A0784

THE DESIGN AND USE OF FORMS FOR THE STUDENT EVALUATION OF TEACHING

Michael Scriven
University of Western Australia[1]

INTRODUCTION Questionnaire design is not an exact science, particularly in the area of teaching evaluation, and the form we use here at the Centre for Tertiary Education Studies, like any other, involves many compromises. The latest version of the CTES ("see-tease") form is appended. We greatly appreciate criticisms and suggestions for improvement and have regularly incorporated them during the form's fifteen years of development[1]. However, a good many of the suggestions we now receive are difficult to accommodate because of the pressure of other design considerations with which the critic may not have been familiar. Hence it seems useful to set out some notes on the underlying design problems that lie behind constructing this or any such form, especially for those departments or individual staff who are thinking of designing their own forms, a desirable exercise.

CONTEXT OF USE Student rating forms can only tell us about certain aspects of teaching. For example, short of extreme cases, students cannot reliably tell us about the instructor's mastery of subject matter, or about the instructor's work on curriculum committees, which sometimes make up an important part of his or her contribution to teaching. However, students are in an excellent position, which also happens to be the best position, to judge a number of matters, ranging from simple observation (e.g., attendance/punctuality, whether the books supposedly on reserve were available, and whether the overheads, if used, were legible), to the vital matter of whether topics were explained so that the students could understand them. For a full evaluation of teaching, though, we have to add peer judgements of content, the Head of Department's (or other relevant experts') judgement of load, subject matter expertise, and of out-of-class contributions to teaching, and usually some independent documentation (e.g. teaching materials that have been published and reviewed).

USES This form is designed as a *general-purpose, first-stage* form for the evaluation of a number of matters related to teaching. (It does not address general questions about the degree program in which the particular course is embedded; that's a matter for program evaluation.) It is *general purpose* because it asks about some matters that are of interest for several purposes. In the teacher-evaluation area, these include personnel decisions (so-called summative evaluation—e.g. the overall rating), self-improvement or other-aided development (sometimes referred to as formative eval-

[1] Many thanks to Sally Clarke for a great deal of help with the form revisions and for comments on an earlier draft; to the Student Guild of the University of Western Australia for their massive funding of a multi-form mailing to all students, their general enthusiasm for the teaching awards project, and for many useful suggestions; to John Hattie for comments and implementation trials; and to Vice-Chancellor Robert Smith and Deputy Vice-Chancellor Doug Clyde for support of the Centre and the forms development effort.

uation—e.g. student comments on the readability of overheads) and other legitimate consumer concerns (here called descriptive evaluation—e.g. comparative workload). There's a question on the adequacy of facilities, which sometimes generates results of interest to the Buildings and Grounds Committee. In the content-evaluation area, relevant to both unit and course evaluation, there are indicators of perceived value and adequacy (questions 1 and 2); these are of interest to the department or dean as well as to the instructor, and often have little to do with the instructor's range of choices. Of course, the problem may simply be that the instructor is not doing a good job of *explaining* why the course is valuable, or why certain matters were omitted; in either case, something needs to be done.

There are, as mentioned, various compromises, due to the brevity of the form and other considerations. For example, some students like to know about aspects of teaching style (is most of the class time spent on discussion?; does the instructor exhibit enthusiasm for the subject?). However, selecting a few dimensions of style from the scores that are of interest to some students, and/or have some research support (which suggests they are particularly successful styles in some situations), would incorrectly suggest greater importance for the aspects listed. Since style data also has the problem that it contaminates personnel evaluation[2], we omit it.

The other reason for calling this a general-purpose form is that it works quite well—that is, it does something useful—*with any type of class*. Of course, there are further questions that apply to special teaching situations which it does not ask (e.g., clinical teaching, where the question "How does the instructor treat patients?" is pertinent), but these can be brought up in a second-stage form (see below). This form is not intended to ask all questions relevant to all classes, but it does get a scan on more than one hundred common issues where improvement might be needed or excellence demonstrated, plus an overall rating on teaching effectiveness. The latter rating is highly—though probably not completely—independent of special circumstances (lab, class size, subject matter, year, etc.). Although by no means fully comprehensive, this form is good enough *in most cases* to obviate the need for a form that is specific to a particular department or faculty, with the consequent loss of comparability.

In calling this a 'first-stage form' we mean to allow for the legitimacy of a more focussed form at a later stage. Second-stage forms—they are usually subject- or approach-specific, or even fault-specific (aimed to uncover reasons for bad ratings on particular items, e.g. comprehensibility of explanations) can be given later, if necessary, once an attempt has been made to address the points that come up from analyzing the results from the first stage form. There is rarely any virtue in going to a second-stage form until the first-stage results have not only been analyzed, a systematic attempt made to improve, and a re-run of the first-stage form done in order to find out what progress has been made. In fact, it often takes two tries to get to the point where a new source of answers is needed ("They say the lectures are hard to understand, but is it that they are ill-prepared, or are my explanations less good than they could be?"). Even then, going to a subject-specific form does not offer much

[2] That is, judges often have style preferences and hence will be affected by information about style although it is illicit to use *any* information about teaching style in personnel evaluation, even if it is known that certain styles tend to be more effective. (This point occurs in "Validity in Personnel Evaluation", *Journal of Personnel Evaluation in Education*, vol. 1, No. 1, 1987, Kluwer Academic Publishers, Boston, pp. 9-23 (copies from CTES on request)).

chance of solving the problem; having a high-rating colleague give a couple of lectures while you watch, after which the students rate him, may be the best move. Or perhaps it may make sense to give a diagnostic pre-test to find out if the students really do have the knowledge they are supposed to have. However, the dedicated teacher will eventually want to use a second-stage form, so as to find out how well they are doing with the matters that are specific to the subject or teaching mode.

When they are used, second-stage forms should be attached to the back of a first-stage form, so as to keep a finger on the pulse of the overall ratings you have previously measured. After all, the standard form will normally be the one on which the improvement has to show up eventually, since it's a little hard for a promotions committee to tell how good you are from a form on which they have no norms.

There are thus several reasons for preferring a two-stage rather than a one-stage process. First, doing so makes it possible to keep the basic form short, which improves return rates and reduces class impact, the problems of stereotyped responses, etc. Second, as mentioned, a first-stage form can be standardized across the campus (or across multiple campuses). Departments often make the mistake of thinking that getting their particular interests included in the form they normally use should be their first priority. However, this approach can do a serious disservice to those of their staff who are going up for promotion since the interdepartmental review committee will then be faced with data on which there are no interdepartmental norms. A department's special interests—lab classes, clinics, team teaching, etc.—can easily be addressed by attaching a second page to the standard form, as long as one remembers not to let the results on that distract you from major issues revealed on the basic form. It is often worth trying to get some agreement with other departments in the same discipline, at comparable colleges, about the supplementary form, so that some further baseline data is available.

The third reason for using a two-stage process is that it avoids the suggestion that everyone needs to improve their score on all scales, a suggestion that is often implicit in the usual 'rate everyone on 20 (or 50) dimensions' approach[3]. By contrast, we encourage the 'if it's working, don't fix it' approach, which begins with the view that someone with a good score on the overall rating (B+ or better) should not lose any sleep, even if anxious for a promotion. B+ is an 8 and our best estimate of the mean score on this form is 6.75. If they are putting in a good deal of time on research and/or service of various kinds, they can continue to do treat those as the areas

[3] When a staff member sees a print-out that shows they are well down from the top on, say, scales 12, 15, and 23, (while appearing to do quite well on the others), they tend to think that these are the scales on which they should work in order to improve. This is often completely wrong. These are typically scales where the medians are low (perhaps through some artifact of the wording of the question) or where improvement is extremely difficult, or where improvement does not lead to significant gains in overall rating. Because we do not have reliable norms on such scales, it is particularly inappropriate to use a form incorporating them. The fallacy is compounded by using the common 'Likert scale' which spans the "Strongly agree" to "Strongly disagree" spectrum. One has absolutely no idea whether a score of 7 on a Likert scale alongside the item "The explanations given in class are usually clear" is good news or bad news; nor does the median resolve all doubts. But if 30% of the class identifies your explanations as "particularly in need of improvement", you receive a much clearer message. Of course, you still have to put it together with your knowledge about how many of the class lacked the appropriate pre-requisites. These responses on Question 5 must never be taken as absolute indicators of inadequate performance. They are intended to be helpful in suggesting possible directions for improvement.

where they are going to gain the points for a promotion. Or they can put some of that time into teaching (our teaching award winners score 10 on the 11 point scale), because there is still plenty of headroom for them to explore.

Second, we restrict the teaching-specific responses to matters that are identified by the customer as *particularly* needing improvement (or as *particularly commendable*). Hence, there are quite often very few responses, whereas they spill out all over the page when there's a problem. This approach is particularly appropriate in a situation where we know essentially nothing about interaction effects[4]. Since marks on the items in question 5 are just suggestions for improvement—or commendations—that seemed to the student to be 'especially worth mentioning', they cannot be interpreted as scales on which everyone should get a high score. If everything is commendable—and in fact commended in the overall rating—then nothing may be seen as *especially* commendable. Hence students will sometimes give a very good overall rating with hardly any indication of 'especially commendable' sub-items; the overall rating pre-empts the need to spell out the details.

The fourth reason for restricting efforts to a simpler form rather than one which would try to cover all interests of all parties is to keep down the CTES staff load. There are other important matters that deserve staff attention. While we can make a few suggestions about second-stage forms that are submitted for comment, we haven't the resources to design them or process them.

DESIGN SPECIFICS

THE FIRST FOUR QUESTIONS These questions have two roles. First, they serve the face-valid role of gathering information that is of value to various audiences, including fellow-students, department members, and deans. Second, they serve as 'deflectors'—opportunities for the student to comment on matters that sometimes threaten to overwhelm their ratings of the teaching. These matters have to be attended to quickly so that those feelings will not contaminate comments on later matters.

The first question asks whether the course content seemed valuable, in terms of whatever values seem important to the student. We avoid the term 'relevant' but it is worth recalling that massive complaints about relevance were one of the issues that fuelled the student revolts of the sixties. This question brings out the worst in faculty members, many of whom are furious at the very suggestion that students should be asked whether they can see any value in a course. A poor score on this scale[5] obviously doesn't prove that the subject *is* worthless, but it may suggest that the teacher might spend a little more time on explaining the value of the unit, as s/he sees it. Once in a while it may flag the instructor or the department about a real problem over a low level of significant content sometimes because the entry level of knowledge has crept up. Some respondents ask "Valuable for what?"; or "Do you

[4] That is, the extent to which efforts to improve on one scale lead to deterioration on others. This is serious with respect to style factors, less so with respect to clearly discrete components of teaching, which we try to emphasize. For example, it is not likely that efforts to improve the legibility of blackboard writing will undermine one's good rating on helpful hand-outs. But increases in the time spent o previewing/reviewing, a common style preference,cut into the time available for discussion or lecturing and may therefore result in a net loss of learning.

[5] 'Low' doesn't mean 'near the bottom of the scale'. It's a pretty serious complaint that one-third of the material lacked apparent value, and that corresponds to the third rating from the top of the scale.

mean 'valuable in preparing for the exams'?". If students lack any over-riding set of values that drive them in evaluating the content of the course (vocational, social, ethical, etc.), they are not the group about whom we are primarily concerned at this point. Nevertheless, this kind of write-in response does suggest that the issue of the value of the course has not been addressed.

The second matter—addressed by the second question—addresses the converse concern that, whether or not everything covered appeared to the students to be of value, too much was *not* covered (of what they see as their needs and/or expectations). Again, the responses are often a matter of considerable interest to many departments as well as to many staff. If there is a list of missing items attached here by the students *and* the load is rated as light in Question 3, then there is an argument for enriching the content. If there's a good deal missing that students feel they need, but the load is already rated as heavy, there may be a need for another unit, perhaps even a required unit, in this general area. An extreme case that deserves attention is the case of not covering what was (seen as) promised.

The third matter—and question—relates to work-load. Students in some classes feel desperate about overload and will use any question to express that concern by criticizing whatever it is they have the chance to criticize; it is crucial to give them a question addressed to the specific concern. And it is often thought to be useful feedback to instructors. Of course, it is often not a matter on which the instructor should be praised or faulted. That's the reason to get it out of the way before getting down to the teaching.

In the fourth question we pick up the bitter complaints about lecture theaters where 20 students are sitting in the aisles throughout the year, or where students whose schedule makes it impossible to get to the lab early never get their hands onto the better equipment or computers. Again, this is useful information and it is, we think, vital that the affect tied to it be released before the student starts taking it out on the teacher.

GENERAL FEATURES OF THE FORM. A number of these have already been mentioned and explained. They include:

(i) the use of deflectors, (ii) brevity, (iii) the 'black marks' approach, (iv) the use of cues instead of questions, (v) salience scoring, (vi) the avoidance of questions about style variables (referred to here as 'de-styling'), (vii) the avoidance of questions inappropriate to the student's expertise, (viii) the large number of questions on practical aspects of delivery of instruction, (ix) the strong explicit emphasis on the overall rating, (x) the specific choice of wording on the overall question and *(xi) the skewed rating scale on the overall question.* Each is responsive to a particular kind of problem that arises with common approaches to the design of student rating forms.

Deflectors Note that Questions 1 and 2 are conceptually distinct. For example, a high score on 1 and a low one on 2 would suggest to the department the need for more courses in the area; the reverse suggests cutting down the length of a course.

Brevity Considerable effort was expended in keeping the form to a single page in the course of scores of design passes. Our latest version, completed only as this goes to press, has abandoned this mainly because of the problems for the key-punchers; those doing manual scoring can still get it on one page. We have also incorporated a number of new cues, based on an analysis of the write-in comments on several thousand forms done under no time pressure. With the expanded form we shall probably have to go over the 10 minutes it used to take us, in classes of up to a hundred, but not by more than four or five minutes. Many comments on multi-page forms *that require responses on all criteria* complain about their length, and reduced response rates (on the form submitted by a given student) speak louder than comments. Experience suggests that most of the data from long forms is not used, and running them wastes a great deal of paper, staff time, student time and computer time; hence money. It's also relevant that restricting the printing to one side of a page leaves plenty of room for comments on the back. The other reasons for brevity are: (i) to minimize the disruption of teaching and learning; (ii) to avoid the problems of poor (overall) return rates and (iii) of stereotyped responding[6], when students react against the effort involved in filling out long forms, especially if they receive a great many of them; (iv) to leave plenty of time and energy for departments and instructors to add their own form on a second page. (Be sure that *you* have the time and energy to do the analysis of the results on the second form!) After you've analyzed the write-in comments of a few thousand students who have been encouraged to suggest improvements, as we now have, you will be inspired to move the most frequently made comments on to the list of cues, thereby reducing the time spent on reading free-form comments. Doing this does steal space from the free comments, but it also reduces their frequency. (Leftover comments are often useful for second-stage forms, and for program evaluation forms.) In any case, the trick is to get the respondents to tell us which of a large number of options (109) they really care about, using one piece of paper, in a short time.

'Black Marks' We separately list the ways in which instructors sometimes fall short of their responsibilities and the ways in which they can excel, rather than incorporating them into a single scale. It is almost certainly the presence of the so-called 'black marks list' which has eliminated the usual headroom problem and brings our mean rating down to 6.75 on the 11 point scale. And the use of any more-or-less comprehensive list of potential sins and virtues tends to improve the validity of the overall ratings since it reminds students of (relevant) events they may have forgotten, or helps them to think about relevant matters they may not have considered.

Cues and salience scoring Cues (or prompts), to which an all-or-none response is requested, are used instead of the usual terms attached to a continuous scale. The first intent is to elicit only salient ('heartfelt') responses, by avoiding pressure on the respondent to answer even when s/he has no real feelings about the item. (The instructions call for marking the box only if the feature mentioned was *particularly* important.) Thus one advantage of the approach is to reduce dilution of significant responses by forced responses. Another advantage is to cut down on respondent load, hence improve the chances of high returns; the respondents only have to put a mark where they feel inclined, and they do not have to think hard about where to mark a continuous scale (something about which they often complain, especially with long forms). There's also a saving in keypunching and number-crunching. It isn't very important that one not get a response on each cue by every student, because the number of students that mention the item (in any class that's big enough to provide a basis for action), provides an alternative and better indicator of real trouble or out-

[6] For example, marking the same point on the scale for all scales on a 20-scale form. It's easy enough to find out that this is occurring (by reversing the sense of some of the questions) but the serious task is to avoid causing it.

standing merit; and you can go to the second-stage form if you need to. Additionally, of course, by using cues we are able to pick up responses on far more details than we could with any single-page form using full scales[7].

Salience scoring and shifting baselines The reduction of response time (and of resentment by instructors) is assisted by the fact that the overall grade sets its own baseline on the items. A high overall rating may be coupled with virtually no marking of cues, meaning that they tended to be uniformly good.

Destyling The use of style research—including valid style research—in summative personnel evaluation involves exactly the same fallacy as racism. It renders invalid as well as probably illegal any decisions based on the use of—or even the seeing of—such data. Data on the performance of duties is: (i) adequate for personnel decisions; (ii) the only data that can legitimately be used for personnel decisions. N instructor has any duty to use any styles—even those that have been, on balance, more successful; any individual instructor may be still more successful using other styles. Instructors can only be judged on the success with which they, as individuals, teach valuable subject matter in an effective and ethical way. In the formative context, if great care is exercised, style data has some legitimate uses. The two main ones are to salvage a disaster case and to check on the success with which an instructor is achieving a particular style which they have decided to adopt. In both cases, this should be done with a second-stage form.

Style data may, however, be of considerable interest to students. This interest may require a student-run evaluation system, with all its difficulties[8]; it's one of the few points on which there seems to be any kind of gap between the student interests in staff ratings and the interest in accountability or professional development . 'Lack of enthusiasm', one of the most common style complaints, are often mentioned in the write-in comments, summaries of which are returned to the teacher. But the teacher who appears bored with the subject to one student may strike another as scholarly, shy, or 'cool'; this is not a good kind of cue to bring into the list that will affect the overall rating. The Attitude to Students cue picks up some style reactions, but it's there because it sometimes serves as the lone locator of trouble or outstanding merit in a teacher who gets no other cue responses; and because there are attitudes that raise questions of ethics rather than style. The most controversial issue is whether 'organization' of presentations is a matter of style or an obligation. It's as well to be cautious about including criteria on which the only person to have a teaching style named after him would have failed. Nevertheless, although Socrates would have failed, there are certainly particular teaching tasks where its absence appears to be a fault. But because its absence is not *universally* a fault (e.g. in tutorials and advanced seminars run as discussions of student-raised questions or just-emerging issues), we exclude it[9]. Close analysis of the cases where organization seems appropriate often

[7] The feedback form we use (for question 5) is a two-way horizontal bar chart, which shows responses on items that are reasonably closely matched in sense (e.g., fair vs. unfair marking of tests) as bars on opposite sides of a vertical axis. Again, this uses about 5% of the space that a set of charts covering all the items would take; and simply providing the median or mean response on scales is very seriously inadequate. On the other five questions we provide the more usual vertical bar charts.

[8] Low return rates, lack of credibility, history of manipulation, technical deficiencies, time costs to students, etc.

[9] Student rating forms often ask about certain specific approaches to organization, e.g. "provides objectives for each class", "follows outline for the term". These are much more rigid and less defensi-

reveals that the instructor is doing what a text should be doing, and can't do it as well. If no available text does it, the instructor might do better by handing out full class notes on which discussion of applications and problems, methods and related material, could then center.

On the other hand, we include 'amount of discussion', which is to some extent a style parameter. But discussion serves other purposes, which go beyond style; it is the students' chance to get further explanations of the part of a lecture or test which they did not understand; it provides them with a chance to get feedback on *their* interpretations; to request more applications to examples, etc. And those other functions bear on the obligations of the teacher, not just on their style of teaching. Still, there is a grey area here.

The avoidance of questions inappropriate to the student's expertise The form has also been pruned, as far as we can tell, of questions which students are not in a position to answer validly, in general, such as questions about the teacher's subject-matter expertise—and judgements about enthusiasm for the subject. Of course, everyone who designs these forms would say the same thing about their forms. The way we have done it may be a little more systematic than is usual. For example, there are clearly some cases where students can tell very well that a new tutor, or a tutor new to this particular unit, does not know the subject matter well. In most cases, they can't. So there should certainly not be a required rating scale on Knowledge of Subject Matter on which staff who are the leading experts will be thrilled to discover that their students only rate them 50% competent. Ask a silly question, you get a silly answer. Moreover, it's an answer which antagonizes staff towards the useful ratings that students *can* provide. There is still the write-in option for the students in the cases where the expertise deficiency is obvious, and they avail themselves of it quite extensively. Now, even if a required rating is out of place, why not a cue for Weak Knowledge of Subject Matter? It's not as bad as a required rating, but it's still inappropriate since the cues are meant to be *clearly* matters that fall within student competence in general, even though some of them call for judgement.

Emphasis on the overall rating Most student rating forms include some kind of overall rating question, and in many hundreds of universities this question is (wrongly) the only measure of teaching merit used for personnel action. (Of course, quality of content and other matters should also be considered.) But is the overall question in any way legitimate? The usual reason for using it is the considerable body of research which shows it to (elicit responses that) correlate strongly with learning gains[10]. Another reason, probably more important, is trying to get a perspective—the student's perspective—on the critical comments. Students can complain about a number of features of a class and check none of the commendable

ble, especially since "objectives" are often taken to mean "behavioural objectives", which excludes perfectly sensible objectives such as "will allow students to raise any problems they encountered in understanding the lectures/reading/tests". It does seem reasonable to expect instructors to make clear their conception of the *function* or *role* of labs/tutorials/lectures in a course, certainly if it deviates from the norm or if there is no clear norm. We cover that under 'defining expectations from students'.

[10] That is, the actual amount learnt by the students; usually obtained from a comparison of pre-test scores with scores on a final examination (the post-test). The pre-test must match the final for content and difficulty, though it may include some other questions to check on prerequisite coverage. Many faculty are amazed to find students in their courses who can pass the final on the first day.

items—and still rate it as a very good class. They are telling you that they think it's worth the extra effort to fix up some deficiencies. And the reverse is also perfectly possible. If the form *doesn't* include an overall rating, quite inappropriate reactions by staff to the micro-ratings are common.

At UWA, some criticism of the overall question has been voiced, based on the fact that the overall rating will reflect very different weighting of the various aspects of teaching, by different students. This is certainly true. But if you don't have a proof that some particular weighting of dimensions is correct, you should surely allow the customers to make that decision. They are entitled to a wide range of different weightings; presumably, we should make an effort to cope with their diversity.

Wording and placement of the overall question The form, and position, of the overall question are very important. Form: no question calling for a comparative rating can validly support the usual personnel decisions (it *can* support an award for the best teacher)[11]. Questions like "Would you recommend this unit to a friend with similar interests?" have been floating around for years, and appeal to many, but are very difficult to interpret (this may be the only unit on the topic although badly taught, etc.). Position: you want to reduce the contamination of the rating by a low opinion of the content or the workload, since that's often not chosen by the instructor (hence questions 1, 2, and 3).

Departments—since they are often interested in recruiting majors—are often keen on questions about how much increase in interest in the subject the teacher has engendered. But that would be an inappropriate issue for someone teaching a required and unpopular course in a professional school, and must also be answered negatively by those already fully committed to the subject; hence the responses to this question are not easy to interpret except in very tightly controlled contexts. Moreover, they are not generalizable—and hence not dear to central administration—since gains in majors by one department are usually at the expense of others.

Skewed scale The key problem is to avoid using the sort of scale—for example, a linear six point A-F scale—which finishes up with (almost) everyone in the top two categories. So we provide more choices at the top end where we need to get some discrimination, while leaving plenty of room at the bottom for clear expressions of unacceptable and marginal ratings.

TEACHING AWARDS We used these forms for teaching awards this year (as well as to support promotion applications, remedial, and perfectionist advising), first sending out 40,000 of them to our ten thousand students in a 'nomination phase', where students were asked to 'write up' their instructors, with especial care if the instructors were particularly good or bad. We used this (bi-modal, relatively low response-rate) data to identify some outstanding candidates (the 'students' candidates'); and asked department heads to nominate others (the 'heads' candidates'). Then we went into the classes of nominees—with their permission, of course—and got (virtually)100% returns from the students present. To this we added ratings by heads or their consultants on knowledge of subject-matter, contribution to out-of-class teaching-related efforts, and size of teaching load. We then checked for large differences between faculties, first-year and upper-division class-

[11] Courts may not be up on fancy talk about construct validity, but they know face-invalidity when they see it. Since it's obvious that the worst teacher on the faculty may be quite good, it's useless to ask questions from which you will only be able to rank faculty.

es, etc., but found none. The half-dozen winners, with their almost-perfect 10s, stood out clearly (a unanimous vote by a distinguished selection committee), and came from across the faculties, ranks, years, and class-sizes[12]. It didn't hurt the legitimacy of the exercise at all that the incoming, elected, Chair of the Academic Board was not only a winner, but the only one nominated by both routes.

[12] Women were 'over-represented' amongst the finalists, and (slightly) under-represented amongst the winners.

NOTES

[1] The history of this form began with a report commissioned by the Vice-Chancellor at Berkeley ("The Evaluation of Teaching at Berkeley", 64 pages, Scriven, 1973, published in UC DATABASE (online), 1978.) That report recommended the use of either a standard form or at least a standard overall question (corresponding to Question 6 on the present version) to provide baseline data. The standard overall question was then required for all faculty evaluation at Berkeley; some departments used the whole form. Other campuses in the UC system followed suit to varying degrees, as did other systems subsequently. Many others had been using student rating forms for a long time (e.g. some departments at Minnesota required the use of the—supposedly voluntary—campus form from 1951). A program of some twenty publications on the evaluation of teaching over this period has led to a number of adoptions of the form or distinctive features from it as well as modifications and improvements to the form. Three of the publications might be of interest, since they further develop some of the procedures and concepts in this memorandum:

"The Validity of Student Ratings", in ***Instructional Evaluation,*** September, 1988, pp. 5-19, 1988.

"Summative Teacher Evaluation", in ***The Handbook of Teacher Evaluation,*** Jason Millman, ed.; Sage, 1981, pp. 244-271.

"The State of the Art in Tertiary Teacher Evaluation" ***Higher Education Research & Development,*** October 1988. (Preprints available from CTES.)

STUDENT RATING FORM

(Centre for Tertiary Education Studies, The University of Western Australia)

Name of Teacher:______________________________ (1-30) Unit Number:____________ (31-37) Date:__________ (38-41)
Last name Initials

Class Time:_______ (42-45) Day:______ (46) Bldg/Rm.___________/____ (47-52) (53)Tut'l(1)/Lab(2)/Lect(3)/S'm'r(4) (circle one)

It's only essential to circle five ratings, one for each of the questions 1, 2, 3,4 and 6.... **Your ratings** will be used to improve teaching and promotions.... **It's best** to read all the questions before answering any.... **Use** the bottom of the page for any comments that won't fit elsewhere.... **Try to avoid** influence from your personal likings (for subject or teacher).... **Suggest** improvements to this form, too.

——THE CONTENT——

1. How much of the material covered seemed valuable, if you feel you are able to tell now? (Circle or tick one of the ranges of numbers, e.g. 44–54)

(56) 0 – 10%(0) 11 – 21%(1) 22 – 32%(2) 33 – 43%(3) 44 – 54%(4) 55 – 65%(5) 66 – 76%(6) 77 – 87%(7) 88 – 100%(8)

Mention here, or at the bottom of the page, any part that seemed not to be valuable....

2. How well did the unit cover what you felt you needed (and could reasonably expect)? (Circle or tick one of the ranges of numbers, e.g. 44–54)

(57) 0 – 10%(0) 11 – 21%(1) 22 – 32%(2) 33 – 43%(3) 44 – 54%(4) 55 – 65%(5) 66 – 76%(6) 77 – 87%(7) 88 – 100%(8)

What was missing? (58)Tick here if the missing things had been promised, in words or print......(1)

——THE LOAD——

3. How heavy a load did this unit represent (compared to other units this year)? (Circle one response in the first line.)

(59) Much lighter(1) Lighter(2) Average(3) Heavier(4) Much heavier(5)

——THE FACILITIES——

4. How are the facilities?

(60) Overcrowded (1) .... Gets too hot (2) .... No ventilation (3) .... Uncomfortable seats (4) ....
Inadequate lighting (5).... Unsafe practices (6).... Insufficient access to equipment (7)....
Late classes cause danger or inconvenience (8)....

OTHER. . .

After completing Questions 5 and 6 on Page 2 put any other comments in the space below.

THE TEACHING

5. With respect to what was covered, **how well was it taught**? (Read the following lists and mark any appropriate boxes as you go, then enter an overall grade on Question 6.)

5A: Any aspects that you felt particularly needed improvement?

61 □ **Classroom control**
62 □ **Explanation of what you're expected to do**
Lecturing
63 □ little relevance to course
64 □ too similar to the text
65 □ too theoretical
66 □ drifts off point too much
67 □ not enough on theory or general methods
68 □ speech is hard to understand
69 □ assumes too much prior knowledge,
70 □ unclear explanations
71 □ too easy
72 □ usually too fast
73 □ too much jargon
74 □ boring
75 □ legibility of board writing
76 □ too rushed towards the end
77 □ never pulls it all together
OTHER . . .

Overheads
78 □ hard to read
79 □ too many
80 □ too few
81 □ irrelevant
82 □ shown too fast
Discussion
83 □ not enough
84 □ too much
85 □ poor quality
86 □ not reponsive to questions asked
88 □ not well controlled
Hand-outs
89 □ quality
90 □ quantity
92 □ not enough practice questions
Texts
92 □ too many
93 □ too costly
94 □ not available
95 □ poor quality or relevance

Reserve Materials
96 □ too late
97 □ missing
Assignments
99 □ not enough
100 □ too many
101 □ unclear
102 □ inadequate comments
103 □ slow or no return
104 □ **Demonstrations**
105 □ **Field trips**
106 □ **Labs**
Testing
107 □ not enough
108 □ too much
119 □ poor coverage
110 □ unfair marking
111 □ poor explanation of grading
112 □ poor supervision of tests
113 □ **Punctuality**
114 □ **Frequently cancelled**
115 □ **Running well over time**
116 □ **Running well under time**
117 □ **Availability** of lecturer or tutor
118 □ **Attitude** to students

5B: Any aspects that you thought were particularly commendable?

120 □ **Classroom control**
121 □ **Explanation of what you're expected to do**
Lecturing
127 □ excellent voice
129 □ clear explanations
2 □ very interesting material
3 □ highly readable board writing
5 □ extremely valuable
Overheads
6 □ very readable
9 □ very helpful
OTHER . . .

Discussion
10 □ ample
14 □ open
15 □ well managed
Handouts
16 □ quantity
17 □ quality
21 □ **Texts**
Reserve Materials
24 □ very helpful
Assignments
27 □ clearly defined
28 □ useful comments

29 □ **Demonstrations**
30 □ **Field trips**
31 □ **Labs**
Testing
32 □ right amount
34 □ good coverage
35 □ fairly marked
37 □ good explanations of grading
38 □ properly supervised
42 □ **Availability** of lecturer or tutor
43 □ **Attitude** to students
Flexibility
44 □ adjusted to student needs

6. **OVERALL: In the light of 5A and 5B, how would you rate the teaching?** (Circle one grade: B+ is one grade. Do not circle the verbal descriptions.)

(44-45) F(1) D(2) C−(3) C(4) C+(5) B−(6) B(7) B+(8) A−(9) A(10) A+(11)