Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 6, May 2007 ISSN 1531-7714

RateMyProfessors.com versus formal in-class student evaluations of teaching

Theodore Coladarci & Irv Kornfield
University of Maine

Using data for 426 instructors at the University of Maine, we examined the relationship between RateMyProfessors.com (RMP) indices and formal in-class student evaluations of teaching (SET). The two primary RMP indices correlate substantively and significantly with their respective SET items: RMP overall quality correlates r = .68 with SET item, *Overall, how would you rate the instructor?*; and RMP ease correlates r = .44 with SET item, *How did the work load for this course compare to that of others of equal credit?* Further, RMP overall quality and RMP ease each correlates with its corresponding SET factor derived from a principal components analysis of all 29 SET items: r = .57 and .51, respectively. While these RMP/SET correlations should give pause to those who are inclined to dismiss RMP indices as meaningless, the amount of variance left unexplained in SET criteria limits the utility of RMP. The ultimate implication of our results, we believe, is that higher education institutions should make their SET data publicly available online.

Student evaluations of teaching (SET) are part and parcel of the university classroom experience. This phenomenon can be traced back to the seminal work of Purdue University psychologist Herman Remmers, who, in the 1920s, pioneered the use of rating scales for evaluating instructors. Remmers and his associates arguably developed the first student evaluation form (Remmers, 1927), and they dominated SET research through the 1950s (Centra, 1993, pp. 49-51).

Today, SET are used formatively to effect instructional improvement as well as summatively to inform decisions regarding promotion and tenure. There is a wealth of empirical research on the reliability and validity of SET, a literature that has been summarized well by others (e.g., Algozzine et al., 2004; Cashin, 1995; Costin, Greenough, & Menges, 1971; D'Apollonia & Abrami, 1997; Greenwald, 1997; Marsh & Roche, 1997; McClatchy, 1997; McKeachie, 1997; Wachtel, 1998). Although faculty sentiments regarding the utility of SET are far from consentaneous (e.g., Nasser & Fresko, 2002), we agree with the position articulated by Centra (2003, pp. 495-496):

> *"No method of evaluating college teaching has been researched more than student evaluations, with well over 2,000 studies referenced in the ERIC system. The preponderance of these study results has been positive, concluding that the evaluations are: (a) reliable and stable; (b) valid when compared with student learning and other indicators of effective teaching; (c) multidimensional in terms of what they assess; (d) useful in improving teaching; and (e) only minimally affected by various course, teacher, or student characteristics that could bias results."*

Or in the more parsimonious words of McKeachie (1997, p. 1219): "student ratings are the single most valid source of data on teaching effectiveness." This is not to say that SET are unimpeachable (e.g., see Johnson, 2002). Rather, when instruments are properly constructed and the resulting data thoughtfully considered, SET can be an important source of information for both improving teaching and informing personnel decisions.

Relative to the long history of SET, the online publication of student ratings of college instructors, such as RateMyProfessors.com (RMP), is a recent development. Established in 1999, RMP allows students to post 5-point ratings on individual instructors with respect to "helpfulness" ("Is this professor approachable, nice and easy to communicate with? How accessible is the professor and is he/she available during office hours or after class for additional help?"), "clarity" ("How well does the professor teach the course material? Were you able to understand the class topics based on the professor's teaching methods and style?"), and "easiness" ("Is this class an easy A? How much work do you need to do in order to get a good grade?"). Helpfulness and clarity are averaged by RMP to form "overall quality." Students also can rate their prior interest in the class, indicate whether they believe the instructor is "hot" or "not hot" (in reference to instructor "appearance"), and provide general comments.

Averaged across student posts, these ratings are summarized by instructor in two locations on the RMP site. First, on the instructor's "scorecard" one finds the total number of posts, average easiness, average helpfulness, average clarity, and average overall quality. A hotness total also is provided, which is the sum of hot ratings (coded +1) and not-hot ratings (coded -1). Where a student opts not to rate the instructor on this dimension, a value of zero is figured into the hotness calculation. A hotness total greater than 0 indicates that hot ratings outnumber not-hot ratings, by a margin equal to the hotness total. For online presentation, RMP converts negative hotness totals to 0. The second location is the list of all instructors at the institution or in a particular department, where RMP reports for each instructor the total number of posts, overall quality, and ease. If the instructor's hotness total is greater than 0, this achievement is acknowledged by the display of a red chili pepper. [1]

RELATED RESEARCH

Several investigators have examined the statistical associations among the various RMP indices.[2] Felton, Mitchell, and Stinson (2004) downloaded RMP data for professors at the 25 institutions having the most student posts, resulting in a sample of 3,190 instructors representing 65,678 posts; only instructors having at least 10 posts were included. With the instructor as the unit of analysis, these researchers obtained a correlation of $r = .61$ between easiness and quality. This value is not dissimilar to the correlation of $r = .50$ reported by Davison and Price (2006) in their analysis of RMP indices for instructors at Appalachian State University. Interestingly, the easiness/quality correlation in the Felton et al. (2004) study was smaller for the 481 "sexy" instructors (pepper possessors) than it was for the 2,709 "non-sexy" instructors ($r = .46$ and .61, respectively), prompting these researchers to conclude that "the sexier the instructor, the more difficult his or her class can be while obtaining high-marks on student evaluations" (p. 106).[3]

In an analysis involving non-sexy instructors only, Felton et al. (2004) found also that the correlation between easiness and quality grew stronger when based on increasingly more selective subsets of instructors regarding number of posts. For example, $r = .61$ for the 2,709 instructors having at least 10 posts, $r = .68$ for the 454 instructors having at least 30 posts, $r = .76$ for the 106 instructors having at least 50 posts, and $r = .85$ for the 37 instructors having at least 70 posts. Finally, these researchers reported that sexiness, a variable they constructed by dividing the hotness total by the total number of posts (J. Felton, personal communication, February 1, 2007), correlated significantly with both quality ($r = .30$) and easiness ($r = .17$). In a follow-up study (Felton, Koper, Mitchell, & Stinson, 2006), these correlations increased to $r = .64$ and $r = .39$, respectively, when based on a sexiness variable that allowed for the RMP-suppressed negative values (which these researchers obtained from the RMP president).

Riniolo, Johnson, Sherman, and Misso (2006) similarly found a relationship between hotness and quality. These researchers contrasted "attractive" (pepper) and "unattractive" (no pepper) instructors at the four universities having the most RMP posts; only instructors having at least 25 posts were included. Controlling for academic department, Riniolo et al. found that attractive instructors at each university were rated more favorably than their putatively unattractive counterparts—a finding that held for males and females

[1] For example, a red chili pepper would be awarded in each of these quite different scenarios (involving a class of 21 students): (a) all students rated the instructor as hot, (b) 10 students rated the instructor as not hot and 11 students rated the instructor as hot, and (c) one student rated the instructor as hot and 20 students provided no rating whatsoever.

[2] To distinguish between instructor-level and student-level RMP information available online, we use *indices* when referring to the averages reported for the instructor and *ratings* when referring to the contributions of individual students.

[3] There also may be a statistical artifact at play due to restricted variability among sexy instructors—who, relative to their pepperless peers, had higher means on both quality and easiness.

alike. The corresponding effect sizes were considerable, ranging from .68 to 1.32 for males and .95 to 2.33 for females.

As noted above, RMP fashions its measure of overall quality by averaging helpfulness and clarity. Felton et al. (2006) and Kindred and Mohammed (2005) both report a high correlation between helpfulness and clarity (r = .94 and .86, respectively), which provides warrant for the union of these two items as a composite.

Although we focus above on the intercorrelations among RMP indices, some researchers also folded student comments into the mix. In their multi-institution study, for example, Kindred and Mohammed (2005) coded posted comments according to various themes (e.g., instructor "intelligence," "competence," "personality") and, in a student-level analysis, then correlated RMP ratings with variables derived from these codings. For instance, instructor competence correlated r = .84 and .85, respectively, with helpfulness and clarity ratings. Taking a somewhat more qualitative tack, Felton et al. (2004) categorized instructors according to the quadrant they occupied in the easiness-quality scatterplot and then extracted illustrative comments regarding such instructors. For example, a high-quality low-easiness instructor drew the following student comment: "This class is a lot of work, but it's certainly not impossible. He really knows his stuff and if you ask for help, you'll get it. One of the best professors I've had yet." In contrast, a student posted this comment for a low-quality high-easiness instructor: "Easy grader, lectures are boring, and you don't really learn anything."

PRESENT STUDY

From the modest RMP research literature, we see that instructors rated more highly on easiness also are rated more highly on overall quality in comparison to their more-difficult counterparts. Further, instructors deemed hot have somewhat higher ratings on both overall quality and easiness when compared to those who do not enjoy this distinction.

To our knowledge, no one has examined the correspondence between RMP indices and SET—i.e., the formal, in-class student evaluations of teaching solicited at the instructor's institution. Knowing the concurrent validity of RMP indices in this regard would be valuable information indeed. With over 6,750,000 ratings of over one million instructors across more than 6,000 schools, RMP enjoys considerable use. Anecdotal accounts as well as more systematic study (e.g., Davison & Price, 2006; Kindred & Mohammed, 2005) suggest that students consult RMP to inform course-taking decisions. Such consultation is warranted if RMP indices were shown to correlate highly with SET. On the other hand, using RMP in this manner would be ill-advised if there were poor correspondence between RMP indices and SET. In short, the present study is a first step in throwing empirical light on this concern.

Using SET data from the University of Maine (UMaine), we address two fundamental questions: First, how do the RMP indices for an instructor correlate with the instructor's SET ratings? This speaks to the concurrent validity of RMP indices, where the institution's SET serve as the criterion variables. Second, does the general magnitude of these concurrent validity coefficients depend on the number of RMP posts an instructor has? In other words, are RMP indices equally trustworthy whether based on sentiments of the few or the many? Our primary interest is in the overall quality and ease RMP indices, although we examine the pepper distinction as well.

METHOD

Data Sources

For each UMaine instructor present on the RMP site in mid-December 2006, we recorded the overall quality and ease indices, the presence (1) or absence (0) of a red chili pepper, and the number of posts. We included adjunct and fulltime instructors alike, but excluded teaching assistants and graduate-level instructors. We did not impose an inclusion criterion regarding minimum number of RMP posts. The resulting data file was given to the Associate Director of the UMaine Office of Institutional Studies, who added SET data for each instructor.

The 29 items on the UMaine SET form each rests on a five-point scale (see Appendix A). For the present study, we selectively reversed item scales so that higher values always represent a more desirable characteristic (greater clarity, fairer grading procedures, much intellectual discipline required, etc.). The two exceptions are Item 17 ("How was the pace at which the materials in the course were covered?"), where 1 = "too fast" and 5 = "too slow," and Item 20 ("How did the work load for this course compare to that of others of equal credit?"), where 1 = "much heavier" and 5 = "lighter." For each instructor, the Associate Director provided each of the 29 items averaged across all undergraduate courses the instructor taught between spring 2000 and spring 2006 (to correspond roughly to the operation of RMP). He then removed identifying information from the records and forwarded the merged database to us for analysis. We eliminated one instructor whose number of postings exceeded the total enrollment, resulting in a final data base comprising 426 instructors.

Analyses

We conducted several analyses to examine the concurrent validity of the primary RMP indices—overall quality and ease—as well as to explore the correlates of the dichotomous pepper index. First, we correlated each of the three RMP indices with Item 13 ("Overall, how would you rate the instructor?") and Item 20 ("How did the work load for this course compare to that of others of equal credit?")—the SET items we believed were most relevant to RMP overall quality and RMP ease, respectively. Although Item 13 and Item 20 are our primary item-level interest, for their descriptive value we also report correlations between the RMP indices and the remaining 27 SET items. We also correlated the RMP indices with the three orthogonal factors that obtained from our factor analysis, described below, of the SET item data. Second, we conducted multiple regression analyses to determine whether the strength of association between an RMP index and SET criterion variable changes when the two remaining RMP indices are statistically held constant. Third, by calculating the RMP/SET correlations separately for instructors having relatively few versus relatively many RMP posts, we determined whether the strength of association between the RMP indices and SET criteria is related to the number of RMP posts on which the indices are based.

RESULTS

We began by conducting a principal components factor analysis of the SET item data, aggregated by instructor across courses and years. A three-factor solution, using varimax rotation, accounted for 76.34% of the total variance. The first factor, which we call Instructor (47.11%), is a general factor largely capturing perceptions of the instructor and instruction. For example, the highest loading item on this factor is Item 13, "Overall, how would you rate the instructor?" (1 = *below average*, 5 = *excellent*). Loading almost equally highly is Item 22, "What is your overall rating of this course?" (1 = *poor*, 5 = *excellent*). The second factor, which we have labeled Assessment (19.56%), reflects sentiments regarding the instructor's assessment practices. Here, the highest loading item is Item 29, "Overall, how would you rate the examination procedure?" (1 = *poor*, 5 = *excellent*). Finally, we call the third factor Facile (9.67%), and it reflects the perceived ease of the course. One of the highest-loading items on this factor is Item 20, "How did the work load for this course compare to that of others of equal credit?" (1 = *much heavier*, 5 = *much lighter*); another is the negatively loading Item 21, "How much intellectual discipline was required in this course?" (1 = *very little*, 5 = *very much*).

Table 1 reports the means and standard deviations for the 29 UMaine SET items, grouped according to the primary factor on which they load and in descending order of magnitude. Correlations between the RMP indices and all SET criterion variables are presented as well.

Table 1. Means, Standard Deviations, and RMP/SET Correlations (*n* = 426).

		Correlation with RMP index[a]		
SET Criterion	*M* (*SD*)	overall quality	ease	pepper
Instructor [b]	0 (1.00)	.57	.10	.20
Item 13 Overall, how would you rate the instructor? (*below average . . . excellent*; .91[c])	4.30 (.44)	.68	.22	.26
Item 22 What is your overall rating of this course? (*poor . . . excellent*; .89)	4.00 (.44)	.62	.19	.23
Item 6 How concerned was the instructor for the quality of his or her teaching? (*unconcerned . . . very concerned*; .88)	4.25 (.39)	.65	.18	.22
Item 4 How clearly did the instructor present ideas and theories? (*often unclear . . . very clear*; .88)	4.12 (.46)	.65	.23	.23

Table 1 (continued). Means, Standard Deviations, and RMP/SET Correlations (*n* = 426).

		Correlation with RMP index		
SET Criterion	*M* (*SD*)	overall quality	ease	pepper
Item 16 Did you develop significant skills in the field as a result of taking this course? (*very little . . . very much*; .86)	3.90 (.44)	.57	.08	.18
Item 2 How clearly were the objectives of the course presented? (*unclear . . . very clear*; .85)	4.23 (.40)	.64	.21	.23
Item 3 How enthusiastic was the instructor about the subject? (*very little . . . very much*; .84)	4.58 (.32)	.48	.09	.17
Item 12 How genuinely concerned was the instructor with students' progress? (*unconcerned . . . very concerned*; .84)	4.18 (.41)	.61	.24	.21
Item 11 Did the instructor inspire confidence in his or her knowledge of the subject? (*little . . . very much*; .84)	4.58 (.32)	.56	.10	.16
Item 14 Were class meetings profitable and worth attending? (*not usually . . . always*; .82)	4.16 (.45)	.55	-.04	.19
Item 7 How orderly and logical were the instructor's presentations of the material? (*not at all . . . very much*; .81)	4.22 (.40)	.63	.18	.23
Item 15 How would you rate the subject matter of this course? (*uninteresting . . . very interesting*; .79)	4.00 (.50)	.44	.10	.18
Item 8 How open was the instructor to other viewpoints? (*often closed . . . very open*; .78)	4.38 (.34)	.63	.29	.26
Item 9 Did the instructor show respect for the questions and opinions of the students? (*rarely . . . always*; .78)	4.57 (.29)	.63	.29	.24
Item 5. How much were students encouraged to think for themselves? (*very little . . . very much*; .77)	4.25 (.36)	.42	-.01	.15
Item 19 Were students required to apply concepts to demonstrate understanding? (*very little . . . very much*; .73)	4.22 (.34)	.42	-.08	.11
Item 1 How prepared was the instructor for class? (*often unprepared . . . well prepared*; .71)	4.49 (.33)	.53	.05	.14
Item 10 How often were examples used in class? (*rarely . . . always*; .70)	4.56 (.28)	.53	.18	.16
Item 18 What is the overall rating of the primary textbook(s)? (*poor . . . excellent*; .54)	3.60 (.39)	.32	.12	.10

Table 1 (continued). Means, Standard Deviations, and RMP/SET Correlations (*n* = 426).

		Correlation with RMP index		
SET Criterion	*M* (*SD*)	overall quality	ease	pepper
Assessment	0 (1.00)	.35	.28	.14
Item 29 Overall, how would you rate the examination procedure? (*poor . . . excellent*; .78)	3.99 (.37)	.57	.38	.24
Item 28 How fair were the grading procedures? (*unfair . . . completely*; .78)	4.28 (.34)	.54	.33	.22
Item 25 Did the instructor let you know what he or she expected on tests and assignments? (*not clear . . . very clearly*; .77)	4.09 (.41)	.56	.37	.24
Item 26 Did exams reflect the important aspects of the course? (*very little . . . very much*; .76)	4.21 (.32)	.55	.22	.23
Item 27 How clear were examination questions? (*unclear . . . very clear*; .73)	4.04 (.37)	.53	.34	.23
Item 23 How promptly were assignments and tests returned? (*too slow . . . very prompt*; .71)	4.10 (.49)	.27	.12	.06
Item 24 Could tests be completed in the allotted time? (*rarely . . . always*; .70)	4.39 (.36)	.32	.28	.12
Facile	0 (1.00)	.08	.51	.10
Item 21 How much intellectual discipline was required in this course? (*very little . . . very much*; -.84)	3.89 (.35)	.17	-.42	-.03
Item 20 How did the work load for this course compare to that of others of equal credit? (*much heavier . . . lighter*; .83)	2.71 (.41)	.10	.44	.10
Item 17 How was the pace at which the materials in the course were covered? (*too fast . . . too slow*; .81)	2.76 (.19)	.12	.44	.08
	M (*SD*):	3.69 (.90)	3.10 (.80)	.17 (.38)

[a] Correlations of $|r| \geq .10$ are statistically significant ($p < .05$, two tailed). [b] Factor. [c] Factor loading.

.

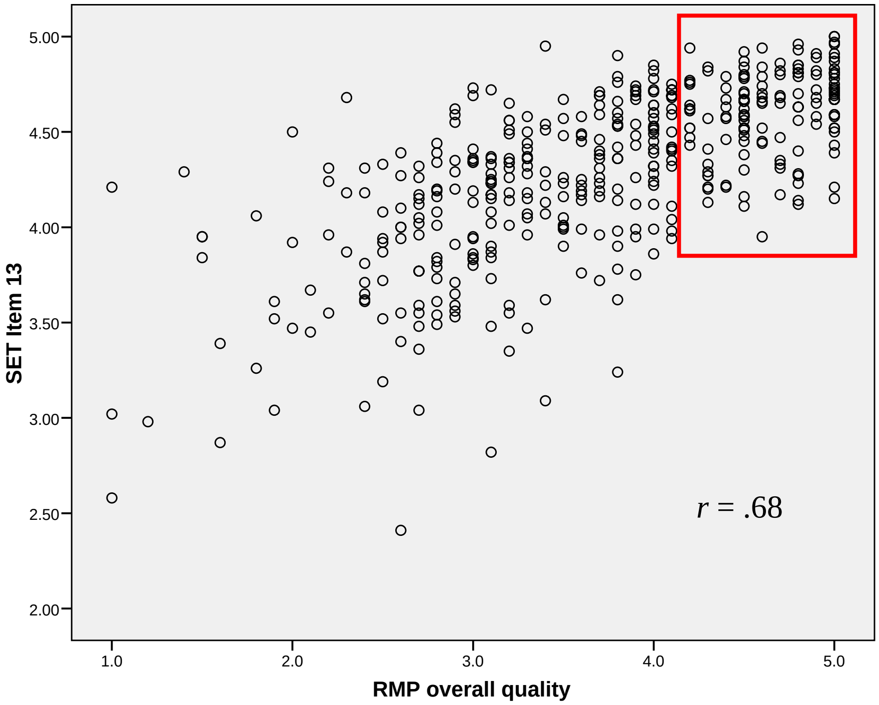


Figure 1. The relationship between RMP overall quality and SET Item 13 (*n* = 426), with area of high concordance highlighted.

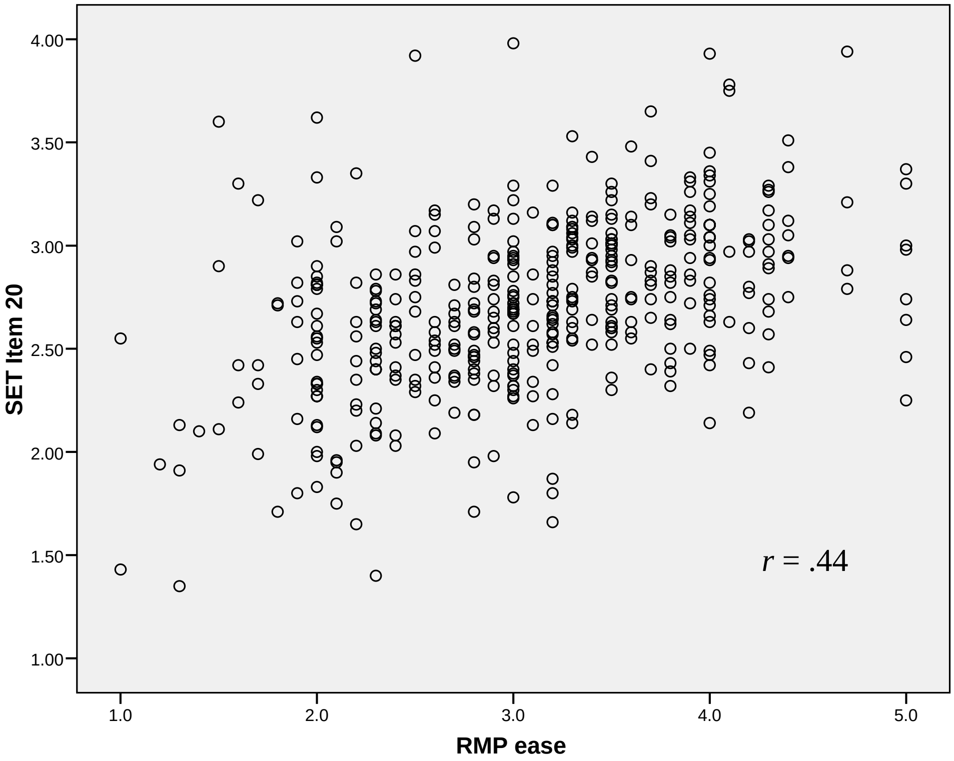


Figure 2. The relationship between RMP ease and SET Item 20 (*n* = 426).

Correlations Among RMP Indices

Before presenting the primary results of this investigation, we briefly consider the correlations among the RMP indices. The two primary indices, RMP overall quality and RMP ease, correlate $r = .40$ ($p < .001$). Consistent with the RMP findings of Felton et al. (2004) and Davison and Price (2006), instructors who are rated more highly in overall quality also are seen as being somewhat easier. Further, the dichotomous pepper index is weakly associated with ease ($r = .17$, $p = .001$) and somewhat more substantively associated with overall quality ($r = .34$, $p < .001$). Regarding this latter finding, the means on RMP overall quality for the pepper and pepperless instructors are, respectively, 4.37 ($SD = .62$) and 3.55 ($SD = .89$), a difference that is statistically significant ($t[424] = 7.55$, $p < .001$) and equivalent to an effect size of +.97. Thus, pepper instructors are roughly one standard deviation higher on overall quality compared to their pepperless peers. This general finding, too, is consistent with RMP results reported by others (Felton et al., 2004, 2006; Riniolo et al., 2006).

Correlations Between RMP Indices and SET Criterion Variables

How do RMP indices correlate with the UMaine SET criteria? RMP overall quality correlates $r = .68$ ($p < .001$) with its SET item counterpart, "Overall, how would you rate the instructor?" (Item 13). As the scatterplot in Figure 1 reveals, there is particularly high concordance for instructors at the upper end of each measure (see superimposed box). In contrast, the data points in this figure fan out markedly for remaining values of RMP overall rating.

RMP ease correlates $r = .44$ ($p < .001$) with its SET item counterpart, "How did the work load for this course compare to that of others of equal credit?" (Item 20). Statistical significance notwithstanding, this positive association betrays considerable scatter (see Figure 2).

We find further convergence when SET factors serve as the criteria: RMP overall quality correlates $r = .57$ ($p < .001$) with Instructor, and RMP ease correlates $r = .51$ ($p < .001$) with Facile. In this light, of course, it should not be surprising to observe in Table 1 that RMP overall quality demonstrates, with few exceptions, high correlations with Instructor-clustered items, as does RMP ease with respect to Facile-clustered items.

There is evidence of discriminant validity as well: RMP overall quality is unrelated to Facile ($r = .08$, $p = .115$) and its association with Item 20 is negligible ($r = .10$, $p = .041$). And while RMP overall quality correlates significantly with Assessment ($r = .35$, $p < .001$), here too the strength of association is less than that with either Item 13 (.68) or Instructor (.57). Similarly, RMP ease correlates less with Item 13 ($r = .22$, $p = .001$), Instructor ($r = .10$, $p = .035$), and Assessment ($r = .28$, $p < .001$) than it does with either Item 20 (.44) or Facile (.51).

As for the RMP pepper index, this dichotomous variable demonstrates a statistically significant, albeit weak, relationship with Item 13 ($r = .26$, $p < .001$) and Instructor ($r = .20$, $p < .001$); its relationship is smaller still with Item 20 ($r = .10$, $p < .05$), Facile ($r = .10$, $p < .05$), and Assessment ($r = .14$, $p < .01$). An examination of the remaining pepper-related correlations in Table 1 confirms that this distinction bears little relationship to the SET criterion variables.

Multiple Regression Analyses

As seen above, the three RMP indices are intercorrelated. Does the strength of association between an RMP index and SET criterion variable change when the two remaining RMP indices are statistically held constant? Here, we focus on the criterion variables Item 13, Item 20, Instructor, Assessment, and Facile, each of which was regressed on the three RMP indices. Intercorrelations among all variables are shown in Table 2; the standardized regression coefficients and semipartial correlations (*sr*) for each of the five equations appear in Table 3.

RMP overall quality and RMP ease continue to demonstrate association with their respective SET criterion variables even when statistical control is exercised (Table 3). With RMP ease and the pepper index held constant, RMP overall quality is related to both Item 13 ($\beta = .70$, $p < .001$) and Instructor ($\beta = .63$, $p < .001$). By squaring the semipartial correlation for RMP overall quality, one obtains the percentage of variance in each dependent variable that is explained by RMP overall quality alone: 37% and 30%, respectively, for Item 13 and Instructor. Similarly, RMP ease is related to both Item 20 ($\beta = .48$, $p < .001$) and Facile ($\beta = .56$, $p < .001$) when RMP overall quality and the pepper index are held constant. Here, RMP ease explains 19% and 27% of the variance, respectively, in Item 20 and Facile. RMP overall quality and RMP ease each remains associated with Assessment, although the magnitude of association is rather modest: $\beta = .28$ ($p < .001$) in the case of overall quality, and $\beta = .17$ ($p = .001$) for ease.

Table 2. Intercorrelations among variables in multiple regression analyses (*n* = 426).

	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) RMP overall quality	.40	.34	.68	.10	.57	.35	.08
(2) RMP ease		.17	.22	.44	.10	.28	.51
(3) RMP pepper			.26	.10	.20	.14	.10
(4) Item 13				.05	.91	.36	.02
(5) Item 20					-.01	.15	.83
(6) Instructor						.00	.00
(7) Assessment							.00
(8) Facile							

Note. Correlations of $\left| r \right|$ ≥ .10 are statistically significant (*p* < .05, two tailed).

Table 3. Multiple regression results (*n* = 426).

SET Criterion	RMP Indices	β	*sr*	*p*
Item 13 (Adj. R^2 = .46)				
	overall quality	.70	.61	< .001
	ease	-.07	-.07	.067
	pepper	.03	.03	.473
Item 20 (Adj. R^2 = .20)				
	overall quality	-.12	-.10	.021
	ease	.48	.44	< .001
	pepper	.06	.06	.174
Instructor (Adj. R^2 = .35)				
	overall quality	.63	.55	< .001
	ease	-.15	-.14	< .001
	pepper	.01	.01	.876
Assessment (Adj. R^2 = .14)				
	overall quality	.28	.24	< .001
	ease	.17	.15	.001
	pepper	.02	.01	.754
Facile (Adj. R^2 = .27)				
	overall quality	-.17	-.15	< .001
	ease	.56	.52	< .001
	pepper	.07	.07	.114

The statistically significant, if weak, association between the pepper index and SET criterion variables, noted above, disappears altogether when RMP overall quality and RMP ease are held constant.

Though negligible, two additional results warrant mention. First, RMP overall quality is *inversely* related to Item 20 (β = -.12, *p* = .021) and Facile (β = -.17, *p* < .001). That is, once RMP ease and RMP pepper are taken into account, lower RMP ratings of overall quality correspond to slightly higher SET ratings regarding easiness: Easier courses are perceived on RMP to be slightly lower in overall quality. Second, and in a similar vein, RMP ease is inversely related to Instructor (β = -.15, *p* = .001). Thus, with RMP overall quality and the pepper index held constant, higher RMP ratings of ease correspond to slightly lower SET ratings of the instructor: Superior instructors are perceived on RMP to be slightly more difficult.

Number of RMP Posts

For these 426 instructors, the number of posts ranges from 1 to 95, with a mean of 14.40 (*SD* = 15.52) and a median of roughly 9. Does the magnitude of association between RMP indices and SET criterion variables depend on the number of RMP posts the instructor has? In other words, are RMP indices more trustworthy when based on more posts? We approached this question descriptively by examining RMP/SET correlations for two subgroups of instructors: those above, versus those below, the median number of RMP posts. In an exploratory spirit, we also formed subgroups in reference to the mean number of RMP posts, which, given the positive skew of this distribution, provides a more extreme upper group on this variable.

We limited these analyses to RMP overall quality, RMP ease, and the SET criterion variables Item 13, Item 20, Instructor, and Facile. Although we report all 24 resulting correlations, our primary interest is in the correlation of RMP overall quality with Item 13 and Instructor, and the correlation of RMP ease with Item 20 and Facile.

When subgroups are formed in reference to the median number of RMP posts, the difference in RMP/SET correlations from one group to the other is inconsistent (see Table 4). For example, the correlation between RMP overall quality and Item 13 is somewhat smaller when based on instructors having more than nine posts (.70 vs. .67), the correlation between RMP ease and Item 20 is somewhat larger (.42 vs. .48), and comparable results obtain with respect to RMP overall quality and Instructor (.57 vs. .58) and RMP ease and Facile (.50 vs. .51). However, a clearer pattern emerges when subgroups are based on the mean number of RMP posts, where we see that each RMP/SET correlation is larger for instructors having more posts. To be sure, the difference is negligible in the case of RMP overall quality and Item 13 (.68 vs. .71), although somewhat larger for RMP overall quality and Instructor (.56 vs. .61) and larger still for RMP ease and Item 20 (.40 vs. .57). It is their consistency in direction, not absolute magnitude, that makes these differences noteworthy.

DISCUSSION

Before discussing these results, we briefly acknowledge the limitations of the present study. First and foremost, we used data for a single institution. Although we have no reason to believe that the import of our results is limited to the University of Maine, one nevertheless must be cautious in making generalizations to other institutions. On a related note, replication studies are needed: It remains to be seen whether similar results are obtained at institutions that differ from ours in size, RMP participation, SET statistics (e.g., central tendency and variability), admissions criteria, and other variables that may moderate the RMP/SET relationship. Second, because RMP indices are averaged over years and courses for each instructor, we followed suit in constructing the SET data base. Insofar as an instructor's effectiveness may vary over time and/or courses, aggregating data in this fashion likely resulted in lower RMP/SET correlations than would obtain had we created RMP indices and corresponding SET criterion variables that were time- and course-specific. Third, RMP has no quality-control mechanisms to prevent multiple posts from a student (for the same instructor and course) or, arguably worse, to prevent nonstudents from posting. The RMP slogan notwithstanding ("Where STUDENTS do the grading!"), instructors are known to make RMP contributions of their own—sometimes playfully, sometimes self-servingly (e.g., Montell, 2006). Even when students in fact do the grading, their posts can be provided anytime during—or after—the course. Although the likely effect of these quality-control problems on RMP/SET correlations is unknown, these problems nevertheless compromise the utility of RMP for students and instructors alike.

Table 4. RMP/SET Correlations as a Function of the Number of RMP Posts.

	RMP index		RMP index	
SET criterion	overall quality	ease	overall quality	ease
	≤ 9 posts[a] (*n* = 217)		> 9 posts (*n* = 209)	
Item 13	.70	.16	.67	.29
Item 20	.10	.42	.08	.48
Instructor	.57	.07	.58	.14
Assessment	.35	.17	.37	.47
Facile	.02	.50	.12	.51
	< 14.4 posts[b] (*n* = 283)		> 14.4 posts (*n* = 143)	
Item 13	.68	.14	.71	.45
Item 20	.07	.40	.18	.57
Instructor	.56	.04	.61	.28
Assessment	.34	.21	.42	.54
Facile	-.01	.48	.26	.58

[a] 50.9% of instructors had 9 or fewer RMP posts. [b] The mean number of RMP posts is 14.4.

That said, the two primary RMP indices correlate substantively and significantly with their respective SET criterion variables: RMP overall quality correlates with SET Item 13 ("Overall, how would you rate the instructor?") and the SET Instructor factor, and RMP ease correlates with SET Item 20 ("How did the work load for this course compare to that of others of equal credit?") and the SET Facile factor. Moreover, these associations persist when statistical controls are in place. These results, we believe, should give pause to those who are inclined to dismiss RMP indices as meaningless.

Nevertheless, the RMP/SET correlations leave much unexplained variance in the criterion measures, which limits the utility of RMP relative to formal student evaluations of teaching. For example, the correlation of central interest ($r = .68$ between RMP overall quality and SET Item 13) reveals that over half (54%) of the variation in the SET measure is unrelated to variation in the RMP index. This is captured by the considerable scatter in Figure 1. The clear exception, as the superimposed box in this figure highlights, is found where instructors enjoy a very high RMP overall rating: They invariably are high on SET Item 13 as well. But among the many other instructors, there is poorer agreement between their RMP overall quality and the SET criterion. The pattern of this association suggests that when an instructor's RMP overall quality is particularly high, one can infer that the instructor "truly" is regarded as a laudatory teacher. However, there is considerable uncertainty about the instructor's true status, as measured by SET, when RMP overall quality is anything less than stellar. In a quick tabulation not reported above, for example, we found that all UMaine teaching award recipients had SET Item 13 means of roughly 4.0 or higher. However, only half of the awardees fell in the superimposed box in Figure 1. In short, the inevitability of many false negatives serves as an important cautionary note for RMP users.

As for the RMP pepper index, the weak association between this dichotomous variable and the primary SET criterion variables disappears altogether when the remaining RMP indices are held constant. Unlike RMP overall quality and RMP ease, then, the presence or absence of a pepper is unrelated to SET criteria: It contributes nothing of substance with respect to formal in-class student evaluations of teaching. In this light, the red chili pepper is a frivolous detraction that compromises the credibility of RMP. This conclusion echoes the findings of Kindred and Mohammed (2005, p. 11) regarding RMP use among their focus-group participants:

> *"The chili peppers are generally disregarded; students reported that they do not place importance on whether or not a teacher is regarded as 'sexy.' One student summarized this idea by stating, 'I think the hot tamale thing kind of takes away from the credibility of the site. If you're looking for a professor, obviously their level of attractiveness isn't really a top priority.'"*

Finally, correlations between the two primary RMP indices and their respective SET criterion variables are consistently larger when based on instructors falling above, versus below, the mean number of RMP posts. Although these differences are small and were only examined descriptively, this finding suggests that RMP indices may be more trustworthy when based on many posts. Nevertheless, we confess our surprise at not obtaining *much* smaller RMP/SET correlations among instructors having relatively few posts, insofar as the students providing these RMP ratings should be less representative of the population of students providing the SET ratings. This is not the case with our results. It is for subsequent research to explain this counterintuitive finding.

Implications

We believe that our results carry at least two policy implications. The first we offer with some ambivalence; the second, with firm resolve.

First, and predicated on the belief that RateMyProfessors.com is not going to go away, higher education institutions should consider encouraging their students to post ratings and comments on RMP. If a large proportion of an institution's student body were to regularly and responsibly contribute to RMP, the potential value of that information to the institution would only be enhanced. The emphasis, however, must be on *responsible* contributions. For example, the institution could stress to its students the importance of providing RMP ratings for each course taken, and posting comments that are both constructive and respectful. Appealing to students' sense of decency and fair play, furthermore, the institution could endeavor to discourage students from rating the hotness of the instructor.

This first implication, to be sure, is complicated by the aforementioned quality control problems that beset RMP. These problems doubtless will remain (although probably to a lesser degree), even for an institution that genuinely promotes regular and responsible RMP participation of its students.

Quality control notwithstanding, there also is no reason to believe that an institution's RMP participation rate could match, or even approach, the level of student participation in that institution's SET process. Further, the few RMP items on which students rate instructors typically pale in comparison to the many items, and underlying dimensions, found on SET forms. In short, RMP inevitably provides an unsatisfactory representation of both an institution's student body and the desired construct.

Hence, our second policy implication: Higher education institutions should make their SET data publicly available online. Although students doubtless would applaud this move, many faculty would oppose it because of genuine concerns about privacy and the negative consequences that published SET data may bring (e.g., see Howell & Symbaluk, 2001). But privacy is a thing of the past in the age of RMP, MySpace, and the like. Moreover, by not making SET data available to students, the negative consequence is greater still: Students will rely on what is publicly available. In light of our results, this inevitably will mischaracterize the true standing of many instructors as measured by formal student evaluations of teaching.

References

Algozzine, B. Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., et al. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching*, *52*, 134-141.

Cashin, W. E. (1995). *Student ratings of teaching: the research revisited.* IDEA Paper No. 32. Center for Faculty Evaluation & Development, Division of Continuing Education, Kansas State University. Retrieved February 1, 2005, from http://www.idea.ksu.edu/papers/Idea_Paper_32.pdf

Centra. J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco: Jossey-Bass.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*, 496-518.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, *41*, 511-535.

D'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*, 1198-1208.

Davison, E., & Price, J. (2006, August). *How do we rate? An evaluation of online student evaluations*. Unpublished manuscript. Retrieved January 22, 2007, from http://www1.appstate.edu/~pricejl/TEACHING/methods/RMP_8_06.pdf

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: the relations between perceived quality, easiness, and sexiness. *Assessment & Evaluation in Higher Education*, *29*(1), 91-108).

Felton, J, Koper, P. T., Mitchell, J. B., & Stinson, M. (July, 2006). Attractiveness, easiness, and other issues: Student evaluations of professors on RateMyProfessors.com. Retrieved February 1, 2007, from http://ssrn.com/abstract=918283

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*, 1182-1186.

Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology*, *93*(4), 790-796.

Johnson, V. E. (2002). Teacher course evaluations and student grades: An academic tango. *Chance*, *15*(3), 9-16.

Kindred, J. & Mohammed, S. N. (2005). "He will crush you like an academic Ninja!": Exploring teacher ratings on Ratemyprofessors.com. *Journal of Computer-Mediated Communication*, *10*(3), article 9. Retrieved October 12, 2006, from http://jcmc.indiana.edu/vol10/issue3/kindred.html

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist*, *52*, 1187-1197.

McClatchy, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, *52*, 1218-1225.

Montell, G. A. (2006, September 27). The art of the bogus rating. *The Chronicle of Higher Education*. Retrieved November 22, 2006, from http://chronicle.com/jobs/news/2006/09/2006092701c/careers.html

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluations of college teaching. *Assessment & Evaluation in Higher Education*, *27*(2), 187-199.

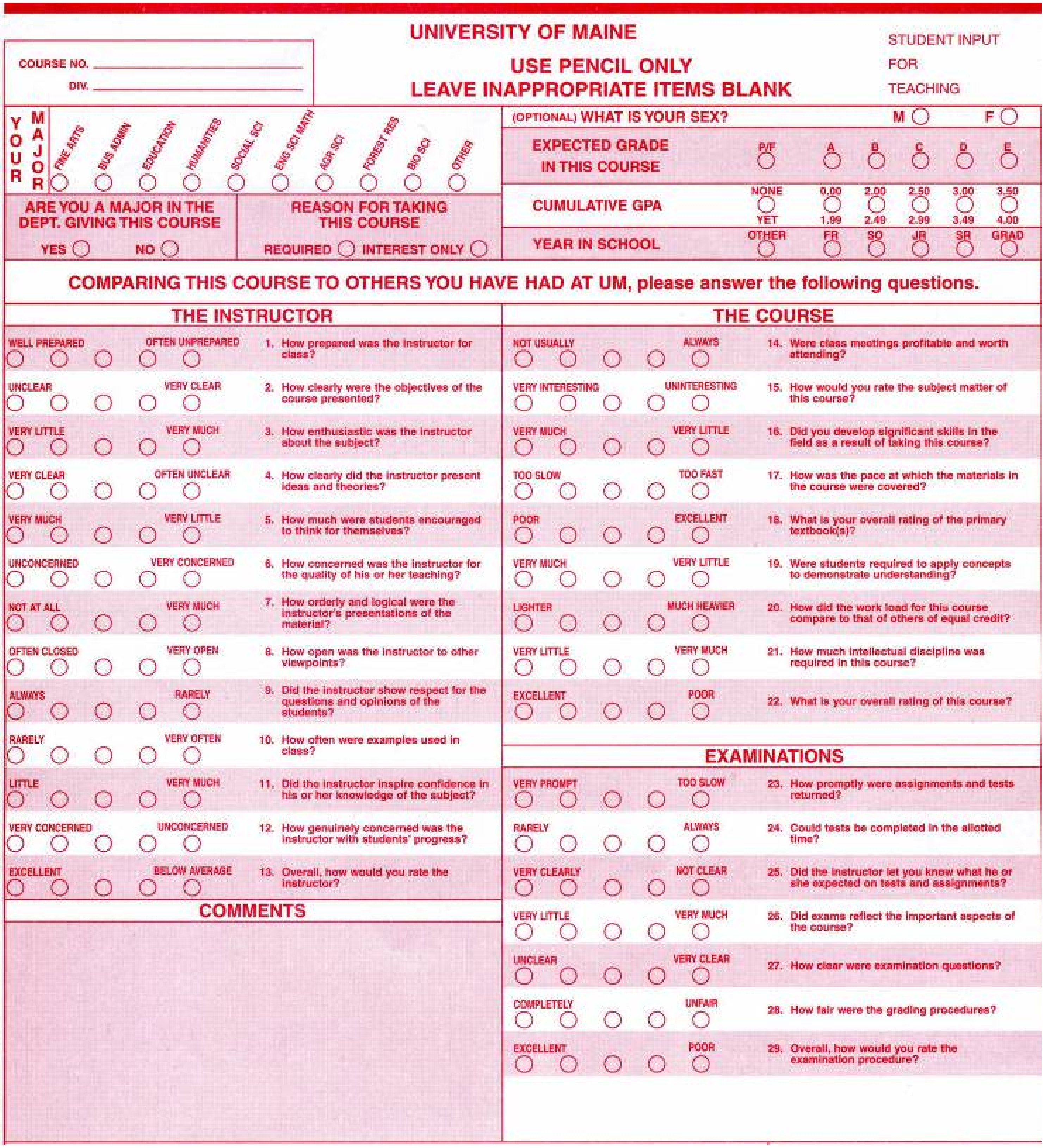
Remmers, H. H. (1927). The Purdue Rating Scale for Instructors. *Educational Administration and Supervision*, *6*, 399-406.

Riniolo, R. C., Johnson, K. C., Sherman, T. R. & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher evaluations? *The Journal of General Psychology*, *133*(1), 19-35.

Wachtel, H. K. (1998). Student evaluation of college teaching: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191-211.

Appendix A

University of Maine SET form



UNIVERSITY OF MAINE

USE PENCIL ONLY
LEAVE INAPPROPRIATE ITEMS BLANK

STUDENT INPUT
FOR
TEACHING

COURSE NO. ____________________
DIV. ____________________

YOUR MAJOR: FINE ARTS ○ BUS ADMIN ○ EDUCATION ○ HUMANITIES ○ SOCIAL SCI ○ ENG SCI MATH ○ AGR SCI ○ FOREST RES ○ BIO SCI ○ OTHER ○

ARE YOU A MAJOR IN THE DEPT. GIVING THIS COURSE
YES ○ NO ○

REASON FOR TAKING THIS COURSE
REQUIRED ○ INTEREST ONLY ○

(OPTIONAL) WHAT IS YOUR SEX? M ○ F ○

EXPECTED GRADE IN THIS COURSE	P/F ○	A ○	B ○	C ○	D ○	E ○
CUMULATIVE GPA	NONE ○ YET	0.00 ○ 1.99	2.00 ○ 2.49	2.50 ○ 2.99	3.00 ○ 3.49	3.50 ○ 4.00
YEAR IN SCHOOL	OTHER ○	FR ○	SO ○	JR ○	SR ○	GRAD ○

COMPARING THIS COURSE TO OTHERS YOU HAVE HAD AT UM, please answer the following questions.

THE INSTRUCTOR

WELL PREPARED	○	○	○	○	○	OFTEN UNPREPARED	1. How prepared was the instructor for class?
UNCLEAR	○	○	○	○	○	VERY CLEAR	2. How clearly were the objectives of the course presented?
VERY LITTLE	○	○	○	○	○	VERY MUCH	3. How enthusiastic was the instructor about the subject?
VERY CLEAR	○	○	○	○	○	OFTEN UNCLEAR	4. How clearly did the instructor present ideas and theories?
VERY MUCH	○	○	○	○	○	VERY LITTLE	5. How much were students encouraged to think for themselves?
UNCONCERNED	○	○	○	○	○	VERY CONCERNED	6. How concerned was the instructor for the quality of his or her teaching?
NOT AT ALL	○	○	○	○	○	VERY MUCH	7. How orderly and logical were the instructor's presentations of the material?
OFTEN CLOSED	○	○	○	○	○	VERY OPEN	8. How open was the instructor to other viewpoints?
ALWAYS	○	○	○	○	○	RARELY	9. Did the instructor show respect for the questions and opinions of the students?
RARELY	○	○	○	○	○	VERY OFTEN	10. How often were examples used in class?
LITTLE	○	○	○	○	○	VERY MUCH	11. Did the instructor inspire confidence in his or her knowledge of the subject?
VERY CONCERNED	○	○	○	○	○	UNCONCERNED	12. How genuinely concerned was the instructor with students' progress?
EXCELLENT	○	○	○	○	○	BELOW AVERAGE	13. Overall, how would you rate the instructor?

COMMENTS

THE COURSE

NOT USUALLY	○	○	○	○	○	ALWAYS	14. Were class meetings profitable and worth attending?
VERY INTERESTING	○	○	○	○	○	UNINTERESTING	15. How would you rate the subject matter of this course?
VERY MUCH	○	○	○	○	○	VERY LITTLE	16. Did you develop significant skills in the field as a result of taking this course?
TOO SLOW	○	○	○	○	○	TOO FAST	17. How was the pace at which the materials in the course were covered?
POOR	○	○	○	○	○	EXCELLENT	18. What is your overall rating of the primary textbook(s)?
VERY MUCH	○	○	○	○	○	VERY LITTLE	19. Were students required to apply concepts to demonstrate understanding?
LIGHTER	○	○	○	○	○	MUCH HEAVIER	20. How did the work load for this course compare to that of others of equal credit?
VERY LITTLE	○	○	○	○	○	VERY MUCH	21. How much intellectual discipline was required in this course?
EXCELLENT	○	○	○	○	○	POOR	22. What is your overall rating of this course?

EXAMINATIONS

VERY PROMPT	○	○	○	○	○	TOO SLOW	23. How promptly were assignments and tests returned?
RARELY	○	○	○	○	○	ALWAYS	24. Could tests be completed in the allotted time?
VERY CLEARLY	○	○	○	○	○	NOT CLEAR	25. Did the instructor let you know what he or she expected on tests and assignments?
VERY LITTLE	○	○	○	○	○	VERY MUCH	26. Did exams reflect the important aspects of the course?
UNCLEAR	○	○	○	○	○	VERY CLEAR	27. How clear were examination questions?
COMPLETELY	○	○	○	○	○	UNFAIR	28. How fair were the grading procedures?
EXCELLENT	○	○	○	○	○	POOR	29. Overall, how would you rate the examination procedure?

Citation

Coladarci, Theodore & Irv Kornfield (2007). RateMyProfessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research & Evaluation*, 12(6). Available online: http://pareonline.net/getvn.asp?v=12&n=6

Note

The authors are indebted to Phil Pratt, who assembled the de-identified SET database for us and, further, provided insightful comments and suggestions throughout the investigation. We also are grateful to Jim Felton, Pete Kroper, and Janet Spector for their thoughtful feedback on an earlier draft of this article, Maryhaven for support, as well as to the anonymous reviewers for their constructive recommendations.

Authors

Authors are listed alphabetically. Coladarci (coladarci@umit.maine.edu) is Professor of Educational Psychology at the University of Maine, where Kornfield (irvk@maine.edu) is Professor of Biology and Molecular Forensics. Correspondence regarding this article can be directed to either author.