University of Washington
Office of Educational Assessment

Gerald M. Gillmore, PhD., Director
Nana Lowell, PhD., Assistant Director

94-4

The Effects of Course Demands and Grading Leniency on Student Ratings of Instruction*

Gerald M. Gillmore
Anthony Greenwald

March, 1994

The Office of Educational Assessment is an agency of the University of Washington which provides a variety of services related to educational research and assessment. The following are programs within which these services are provided:

- Institutional Research
- Student Outcomes Assessment
- Instructional Evaluation
- Test Scoring and Analysis
- Standardized Testing

* This report was also delivered at the American Educational Research Association Convention, New Orleans, April, 1994.

453 Schmitz Hall, 1400 NE Campus Parkway, University of Washington, PB-30, Seattle, WA 98195
(206) 543-1170 / Fax:(206) 543-3961 /E-mail: oea@u.washington.edu

The Effects of Course Demands and Grading Leniency on Student Ratings of Instruction

Gerald M. Gillmore
Anthony Greenwald

University of Washington

March 3, 1994

Overview

The major purpose of the research reported below was to better understand the effects of grades and measures of course difficulty on student ratings of instruction. Research was conducted in four stages, each of which obtained data from regular University of Washington (UW) classes whose faculty volunteered to participate. The initial studies used the optional items portion of the UW Instructional Assessment System (8 items), first in psychology classes and then in campus-wide classes, to develop items to better measures of course work load and grade expectations. The third study also used the optional items portion of the UW Instructional Assessment System. Its purpose was to make a beginning in analyzing the relationships among the variables mentioned above. Finally, this research, plus other considerations described below, led to the development of a new student rating form (Form X) that was used experimentally in 337 UW classes fall quarter, 1993.

While the prior research was instrumental in the development of Form X, essentially all findings were validated and extended by those deriving from the use of Form X.[1] Hence, the subsequent presentation focuses these latter results.

Impetus for the Development of Form X

A copy of Form X is found in Appendix A. A number of reasons precipitated the development of this form and influenced its design. The first two are particularly germane to the research presented here; others will be mentioned for completeness.

Course difficulty. A memo was sent to UW faculty users soliciting their complaints about student ratings in general and about the current system under which student ratings were collected in specific. Faculty were also asked for suggestions for improvement of the system. Qualitative analyses of the written responses of 130 faculty revealed several themes. One of these was that many faculty expressed the view that the way to achieve high ratings is by demanding little work from one's students and by grading leniently[2].

[1] There is one exception which will be described subsequently.
[2] Gillmore and Mahmood, Faculty Suggestions for Instructional Assessment System Revisions, OEA Report, *92-3*, 1992.

This view, in one form or another, is quite common among faculty and has been documented elsewhere. In terms of demanding little or much from one's students, the literature has been quite consistent in showing little or no correlation between some measure of students perception of course difficulty and the rating given the course. For example, Cashin and Sixbury[3] correlated results of the IDEA system from over 100,000 classes. The correlation between "Difficulty of the subject matter" and their overall composite course rating was only .08, but with more difficult courses being rated higher. However, another IDEA item, "I worked harder on this course than most courses I have taken" correlated .44 with the overall composite measure, indicating that working harder in a class was positively related to higher ratings. Thus, we are faced with the puzzling question: why do many faculty feel that being demanding in a course causes lower ratings by students? Items were included on Form X to try to understand this conundrum, including ones focusing on time spent on the class, student effort, and the intellectual challenge of the course.

Grades. Research has been consistent in finding a positive correlation between grades (expected or actual) and ratings, with either the student or the course as the unit of analysis.[4] The debate with regard to grades and student ratings is not about the presence of a relationship, but an explanation for it. In brief, do high grades cause good ratings (i. e., is the promise of good grades a biasing factor?), or is learning a third variable that is positively related to both grades and ratings (i. e., classes that foster more learning both get higher ratings and offer higher grades)? Form X included an item on the grade that students expected[5] and an item on the magnitude of that grade relative to what the student gets in other classes.

Learning. The IAS was developed in 1973 and has not undergone substantial change since. In the 70's greater stress was placed on teaching process and, hence, only one item on the IAS directly relates to student learning. Reflecting greater emphasis on student learning outcomes in the current era of assessment, Form X includes 7 items on student learning, each focusing on a different learning outcome.

Halo Effects. Items of the IAS, like many such forms, are highly intercorrelated across classes. A contributing reason may be that all items are in the same format (Responses Excellent through Very Good.) Form X includes different sections of items each with different formats. Further, items in the first section asked students to make judgments about the frequencies of various teacher behaviors rather than making direct value judgments. Items of the second section asked about quite distinct learning outcomes. It was hoped that these changes would reduce the high inter-item correlations.

Leniency. It is well-known that students tend to use the upper end of the scale to rate faculty. This tendency is problematic mainly in regard to comparisons among faculty. First, the scale under which these comparisons are made is artificially shrunk. Discriminations are often made based upon small scale differences. Secondly, faculty

---

[3] IDEA Technical Report # 7, Center for Faculty Evaluation and Development, Kansas State University, 1992

[4] Feldman, K. A., Grades and College Students' Evaluation of Their Courses and Teachers, Research in Higher Education, 4, 1976, 69-111.

[5] UW uses the 4.0 grading system. Form X provides 14 non-overlapping categories.

find the association of high absolute ratings with low relative (normative) ratings to be confusing and upsetting.

Three changes were made to atternpt to ameliorate the tendency toward students' overuse of the high end of the scales. First, the first set of items were based on frequency of behaviors rather than a straight evaluative scale. Secondly, a number of items were asked in terms of "relative to other courses you have taken". In theory the average responses to these items should hover around the scale midpoint. Finally, seven point scales were used for the items, rather than the six point scales of the IAS forms.

Method

The first study in the series was performed in psychology classes over several quarters during the 1991-92 academic year and was used to formulate hypotheses about the relationship between work load, grading standards and ratings and how each might be accurately measured. This study was followed by a campus wide data collection effort Summer Quarter of 1992, the purpose of which was to extend the sample and further investigate item wordings. In both of these studies, faculty who had requested use of the University of Washington (UW) Instructional Assessment System (IAS) were asked to use an additional eight items for research purposes. In the latter study, ninety-nine faulty volunteered and were supplied with an overhead transparency containing the additional items. For this study, two versions of the items were used; half the faculty received each version by random selection. Following the format of the form, the eight additional items added were of two types -- four used a 6-alternative response format (Excellent to Very Poor) and four using a 10-alternative format (0 to 9).

Subsequently, all 592 faculty who requested forms from the UW Instructional Assessment System (IAS) fall quarter, 1992, and who indicated that their class contained 30 or more students were contacted by mail and asked to participate in the next phase of the research. The letter explained the purpose of the study and enclosed an overhead transparency containing eight additional items, consistent with the space and format of the IAS forms. Participation required administration of the additional items to students as part of the regular administration of the IAS forms. The final number of courses with usable data was 262, representing 44% of the 592 who met the inclusion criteria. All faculty received the same version.

One of the additional items added to the forms in the study was:

> The grade I receive in this class is likely to increase my grade point average (Agree Strongly [5] to Disagree Strongly [0])

For future reference we term this item *Relative Grade.* In addition, for each participating class, we determined the average GPA of the students in the class and the average grade given in the class. The difference between these two values we termed the *grade inflation factor.*

The results of the two pilot studies, as well as other factors mentioned above, led to the development of Form X over the summer of 1993. Fall Quarter, 1993, a memo was sent

to all tenure track faculty and to all academic departments inviting use of the new form in their classes. The form was described as experimental. Faculty from 337 classes agreed to use the form, that accounted for 12% of the classes rated by instruments of the Office of Educational Assessment. Those using the form were representative of the campus in terms of academic unit, rank (including graduate teaching assistants), and course level. We have no data on whether the users were representative in terms of teaching competence.

Results

Generally, the results of the three phases of this study were consistent. In this paper, the results to be presented are taken completely from the 337 classes in which Form X was used, with the following exception. In phase 2, we found that correlation of the students' rating of their relative grade (The grade I receive in this class is likely to increase my grade point average) with the grade inflation factor (average grade given minus average student GPA) was .70. This is encouraging in suggesting that the self-report measure (relative grade) can serve as a proxy for the more difficult to obtain archive-based measure of grade inflation.

In Form X, the wording of the relative grade item was changed somewhat to fit the characteristics of the form (Relative to other classes you have taken, do you expect your grade in this course to be: [Much Greater to Much Less]. However, we submit that the item in this form can be considered an index on grading inflation. We did not, however, collect new grade data test the relationship do to the extensive time it required.

Reliability. Inter-rater reliability coefficients for the items of Form X are presented in Table 1. These coefficients were computed using intraclass correlations[6,7]. The table presents reliabilities for the average sized class and for classes of size 1, 10, 20, and 50. (The reliability for any sized class can be determined through use of the Spearman Brown formula).

One can note that the reliabilities tend to be quite high for all items except the following three:

> 22. Your involvement in the class . . .
> If you had it to do over, and this course was optional for your program, would you enroll in it:
>
> 24. If a different instructor taught it?
> 25. Regardless of who taught it?

However, even these items showed reasonable reliability for larger classes.

---

[6] Ebel, R. L., Estimation of the reliability of ratings,. Psychometrica, 1951,*16,* 407-424.

[7] The formula for the reliability of an average class is: (F-1)/F where F is determined by computing a one-way analysis of variance with differences among class means as the source of between groups variance and students within classes as the source of the within groups variance.

Table 1

Inter-Rater Reliabilities of the Items of Form X

Items	Reliability Coefficients Class Size				
	Average	1	10	20	50
1. Explain	0.90	0.24	0.76	0.86	0.94
2. Rephrase	0.89	0.23	0.75	0.85	0.94
3. Interest	0.91	0.26	0.78	0.88	0.95
4. Organiz	0.89	0.23	0.75	0.86	0.94
5. Participate	0.93	0.32	0.82	0.90	0.96
6. Expected	0.86	0.18	0.69	0.81	0.92
7. Help	0.81	0.14	0.62	0.77	0.89
8. Readings	0.84	0.16	0.66	0.80	0.91
9. Grades	0.86	0.20	0.71	0.83	0.92
10. Feedback	0.87	0.20	0.72	0.83	0.93
11. Goals	0.83	0.16	0.66	0.80	0.91
12. Facts	0.84	0.17	0.67	0.80	0.91
13. Apprec	0.85	0.18	0.68	0.81	0.91
14. Reading	0.81	0.14	0.61	0.76	0.89
15. Writing	0.84	0.16	0.66	0.80	0.91
16. Prob Solv	0.81	0.14	0.62	0.77	0.89
17. Real World	0.85	0.18	0.69	0.81	0.92
18. Intel Dev	0.83	0.16	0.65	0.79	0.90
19. Rel Grade	0.80	0.13	0.60	0.75	0.88
20. Challenge	0.86	0.19	0.70	0.82	0.92
21. Effort	0.86	0.19	0.70	0.82	0.92
22. Involvement	0.69	0.08	0.45	0.62	0.80
23. Same Inst	0.89	0.23	0.75	0.86	0.94
24. Diff Inst	0.69	0.08	0.45	0.62	0.81
25. Regardless	0.63	0.06	0.39	0.56	0.76
26. Tot Hrs	0.91	0.29	0.80	0.89	0.95
27. Good Hrs	0.89	0.23	0.75	0.86	0.94
28. Exp Grade	0.87	0.21	0.72	0.84	0.93

* See Appendix A for precise items text.

Four items whose high reliabilites was somewhat surprising were:

19. Do you expect your grade in this class to be:
26. On average, how many hours have you spent per week on this class . . .
27. From the total average hours above, how many do you consider were valuable for advancing your education?
28. What grade do you expect in this class.

Evidently, for these items there was substantial agreement of students within classes relative to between class differences. This results suggest that the classes in the sample exhibited systematic differences in the students' estimate of their workload[8], valuable hours, relative grade and expected grade.

Inter-item Correlations (Scales). The first set of items on Form X are all intended to be descriptive of the course and particularly of the instruction. The intercorrelations of these items, across class medians[9], are presented in Table 2. One can notice in this table that the correlations among the 11 items tend to be high, but with exceptions. The internal consistency reliability (Cronbach's Alpha) was .945, indicating that these items form a reliable scale for the analyses that follow. Furthermore, a principle components factor analysis revealed that 65.7% of the total variance of this set of items was associated with the first factor. The eigenvalues of all other factors were very small.

Table 2

Correlations among Course Description Items

Item	Explain	Rephr	Intrst	Organ	Partic	Expted	Help	Readg	Grade	Feedbk	Goal	Mean	St Dev
1. Explain												5.9	0.69
2. Rephrase	92											6.0	0.69
3. Interest	72	73										5.7	0.88
4. Organiz	74	68	58									5.9	0.77
5. Participate	55	63	61	29								6.2	0.81
6. Expected	77	75	57	66	58							6.0	0.75
7. Help	59	62	51	45	52	61						6.2	0.63
8. Readings	46	46	61	44	37	50	42					5.8	0.79
9. Grades	65	66	57	50	55	76	63	49				6.1	0.69
10. Feedback	63	67	58	48	58	68	66	50	72			5.8	0.87
11. Goals	71	74	68	57	60	75	66	62	79	82		5.9	0.71
SUM	86	88	82	77	70	87	74	67	81	82	89	65.2	6.68

Alpha = 0.945

[8] Converting the workload items to hours per credit resulted in similar reliability estimates.
[9] Medians were used rather than means because Form X reports medians back to faculty. Previous (unpublished) research on student ratings results by Gillmore has shown the correlation between the two measures of central tendency is over .99.

One can also note that the correlations between certain pairs of items were considerably smaller than others (e.g., r = .29 between "Student participation was encouraged" and "Class sessions were well organized"). Thus, students' responses were not completely dictated by a good course-bad course "halo" effect.

The intercorrelations among the seven learning items are presented in Table 3. These correlations tend to be uniformly high. Cronbach's alpha for this set of items was .949 and 77.4% of the variance was associated with the first principle components factor. Thus, the seven learning outcome items can be considered a scale. In terms of reducing a halo effect, these results were particularly disappointing. One can note that the learning outcomes span a broad space, from learning the facts and concepts of a field to general intellectual development to problem solving to developing an ability to express oneself. One would expect different courses to emphasize different outcomes. Yet, students tend to see them along a single dimension.

Table 3

Correlations among Learning Outcome Items

Items	IntDev	Apprec	Facts	Read	ProbS	Write	RealW	Mean	St Dev
18. IntelDev	100							5.8	0.73
13. Apprec	87	100						6.0	0.73
12. Facts	82	80	100					5.8	0.71
14. Reading	73	72	81	100				5.7	0.66
16. ProbSolv	75	72	80	81	100			5.6	0.69
15. Writing	67	62	68	79	75	100		5.4	0.80
17. RealWorld	68	70	68	69	71	64	100	5.8	0.78
SUM	90	88	91	90	90	84	83	40.0	4.47

Alpha = 0.949

Furthermore, the correlation between the average of the descriptive items and the average of the learning outcomes items was .86. A factor analysis over class medians of both sets of items yielded a first factor accounting for 66% of the variance. The item, "If you had it to do over again and this course were optional for your program, would you enroll in it: If the same instructor taught it" correlated with the total descriptive rating at .80 and with the total learning rating at .78. This item, too, would serve well as a general evaluative item.

Leniency. About 99% of the medians across classes were above the natural scale midpoint for both the Total Descriptive rating and the Total Learning rating. In Figure 1, the average medians, as well as the average 25th and 75th percentiles are presented for the descriptive items. The same data are presented in Figure 2 for the learning items. Clearly, one can see from these high values that students were very lenient judges of their teachers. Indeed, the learning items asked for ratings relative to other classes taken. While it is possible that the particular teachers who chose the experimental version were above average teachers, it is very unlikely that they were as much above average as the average medians indicate.

Figure 1
Average Quartiles: Descriptive Items

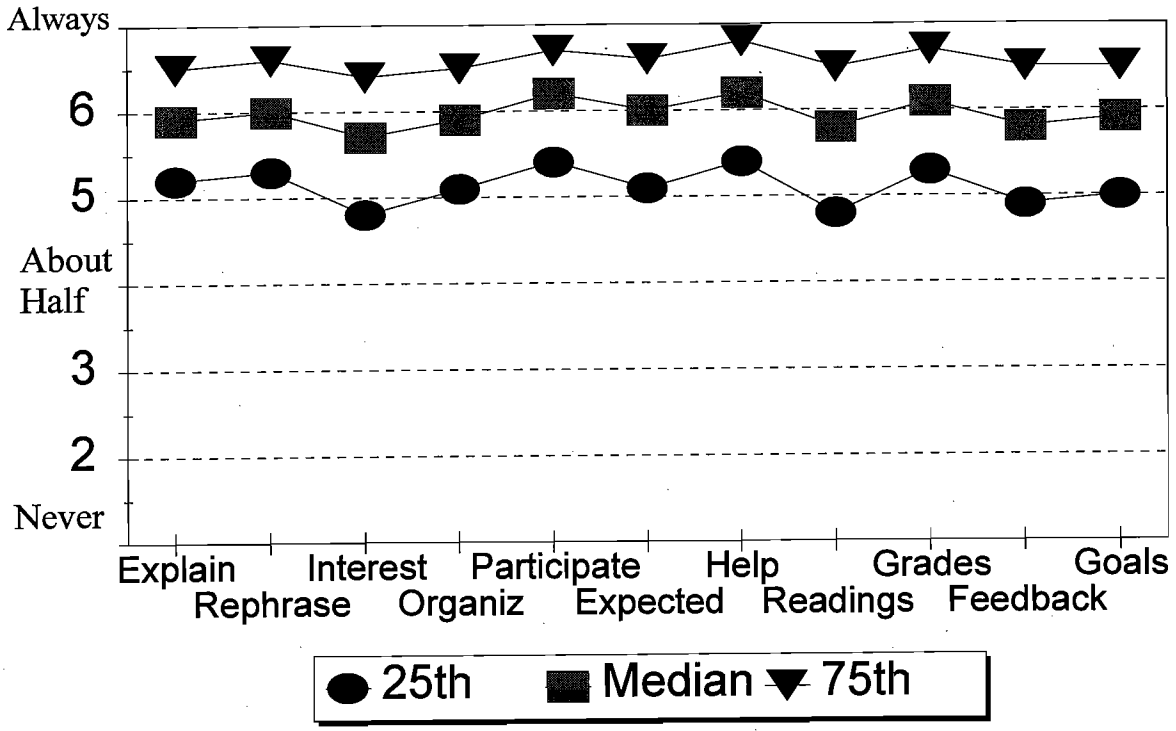
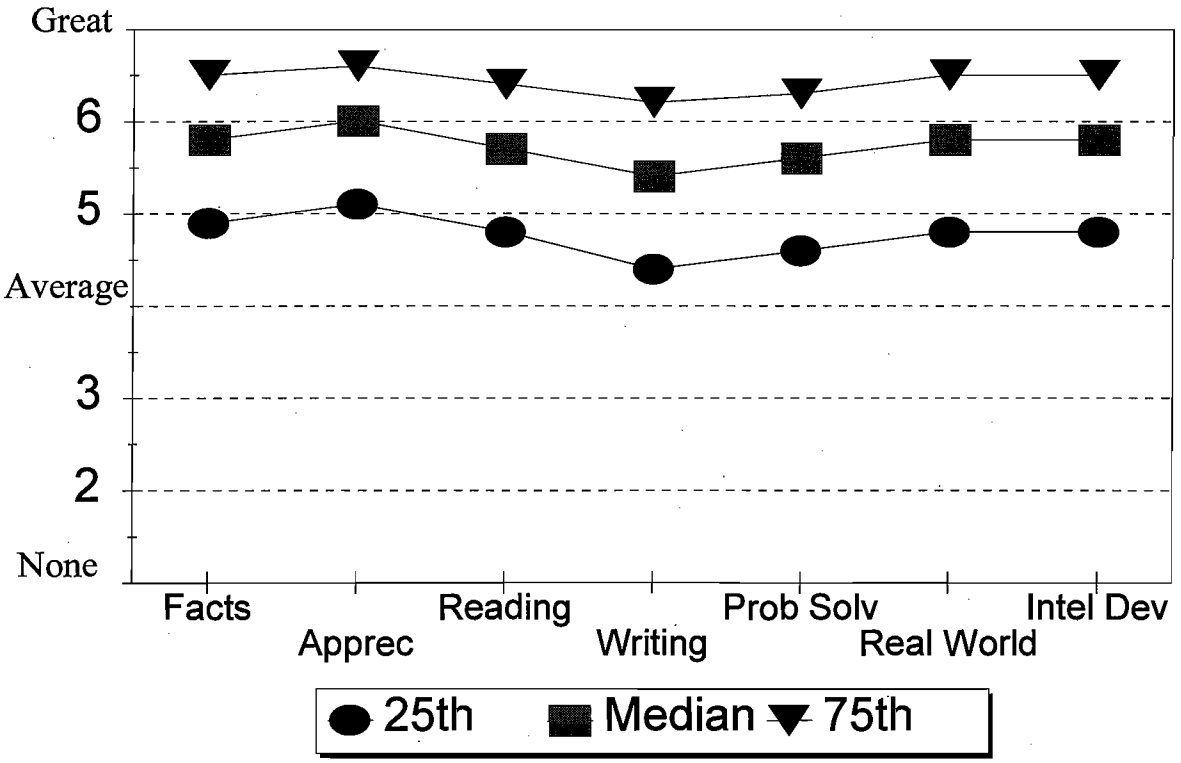
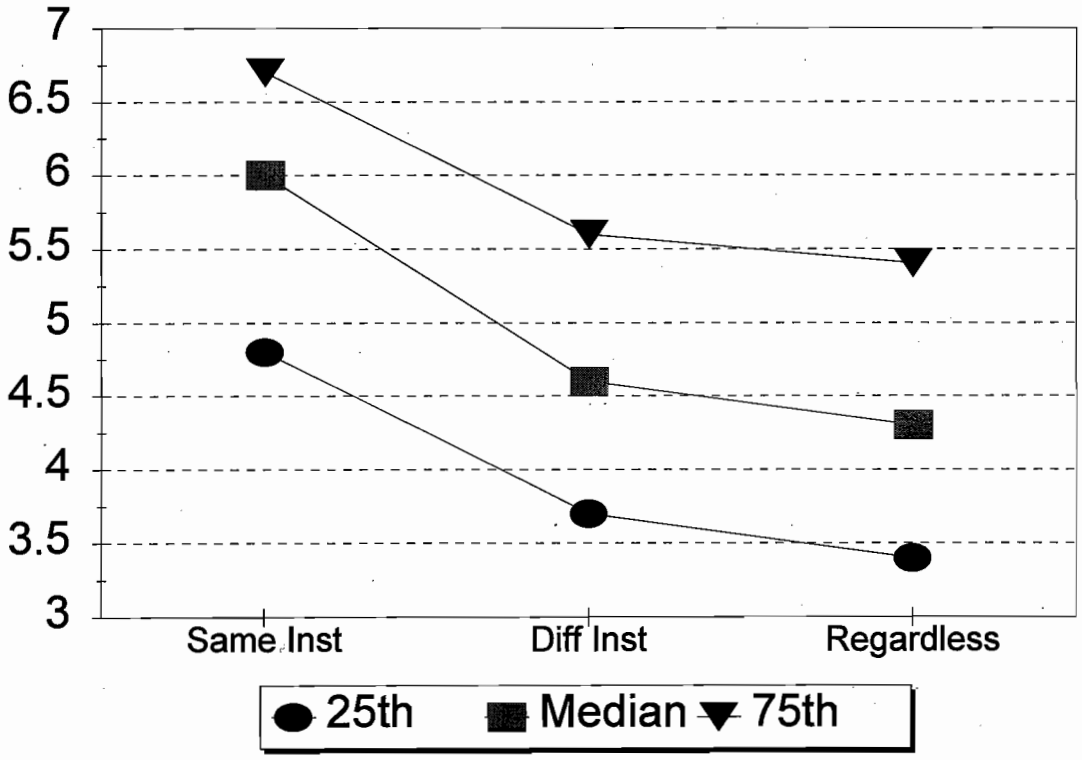


Figure 2
Average Quartiles: Learning Items



Student leniency can be illustrated in another way. The form contains three questions concerning whether the students would enroll again if they had it to do over -- with the same instructor (23), a different instructor (24), and regardless of the instructor (25). In Figure 3, one can note that the median for the same instructor is considerably higher than that for a different instructor and that for regardless of the instructor. In 83% of the class rated, the median for the same instructor was greater than the median for a different instructor. It seems to be a clear case for that for students "the devil you know being better than the devil you do not know. In other words, students tend to prefer an instructor they have experienced over one that they have not experienced.

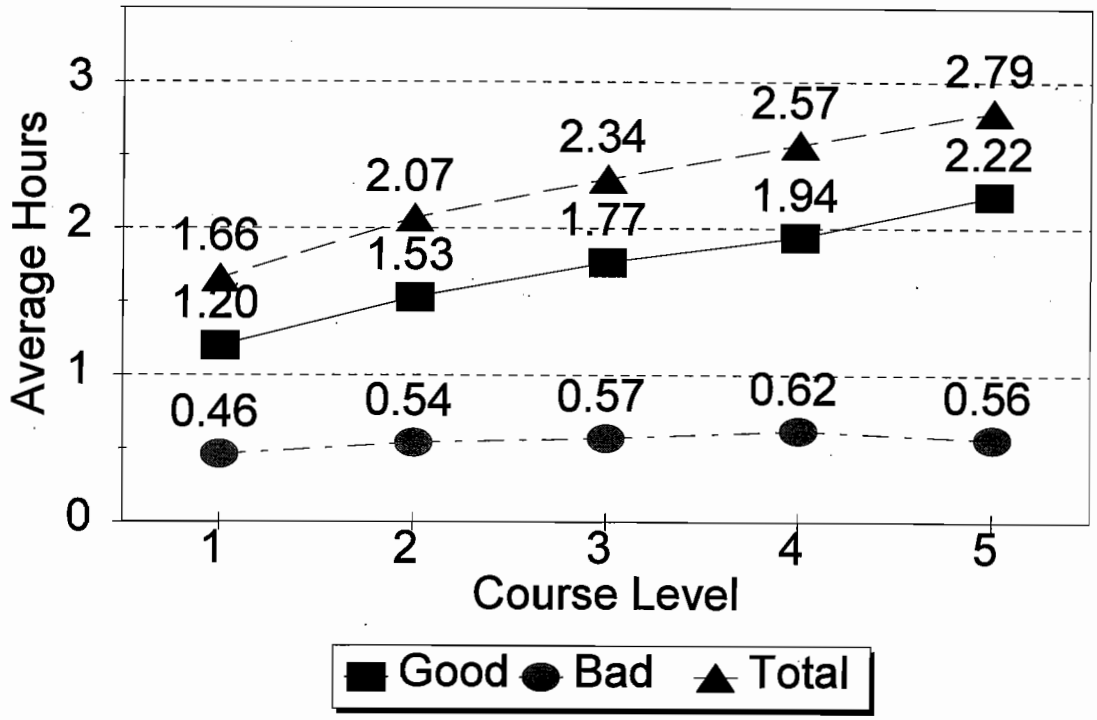
Figure 3
Aver. Quartiles: Take Course Again If:



Hours spent. Form X contains a pair of items relating to the time spent on the course. The first item (26) asks for total time per week, on average, spent attending classes, doing readings, reviewing notes, writing papers, and any other course related work. The second item (27) asks for the number of that total per week that were valuable in advancing your education. Thus, we have total hours, valuable or good hours, and bad hours, which is the difference between the two.

In Figure 4, the three averages per credit hour of the course are presented by course level. Overall, students considered about 75% of their total weekly hours to be spent valuably. However, one can see that average bad hours are essentially constant over course levels, while good hours and total hours increase. One can also see that the two hours outside of class for every hour in class rule of thumb is violated, on average, even in graduate courses.

Figure 4
Average Hours Per Credit



The correlation of total hours per credit with Total Descriptive rating, over class medians was .03. The correlation of total hours per credit with Total Learning rating, over class medians was .07, thus replicating the lack of correlation of ratings with difficulty or workload that has been found in numerous studies. *However, if one correlates the ratio of valuable hours to total hours, a very different result emerges.* This latter variable correlated .61 with Total Descriptive rating and .62 with Total Learning rating. What appears to be related to ratings is not how much time is spent, but the extent to which the time spent is seen as valuable.

Grades. Grades were measured in two ways. The expected grade was simply the students estimate of the grade they would receive in the class. The average median expected grade across all classes was a B+ -- leniency operates both ways. Relative grade was the students judgment as to whether their grade in the given class would be greater or less, relative to other classes. Students tended to be optimistic with an average median across classes of 4.9 on a 7 point scale.

The correlation between the two grade items was .52. Relative grade correlated with the Total Descriptive rating .25 and with the Total Learning rating .32. For expected grade the correlations were .34 and .38, respectively. These correlations are consistent with what has been reported in the literature. In the next section, we present regression results to try to better understand the nature of this relationship.

Predicting ratings. In this last section we pose the question, of the variables we have available, that in combination best predict general ratings. We have particular interest in the effects of grades, either absolute or relative, when other measures are taken into account. For this analysis, two dependent measures are used: the Total Descriptive rating and the Total Learning rating. The unit of analysis is again classes and the data used are medians. As was reported above, the two dependent measures are highly correlated (r = .86).

In Table 4, the variables that went into the prediction equation are listed. For each, the simple correlation with the two dependent variables is listed as well as the beta weight of the regression equation. Beta weights that are significant are printed in bold.

Table 4
Multiple Regression Results

	Total Description R Square = .52		Total Learning R Square = .57	
Variable	Simple Correlation	Beta Weight	Simple Correlation	Beta Weight
Valuable\Tot Hrs	0.62	**0.48**	0.61	**0.44**
Challenge	0.35	**0.33**	0.40	**0.25**
Relative Grade	0.25	**0.18**	0.32	**0.27**
Involvement	0.24	0.06	0.36	**0.14**
Credits	-0.09	0.01	-0.07	0.00
Class size	-0.16	-0.04	-0.19	-0.06
Class level	0.14	0.07	0.15	0.08
Effort	0.14	-0.08	0.24	0.03
Expt. Grade	0.34	0.07	0.38	0.07
Faculty Rank	0.01	0.07	0.03	0.00
Tot Hrs/Cred	0.03	-0.08	0.07	**-0.14**

* Beta weights in bold are significant (P < .05)

The variables entered into the equations predicted over half of the total variance for both dependent variables. One can see that three variables are significant for both dependent variables: the ratio of valuable hours to total hours spent on the class, relative grade, and the intellectual challenge that the course provided. Of these, the former was the most important predictor, while the latter two were about equivalent in weight, with challenge perhaps being a little stronger. Total Hours per Credit and Student Rating of Involvement in the Course were also significant for Total Learning. However, their beta weights were relatively small, and in a stepwise regression, addition of neither of these variables added significantly to the multiple correlation. Neither class size, level, and credits nor faculty rank were significant as predictors.

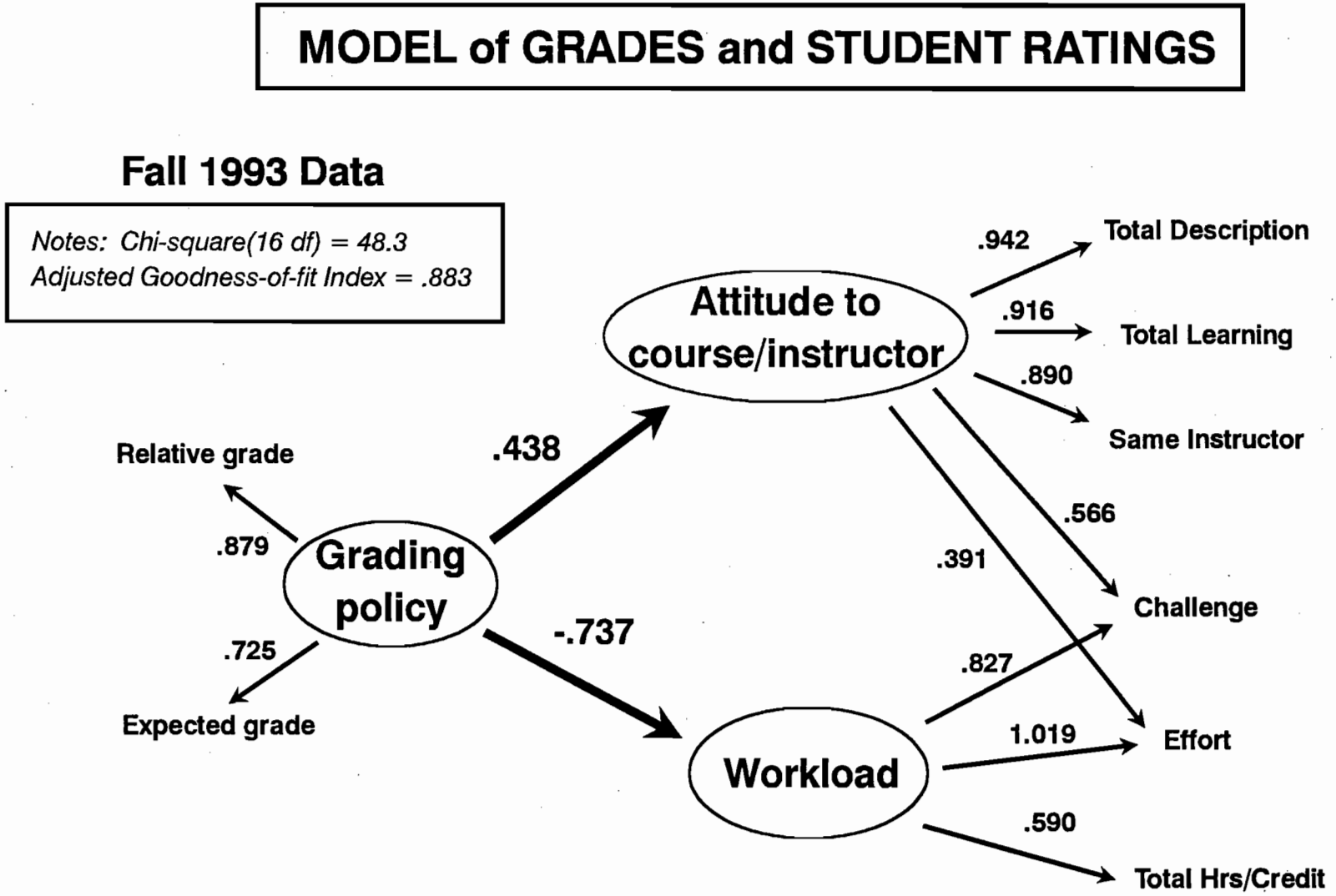
It should be noted that correlations exist among certain predictors. In particular, if relative grade is excluded from the predictors, then a significant beta weight is associated with expected grade. Likewise, if challenge is excluded from the list of predictors, a significant beta weight is associated with ratings of the amount of effort required to succeed. Thus, we can say that overall ratings are predicted by a combination of the ratio of valuable hours to total hours, grades, and the challenge or effort needed to succeed in the course. For all practical purposes, other variables do not add significantly to this prediction.

Causal Modeling. Finally, we end with a presentation of an initial attempt to apply causal modeling techniques to the data. Using the CLAIS module in SAS, three latent variables were defined. Grading Policy was defined by Relative grade and Expected grade. Workload was defined by Challenge, Effort, and Total hours per credit. Course/Instructor Attitude was defined by Total Description, Total Learning, Same Instructor, and by Challenge and Effort. We hypothesized for the purposes of this model that Grading Policy might affect both Course/Instructor Attitude and Workload. However, we know from research that Workload and the Attitude Toward the Instructor and Course are not causally related and therefore did not include a correlation between these constructs.

Unfortunately, at this point we do not have a measures of teaching effectiveness which are unconfounded with students' course/instructor attitude. Furthermore, the model cannot clarify whether the ratio of good hours to total hours is an antecedent or a consequent of the Course/Instructor Attitude because we do not have a second measure of this concept; thus, we did not include this variable in the model.

The results are presented in Figure 5. Even though the quality of the fit is a little less than that for which one would hope, in this preliminary analysis there is a positive link from Grading Policy to Course/Instructor Attitude. This result is consistent with the regression results presented above. There is also a strong negative link from Grading Policy to Workload, which is not predictable for prior analyses. Two hypotheses come to mind to explain this relationship. First, students are reacting to a relationship that exists -- faculty who are more rigorous in their grading standards are also more rigorous in the workload demands they place on students. Second, students are interpreting workload from the perspective of their perceived success -- they interpret expected poor grades as an indication of a hard demanding course.

Figure 5


MODEL of GRADES and STUDENT RATINGS
Fall 1993 Data
Notes: Chi-square(16 df) = 48.3
Adjusted Goodness-of-fit Index = .883
Attitude to course/instructor
Grading policy
Workload
Relative grade
Expected grade
.879
.725
.438
-.737
.942
.916
.890
Total Description
Total Learning
Same Instructor
.566
.391
Challenge
.827
1.019
Effort
.590
Total Hrs/Credit

Implications

This report began with a list of five factors providing the impetus for the development of Form X. Each of these will be revisited with comments concerning what we have learned form the research reported above.

Leniency. Median ratings tended to be bunched up at the top of the scale. Essentially all of the class medians for the Total Descriptive rating and the Total Learning rating were greater than the scale midpoints. Even items in which students were asked to respond relative to other courses almost always were rated above the scale mid-point. In other words, nearly everyone was rated above average -- a dramatic instance of the Lake Wobegone effect. Students also preferred the instructor they had over a different instructor, if they were to take the course over, in 83% of the classes in the sample. There is no question that the tendency for students to be lenient in their ratings was replicated in the results of this research. *Halo Effects.* A halo effect was strongly in evidence in the correlations among the first two sets of items on Form X. It appears that students have a general course/instructor attitude and that this attitude influences their ratings. There are three comments that can be made. First, to some unknown extent some of what appears to be a halo is surely reflection of a reality that good teachers (from the students' point of view) tend to do everything well, while poor teachers tend to do everything poorly.

Second, the correlations among class medians may obscure differences in the patterns of perceptions of individual students. Furthermore, in the descriptive items, there are exceptions to the high inter-item correlations (e.g., the relatively low correlation between course organization and student participation [r = .29]). It is a mistake to believe that all of the ratings in the first two sections are explained by this course/instructor attitude.

Finally, insofar as there is a general course/instructor attitude, it is essential that we understand what is and is not affected by that attitude, as opposed to what is affected by specific elements of the course and instructor. For example, judging by the correlations among the descriptive and the learning items, it appears that student responses are influenced by a general evaluative attitude. In fact, in this report, we have essentially operationalized the attitude as the sum of the items. Equally important, we need to better understand the direction of causality in order to properly interpret our results. In the current study, this issue is best illustrated by items such as the total hours spent on the course and the proportion of those that are considered valuable, the challenge of the course, and the effort required in the course. The regression equations that were presented implicitly assume that these items influence the general evaluation and not vice versa (as well as the assumption that the items validly relate to components of the course that they describe). Further research and further analyses under casual models will be necessary to better understand the extent and nature of the halo effect. In the preliminary model that was presented, we were not able to directly address this issue because we had no measures that separated teaching effectiveness from the course/instructor attitude.

Learning. Seven items presumably measuring different dimensions of learning were a part of Form X. These items taken together formed a nice scale with a very high internal consistency. On the downside, given the variety of classes that were rated and their goals, it was surprising and somewhat disappointing to see the high inter-item correlations. These high correlations shake ones confidence in students' ability to differentiate the outcomes described by the items. Thus, while the seven items taken together form a good scale of student perception of their learning, the accuracy of each individual item is suspect because it seems strongly colored by the general evaluation attitude.

Course difficulty. This study replicated the finding of little or no relationship between course difficulty, as indexed by hours spent on the course, and average ratings. Early in this report we posed the question of why many faculty feel that difficult courses get lower ratings when research evidence indicates otherwise. The strong relationship of average ratings with the ratio of valuable hours to total hours (or the strong negative relationship of average ratings with the ratio of bad hours to total hours) suggests an entertaining hypothesis, going beyond these data, to explain this apparent inconsistency. The data show that demands students perceive as valuable for learning improve ratings while demands that students perceive as not valuable for learning depress ratings. But faculty see all of their demands as valuable. Clearly, not all students see it that way, and classes can be reliably differentiated on the basis of how students, on average, see the mix. The faculty who see leniency as the path to high ratings may be those whose perceptions of the value of class activities differ from those of their students. For example, if one adds more work, but the students see this additional work as "busy work", ratings will surely suffer.

According to this hypothesis, faculty may do well to give careful consideration to the perceived value to their students of what they are assigning and what they are doing in class. They may also be wise to explain to students why these activities and assignments are worthwhile. Furthermore, some faculty may be well-served by developing closer ties between assignments and grades. On the other hand, we cannot reject the plausibility of the argument that students' perception of the value of their work is heavily influenced by their general evaluative attitude toward the instructor and course.

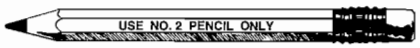
Grades. This study found a correlation between grades, both expected and relative, and average ratings, much as has been found in other studies. The results of the multiple regression argues that lenient grading standards are a positively biasing factor beyond any mediating effects of learning, although absent an independent measure of learning we can not be sure of this conclusion. We make this claim because of the significant beta weight for grades even after the effects of the proportion of valuable hours to total hours and the intellectual challenge or effort were removed. Indeed, according to these results, the way to achieve high ratings is to do things in class and give out-of-class assignments that are viewed by students as valuable, to make the course intellectually challenging, and to promise students high grades or grades that are higher than they are accustomed to receiving. It would be a mistake, however, to conclude that by giving high grades alone one can assure high ratings.

The study found that the student ratings general factor appears to be influenced by three factors: students perceptions of the ratio of valuable hours to total hours, the challenge of the course, and their grades in the course. In terms of arguing for student ratings as an agent for improved teaching, the first two legs are very favorable. No one could argue with intellectually challenging students with educationally valuable experiences. The third leg, however, is troubling. For example, one could hypothesize a cycle of grade inflation -- giving higher grades leads to higher ratings and the averages of both slowly creep upward. Insofar as the relative grade item is a good proxy variable for an index of grading leniency, one might adjust ratings to take that factor into account. Such an adjustment, if done carefully, might serve as a force against grading leniency as well as enhancing the validity of the ratings.

Appendix A

Form X

Instructional Assessment System



Fill in bubbles darkly and completely.
Erase errors cleanly.

FORM **X**

Instructor _____ **Course** _____ **Section** _____ **Date** _____

Completion of this questionnaire is voluntary. You are free to leave some or all questions unanswered.

	Excellent	Very Good	Good	Fair	Poor	Very Poor
1. The course as a whole was:	○	○	○	○	○	○
2. The course content was:	○	○	○	○	○	○
3. The instructor's contribution to the course was:	○	○	○	○	○	○
4. The instructor's effectiveness in teaching the subject matter was:	○	○	○	○	○	○

How frequently was each of the following a true description of this course?

	Always			About Half			Never
5. The instructor gave very clear explanations.	○	○	○	○	○	○	○
6. The instructor successfully rephrased explanations to clear up confusion.	○	○	○	○	○	○	○
7. Class sessions were interesting and engaging.	○	○	○	○	○	○	○
8. Class sessions were well organized.	○	○	○	○	○	○	○
9. Student participation was encouraged.	○	○	○	○	○	○	○
10. Students were aware of what was expected of them.	○	○	○	○	○	○	○
11. Extra help was readily available.	○	○	○	○	○	○	○
12. Assigned readings and other out-of-class work were valuable.	○	○	○	○	○	○	○
13. Grades were assigned fairly.	○	○	○	○	○	○	○
14. Meaningful feedback on tests and other work was provided.	○	○	○	○	○	○	○
15. Evaluation of student performance was related to important course goals.	○	○	○	○	○	○	○

Relative to other college courses you have taken, how would you describe your progress in this course with regard to:

	Great			Average			None
16. Learning the conceptual and factual knowledge of this course.	○	○	○	○	○	○	○
17. Developing an appreciation for the field in which this course resides.	○	○	○	○	○	○	○
18. Understanding written material in this field.	○	○	○	○	○	○	○
19. Developing an ability to express yourself in writing or orally in this field.	○	○	○	○	○	○	○
20. Understanding and solving problems in this field.	○	○	○	○	○	○	○
21. Applying the course material to real world issues or to other disciplines.	○	○	○	○	○	○	○
22. General intellectual development.	○	○	○	○	○	○	○

Relative to other college courses you have taken:

	Much Higher			Average			Much Lower
23. Do you expect your grade in this course to be:	○	○	○	○	○	○	○
24. The intellectual challenge presented was:	○	○	○	○	○	○	○
25. The amount of effort you put into this course was:	○	○	○	○	○	○	○
26. The amount of effort to succeed in this course was:	○	○	○	○	○	○	○
27. Your involvement in this course (doing assignments, attending classes, etc.) was:	○	○	○	○	○	○	○

28. On average, how many hours per week have you spent on this course, including attending classes, doing readings, reviewing notes, writing papers and any other course related work?

○ Under 2	○ 6 - 7	○ 12 - 13	○ 18 - 19
○ 2 - 3	○ 8 - 9	○ 14 - 15	○ 20 - 21
○ 4 - 5	○ 10- 11	○ 16 - 17	○ 22 or more

29. From the total average hours above, how many do you consider were valuable in advancing your education?

○ Under 2	○ 6 - 7	○ 12 - 13	○ 18 - 19
○ 2 - 3	○ 8 - 9	○ 14 - 15	○ 20 - 21
○ 4 - 5	○ 10- 11	○ 16 - 17	○ 22 or more

30. What grade do you expect in this course?

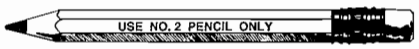
○ A (3.9-4.0)	○ B (2.9-3.1)	○ C (1.9-2.1)	○ D (0.9-1.1)	○ Pass
○ A- (3.5-3.8)	○ B- (2.5-2.8)	○ C- (1.5-1.8)	○ D- (0.7-0.8)	○ Credit
○ B+ (3.2-3.4)	○ C+ (2.2-2.4)	○ D+ (1.2-1.4)	○ E (0.0)	○ No Credit

31. In regard to your academic program, is this course best described as:

○ In your major?	○ A distribution requirement?	○ An elective?
○ In your minor?	○ A program requirement?	○ Other?

Mark Reflex® forms by Pearson NCS MM217972-2 65432 Printed in U.S.A.

ADDITIONAL ITEMS
Use only if directed



1	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
2	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
3	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
4	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
5	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
6	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
7	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
8	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
9	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
10	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
11	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ
12	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪	⓪
①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①
②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②	②
③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③	③
④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④	④
⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥	⑥
⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦	⑦
⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧	⑧
⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨	⑨