A0776

The Validity of Student Ratings[1]

Michael Scriven[2]
University of Western Australia

INTRODUCTION

Student ratings of instruction are widely used as a basis for personnel decisions and faculty development recommendations in tertiary education today, most commonly in the United States. But their validity is still a matter of dispute. Some of the objections come from people who are simply ignorant of the research literature, and these complaints have been admirably handled by Centra,[3] McKeachie,[4] and more recently Aleamoni.[5] Other concerns about validity remain, for example: (i) the general concern that student rating forms ask many questions about matters that students do not appear to be in any position to judge reliably; (ii) the fact that the overall rating of teaching merit by students is only statistically related to learning gains, a concern if one believes that it is inappropriate to use statistical indicators in personnel decisions; or (iii) the concern that the validation studies used to justify appeal to student ratings use questionable indicators instead of the true criterion. (For example, some of them correlate the student ratings with peer ratings of the merit of teachers instead of with superior learning gains.) For these reasons, it seems opportune to look carefully at the whole foundation of the claimed validity of student ratings.

Student ratings of courses and departments. The *general* question of the validity of student ratings brings in their use in other and less controversial contexts, notably in course evaluation, and in the (analogous) evaluation of seminars, workshops, and special lectures or other presentations. Rating departments and colleges brings in some special issues, but nothing that presents novel methodological difficulties comparable to the rating of instruction. Therefore, these other cases are only treated in passing, and the main focus is on the use of student ratings for faculty evaluation.

Nine potential sources of validity for student ratings are distinguished in this paper, although some of them are quite closely related and could be grouped. We list them very briefly now, coming back to them later in more detail.

A. The statistical correlation of student ratings with learning gains.

B. The unique position and qualifications of the students in rating their own increased knowledge and comprehension.

C. The unique position of the students in rating changed motivation towards (i) the subject taught; perhaps also towards (ii) a career associated with that subject; and perhaps also (iii) a changed general attitude toward further learning in the subject area or more generally.

D. The unique position of the students in rating observable matters of fact relevant to competent teaching, such as the availability of copies of the text.

E. The unique position of the students in identifying the presence of teaching style indicators. Is the teacher enthusiastic; does s/he ask many questions, encourage questions from students, etc.?

F. Relatedly, students are in a good posi-

tion to judge–although it is not a matter of direct observation–such matters as whether tests covered all the material of the course.

G. Students as consumers are likely to be able to report quite reliably to their peers on such matters as the cost of the texts, which has no essential bearing on the quality of instruction.

H. Student ratings represent participation in a process often represented as "democratic decision making."

I. The "best available alternative" line of argument.

Only two of these get much attention in the usual discussions–A and B. Contrary to the usual view, one of those two–the strong empirical connection to successful teaching–cannot be used to support personnel decisions, even if the research were impeccable. But it is argued here that combinations of the other eight do provide a secure foundation for the four main uses of student ratings, namely to provide a basis for personnel decisions, staff development, course evaluation, and student information. (These are special cases of the general evaluative tasks of summative evaluation, formative evaluation, and product description.[6])

PRECONDITIONS FOR THE USE OF STUDENT RATINGS

None of these nine possible lines of argument contributes to validity unless the particular rating form used is appropriate for the particular kind of use that is envisaged.[7] Since rating forms vary widely, the usual generalizations to the effect that student ratings are a good indicator of something else (learning gains or teacher merit, for example) are misleading since they rest on the assumption that there is a common property to all such ratings.[8] In terms of the potential sources of invalidity stressed here, most forms (when used in the most common ways) are invalid as a basis for personnel action. For example, many forms used as input for such decisions ask questions that may influence the respondent by bringing in extraneous and potentially prejudicial material, such as questions about the teacher's style or personality, or the appeal of the subject matter.

Another problem with the use of rating forms for summative evaluation is that many of them ask global or overall questions–the ones on which most personnel decisions appear to be based, in the writer's experience–which cannot (except accidentally) generate the information needed for such conclusions. Common examples of this mistake include the use of forms whose "key question" asks for: (i) comparisons of teachers; (ii) whether one "would recommend the course to a friend with similar interests;" or (iii) whether "it's one of the best courses" one has had. All are face-invalid and certainly provide a worse basis for adverse personnel action than the polygraph in criminal cases.[9]

This means most of the usual forms are at best defensible for formative evaluation, not for summative, and there are serious problems about their use for formative evaluation, centering around the usual procedures for interpretation. For example, it is often recommended that improvement should take place on a number of dimensions concurrently. This assumes that these dimensions are causally independent of other dimensions on which the performance was good or better; at least it assumes they are not mutually inhibitory. Since that is certainly not true in general, there is an obvious risk. It's not helpful to an anorexic patient to recommend that they put on weight by eating more and exercising less if the only way they can build an appetite is by exercising.

One rating form can validly perform more than one function, but the possibility of interactions between the functions, even when they are not completely incompatible,[10] and the lack of evidence on the size of such interactions, makes it overoptimistic to assume that we yet have a combination form that can perform all functions. Hence, it is unlikely that all of the considerations or lines of argument to be presented here, each of which appears to validate certain uses and types of student ratings, can safely be invoked in support of a particular form.

There are also pragmatic considerations (logistical, political, economic, psychological)

affecting the design of forms, which are crucial for *validity* and count against multi-functionality. Traps include: (i) using forms that are so long that students do not fill them all in,[11] (ii) overkill, as in (officially) rating every course every year, which leads to students "turning off" on the forms unless they are very brief; (iii) using forms that do not include the questions students want answered about courses they are considering taking, thus creating resentment and a lack of willingness to take trouble with filling in the forms;[12] (iv) using forms that do include questions students suspect will be used to discriminate against their comments[13] or against them personally;[14] (v) using forms with inadequate head-room for the best teachers' performance to show up; and (vi) using forms that are strongly biased towards favorable comments (or unfavorable comments).

Again, none of the lines of argument here will support the use of student rating results that are obtained from poorly administered tests, poorly controlled data collection, and poorly implemented use of the results. Among the common errors of these types that undermine validity, not sharply distinct from the preceding, are: (i) the absence of adequate demonstrations to students of the importance attached to their ratings;[15] (iii) the use of instructors to collect and turn in forms rating their own instructional merit;[16] (iv) lack of controls over pleas for sympathy or indulgence by the teacher in advance of the distribution of the forms;[17] (v) allowing inadequate time for completion of the forms; (vi) providing rewards for racing through the form (e.g. by having them filled out at the end of a class and allowing students to leave as soon as they turn them in); (vii) lack of control against political, religious or gender conspiracies to damage the teacher;[18] (viii) failure to pre-announce the day on which forms will be distributed;[19] (ix) failure to ensure an acceptable return rate;[20] and (x) distributing forms too early or too late in the course.[21]

Since the validity of student rating forms is just as dependent on the techniques and contexts of their administration as on the intrinsic merit of the form, and since few or no studies meet the conditions on proper administration mentioned here, one might suppose that most conclusions so far should be regarded as rather speculative. However, some of them get a new lease on life via the alternative routes to justification to be discussed shortly.

Equally strong warnings apply, of course, to errors of data processing and interpretation. Examples include: (i) the use of averages alone, without regard to the distribution;[22] (ii) failing to set up appropriate comparison groups so that the usual tendency for ratings to be higher in graduate professional schools can be taken into account; (iii) treating small differences as significant, just because they are statistically significant; (iv) use of factors based on factor-analysis without logical/theoretical validation; and (v) ignoring ceiling/floor effects. The most serious error of interpretation, of course, is: (vi) that of supposing the ratings can carry the whole load of *either* formative or summative evaluation; see the following paragraph.

Student ratings are one important source of data for the evaluation of teaching merit, but neither necessary nor sufficient. In particular, student ratings should not be used in the absence of some evaluation of: (i) the academic quality of content, best rated by subject matter specialists; (ii) the justification of the content against the eventual (and sometimes the immediate) needs of the students and society; (iii) the quality of test construction–best rated by measurement specialists, though there should be some faculty in the same department who have the necessary minimum competencies; (iv) the use of valid and ethical grading practices, by those with some understanding of proper practice in these areas; (v) knowledge of procedures for handling emergencies; (vi) the professionalism of the individual's attitude to teaching (evidenced by attendance, efforts to improve teaching, response to criticism, etc.); and (vii) evidence of out-of-class teaching-related activities, such as work on curriculum committees, turning in grades on time, assisting other teachers when needed, etc. (A complete list of what ought to be covered has been provided elsewhere.[23]

Similarly, student ratings form an essential part of the data for the evaluation of courses, workshops, degree programs, etc., but they cannot carry the whole burden. It is essential to look at the data on such parameters as needs, demand, opportunities for symbiosis, content and costs, and estimate the probable longevity of all of them. These are covered quite well in the program evaluation literature and will not be discussed further here.[24]

Now there are some approaches to faculty evaluation that appear to offer the possibility of providing not only a useful supplement to student ratings, but an alternative that completely excludes them. It is essential to consider these, because one of the main arguments one hears for the legitimacy of using student ratings is that, despite certain problems, they are unavoidable. Conversely, it is often argued that there is no need to use them at all, because one or a combination of the following is a satisfactory or superior alternative.

ALTERNATIVES TO THE USE OF STUDENT RATINGS— ALTERNATIVE (A): GAIN SCORES

While it should not be claimed that ratings data are absolutely superior to data about learning outcomes or gains, the problems with using terminal or gain scores are often seriously underestimated.[25] To begin with, we can draw no conclusions about the merit of the teacher from the raw gain scores of the students, as such, and the problem is to identify what else we need in order to extract a conclusion about teacher merit from those scores. To avoid the "Harvard fallacy"[26] one must at least have comparative data on the post-test (or gain scores) of comparable groups dealing with the same material and taught by other teachers. This is possible with large multi-section introductory courses (and in large primary schools) but not in general. (Of course, it must also be true that the students were examined by means of a test that is set, administered, and marked by someone other than the teachers, or, as a somewhat unsatisfactory substitute, by the teachers collectively. And the test should not be set until after the end of the teaching period.)

Getting gain-score data from other institutions with similar courses is just possible in the case of extremely tightly structured curricula, as in accounting courses aimed at the national exams, but even there the results do not reveal which slice of the "value-added" components is due to the teacher rather than the peers or the library, etc. We would probably be better off with student ratings since students can at least estimate and discount, whereas we cannot in practice factor out the contribution of peers and school resources. Merely having an experienced teacher of similar students look at the gain scores (or just the final exam results) is far too weak an approach, since we have no way to tell how good the experienced teacher is, let alone how objectively s/he can judge the results of others.

We may do better if we use gain scores from earlier years, when other teachers taught the same subject at the same institution, but then we face the need to factor out changes in facilities, curriculum, other staff, and, on the student side, changes in demographics and general motivation. (In the case of self-evaluation–an important case of formative evaluation–the comparison with one's own performance in previous years is very important and not so fraught with problems as comparisons with others.[27])

But the problem is more serious, because all we can hope to get from studies meeting *all* the above conditions are comparative ratings, and those are still not the criterion-referenced ratings we usually need for personnel evaluation.[28]

In the more usual cases where we lack the comparative data, the student ratings loom as more important–though still not sufficient–unless there is another approach we can use. One possibility still appeals to many: the idea of using someone besides the students, some skilled observer, who can better determine how well the class content is being conveyed to the students. We'll argue below in some detail that there is no possibility of an observer doing this even *nearly* as well as students, certainly at the secondary or post-secondary level.[29] But even at this point, it

cannot seem highly plausible to argue that a non-student is more likely than a student to judge how well the material is explained to students, or that one person will be less likely to err than many.

Hence the *faute de mieux* argument, that there is simply nothing better, begins to look more attractive. This is, after all, a line of argument we use in selecting students for admission to and retention in tertiary institutions; the tests we use are by no means ideal, just the best we can manage in the time and with the evaluators available. However, there are serious problems about the use of any indicators with merely statistical validation in personnel evaluation, including the evaluation of students, and we have to understand that legitimacy depends not just on showing that certain data is the best available, but also that it is the best that could be obtained with the *maximum feasible* time, money, and expertise available. The problem with gain scores is that they are *usually* not going to do the job even fairly well, and they *virtually never* provide what we need, conceptually speaking, since they only yield *comparative* rankings.[30] But, in the case of a large number of sections of a stable first-year English course, for example, they may be the best evidence of all.

The preceding arguments apply, with some obvious modifications, to the evaluation of courses or programs using outcome data.

ALTERNATIVE (B): VISITING OBSERVERS

The only other hope for a general solution is the visiting observer, either a colleague or a specialist. But visitors are badly placed for many reasons: (i) the time they are present fails to meet minimal standards for sample size,[31] (ii) they distort what they observe by their presence, to an unknown and essentially immeasurable degree; (iii) under some arrangements, their visits are foreshadowed and hence, for a second reason, do not provide a typical sample; (iv) they often bring in personal prejudices of an unknown magnitude (if they know the evaluee outside the classroom), or (v) they bring in prejudices in favor of some teaching style or personality type that have no absolute merit for personnel decisions; (vi) they are socially protected from access to certain kinds of data to which students have access (e.g. they are unlikely to be subject to sexual harassment); and (vii) they lack some of the crucial cognitive qualifications to judge the adequacy of the explanations, as discussed in the next section.[32]

These "crucial cognitive qualifications" consist of estimating the cognitive resources of the whole student group, since it is only in terms of those resources that one can judge whether the presentation is suitable or successful in creating comprehension. No one person can *match* those resources, for logical reasons, so the question is whether any one person can *estimate* them accurately from an observation or two that may be very atypical. The visiting evaluator is, by virtue of age, experience, and social milieu, almost certain not to match the cognitive background of even one of the students and certainly lacks the same motivations (which significantly affect what one sees). Student ratings, by contrast, are done by those for whom these mismatches do not exist. Students are *necessarily* better judges than visitors in these two respects: in respect to the logical point, and in most of the other respects in which visitors are handicapped.

Even a teacher with extensive experience in teaching the same subject to similar students can only make estimates of comprehension during a visit–estimates that are of dubious value for the students that remain silent–whereas the students themselves can make direct observations of it. If the visitor moves back to the still more indirect approach of judging the level of *presentation* against what they have found to be successful, the whole matter of validity dissolves into the question of the merit of their own performance and their own reliability in making judgements of this kind (two different matters, neither of them tested in any studies I have seen).[33]

Presumably the persistent myth about the value of having peers visit classes to judge teaching merit derives mainly from the argument that students can't be regarded as capable

of making judgements about content and *hence* can't be competent to judge teaching. This argument presupposes that one is constrained to evaluate on the basis of one kind of input, and that variations in content quality are more important than variations in process quality (rare, because the text provides a large common core). The commonsensical process is to use both inputs, with peers judging content and students judging process. (Of course, the peers should not think that visits are an adequate basis for judging content.) The view argued for here is that input from more than one source is essential and that students are best placed to provide one of the key inputs. Of course, we also need to look at content, which is just as much a necessary condition for success as the conveyance of comprehension, but visiting a class or two, of unknown representativeness, is no way to do it.[34]

Even if we accepted the view that only one group can do the evaluation, and that it should be peers, it's doubtful whether having them visit a classroom to judge pedagogical skill is better than nothing. Given the sources of error and bias, we might do better with nothing, as indeed was the modal practice for decades in the tertiary sector. But students aren't nothing, although one suspects that they must have been rated very close to zero–as judges–in order to think that a casual visit by an elderly man was worth more than their considered collective opinions. The error of having faculty replace students' input on this dimension is at least as bad as having students replace faculty input on quality of content.

Do the students perhaps lack an essential ingredient of maturity which the middle-aged think they have acquired? (One might as well restrict discussions to those of at least middle age since it is they who control the decisions.) It is surprising how many faculty members will quote the views of their own children when they happen to be enrolled in a class at school or college taught by a teacher under consideration, while not favoring student ratings. But the children of faculty are not even arguably more mature than the children of farm workers, clerks or mechanics. Listening to the discussions in faculty senate meetings, it is surely difficult to argue that all middle-aged faculty are mature in any evaluative sense, and it seems implausible to argue that a higher proportion of them are capable of a fair judgement of their colleagues than students are of their teachers. With suitable precautions, the claim of immaturity hardly seems to offset the considerations pointing in the other direction.

Students are thus, it appears, the best-placed judges of at least *prima facie* pedagogical competence–the teacher's ability to explain things to them–and their ratings are not only indispensable when we cannot meet the conditions for interpreting gain scores, but usually valuable even when we can, because they cover more than just the learning. (Of course, we have to meet the previously mentioned conditions on the rating forms or interview structures and on the process of using them.) Now it is time to look more closely at the lines of argument that might validate certain specific types of–and uses for–student ratings.

POTENTIAL FOUNDATIONS OF VALIDITY FOR STUDENT RATINGS

The general logic of inquiry here is to locate *types of information*, if there are any, that meet two conditions: (i) they are important and preferably essential for the evaluation of teaching, or the description of certain aspects of it, and (ii) they are matters which the student is well-placed (preferably best-placed) to determine.

The first of the potential grounds illustrates a different kind of possibility, the possibility of *indirect* evidence for the validity of student rating. This kind of evidence supposedly validates it as an *indicator* of teaching merit, rather than as direct evidence of that merit.

In describing such types of information here, we expand on the earlier description somewhat and introduce a name for the epistemological role of the students (or the indirect measure) that invokes an analogy with some other situations to suggest a source of possible validity. The types of information include:

A. The statistical correlation of student ratings with learning gains (the "**surrogate**" role of ratings).

B. The unique position and qualifications of the students in rating their own increased knowledge and comprehension (the "privileged access" analogy; we will use the term "**cognitive witness**" for the student role here). Derivatively, they are in a unique position to rate the extent to which texts, class handouts, and instructor feedback–on their assignments, class performance and tests–was enlightening.[35]

C. As for B, but with respect to changed motivation towards (i) the subject taught; perhaps also towards (ii) a career associated with that subject; and perhaps also (iii) with respect to a changed general attitude toward further learning (the "**affective witness**" role). Obviously, too, students are well placed to be witnesse to (iv) whether they enjoyed the course, and this is of some relevance to course/program/department evaluation, especially in recruiting terms. And it is relevant to the wish of other students to pick enjoyable courses, *mutatis mutandis*. (It may even have some relevance to instructor evaluation, as a marginal consideration.)

D. As for B, but with respect to simple matters of fact relevant to competent teaching. There are two categories of such matters of fact, one relevant to summative evaluation (because it relates to the performance of duties or undertakings) and the other to formative evaluation, covered in E below. Examples of summatively relevant facts include: whether allegedly necessary texts were in fact invoked; whether books allegedly on the library reserve shelf were actually there; whether the handwriting on the board was readable at the back of the lecture theatre; whether the lecturer's speech was comprehensible there; whether student grades were merely reported to them or explained, etc. Although more likely to be controversial, the same applies to matters such as the use of racially derogatory epithets and other improper discriminatory behavior such as sexual harassment. (The "**eyewitness**" role: sub-category, "performance of duties" eyewitness.)

E. The second eyewitness category of factual observations, which the students are uniquely placed to make, concerns teaching style indicators. Is the teacher enthusiastic; does s/he ask many questions; does s/he encourage questions; does s/he look at you when talking; does s/he use examples which are relevant to current issues, etc. Responses to these questions not only cannot be used for summative evaluation, but should not really be present on any form used for summative evaluation. They may, however, appear on a form used for formative evaluation, on which a remedial plan is to be based. And they are often of considerable interest to the student-consumer. (The "**eyewitness**" role: sub-category, "style" eyewitness.)

F. Relatedly, students are in a good position to tell whether tests covered all the material of the course; whether prior courses–and reading–said to be prerequisite were in fact necessary; and whether the course was as promised in the first few lectures or handbook. (These relate to duties or undertakings, to both of which, unlike the adoption of a particular style, the teacher is obligated.) Since a modest degree of interpretation is involved in judging whether these things occurred, students will probably be more fallible in making them. Still, they are not only well placed to make such judgements, they are better placed than anyone else, *and* it is the kind of judgement they have had experience in making in many other courses. (This involves some aspect of the "**well-placed witness**" role and some of the "**expert witness**" role.)[36]

G. The role of the students as consumers with certain tastes and preferences, reporting to their peers. This makes relevant not only such questions as whether the teacher favors a discussion approach or stays with straight lecturing (already covered in E), but also questions about the cost of the texts and whether the student enjoyed the course or certain aspects of it. Although the student has the status of eyewitness or privileged access on such matters, the matter itself is not connected to the merit of teaching in the direct way it is in categories B, C, and D. These matters partly involve **product description** and partly the "**market survey**" role of a questionnaire (finding out what the consumer likes or doesn't like about the product.) It would be naive to think that such considerations are not

relevant to the value of a teacher to an institution, even if that is demarcated from merit.[37] And they are often regarded as appropriate in course evaluations, largely because such evaluations are done for marketing reasons.

H. The "participation in democratic decision making" conception of student ratings; this invokes what we might call the "**voter**" role of the student and the "**ballot**" role of the questionnaire.

I. The "**best available alternative**" line of argument, already discussed briefly.

Notably absent from the list is the "**recollections in tranquility**" type of student rating often put forward to justify one kind of student rating, sometimes by people who are not willing to accept any other kind. This is the idea that student ratings would be a valid basis for evaluation if we waited ten or twenty years to get them, by which time the correct perspective on the importance of certain teachers will have emerged. This is a naive view, even conceptually; memory may distort or omit as much as it clarifies, and in any case it refers to teachers as they were, not as they are. Logistically, it is shaky; a huge effort at one institution achieved an 8% return rate.[38] One would be reluctant to initiate personnel action based on such a sample. And once we get near the arguably relevant region (inside five years, say), we find convergence with current ratings.

SPECIFIC ARGUMENTS FOR EACH APPROACH

(A) It's common to suppose that the first of these considerations–the existence of a positive correlation between high student ratings and high learning achievement–is extremely important. Indeed, it is sometimes said that unless it were true, there really could be no justification for the use of student ratings in personnel decisions. The opposite is the case. Contrary to the intuitions of many people, this argument cannot be used at all in establishing the legitimacy of using student ratings, and it would not matter in the least if no such correlation existed. Additionally, of course, there are all the problems of getting from the learning gains to conclusions about the real merit of the teaching.

The reason for this conclusion is that one can essentially never use statistical correlates of merit in personnel decisions affecting staff. You can't fire people for exhibiting a characteristic known to be associated with poor performance, or hire them for the opposite, not even if the correlation coefficient is 1.00. To do so involves guilt by association (or "stereotyping"), and is the essential flaw in racist or sexist practices. You can only proceed against (or for) an individual on the basis of data about that particular individual's performance on, and only on, what are demonstrably duties of the particular job. I have set out the reasons for this view at length elsewhere and will not repeat them here.[39]

Now we need to look at the other possible foundations for the validity of student ratings to replace the one we have just dismissed. We will first examine the "epistemological" grounds of validity, considerations B through F. Then we'll look at the special cases of H and I.

(B through F) Student ratings of certain crucial aspects of teaching merit are potentially valid, and are superior to the ratings of visitors to the class, simply because the students are in the best position to report on several matters that are crucial to evaluation, and because they are not lacking, to a fatal degree, any relevant wisdom or skill. For personnel action, developmental planning, and consumer recommendations, one needs to know whether the material covered is being well explained, and that is what the student is reporting on in the case of questions that ask for a rating of the kind of things covered in B. (For brevity, in discussing summative evaluation, we will focus on the role of B, allowing the reader to make the appropriate adjustments to the arguments for C, D, and F.) It is surely from the privileged position of the students in reporting on how well the instructor has explained the subject matter to them, and the centrality of that consideration in their overall rating, that the high positive correlation with learning gains emerges.

Student input is just as crucial for formative evaluation, which involves D and E as well as B and C. Once one shifts emphasis from the idea that teaching improvement should be focused on improving teaching style, to a primary focus on performing duties, using whatever style suits the teacher, then the data from D and F become crucial. And the same case can be made, even more strongly, for the **product description** questions illustrated in F, which yield data of value to students "shopping around," although they are not merit indicators. But there are many institutions today where getting enrollments is a consideration that must enter into the evaluation of the teacher.

(H) We now come to the ugly duckling, which appears to suggest that the student's ratings are valid by analogy with the voter's ballot. This general line of argument is usually greeted with the scorn appropriate for the idea that the students' vote should be used to determine the answers to problems in mathematics. Let us begin by conceding that this consideration must be regarded as playing a very different kind of role from the others; it is sociological/psychological/political rather than epistemological. But, even if it is not put forward as bearing on validity, it can be unpacked into substantial *additional support* for the use of student ratings. There are several associated sub-arguments here, whose force varies greatly from student to student, campus to campus, and time to time. They range from reasonably plausible hopes, through frequently reported student views, to clear indications: (i) To have students involved in the process of evaluating teachers contributes to avoiding the "students vs. administrators" set, and the "teachers as oppressors against whom students are powerless"[40] set, with consequent benefits for campus atmosphere and hence–one may hope–motivation, reduction of trouble, etc. Part of this process involves the reduction of alienation or *anomie*, and increasing the student's sense of part-ownership of the institution and of his or her own destiny; (ii) The use of student ratings can lead, if there is some appropriate associated discussion, to a much better understanding by students of the difficulties involved in grading students; (iii) It can have a salutary effect on the attitude towards, and grading of, students by teachers, although there are certainly dangers as well as benefits in this area (one of which has been mentioned before–the special pleading problem); (iv) It can, students have volunteered, have a marked benefit on their own self-esteem ("I am not just a victim of criticism, I am also a critic"); (v) It clearly gives some substance to the notion of student involvement in governance, for which student governments often cry in vain.[41] These effects appear to–and should in any case be explored in case they do–lead to a cumulative improvement in the care taken with filling out the student rating forms. And this increases validity.

So H is not a poor relation, except possibly with respect to the degree of confidence one can have in the connections involved. It does not just contribute to campus spirit, and so on. It is certainly of the right type to be in this list, and we should be doing some research on the size of its contribution to validity.

If these arguments are sound, it would appear that there are a number of types of questions which can legitimately appear on a rating form, and one must prioritize the local needs in order to pick a selection. It is useful that there is no need to argue over the issue of whether the "one overall question" is in a privileged position in terms of the empirical evidence, an issue which has been the subject of the interesting recent discussions cited earlier. However, the One Big Question–something like "Overall, how would you rate the instructor on a scale of Excellent/Good/Acceptable/Unsatisfactory/Very Poor"–remains in a privileged position in terms of what is logistically feasible for the summative process (which includes one function of the "annual review" process). It is rare for that process to be capable of handling more than one result, and indeed hard for it to handle more than one number (exceptional cases apart), usually the mean or median of the ratings on the One Big Question. In my experience, it is possible and highly desirable to get a general purpose form down to one page with four questions and a few prompts (50 or so),[42] but how to do that and what to do with the results is another story, one that cannot be the same for all situations.

CONCLUSIONS

There are two significant but very limited places for the use of expert visitors giving style advice, namely for remediation suggestions to a teacher who has proved to be very unsuccessful and asks for help, and for suggestions to those new teachers at a loss for workable approaches. But in order to prove that a teacher *is* unsuccessful, or successful, we would almost always have to use student ratings, we can virtually never use visitors, and we can never use style considerations. Moreover, valid recommendations for teacher improvement can get started very well just by using student ratings on aspects of teaching that are not related to style (audibility, readability of overheads, utility of comments on assignments, adequacy of explanations or materials given out in class etc.). However, one should not carelessly pass up any port in a storm, and it is arguable that someone who must improve should *inter alia* call for style analysis and expert observation, preferably including videotaping (so that pre- and post-change performance can be seen). Where style suggestions promise solutions to desperate problems ("stand next to the student that initiates trouble"), they should be tried for minimax reasons; the objections begin with minimax ends and we are aiming for optimization of teachers who are not in trouble.

A more dangerous use of the visitor is thus when the teacher is not demonstrably bad, but simply desires to improve. The risk is then that trying to upgrade style, piecemeal or holistically, will interfere with something else that is working very well. Again, it is much more sensible to begin by trying to upgrade non-style performance, and it is arguable that one should do only that since then at least the gains from every change seem certain. Even there, however, there is still some risk in ignoring the Master Mechanic's Rule, "If it works, don't fix it."

The visitor to the classroom of a good teacher who has asked for help in improving, *if very specifically trained*, can be useful in observing and reporting on the evaluable non-style features of the teacher's performance, such as content quality, coverage of material promised at the beginning of a session, firmness of treatment, audibility at the back of the room, and discharge of minimum obligations to teach for the benefit of all students. However, almost all trained people in this area have been trained to focus on and comment about *style* because they are usually from preservice or traditional remedial programs.[43]

Surely such a visitor might also venture some suggestions, such as recommending that the students be given more time to respond to questions. The danger of moving into the area of recommendations is the unknown nature of the interactions. There is a risk of unknown size that one may spoil a good wine by tinkering with one feature that seems less-than-ideal in isolation. Still, common sense suggests that, if appropriate cautions are stressed, some suggestions can be made. After all, the teacher can obtain enlightenment even from looking at videotapes of her or his own performance, and can surely sometimes benefit from the point of view of another–especially an experienced–commentator. While it is surely not *improper* to respond to the request for such suggestions, we are now at a point where we should be *extremely cautious* about doing so.

An adviser, not visiting the classroom but looking at materials, tests, grading practices, an outline of topics and activities, and the comments on student work that are being provided, can cover another large slice of evaluable aspects of teaching and is much less likely to slide over the line into style-hyping.

In any case, the data from student rating forms is crucial in planning an overall remedial program, because there are aspects of style that only show up in the long haul and in the absence of visitors.[44] Those aspects may need a higher priority than anything the visitor sees. So there is at most a very small role for the expert (or inexpert) visitor in formative evaluation, and much less in summative evaluation since the most obvious features of the teaching are style characteristics that cannot be used at all.[45] In both these cases, and in providing product descriptions, the student ratings are absolutely crucial. Hence student ratings are not only *a* valid, but often *the only* valid, way to get much of the information needed for most evaluations–and for many useful

descriptions–of teaching. (And the same applies to the evaluation of courses, departments, and programs, where student input is rarely protested although it is also rarely well-designed.)

The validity of student ratings is not the only reason for using them. Their use has potential reactive effects on, for example, the quality of instruction.[46] Properly arranged, these effects should be, on balance, positive–and significant in size for most instructors, although some flurry of negative effects may persist.[47] As the literature makes clear, it certainly can not be assumed that the mere use of student ratings will automatically produce benefits in instruction, regardless of the quality of the form or of its administration and interpretation. But in the end, good teachers, no less (and probably more) than appliance manufacturers, can benefit substantially from, and recognize that they benefit from, appropriate evaluation by those they try to service. If teachers do not benefit as a result of this evaluation, one must keep in mind the possibility that the fault lies in them rather than in the expectation that student ratings will produce improved teaching. There are always two ways to improve teaching: get the teachers to improve, or replace them with better teachers. Of course, one should *also* keep in mind the possibility that the teachers are doing about as well as is feasible in the circumstances, and hence have little or no room to improve. The system of faculty evaluation should be designed to determine whether this possibility applies.[48]

It has been argued here that educational administrators interested in the improvement of instruction (whether by improving courses themselves, or the performance or the composition of the faculty)–and instructors and students with the same interest–will benefit from the use of a sound system of student ratings. In fact, it has been argued that student ratings are essential as well as valid for this purpose. Their use is also likely to be beneficial in other important ways and they are inexpensive to acquire and process. But it may be appropriate to end this paper by reminding the reader that to leave them out of a system for faculty or course evaluation is no worse a sin than to include only them.

Endnotes

[1] Thanks to John Hattie, Barbara Davis, and Robert Cannon, for valuable suggestions on earlier drafts. It should not, however, be assumed that they agree with the main theme or any part of the final version.

[2] Or P.O. Box 69, Pt. Reyes, CA 94956; phone (415) 663–1511.

[3] J. A. Centra, *Determining Faculty Effectiveness*, 1979, Jossey-Bass.

[4] W. J. McKeachie, *Teaching Tips*, (Eighth Edition), 1986, D. C. Heath.

[5] L. M. Aleamoni, "Student Rating Myths vs. Research Facts," *Journal of Personnel Evaluation in Education*, vol. 1, no. 1, 1987, pp. 111–119.

[6] Description is included here as part of the task of evaluation, instead of being contrasted with it (as is usual), for two reasons. First, the reality is that most descriptions involve some evaluative language, at some level of analysis (e.g., "speaks clearly"). Second, it often takes a good part of a serious evaluative investigation before one can describe a product (person, proposal, program, etc.) correctly and objectively, even in non-evaluative language. Description is also the main function of student ratings used for research purposes, for example in research on the most common ways in which teachers administer discipline.

[7] We talk throughout of "rating forms" for simplicity, but one might gather student opinion from interviews instead. The great loss is absolute anonymity, but some experiments with interviewing graduating seniors at Berkeley (Robert Wilson, personal communication) clearly showed that important results can be obtained in that way (notably, the existence of otherwise undetected superteachers).

[8] The existence of a positive correlation (even a correlation of 1.0) between the scores on several forms does not show the presence of a common property; there must also be logical or theoretical grounds for the identification and usually also further factual evidence for it. See "Fallacies of Statistical Substitution" by the present author in *Argumentation*, 1987, pp. 333–349, D. Reidel.

[9] (i) Comparative ratings of teachers can't support the claim that the worst are bad or the best good, without which conclusions few personnel actions can be supported. (ii) Friends with similar interests may be committed to other programs because of different career choices; hence no recommendation would be appropriate. (iii) Course preferences are not usually relevant when evaluating teachers.

[10] As in the case mentioned above, where style assessment, which has a limited use in formative evaluation, completely confounds the use of a form for summative purposes (personnel action).

[11] Apart from validity, long forms massively increase processing costs and raise serious problems of dilution of impact.

[12] Common examples of questions to which students want to know the answer include: (i) how heavy the work load is compared to other courses; (ii) whether grading is easy; (iii) whether it is really necessary to buy all the "required" texts; (iv) what style of teaching is employed (discussion vs. lecturing, for example); and (v) whether the course is "relevant" or "too academic." In the limit, failure to attend to these concerns forces student government to set up a duplicate system, with consequent waste of resources, especially classroom time, and leads to refusals to fill in the official form, or fatigue effects from doing both.

[13] The usual ones concern: (i) expected grade in this course; (ii) overall grade in school to date; (iii) whether the student has been required to take the course. These are inappropriate, partly because the evidence is that the answers are not reliable (from comparing the registrar's figures with the class reports and also from asking students to say, anonymously, whether they lie on such questions), but mainly because they encourage faculty in the entirely improper response of disregarding the complaints of the weaker students.

[14] The obvious example is requesting their names. One still hears faculty arguing that if students haven't the courage to sign their evaluations, the evaluations should not be taken seriously. This is reminiscent of dictators who say their door is always open for dissidents who wish to complain. But it is also extremely unprofessional unless one has very well-thought-out ways to ensure that one's grading or letters of recommendation will not be influenced by complaints or praise *and* has proven to the satisfaction of the students that those procedures will be enforced. I've yet to find someone who meets those conditions.

[15] It's not enough just to *say* this. To *mean* it entails, among other things, that: (i) there has to be space on the form for suggestions about how to improve the forms and the evaluation process; (ii) student government must have some input into procedures and content; (iii) the results are at least sometimes used to improve the course whose members are asked to fill in the forms, e.g. by running a mid-term version for improvements as well as the end-of-term one for the record; (iv) the students are informed about how their ratings are weighted in the faculty evaluation process, and student government is assisted in verifying any such claims. Absent a policy which addresses these considerations, one must face the problem that students have good reasons for running their own ratings system. This involves a considerable duplication of production resources, of class time, and a reduction of student interest. It may involve open hostility.

[16] There is no great reason to object to teachers *distributing* the forms and appointing a student to take them in to the department head or, better, central office, as long as students are informed in some other way about the importance of the process. But having department secretaries or the staff of a support center do the whole business is in general preferable. If teachers are to do it, students must be independently counseled and asked about possible abuses of the system–perhaps by means of some remarks on the form itself–and provided with space on the form to register a belief that there has been an attempt to use inappropriate influence; sob stories about family circumstances are one of the problems. It is usually not necessary to ask staff to absent themselves from the room while the forms are being filled in, but it is one way to keep students informed about changes in the process, and to provide them with a chance for asking questions.

[17] There must, of course, be a warning to faculty that this is unprofessional behavior that will be treated very seriously in personnel evaluation.

[18] This was a serious problem in Berkeley in the late sixties, when some of the radical left got control of the student evaluation process. There are various ways to detect and control such conspiracies, but the mere possibility of them is a strong reason for allowing appeals against student ratings.

[19] In a sophisticated system, there could be reasons for unannounced visits, but in general it is better to encourage the presence of those who wish to put in rating forms.

[20] This requires that the evaluators: (i) know the *current* enrollment figure: (ii) ensure that those present do fill out the form, if they have the authority to do this; (iii) do something about absentee ballots. My inclination is not to accept less than 95% return rates, and I have found this to be achievable; certainly anything below 80% is very hard on validity.

[21] Too early rules out reactions to grading procedures and feedback; too late loses input from those who drop the course because they think it's bad. I prefer the following arrangement. Distribute forms on the first day, with envelopes, and request that those dropping the course fill them in before or after doing so, using campus mail to return them. Request returns after the mid-term tests have been handed back and use them to improve the course. Arrange for the final to be given on the last day of class. Require attendance at the time scheduled for the final in the exam period as a condition to getting a grade. At that session, return the exams for immediate study, with comments, or at least a demonstration or examples of answers that would have received an A. Allow questions and then protests about the grades. Collect the exams for the archives and distribute the rating forms. Collect the forms, and check off that every student has turned one in, as well as returned an exam. This procedure has a salutary effect on any headroom problems, but it is also a good way to avoid wasting the final exam as a learning experience.

[22] Bi-modal distributions tell a very different story from bell-shaped curves with the same mean.

[23] "Duties-Based Teacher Evaluation" in Vol. 1, No. 3 of *Journal of Personnel Evaluation in Education*, 1987.

This list readily generates student rating forms to match the parts of it that students can rate and converts easily to cover tertiary duties.

[24] It might be mentioned that one hardly ever sees tertiary institutions using the hard-won discipline of program evaluation to evaluate their programs. Even the use of alumni to help with the needs assessment is extremely rare. This is a replay of colleges not using what was long known about how to evaluate teaching. The track record is quite different on admissions; presumably a cynic would say the explanation is that the exercise does not involve looking critically at themselves.

[25] I do not here discuss the very serious technical objections to the use of gain scores that have been raised by measurement specialists, including, for example, the problem of regression to the mean, and special problems of reliability and validity that arise with any tests. The problems discussed here are conceptual problems that apply even if the technical problems can be dismissed, as is possible in cases of massive gains.

[26] The "Harvard fallacy" is the fallacy of supposing that the teaching at Harvard must be good because its graduates do very well in later life in proportion to their numbers. All that one can infer from that data is that Harvard does not inflict permanent brain damage–usu*ally*. The rest of the trick lies in selecting a talented entering class and not getting in the way of their use of the library, labs, peer tutoring, and family influence or brand-name reverence. The contribution from the faculty, if any, is the residue after factoring out the non-faculty influences on the academic side, and the effect of the "old boys network" and brand recognition on job selection and promotion. While Harvard is demonstrably a great university, it is certainly not demonstrably a great teaching university, just a superbly equipped one.

[27] Self-evaluation is, of course, not evaluation without input from others, but evaluation that is self-initiated and directed; the use of anonymous student ratings is a *sine qua non* of any self-evaluation by teachers. Naturally, any systematic process for the evaluation of faculty should reward serious self-evaluation and systematic self-development based on it. In teaching, as in any profession, the combination of these two practices are the hallmarks of professionalism, the minimum standards for social tolerance of the practitioner.

[28] That is, all the teachers being compared may be weak, or the worst of them may be very good and the others better, etc. Student ratings, especially in upper secondary and post-secondary contexts, are based on a much wider range of comparisons than this, which provides a better approximation to criterion-referencing.

[29] It is hard to see why student evaluations, if supported by a substantial effort at prior training for the students (something of considerable educational merit in its own right), would not be sound well down into lower primary. But we need more experimentation in that area, as well as legitimation by leadership use.

[30] There are also conceptions of teaching that make them less important, perhaps–in extreme cases–completely inappropriate, for example, the conception of teaching as creating a climate for learning rather than as transmitting it.

[31] In a common situation, there are one or two visits, usually for less than a complete period, out of 30–100 class meetings. Given the way in which individual class meetings vary, as every instructor knows, both for idiosyncratic reasons and as the term goes on, as the first or the final test looms or as the topics vary in interest, this cannot be thought of as an adequate sample. One should also take into account the way in which visitors' ratings can change as they come to "see through" the teacher's style, a process which may continue over a large number of visits (Studies at the University of California at Davis make clear that this effect can be very substantial, and I know of none that found it to be small; Wilson et al. *College professors and their impact on students*, Wiley, 1975).

[32] There is some reduction of impact of these criticisms if we videotape all sessions of a course and select a substantial random sample of these to evaluate. But the cost and connotations of this approach are worrisome, and we lack experience with it.

[33] There are some extreme cases where the visitor can make a reliable judgement of pedagogical skill. The validity of these judgements is skewed along the merit axis; it is easier to identify deep trouble than great merit. But, of course, the students can make the same judgement with the same or greater validity, so the visit is unnecessary.

[34] The teaching materials and test or project work done by the students will better serve that purpose.

[35] Students are similarly in a uniquely strong position to rate the presence of immediately identifiable benefits from the material and skills acquired from a teacher, but this is arguably not crucial in evaluating teaching merit. However, it can be very useful for formative evaluation, if you are or are considering spending some time persuading the class of the importance of the subject to them. And it is significant for many discussions of a department's curriculum. Hence, it should be considered for inclusion on a "general purpose" form. This is a different matter from rating the eventual or long-term value of the course to the student's e.g. professional needs, about which the students are not in a good position to pass judgement.

[36] It seems sensible to use student rating forms in a two-stage procedure. In the first stage, a good summative-valid form is used, administered in a summative-valid way (security procedures, etc.). Only if someone does so badly on that stage as to jeopardize their job or offend their own sense of satisfaction with the quality of their work should they then move to the use of a second form.

The second form can simply call for a more detailed analysis of the duties (expanding on the type of questions mentioned in D). But it could also–if the teacher preferred–ask the student to answer questions about style (as in E), so that the information provided by the style literature, as to what works best for many teachers, can be invoked. (Thanks to John Hattie for pointing out the legitimacy of category E in this list.)

[37] The usual distinction here is between merit and worth (or suitability). Both are legitimate in the evaluation of faculty, within limits. It is worth and not just merit that leads to a position being advertised in the first place, and to the verdict of redundancy. Worth can sometimes be used, properly, to justify gender preference; and is often used, improperly, to rationalize political discrimination. A longer discussion of it is provided in a paper on teacher selection in the forthcoming *Handbook of Teacher Evaluation, Second Education* (editors, J. Millman et al., Sage, 1988).

[38] The Colorado School of Mines, a prestigious and much-beloved engineering school.

[39] "Validity in Personnel Evaluation" in *Journal of Personnel Evaluation in Education*, vol. 1, no. 1, 1987. In a sense, the genesis of the present article is that the *JPEE* one just cited destroys the usual argument for the use of student ratings, thus creating the need for rethinking their justification. Doesn't this argument invalidate the use of exams for tertiary admission, since it is usually thought that the justification for that procedure is the predictive validity of the results against success in college? The use of what appear to be statistical indicators can be justified in the case of using Public Service Board exams, or tertiary admission testing, by the absence of any other way to get valid evaluations of all members of a very large and very widely dispersed group; by the fact that the exam, if properly designed, is close to being a work sample; and by the fact that it is also a test of prerequisite skills and knowledge. None of these mitigating considerations apply to rating accessible personnel, with accessible track records, using style variables. (Breath-alcohol and polygraph testing raise similar questions; but the first can be given an indirect defense whereas the second cannot.)

[40] This viewpoint was expressed in the late sixties by the radical left in the bitter phrase "the student as nigger."

[41] Provided, of course, that the results of the student evaluations *do* carry weight in the decisions made. A monitor from student government on the committee is desirable here, and appropriate controls of anonymity are possible.

[42] A prompt is simply a hint as to something that the respondent *may* wish to underscore as significant, or comment on, or simply take into account when selecting an answer to the One Big Question, but which does not *require* a response. We get more than 50 prompts in small print on our one-page four-question rating form, but the average time to complete is still around 3 minutes compared to 10–15 for a 50-question form. Other advantages are: (i) coding for summative evaluation is simpler (and it's no more difficult for formative evaluation); (ii) the integration of multiple considerations is done by the respondent, not by the evaluator who lacks good reasons for any particular relative weighting; (iii) it uses relatively little paper, time, and computer processing. Readers are welcome to a copy of the form we use upon request.

[43] Which are based on the idea that if someone is not doing well at teaching, they must be "going about it the wrong way," i.e., they need to have their *teaching style* improved. The correct approach would be to see, first, if their discharge of *teaching obligations* needed improvement. There are many such obligations that need improvement in most teachers and can easily be improved as *one can immediately show*. You can't get good agreement between two independent observers as to the best style, but even if you could, the size of the benefits are not demonstrable.

[44] Bad temper is an obvious example, but excessive repetition, reading from texts, and the failure to ask questions except in the presence of visitors are others.

[45] Certainly, since the visitor cannot avoid seeing the style features and, hence, cannot guarantee not being influenced by them, visitors cannot be used for input on personnel decisions.

[46] As well as possible benefits for campus morale, as already mentioned.

[47] The usual problem of initial faculty opposition to weighting student ratings sometimes turns into the opposite one; after a few years' use, there is a tendency towards overweighting student evaluations. A major source of benefit comes from improved faculty self-evaluation, resulting from the need to face up to and discuss the student ratings of their work, which in other institutions will only occur to those who actively and independently undertake to get their performance rated by students.

[48] Some suggestions about such a system are provided in "Summative Teacher Evaluation" in *Handbook of Teacher Evaluation*, ed. J. Millman, (Sage, 1981).