Shaping Online Discussion: Assessment Matters

Karen Swan, Jason Schenker, Stephen Arnold, Chia-Ling Kuo

The purpose of this study was to empirically test the common-sense notion that the quality of online course discussions can be shaped by how discussion is assessed. A quasi-experimental design was used to compare discussion activity among 8 sections of an undergraduate class in educational technology at a Midwestern public university. All sections were required to participate in five identical discussions which collectively counted for 10% of their final grade. Two instructors each taught 4 sections of the course, 2 of which were given quality criteria for discussion participation, two of which were not. Results reveal that students in the criteria group responded significantly more often and at greater length to their classmates, and that they read significantly more of their classmates' postings. In addition, discussions in the criteria classes evidenced more posts, more threads, and a greater depth than did discussions in the classes given no discussion criteria.

Theoretical Framework

While assessment is often equated with tests, exams, and evaluation, the term assessment is often used more broadly in education to include its application to learning, as in the following two definitions:

> *"Assessment is defined as the systematic basis for making inferences about the learning and development of students. More specifically, assessment is the process of defining, selecting, designing, collecting, analyzing, interpreting, and using information to increase students' learning and development."* (Erwin, 1991, p. 19)

> *"Assessment is an ongoing process aimed at understanding and improving student learning. It involves making our expectations explicit and public; setting appropriate criteria and high standards for learning quality; systematically gathering, analyzing, and interpreting evidence to determine how well performance matches those expectations and standards; and using the resulting information to document, explain, and improve performance."* (Angelo, 1995, p. 7)

Indeed, value in any instructional system comes from assessment: what is assessed in a course or a program is generally associated with value; what is valued becomes the focus of activity. The link to learning is direct. Instructors signal what knowledge skills and behaviors they believe are most important by assessing them. Students quickly respond by focusing their learning accordingly.

Many theoretical and empirical analyses emphasize the importance of active participation and collaboration among students in promoting the effectiveness of online learning (Benbunan-Fich, Hiltz, 1999; Hiltz, 1997). Indeed, researchers have found that successful online discussion is directly linked to its being assessed (Hawisher, Pemberton, 1997; Jiang, Ting, 2000; Swan, 2001; Swan, et al., 2000). Simply put, this means that to encourage online discussion, one must grade it, and discussion grades must count for a significant portion of final course grades. Some online educators, however, believe that to get the most out of discussions, however, instructors must go further and assess individual discussion postings for the qualities they believe most important (Swan, Shen, Hiltz, 2006).

The research reported in this paper tests the premises of that assumption: the common sense notion that the quality of online course discussions can be shaped by how discussion participation is assessed.

Methods

A quasi-experimental design was used to compare discussion activity among 8 sections of an undergraduate course in educational technology at a Midwestern public university. The educational technology course is required of all students enrolled in teacher preparation programs and is generally taken at the beginning of those programs. Seventy-eight percent of the students enrolled agreed to participate in the study, giving a sample size, after drops, of 119 students (97 female, 22 male), most (96%) of whom were in the 18 to 24 year old range.

While the educational technology is taught face-to-face in a computer laboratory, students in all sections of the course were required to participate in five identical discussions (corresponding to five of the six sections of the course) in the Spring 2006 semester. Participation in online discussion counted for 10% of students' final grade. Two instructors each taught 4 sections of the course. One instructor (Instructor 1) was a white, native English speaking male; the other (Instructor 2) was an Asian, non-native English speaking female. Both were young, beginning professionals. Each instructor's class sections were randomly assigned to one of four treatment conditions.

In the first condition (1), students were told that they had to participate in the discussion and that it would count for 10% of their final grade; in the second condition (2), students were additionally told that they should post one original comment and respond to two of their classmates' comments in each discussion to receive full participation credit. For analysis purposes, these two conditions were collapsed to create a "No Criteria" condition. In the remaining two conditions, students were also given quantitative participation requirements but additionally provided either criteria (condition 3) centered on clear statement and defense of a position, or rubrics (condition 4) for assessing discussion postings on three criteria: the relevance of the posting to previous posts and the discussion topic, its originality, the quality of the writing. These two conditions were collapsed to create a "Criteria" condition for analysis purposes. The collapsed Criteria and No Criteria conditions thus compare the difference between discussions assessed for quality and those that are not. Figure 1 summarizes that comparison.

Figure 1. Group Difference Statistics for Number of Responses Across Modules

No Criteria	Criteria
Condition 1 Participation counts for 10% of the final grade	Condition 3 Participation counts for 10% of the final grade + 1 initial post & 2 responses /discussion required for full credit + grading criteria: postings must clearly state a position defended with at least 1 example from readings or experience
Condition 2 Participation counts for 10% of the final grade + 1 initial post & 2 responses /discussion required for full credit	Condition 4 Participation counts for 10% of the final grade + 1 initial post & 2 responses /discussion required for full credit + grading rubrics which assess postings on 3 criteria: relevance, originality & quality of writing

Overall, 66 students in four course sections were in the criteria condition, while 53 (in four sections) participated in the No Criteria condition.

Data Sources & Analysis

Data collected from all participating students included the number and length of initial postings, the number and length of responses, and the total number of posts read for each of the five required discussions. These data were compared between conditions and instructors using multivariate analysis of variance and non-parametric statistics.

Data concerning the discussions themselves (total posts, total threads, average thread depth) was also collected and compared descriptively between conditions.

Results

Number and Length of Initial Posts. The number of initial posts by students within each module varied from a minimum of 0 to a maximum of 2. Across all modules, students varied from a total of 0 posts to 6 posts. Chi-square tests were conducted to determine if groups presented with a criteria differed from those who were not. The results indicate that, for all five modules, the conditions did not differ significantly on initial posts. The conditions also did not significantly differ on total initial posts across modules, nor were there differences in number of initial posts between students taught by different instructors.

Multivariate analysis of variance was conducted to compare treatment groups and instructors on the length of the initial posts (i.e., total word count across all responses for each module). The multivariate test indicated no significant difference between conditions (Wilks' Lamda = 0.963, $F_{(5, 111)}$ = 0.844, p = 0.522, eta^2 = 0.037) across all modules, but the main effect for instructor was statistically significant (Wilks' Lamda = 0.747, $F_{(5, 111)}$ = 7.534. p = 0.000, eta^2 = 0.253). The students enrolled in the class taught by Instructor 1 wrote significantly longer posts than students in the class taught by Instructor 2 on modules 2-5. This result, which is mirrored in findings on other variables, may indicate a non-native speaker instructor effect that definitely deserves further investigation. However, the interaction between instructor and group was also not statistically significant (Wilks' Lamda = 0.909, $F_{(5, 111)}$ = 2.232, p = 0.056, eta^2 = 0.091), suggesting that instructor effects did not apply differentially across conditions..

Number and Length of Responses to Initial Posts. The total responses to initial posts on any given module generally ranged from a minimum of zero to a maximum of six (however, in module five, one participant made sixteen responses). Combined across all modules, the number of responses ranged from zero to twenty-nine. Chi-square tests were conducted to determine if the number of posts differed by condition (Table 1). On modules one, two, and three, the Criteria Group posted significantly more responses than the No Criteria Group. However, the groups did not significantly differ on modules four and five.

Table 1. Group Difference Statistics for Number of Responses Across Modules

Module	Chi^2	df	p	Eta
1	16.355	3	0.001	0.364
2	17.733	5	0.003	0.276
3	14.883	6	0.021	0.176
4	5.968	5	0.309	0.156
5	3.736	5	0.588	0.089

Combined across all modules, the point-biserial correlation between group membership and number of responses was 0.305, p = 0.001, indicating that the Criteria Group made significantly more responses overall than the No Criteria Group. Figure 2 illustrates those differences. It shows that students not given assessment criteria averaged ¾ as many responses per module as students given assessment criteria. Notice also that students averaging the greatest number of responses per module were those assessed for clear position statements supported by examples (Condition 3).

3

Figure 2. Comparisons of Average Number of Responses by Group & Condition
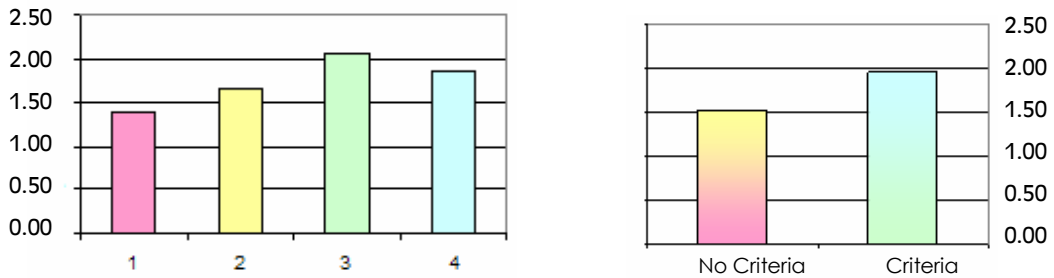
Multivariate analysis of variance was again conducted to compare conditions and instructors on the length of the responses for all five modules. The multivariate test indicated that the main effect for treatment group (Wilks' Lamdạ = 0.893, F (5, 111) = 2.633, p = 0.027, eta^2 = 0.107) and instructor (Wilks' Lamda = 0.745, F (5, 111) = 7.540, p = 0.000, eta^2 = 0.255) were statistically significant. Students in the Criteria Group wrote significantly longer responses than students in the No Criteria Group. Student taught by Instructor 1 wrote significantly longer responses than students taught by Instructor 2. However, the interaction between instructor and treatment group was not significant (Wilks' Lamdạ = 0.920, F (5, 111) = 1.916, p = 0.097, eta^2 = 0.080), indicating that the use of assessment criteria resulted in longer responses regardless of instructor.

The average length of responses for the Criteria and No Criteria Groups by module are given in Table 2. Univariate F tests on length of responses reveal that the Criteria Group wrote significantly more words in their responses than the No Criteria Group in modules one through four. However, there was no significant difference between the two groups on module five. In addition, students taught by Instructor 1 wrote significantly more in their responses than students taught by Instructor Two for all modules except module one.

Table 2. Descriptive Statistics for Length of Responses by Group Across Modules

Module	Criteria		No Criteria	
	Mean	SD	Mean	SD
1	87.72	111.86	40.36	82.36
2	134.55	105.90	92.99	62.95
3	159.26	162.40	122.40	128.16
4	137.46	101.94	106.34	94.35
5	178.50	324.77	116.04	82.74

Total Posts Read. The treatment condition and instructors were also compared on the number of posts and responses that students read. A multivariate analysis of variance was conducted, and the results indicated that the main effect for grouping (Wilks' Lamdạ = 0.882, F (5, 111) = 2.973, p = 0.015, eta^2 = 0.118). Students in the Criteria Group read more messages than students in the No Criteria Group These differences are illustrated in Figure 3 which shows that students in Criteria group read on average nearly twice as many messages as those in the No Criteria group. Figure 3 also shows differences between conditions in numbers of messages read. Notice that students who were assessed in terms of the relevance and originality of their posts (Condition 4) outperformed all other conditions, perhaps because they needed to pay attention to others' postings to make sure theirs met these criteria.

Figure 3. Comparisons of Average Number of Responses Read by Group & Condition
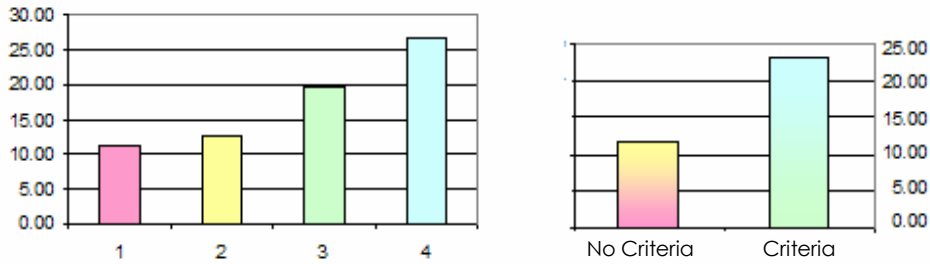
Multivariate analysis of variance also showed a significant main effect for instructor (Wilks' Lamda = 0.896, F (5, 111) = 2.564, p = 0.031, eta ^{2} = 0.104) on the number of messages read. Students taught by Instructor 1 read more messages than students taught by Instructor 2. However, the interaction between treatment group and instructor was not significant (Wilks' Lamda = 0.972, F (5, 111) = 0.630, p = 0.677, eta ^{2} = 0.028), indicating that the effects of assessment criteria applied similarly to both instructors.

The average number of message read for the Criteria and No Criteria Groups by module are given in Table 3. While the descriptive statistics show that members of the Criteria Group read more posts than members of the No Criteria Group across all five modules, univariate tests indicated that the differences were only significant for modules one and four. In addition, students taught by Instructor 1 read significantly more messages than those taught by Instructor 2 for module three only.

Table 3. Descriptive Statistics for Number of Posts Read by Group Across Modules

	Criteria		No Criteria	
Module	Mean	SD	Mean	SD
1	17.53	17.02	8.17	10.52
2	21.53	24.75	15.11	13.59
3	18.89	22.39	14.94	14.63
4	26.65	75.06	8.08	8.38
5	22.79	38.48	15.04	18.71

Descriptive Data Concerning Discussions as a Whole. Because our research question was at its heart concerned with the effects of criteria-based assessment on the quality of online discussion, we also collected data concerned with module discussions viewed as a whole. Data collected included the total numbers of messages posted, the total number of discussion threads, the average number of posts per thread, the average thread depth, and the greatest thread depth for discussions in each of the five modules in every class. Table 4 gives that data collapsed across modules and compared between conditions and instructors.

Table 4: Descriptive Statistics for Whole Discussions Across Modules

	Criteria	No Criteria	Instructor 1	Instructor 2
posts/discussion	58.30	40.55	52.40	46.45
threads/discussion	18.80	15.50	16.90	17.40
avg posts/thread	2.04	1.65	1.99	1.71
avg thread depth	0.98	0.71	0.99	0.71
greatest depth	2.70	1.90	2.65	1.95

These comparisons support the statistical findings. They show discussions in course sections in the Criteria condition had higher ratings on all criteria than course sections in the No Criteria condition, indicating that discussions in that group were somewhat more interactive. Similarly, discussions in classes taught by Instructor 1 had higher ratings on all but one variable (threads/discussion). Interestingly, when viewed at the discussion level, differences between instructors were less than differences between conditions.

Educational Significance

The analyses of student behaviors in online discussions detailed above seem to suggest that students whose discussion behaviors were assessed according to specific criteria were likely to participate more interactively in the discussions than students who were assessed for participation alone. This is shown not only by the descriptive comparisons of whole discussions, but by the fact that statistical differences were found in both responses and messages read, the latter being a measure of virtual interactivity, and not in initial postings. In addition, the lack of interaction effects shows that these differences hold across instructors. The findings have educational significance in that they empirically demonstrate that how online discussion is assessed matters. Future research should investigate how online discussion can be shaped by particular criteria.

The findings also reveal significant differences between students taught by different instructors. This unanticipated finding shows that students taught by a native English speaking male participated more and more interactively than those taught by a non-native speaking female. While a comparison of the two instructors, in this instance, cannot provide any clear determination of such matters, the results do suggest that discussion facilitation may represent previously unidentified difficulties for non-native speakers. This is not an area that has been investigated in the literature and it surely deserves further exploration.

References

T. A. Angelo, *Definition of assessment, AAHE Bulletin,* 7.11.1995.
R. Benbunan-Fich, S. R. Hiltz, *Impact of asynchronous learning networks on individual and group problem solving: A field experiment, Group Decision and Negotiation, 8*, 409-426, 1999.
T. D. Erwin, *Assessing Student Learning and Development*. San Francisco: Jossey-Bass, 14-19, 1991.
G. E. Hawisher, M. A. Pemberton, *Writing across the curriculum encounters asynchronous learning networks or WAC meets up with ALN,* Journal of Asynchronous Learning Networks, 1 (1), 1997 http://www.sloan-c.org/publications/jaln/v1n1/v1n1_hawisher.asp, 01.07.2005
S. R. Hiltz, *Impacts of college-level courses via asynchronous learning networks: some preliminary results,* Journal of Asynchronous Learning Networks, 1 (2), 1997.
M. Jiang, E. Ting, *A study of factors influencing students' perceived learning in a web-based course environment,* International Journal of Educational Telecommunications, 6 (4): 317-338, 2000.
K. Swan, *Virtual interactivity: design factors affecting student satisfaction and perceived learning in asynchronous online courses,* Distance Education, 22(2): 306-331, 2001.
K. Swan, J. Shen, R. Hiltz, *Assessment and collaboration in online learning,* Journal of Asynchronous Learning Networks, 10 (1), 45-62, 2006.
K. Swan, P. Shea, E. Fredericksen, A. Pickett, W. Pelz, G. Maher, *Building knowledge building communities: consistency, contact and communication in the virtual classroom,* Journal of Educational Computing Research, 23(4): 389-413, 2000.

Information about the author

Karen Swan is Research Professor in the Research Center for Educational Technology at Kent State University and a faculty member in the Instructional Technology Program in the College and Graduate School of Education, Health and Human Services. Her current research focuses on online learning, mobile computing and on student learning in ubiquitous computing environments. Dr. Swan has authored several hypermedia programs and co-edited two books, *Social Learning from Broadcast Television* and *Ubiquitous Computing in Education: Invisible Technology, Visible Impac*t. She served as a project director on several large scale grants. She is a member of the Advisory Board for the Sloan Consortium on Asynchronous Learning Networks, the Special Issues Editor for the *Journal of Educational Computing Research,* and Editor of the *Journal of the Research Center for Educational Technology*, and is the chair of the Media, Culture and Curriculum Special Interest Group of the American Educational Research Association.