having more response options, we would expect to see the percentage of maximum points earned to be lower on the pilot form than on the current form. This seems generally to be the case, although ratings on these particular courses aren't as close to the ceiling as those in many courses.

Presumably, using a 6-point rather than a 5-point scale would also mitigate a floor effect, although this is far less frequently seen on FCQs; ratings given to the CHEN and PSCI instructors do not show such an effect, but the BCOR ratings on the current form do (a higher percentage of F's than D's). On the pilot form, the floor effect was eliminated on the course rating and much reduced on the instructor rating.

BCOR	Percent of responses							
	6	5	4	3	2	1	blank	%
Course rating		A	B	C	D	F	blank	max
Pilot form (1-6) (n=308)	3%	18%	24%	22%	19%	14%	0%	53%
Current form (A-F) (n=213)		6%	23%	27%	19%	24%	0%	54%
Instructor rating								
Pilot form (1-6) (n=308)	10%	28%	18%	14%	15%	15%	1%	57%
Current form (A-F) (n=213)		10%	25%	25%	15%	23%	0%	60%

PSCI	Percent of responses							
	6	5	4	3	2	1	blank	%
Course rating		A	B	C	D	F	blank	max
Pilot form (1-6) (n=25)	20%	36%	20%	16%	8%	0%	0%	74%
Current form (A-F) (n=25)		72%	16%	8%	0%	0%	4%	93%
Instructor rating								
Pilot form (1-6) (n=25)	44%	24%	12%	20%	0%	0%	0%	82%
Current form (A-F) (n=25)		96%	4%	0%	0%	0%	0%	99%

CHEN	Percent of responses							
	6	5	4	3	2	1	blank	%
Course rating		A	B	C	D	F	blank	max
Pilot form (1-6) (n=42)	19%	57%	10%	10%	5%	0%	0%	79%
Current form (A-F) (n=35)		43%	46%	6%	3%	0%	3%	86%
Instructor rating								
Pilot form (1-6) (n=42)	67%	29%	5%	0%	0%	0%	0%	94%
Current form (A-F) (n=35)		74%	17%	6%	0%	0%	3%	94%

#### Open-ended comments

In addition to the usual box for comments on the course/instructor, respondents were also invited to comment on the new form – what they liked, disliked, found difficult or confusing, etc. We received comments from 223 of the 375 respondents (59%). These were coded into the following categories:

Category	N	%
Better than previous form	66	30%
Positive comment – no mention of previous form	72	32%
Negative comment – no mention of previous form	27	12%
Worse than previous form	13	6%
Neutral comment	28	13%
Mixed positive and negative comments	11	5%
Not responsive (i.e., comment not about form)	6	3%
Total	223	100%*

\*Individual percentages do not sum to 100 due to rounding.

In sum, 62% of comments were positive, 18% negative, 18% neutral or mixed.

Thirty-nine respondents (10%) also made some specific comment about one or more items, or changes they would suggest. Most frequent among these were the need for a “not applicable” category (8 respondents), and a dislike for the

diversity item (7 respondents). Three who wanted a N/A option specified the diversity item in particular as one needing such an option. Five respondents expressed confusion about the change in rating scales; all five used the workload-middle scale. (Eight students said they did not like rectangles to bubble in, apparently preferring circles or ovals, but the rectangles were just for the pilot form, which was designed on Excel and wasn't intended to be optically scanned. As far as we know, all optically scannable forms have circles or ovals, so this issue is moot.)

Here is a summary of the open responses on the form:

<u>Comment (paraphrase)</u>	<u>N</u>
Confusion about rating scale (workload item)	5
Don't like diversity item (not including needs N/A) . <i>See.</i>	4
Needs N/A response option on diversity item, with associated comment. <i>See.</i>	3
Needs N/A response option, not specified for diversity item	3
Need more questions about instructor	3
Don't like lowest/highest labels, or don't like scale	2
Needs more questions	2
Needs "fairness" question	2
Need middle response option	2
Needs <i>loss</i> of interest question	1
Workload item should have numbers, not ranges	1
Needs question on work in relation to tests, grades	1
Needs question about instructor bias (seemingly meaning political bias)	1
Needs comment section <i>not</i> to be read by instructor	1
Needs question about favorite, least favorite aspects	1
Item on interest before enrolled not useful	1
[Following are unclear or difficult to paraphrase, so are given verbatim]	
"Not enough input"	1
"It can still be modified. Ex: if there were a shorter class, but still the same amount of hours/week, would it benefit even more"	1
"No prof. name"	1

Here are the verbatim comments from the seven students who had negative comments about the diversity item:

- Should have a N/A check on #9, statistics has nothing to do with gender or race.
- Need "N/A" on question #9 and what type of question is this anyways.
- I don't like the question 9. It seems odd as a mandatory question.

- Don't like question 9, irrelevant to class - at least should get the option of NA.
- Question 9 is too wordy.
- Take number 9 off. Race, gender, and ethnic background have nothing to do with learning and that question simply perpetuates segregation and disparity between groups.
- Please remove the question on discrimination. It's pointless and if there's a problem people can write in comments.

#### Audience for questions

Student comments were solicited by the following statement:

Please highlight or comment on any specific aspects or elements of this course, positive or negative, which you feel are important for the instructor to consider. *These comments will go ONLY to the instructor. Please note that you may also send comments directly to the school/college dean or the department chair.* [Italics and capitalization in the original.]

About 60% of students made comments; 59% of BCOR students, 67% of CHEN, and 68% of PSCI. Comments were coded as either directed at the supervisor, at the instructor, or neutral/could not tell. The subject of the comment, especially pronouns, was usually the determinant; comments were considered to be directed to the supervisor if they used such terms as "he," "she," or "the instructor," and to the instructor if they used "you" or had a direct suggestion or request such as "please talk slower."

Despite the instructions, not all comments were directed to the instructor. In fact, 23% were clearly directed to the supervisor. The remaining 77% were either clearly directed to the instructor (9%), or could not be coded with respect to whom the response was directed (68%). These percentages were similar for each course. Whether students did not read the instructions, did not understand them, or ignored them, it seems clear that large numbers of students - nearly 1 in 4 of those who made comments -- want to use the FCQ to direct comments to the instructor's supervisor. Three courses using the current form were unsystematically selected and also coded in the same way; in those courses, 16% of comments were directed at the instructor, 28% at the supervisor, and 57% could not be coded. From this small sample we conclude that the instructions made no difference in how students directed their comments.

Part 5:

**Recommendations for users of FCQ data** (e.g. chairs, department or dean's personnel committees, VCAC, etc. as specifically related to evaluation of faculty (Arabic numerals refer to bibliographic references). FCQ data can also be useful for managers and primary units in evaluating courses over and above evaluating instructors.

I.) FCQ results must not stand alone as the sole or even overwhelmingly dominant information about a faculty member's teaching performance. Multiple measures, which include FCQ data as but one part, are essential. The national literature suggests that FCQ data should probably account for between 20 and 50% of the total teaching evaluation score. (1-7, 10, 11, 12, 13, 17, 33)

II.) FCQ results constitute data only; they are not evaluations until someone places an interpretive evaluation on those data. (10, 11, 12, 13, 17, 21)

III.) Interpretation of FCQ results should be done only in accordance with well-established principles and best practices; individual strategies based on personal experiences by reviewers are generally unsupported by research literature and are likely to do more harm than good. We have heard anecdotal stories of individual reviewers including quite unfounded individual perceptions such as these three erroneous misperceptions: (a) a low 'return' rate of FCQ forms constitutes good evidence of poor instructional quality, (b) FCQ scores are strongly related to student's grades, (c) FCQs reflect important biases according to gender, faculty rank, age, hour of the class, faculty personality, etc. (4, 10, 11, 14, 15, 16, 18-20, 22-24, 25-31, 34-36). There are a few important outside factors which research says do affect FCQ scores and we address them below in section VI.

IV.) Use FCQ data only from classes with at least 10 FCQ scores for undergraduate classes and at least 5 from graduate classes. Better statistical evidence can be found where 2/3rds or more of the class respond. (10, 11)

V.) At every evaluation (even annual ones), employ FCQ data from multiple courses over multiple years. This strategy provides the best mechanism upon which to base generalizations about instructor effectiveness. (10-13, 17, 33)

VI.) There are a few (fewer than most faculty believe) well-documented external sources of bias in FCQ scores and users should be very aware of them. Our first recommendation here is to only compare FCQ data within closely related disciplines and subject matter. The research literature clearly demonstrates that using FCQ-type scores to compare, say, faculty within the Humanities to faculty within Engineering is quite inappropriate. FCQ comparisons within Natural

Sciences only and within the Humanities only constitute good practice; comparing faculty between such areas does not. This is an issue of special concern to campus-wide decision makers such as the Vice Chancellor's Advisory Committee.

Our second recommendation to chairs, deans and other personnel committees is to also restrict comparisons to results within, not across, class level (in particular, upper division, lower division and graduate) as best one can, within the primary unit. This recommendation is generally confounded with class size which does have a modest, but significant, biasing effect on FCQ scores, especially at the extremes (several hundred in class vs 10-20). These biases are best 'adjusted' for at the primary unit level. (4, 10, 11, 14, 15, 16, 18-20, 22-24, 25-31, 34-36)

VII.) Don't 'over-interpret' small numerical differences in FCQ statistics. The summaries are statistical 'estimates' and always have significant error associated. The literature suggests that such data can be extremely useful in sorting results into, say, four or five categories, but no more. These categories might, for example, be qualitatively labeled as poor, fair, good, very good, and outstanding. (10, 11)

VIII.) For evaluation, use only a few global scores. In the newly recommended form, those would be primarily 'Rate the instructor overall' and 'Rate the course overall'; these two measures tend to be very highly correlated but can be used for different purposes. (10-12)

IX.) Develop a set of written explanations of how FCQ data will be interpreted at each level (department, school or college, campus level) and make sure faculty and administrators are routinely and fully informed about those explanations. Then, when doing evaluations of faculty, stick to those interpretations. (10-12, 17, 33)

Part 6:

**The FCQ Process – Recommendations to individual faculty from the Chancellor’s FCQ Advisory Committee.** Arabic numerals refer to the bibliographic references in this report.

I.) Communicate, by words and actions, that faculty FCQs play a major role in how teaching is evaluated on the Boulder campus. By words, it would be quite appropriate to make a general comment to the effect that student opinions constitute a very important part of your evaluations as an instructor and the evaluation of your courses. However, faculty must not make individually focused comments to students such as ‘My tenure, or my promotion, or my annual raise, or my re-hiring, depends on you giving me high FCQ ratings.’ By actions, be prepared to allow students adequate time to fill out the FCQ, including the open ended questions which are to be directed exclusively to you as the instructor. (9,10, 11, 15, 16, 17) Be properly prepared logistically with an appropriate number of pencils, appropriate FCQ administrators who understand the procedure, and a mechanism for delivery of the filled-out forms properly. (10)

II.) Carefully avoid being anywhere near the classroom during the time the students are filling out or turning in their FCQ forms. Students take very seriously any perceived risk to their anonymity. (10, 13)

III.) Do not denigrate the FCQs, or the FCQ process by stating, for example, ‘I never pay any attention to these things.’ (10, 13, 17)

IV.) Ignore isolated, rude or inappropriate comments from a few individual students such as ‘You need breast implants’ or ‘The way you dress is ridiculous’ or comments filled with profanities.

V.) Pay attention to general patterns in the FCQ results such as, for example, a bimodal distribution on ‘how much you have learned’. This would suggest that the methods of the class were successful for some but not successful for other members of the class and your tactics might need reconsideration. (25, 27, 29, 32, 33) We strongly recommend supplementing the FCQs with a custom-designed class survey of students’ perceptions by the 4<sup>th</sup> or 5<sup>th</sup> week of classes through which the instructor specifically seeks student feedback on how to improve teaching effectiveness. Such information should go only to the instructor unless he or she wishes to share it with others.

VIII.) Do not treat FCQs as but one segment of your teaching portfolio and be sure there are other measures and information available to your Chairs, Deans, etc. (10, 13, 27, 28, 33)

Part 7:

## Bibliography

This bibliography aims to sample broadly from the enormous literature on student ratings. Many will contain additional references to specific investigations and results. We recommend that both faculty and evaluators employing FCQ results familiarize themselves with some of this literature. We particularly recommend #7, #10, #11, #21, #25, and #35 for individual faculty; in addition to these, we recommend #4, #6, #31, and #36 for evaluators.

- 1) Abrami, P.C. 1989. How should we use student ratings to evaluate teaching? *Research in Higher Education* 30:221-227.
- 2) Abrami, P.C., d'Appollonia, S. and P.A. Cohen. 1990. The validity of student ratings of instruction: What we know and what we don't. *Journal of Educational Psychology* 82:219-231.
- 3) Aleamoni, L.M. 1981. Student ratings of instruction. In: J. Millman (ed.). *Handbook of Teacher Evaluation*. Sage, Beverly Hills, Ca.
- 4) Aleamoni, L.M. 1998. *Student rating myths versus research facts from 1924 to 1998* *Journal of Personnel Evaluation in Education*, 13(2), 153-166.
- 5) Anonymous. 2004. Advocate Online. *Ratings Myths and Research Evidence*. National Education Association.  
URL: <http://www.nea.org/he/advo99/advo0199/feature.html>
- 6) Bain. K. R. 2004. *What the Best College Teachers Do*. Harvard University Press. Cambridge, Ma.
- 7) Bain, K. R. 2004. *Evaluation of Teaching: Using Student Ratings. Student ratings and the evaluation of teaching; Student ratings for formative (self-improvement) purposes; Limitations of student ratings of teaching*. New York University, Center for Teaching Excellence.  
URL: <http://www.nyu.edu/cte/white.html>
- 8) Cashin, W.E., Norma, A., Hanna, G.S. 1987. *Comparative data by academic field*. IDEA Technical Report No. 6. Center for Faculty Evaluation and Development. Kansas State University. (<http://www.idea.ksu.edu/>)

- 9) Cashin, W. E. 1988. *Student ratings of teaching: A summary of the research*. IDEA paper No. 20. Center for Faculty Evaluation and Development. Kansas State University.  
(<http://www.idea.ksu.edu/>)
- 10) Cashin, W. E. 1990. *Student ratings of teaching: recommendations for use*. IDEA paper No. 22. Center for Faculty Evaluation and Development. Kansas State University.  
(<http://www.idea.ksu.edu/> )
- 11) Cashin, William E. 1995. *Student ratings of teaching: The research revisited*. IDEA paper No. 32. Center for Faculty Evaluation and Development. Kansas State University.  
(<http://www.idea.ksu.edu/> )
- 12) Cashin, W.E. and Downey, R.G. 1992. *Using global student ratings for summative evaluation*. *Journal of Educational Psychology* 84:563-572.  
(<http://www.idea.ksu.edu/> )
- 13) Chism, N.V. (1999). *Peer review of teaching*. Anker Publishing, Bolton, MA
- 14) Cohen, P.A. 1981. *Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies*. *Review of Educational Research* 51: 281-309.
- 15) d'Appollonia, S. and P.C. Abrami. 1997(a). *Scaling the Ivory Tower Part II: Student ratings of instruction in North America*. *Psychology Teaching Review* 6: 60-76.
- 16) d'Appollonia, S. and P.C. Abrami. 1997(b). *Navigating student ratings of instruction*. *American Psychologist* 52: 1198-1208.
- 17) Doyle, K.O. 1983. *Evaluating teaching*. D.C. Heath, Lexington, Ma.
- 18) Feldman, K.A. 1979. *The significance of circumstances for college students' ratings of their teachers and courses*. *Research in Higher Education* 10:149-172.
- 19) Feldman, K.A. 1984. *Class size and college students' evaluations of teachers and courses: A closer look*. *Research in Higher Education* 21: 45-116.
- 20) Feldman, K.A. 1993. *College students' views of male and female college teachers: Part II – Evidence from students' evaluations of their classroom teachers*. *Research in Higher Education* 34:151-211.

- 21) Feldman, K.A. 1997. *Identifying exemplary teachers and teaching: evidence from student ratings*. In: *Effective Teaching in Higher Education: Research and Practice*, R.P. Perry and J.C. Smart (Eds.) Agathon Press, New York, NY
- 22) Greenwald, A. G. 1997. *Validity concerns and usefulness of student ratings of instruction*. *American Psychologist* 52: 1182-1186.
- 23) GreenWald, A.G. and G.M. Gillmore. 1997. *Grading leniency is a removable contaminant of student ratings*. *American Psychologist* 52: 1209-1217.
- 24) Howard, G.S. and S.E. Maxwell. 1982. *Do grades contaminate student evaluations of instruction?* *Research in Higher Education* 16:175-188.
- 25) Kulik, J.A. 2001. *Student ratings: validity, utility, and controversy*. pp 9-26 In: *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* New Directions for Institutional Research, No. 109. Jossey-Bass.
- 26) Marsh, H.W. 1980. *The influence of student, course, and instructor characteristics in the evaluations of university teaching*. *American Educational Research Journal* 17: 219-237.
- 27) Marsh, H.W. 1984. *Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility*. *Journal of Educational Psychology* 76: 707-754.
- 28) Marsh, H.W. 1992. *Students' evaluations of university teaching: A multidimensional perspective*. In: *Higher Education: Handbook of theory and research*. J.C. Smart (ed). Vol. 8, pp 143-233. Agathon, New York, N.Y.
- 29) Marsh, H.W. and J.E. Ware. 1982. *Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretation of the Dr. Fox effect*. *Journal of Educational Psychology* 74: 126-134.
- 30) Murray, H.G., Rushton, J.P. and S.V. Paunonen. 1990. *Teacher personality traits and student instructional ratings in six types of university courses*. *Journal of Educational Psychology* 82:250-261.
- 31) Ory, J.C., Braskamp, L.A. and D.M. Pieper. 1980. *Congruency of student evaluative information collected by three methods*. *Journal of Educational Psychology* 72:181-185.
- 32) Shepard, L. 1993. *Evaluating test validity*. *Review of Educational Research* 19: 405-450.

- 33) Theall, M. 2004. *Faculty evaluation*. The National Teaching and Learning Forum.  
<http://www.ntlf.com/pod/>
- 34) Theall, M., Scannell, N. & Franklin, J. (2000, Spring). *The eye of the beholder: Individual opinion and controversy about student ratings*. Instructional Evaluation and Faculty Development.  
[http://www.umanitoba.ca/academic\\_support/uts/sigfted/iefdi/spring00/matrx.htm](http://www.umanitoba.ca/academic_support/uts/sigfted/iefdi/spring00/matrx.htm)
- 35) Theall, M., Abrami, P.C. and L.A. Mets (eds) 2001. *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* New Directions for Institutional Research, No. 109. Jossey-Bass.
- 36) Theall, M. and J.L. Franklin 2001. *Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction*. pp 45-58 In: *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* New Directions for Institutional Research, No. 109. Jossey-Bass.

Part 8:

**Facsimile of recommended revised form.** Final form will be scannable and will have room for optional questions on the reverse.

About the FCQ

FCQ results are used by students in selecting courses, by instructors to improve their teaching, and by administrators for tenure and promotion decisions. See results at [www.colorado.edu/pba/fcq](http://www.colorado.edu/pba/fcq).

Neither instructors nor TAs for the course may be present at any time during administration of this form.

These forms will not be returned to your instructor until AFTER grades have been posted.

Note: If multiple forms are determined to be from one person, they will be removed.

1.	Estimate the average number of hours <u>per week</u> you have spent on this course for <u>all</u> course-related work including attending classes, labs, recitations, readings, reviewing notes, writing papers, etc.	0-4	5-8	9-12	13-16	17-20	21+
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Lowest			Highest		
		1	2	3	4	5	6
2.	Rate your personal interest in this material before you enrolled.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	Rate the instructor's effectiveness in encouraging interest in this subject.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4.	Rate the instructor's availability for course-related assistance such as e-mail, office hours, individual appointments and phone contacts, etc.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	Rate the intellectual challenge of this course.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6.	Rate how much you have learned in this course.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.	Rate the course overall.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8.	Rate the instructor overall.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9.	<i>Rate this instructor's respect for and professional treatment of all students</i> regardless of race, color, national origin, sex, age, disability, creed, religion, sexual orientation, or veteran status.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Your comments, for the instructor ONLY. Please highlight or comment on any specific aspects or elements of this course, positive or negative, which you feel are important for the instructor to consider. *Please note that you may also send separate comments directly to the school/college dean or the department chair.* This form will be viewed only by your instructor.