تطوير وإثبات صلاحية نظام لتقويم الأستاذ والمقرر في الجامعة الأمريكية في بيروت

كرمة الحسن*

الملخص

تـصف هذه الدراسة كيفية تطوير نظام جديد لتقويم الأستاذ والمقرر في الجامعة الأمريكية في بيروت، من خلال تطوير نسختين: ألكترونية وورقية، ولقد تم تطبيق الأداة خـلال العـام الدراسي ٢٠٠٣ في كل الكليات، وجرت مقارنة النسختين من حيث نسب الإجابة والثبات، وقد تم التأكد من الصدق من خلال إجراء التحليل العاملي، بالإضافة إلى دراسة أثر بعض الخصائص المؤثرة في تقدير الطلاب.

وقد أظهرت النتائج أن النسختين تتمتعان بثبات داخلي عالٍ، ولكن نسب الإجابة في النسخة الألكترونية كانت منخفضة، كما أظهر التحليل العاملي وجود عاملين أساسيين يقيـسان فعالية الأستاذ والمقرر، أما نتائج دراسة العوامل المؤثرة فقد أظهرت تطابقاً مع الأدبـيات حـول الموضوع وأكدت صدق الأداة ، وأوصت الدراسة بالتوقف المؤقت عن استعمال النسخة الألكترونية وذلك لضعف نسب الإجابة، كما قدمت توصيات للقيام بأبحاث مستقبلية لتطوير الأداة.

* مدير مكتب القياس والبحث المؤسسي – الجامعة الأمريكية – بيروت – لبنان.

Development and Validation of an Instructor Course Evaluation (ICE) System at the American University of Beirut

Karma El Hassan*

Abstract

This study describes the development of a paper and an on-line version of a new instructor course evaluation (ICE) system at the American University of Beirut, Lebanon. The ICE was administered in spring and fall 2003 to students in all faculties. Both versions were compared with respect to response rates and reliabilities. In addition, the validity of the paper version was investigated using factor analysis and the effect of certain 'biasing' characteristics on student ratings was examined. The results revealed that both versions had similar high internal consistency reliabilities; however, the response rate on the on-line version was much lower. Factor analysis of responses revealed two factors measuring instructor and course effectiveness. Investigation of the effect of certain 'biasing' characteristics on student ratings was in agreement with those reported in the literature and confirmed the validity of the ICE. Because of low response rate, the on-line version was temporarily discontinued and recommendations for further research were presented.

* **Director Office of Institutional Research & Assessment (OIRA) - American University of Beirut.**

Introduction

Students' ratings of instruction are widely used in universities and colleges as they serve a variety of important practical reasons. Initially, they were used to help students make better course selections. Currently, they are used by faculty to improve their teaching and courses and by administration to make personnel and program decisions. In recent years, there has been an increase of nearly 20% in use of student ratings.

Approximately 86% of liberal arts colleges and 100% of large research universities systematically collect student ratings of instruction (Seldin 1999). Students' evaluations have been extensively studied in education. Most of the research deals with the dimensionality, validity, reliability and generalizability of students' ratings of instruction and the investigation of the potentially "biasing" factors that could affect these ratings. More recently, several studies investigated web-based surveys and their strengths and potential methodological issues (Hmieleski & Champagne, 2000; Roscoe, Terkla, & Dyer, 2002; McGourty, Scoles and Thorpe, 2002; Theal, 2000).

Various validity studies conducted tried to control for the effect of potentially 'biasing' characteristics that may influence student ratings.

Research has demonstrated that some of the instructor variables (e.g., age, sex, teaching experience), student characteristics (e.g., age, sex) and course characteristics (e.g. class size, time of the day) have little or no effect on student ratings (Feldman, 1986, Marsh, 1984). Researchers, however, do report the following relationships:

(1) Students tend to rate courses in their major fields and elective courses higher than required courses outside their majors (McKeachie, 1979, Marsh and Dunkin, 1992).

(2) Ratings in higher-level courses tend to be higher than in lower-level courses (Marsh 1987).
(3) Ratings can be influenced by class size (small classes receive higher ratings), by discipline (humanities instructors tend to receive higher ratings than instructors in the physical sciences) (Marsh et. al, 1992).
(4) Grade expectation affects ratings; students expecting high grades give higher ratings (Howard & Maxwell, 1980; Marsh et. al, 1992).

With respect to on-line administration of the student ratings, research conducted investigated reliability and validity of such administrations, in addition to their effect on response rates, bias in responses, anonymity/confidentiality and representative ness of the data issues (Hmieleski et. al., 2000; Underwood, Kim, & Matier, 2000; Roscoe, et al., 2002).

The primary objectives of this paper are to (1) describe the development of the paper and on-line versions of the new instructor course evaluation (ICE) system at the American University of Beirut (AUB), Lebanon, and (2) compare both versions in terms of reliability, validity, and response rates. In addition, the validity of the paper version was further investigated using factor analysis and the effect of certain 'biasing' characteristics on student ratings was examined.

Method

Development of the Instructor Course Evaluation (ICE) System

The Instructor Course Evaluation System (ICES) was established fall 2001 to serve as a component in evaluating teaching effectiveness at AUB. It is based on student evaluations of teaching and is to be

administered before the end of every semester by the Office of Institutional Research & Assessment (OIRA.) at the University. The Cafeteria Model was selected to provide the underlying framework for the Questionnaire. Accordingly, the Questionnaire includes questions on instructor, on course and on student learning outcomes and development in order to provide information for all stakeholders (deans, department heads, students, and faculty).

The ICE Questionnaire includes the following components:

(1) Student background items covering major, grade-point average, class, required / elective status, expected grade in course, gender, etc.

(2) Core items (19) to be included on all forms. These are generic items that can apply to all courses irrespective of course design or size, and they can be used for normative scores and comparison across courses and over time to show improvement. They cover instructor (10), course (7), and student learning outcomes (2) and they include global evaluation items (3).

(3) Specific items selected by department/faculty (11-12) from item bank depending on type of course (lecture, seminar, lab, studio) and its size. Item bank includes specific items for large lecture courses, for labs/studio/clinical teaching classes, and for discussion classes. In addition, the item bank includes extra items on instructional methodology, student interaction and rapport, feedback and evaluation, assignments and student development. Items will be selected from bank by faculties to supplement core questionnaire depending on type of course and kind of information required (course category).

(4) Open-ended questions focusing on instructor strengths and weaknesses and requesting suggestions for improvement.

Administration of the ICE

The ICE was administered end of fall term 2001-2 in two versions on a pilot basis, an on-line version to students in two faculties (Faculty of Engineering & Architecture, FEA: School of Business, SB) and a paper-based version for students in other faculties. For the on-line version, OIRA in collaboration with PC Support Unit worked on electronic administration of the ICE Questionnaire using the Banner system. Each registered student in every course was contacted before end of term and requested to fill relevant ICEs.

All data was then transferred to an SPSS file and then analyzed and reported. The system was made user-friendly enabling students to complete as many surveys as they wanted at any one time and to stop whenever they want to with the option to return later before final submission of the evaluation. Students also typed in their responses to the open-ended questions and their comments. The system also enabled its administrators to monitor response rate and to automatically send reminder e-mails to students who have not responded yet. Students worried about confidentiality of information but were assured of the security and anonymity of their responses. For the paper-based version, ICE forms were prepared by OIRA, sent to department chairs and administered through departmental graduate assistants. To enhance reliability and validity of obtained results, detailed administration guidelines were provided to graduate assistants. The completed ICE Questionnaires were computer scored and analyzed. Reports were issued to instructor, departmental chair and dean covering the following:

(1) Frequencies and percentages of responses to each item, and means, standard deviations for each item, for each subscale, and for the whole scale.

(2) Percentile ranks as compared to category of courses (e.g. Humanities, Sciences, Social Sciences), faculty, and the university for each item, subscale and for the scale.

Finally, a summary report was prepared by OIRA describing the administration and providing a comparative interpretation of the results by course categories and by faculty, in addition to the problems encountered during administration

Results

Response Rates

Overall response rates for fall and spring 2002 ICE administrations were 67% and 53%, respectively. Table 1 compares rates for paper and on-line versions.

Table (1)

Response Rates for ICE Fall and Spring Administrations

	Fall	Spring
Paper-based	69%	66%
On-line	47%	31%
Overall	67%	53%

Reliability Analysis

Internal consistency reliability revealed very high coefficients for both the on-line and paper versions. Table 2 provides reliability coefficients obtained for each subscale and for the whole ICE for spring and fall 2001-2.

Table (2)

Reliability of Scale and Subscales

	No of items	Spring		Fall/ Paper
		Paper	On-line	
Instructor effectiveness subscale	10	.94	.87	.90
Course effectiveness subscale	7	.90	.90	.83
Whole scale	19	.96	.94	.93

Validity Issues

The construct validity of the ICE paper version was addressed by investigating the factor structure of the scale and the relationship between certain 'biasing' variables and evaluations.

With respect to the biasing variables, independent samples t-test revealed significant gender differences with females ($\underline{M}$ = 4.0, SD = .87)

exhibiting slightly higher means than males (M = 4.1, SD = .80), t = -4.48, p< .00 on all subscales. The significance could be attributed to the large sample size (7500). ANOVA revealed significant differences also by level of students. Freshmen and sophomores had significantly lower learning outcomes evaluations than juniors and seniors, F (7, 7233) = 11.04, p< .00, similarly sophomores had significantly lower course evaluations than juniors and seniors, F (7, 7258) = 5.94, p < .00. As to motivation behind the course, elective courses from major and outside major got significantly higher evaluations than required courses outside major and university requirements, F (5, 7165) = 3.81, p < .00. Grade expectations of AUB students were exceptionally high with 70% of students expecting a grade of 80 or higher.

The correlations of grade expectation with evaluations were weak and negative (-0.18 to –0.22) but significant at the 0.01 levels. Apparently students with higher expectations gave lower ratings. Similar weak and negative correlations were obtained when global items of instructor and course effectiveness (items 10 and 17) were correlated with grade expectations. This finding supports the validity of global ratings of instruction. However, correlations of evaluations with actual grades obtained were low (0.18-0.25) but positive and significant at the .00 level. As to differences between evaluations that are related to course subject, lower ratings were obtained by science and engineering instructors/courses than by humanities/social science courses as revealed by Table 3.

Table (3)

Mean Ratings by Subject

Dimension	Subject				
	Humanities	Social Science	Education	Science	Engineering
Instructor effectiveness	4.0	4.1	4.0	3.9	3.8
Course effectiveness	4.1	3.8	3.7	3.7	3.7
Learning outcomes	3.8	4.0	4.0	3.7	3.9
Additional items	3.9	4.0	4.0	3.9	3.7
Overall instructor	4.0	4.0	3.9	3.8	3.7
Overall course	3.7	3.8	3.7	3.6	3.6
Average	3.9	4.0	3.9	3.8	3.7

As to factor analysis of the whole scale, principal components analysis revealed the existence of two factors accounting for 64% of the variance. The first factor accounted for 57% of the variance and had items that included instructor, course and learning outcome effectiveness. The second factor accounted for 7% of the variance and included items mostly related to course effectiveness. Table 4 reports item loadings on two factors. The moderate correlations among the subscales confirm the above findings. Instructor/course correlations were 0.78, while instructor/learning outcomes were 0.62 and course/learning outcomes were 0.66. Factor analysis of each subscale revealed one factor for each accounting for 65% of the variance of the instructor effectiveness subscale and 57% of course/learning outcomes effectiveness subscale.

Table (4)

Item Loadings on Two Principal Components Factors

Items	Factor 1	Factor 2
4. Communicated his/her subject well.	.832	-.235
10. Rate instructor's overall teaching effectiveness.	.830	-.144
9. Provided helpful feedback.	.810	-.157
5. Stimulated interest in the course subject.	.808	-.121
13. Organization of course was adequate and logical.	.784	8.8E-02
1. Was well prepared for the class.	.780	-.315
12. Covered stated objectives.	.778	2.4E-02
11. Objectives and requirements were clearly presented.	.773	2.9E-02
6. Demonstrated positive attitude towards students.	.762	-.185
7. Encouraged students to thing for themselves.	.762	-.185
17. Rate the overall quality of the course.	.759	.359
2. Was knowledgeable about the subject.	.753	-.332
8. Evaluated work fairly.	.738	-6.3E-02
15. Course was appropriately paced.	.727	.318
19. In this course I have learned something that I consider valuable.	.720	.274
18. This course increased my interest in the subject.	.710	.350
3. Instructor was not readily available for consultation outside of class.	.690	-.251
14. Amount of work required appropriate for the credits received.	.651	.374
16. Course material was **not** too difficult.	.571	.472

Discussion

Response Rates

The overall response rates for the fall and spring administrations are acceptable and within rates reported in the literature (Hmieleski et. al., 2000). However, the on-line version response rate was low and affected the reliable use of the results, as some of the samples were not really representative of all students, and certainly nonrandom. This is in spite of our frequent reminders to students and our requesting the faculty members to keep reminding students to fill out the on-line version and to motivate them to do so.

In addition, the researcher met with student representatives to explain the system, to enhance their ownership of the process and to stress the importance of students' taking it seriously. Some technical problems encountered during fall administration like inability of students to access their surveys from some servers and students' fear of confidentiality may have accounted for low response rate. Anyway, low response rate is a pervasive problem in web-based surveys as experience to date of return rates reveals return rates of 30-40% at best (McGourty et. al., 2002). This problem, in addition to the total absence of control over how and with whom the evaluations were completed, made us resort to temporarily stopping the on-line version of the instructor course evaluation

This is despite the numerous privileges gained by it like obtaining immediate student feedback via automated results, organized typed-in student comments and extensive savings in administrative and paper work.

Reliability of the ICE

Internal consistency reliability estimates of both the on-line and paper versions of the ICE revealed comparable and excellent coefficients for both the whole scale and each of the subscales. This high reliability indicates that traits measured by the scale are similar and highly consistent. It does not indicate the degree of error due to the lack of agreement among different students. For this inter rater agreement should be determined. Similarly, we could not determine stability of ratings over time, as we would need data for several years to investigate this question.

Validity Issues

The basic question concerning validity investigates the extent to which the ICE items measure some aspect of teaching effectiveness. As there is no agreed upon criterion of effective teaching (Cashin, 1995), we will try to collect information that either supports or contests the conclusion that the ICE items reflect effective teaching.

The results of the factor analysis and subscale inter correlations confirmed the construct of the scale. The moderate subscale correlations confirmed that the scale is measuring interrelated traits yet they are separate. The two-factor structure confirmed that these traits measure instructor and course effectiveness. Similarly, the subscale factor structure confirmed the unitary trait measured by each subscale. These results confirm findings in the literature that construct of teaching effectiveness is multidimensional (instructor, course, learning outcome and within each probably several dimensions) and that these dimensions are related and reflect a large homogeneous general trait underlying all "effective teaching" (El Hassan, 1995).

As to the results of investigating the effect of certain 'biasing ' variables on student ratings using the ICE, most of the findings are in agreement with those in the literature and confirm the validity of the ICE. The significantly lower course and learning outcome ratings by freshmen and sophomore students is confirmed by the literature (Aleamoni, 1981, Cashin, 1995, Braskamp & Ory, 1994) and could be attributable to class size and to interest of the students. In general, lower level classrooms are larger and students have less chance for taking courses of interest to them. Both of these variables are known to lead to lower ratings of instruction (Centra, 1993; Braskamp et. al., 1994).

Similarly, the lower ratings obtained by university requirements and required courses outside major are attributed to lower interest of students in these courses. Research extensively documents the positive effect of student motivation to take a course on ratings. Instructors are more likely to receive higher ratings in classes where students had prior interest in the subject matter and lower ratings are expected when the course is taken as a major requirement or a general education requirement (Marsh & Dunkin, 1992; Cashin, 1995).

As to lower ratings found in science/engineering courses than humanities/social sciences courses, the literature strongly documents these differences (Cashin, 1990; Centra, 1993; Feldman, 1978; Hoyt & Perera, 2001). However, the reasons for these differences are not clear yet (Cashin, 1995). They could be attributable to the fact that these courses are more quantitatively oriented and today's students are less competent in those skills.

The investigation of other 'biasing' or controlling variables revealed some discrepancy with literature findings. For example, no

gender differences were identified in student ratings (Centra, 1993; Feldman, 1993) however, some significant differences were reported in ICE ratings. Reason could be attributable to large sample size, (N=7,500); it is likely that very small differences in mean responses (0.1 or less) were found to be significant. For example mean instructor effectiveness scores for males and females were 4.0 and 4.1, respectively. Alternatively, the difference could be attributable to gender of student/gender of instructor interaction that was not investigated in this study. Feldman, 1993 reports that female students tended to rate female teachers higher, and male students rated male instructors higher.

The low positive correlation between student ratings and obtained grades is confirmed in the literature, however, the grade expectation inverse correlation with student ratings is not supported (Braskamp et. al., 1994; Marsh & Dunkin, 1992). The literature found a positive, but low correlation (0.10-0.30) between student ratings and expected grades. Student's obtaining/expecting high grades give higher ratings. This study reported a low but negative correlation between subscale means and expected grade. Similarly, global instructor and course ratings produced similar low but negative correlations. Reason could be because a high percentage (70%) of courses evaluated were not elective courses and thus student motivation was reduced and accordingly, the inverse relationship. Grading leniency has been proposed as a possible explanation for the positive relationship (Cashin, 1995). AUB is well known for its strict grading policy and relatively lower grades so this could be another possible reason for this inverse relationship.

Conclusion and Recommendations

This study has described the development and validation of a new instructor evaluation system at the American University of Beirut. The results provide evidence for the reliability and validity of the instrument. Despite the low response rate on the on-line version and its temporary discontinuing, work will continue on improving the conditions for future administrations in terms of enhancing student motivation, assuring privacy and confidentiality and improving the technical aspects.

A significant step in this regard has happened this year with the publishing of departmental and faculty ICE ratings on OIRA website for students and faculty information. Sharing of the results will give a signal to the seriousness of the process and will enhance the partnership involved in teaching/learning between students, faculty and the administration. Furthermore, interpretation of the evaluations should take into consideration student class level, course status (required or elective) and discipline. The fear of grade inflation and 'popularity contest' due to the use of student ratings is unjustified as evidenced by low correlation between mean ratings and grades attained.

Further research should continue working on the validity of the ICE, as validation is an on-going process. Further reliability investigation should concentrate on the stability of the ICE over different administrations and on the generalizability of the evaluations across different courses and instructors. Also, future research should re-examine grade expectation and gender relations with evaluations to determine if they are real or sporadic.

References

- Aleamoni, L. M. (1981). Students ratings of instruction. In J. Millman (Ed.), Handbook of teacher evaluation (pp. 110-145). Beverly Hills, CA: Sage.
- Braskamp, L. A., & Ory, J. C. (1994). Assessing faculty work: Enhancing individual and institutional performance. San Francisco: Jossey-Bass.
- Cashin, W. E. (1990). Students do rate different academic fields differently. In M. Theal, & J. Franklin (Eds.), Student ratings of instruction: Issues for improving practice: New Directions for Teaching and Learning, No. 43 (pp. 113-121). San Francisco: Jossey-Bass.
- Cashin, W. E. (1995). Student Ratings of Teaching: The research revisited. IDEA Paper No. 32. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.
- Centra, J. A. (1993). Reflective faculty evaluation: Enhancing teaching and determining Faculty effectiveness. San Francisco: Jossey-Bass.
- El Hassan, K. (1995). Students' ratings of instruction: Generalizability of findings. Studies in Educational Evaluation, 21, 411-429.
- Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers; what we know and what we don't. Research in Higher Education, 9,199-242.

- Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. Research in Higher Education, 24, 129-213.
- Feldman, K. A. (1993). College students' views of male and female college teachers: Part II-Evidence from students' evaluations of their classroom teachers. Research in Higher Education, 34, 151-211.
- Hmieleski, K. & Champagne, M. V. (2000). Plugging in to course evaluation. Technology Source, September /October. Retrieved April4, 2003 from http:// ts.mivu.org/default.asp?show=article&id=795.
- Howard, G. S., & Maxwell, S. E. (1980). The correlation between student satisfaction and grades. A case of mistaken causation? Journal of Educational Psychology, 72, 810- 820.
- Hoyt, D. P. & Perera, S. (2001). Are quantitatively-oriented courses different? IDEA Research Report No.3. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, potential biases, and utility. Journal of Educational Psychology, 76, 707- 754.

- Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-388. Marsh, H. W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart (Ed.) Higher Education: Handbook of theory and research (Vol. 8, pp. 143-233). New York: Agathon.
- McGourty, J., Scoles, K., & Thorpe, S. (2002, June). Web-based student evaluation of instruction: Promises and pitfalls. Paper presented at the 42nd annual forum of the Association for Institutional Research, Toronto, Canada.
- McKeachie, W. J. (1990). Student ratings of faculty: A reprise. Academe, 65, 384-397. Roscoe, H. S., Terkla, D. G., & Dyer, J. (2002, June). Administering Surveys on the Web: Methodological issues II. Paper presented at the 42nd annual forum of the Association for Institutional Research, Toronto, Canada.
- Seldin, P. & Associates (1999). Changing practices in evaluating teaching. Bolton, MA: Anker.
- Theal, M. (2000). Electronic course evaluation is not necessarily the solution. Technology Source, November /December. Retrieved April 4, 2003 from http://ts.mivu.org/default.asp?show=article &id=823.

- Underwood, D., Kim, H., & Matier, M. (2000). To mail or to web: Comparisons of survey response rates and respondent characteristics. Paper presented at the 40th annual forum of the Association for Institutional Research, Cincinnati, OH.

تاريخ ورود البحث : ٢٠٠٤/١٢/٢٧م
تاريخ ورود التعديلات : ٦/٢٧ /٢٠٠٥م
تاريخ القبول للنشر : ٦/٢٨ /٢٠٠٥م