

## **Analyzing Assessment: Case Studies at the Individual and Department Level**

Douglas J. Gillan, Peter Foltz, and C. Lausanne Renfro-Fernandez

Department of Psychology

New Mexico State University

[gillan@crl.nmsu.edu](mailto:gillan@crl.nmsu.edu), [pfoltz@crl.nmsu.edu](mailto:pfoltz@crl.nmsu.edu), and [crenfro@nmsu.edu](mailto:crenfro@nmsu.edu)

Instructional design requires rigorous empirical feedback from students to facilitate the redesign of instructional systems. Two case studies from the NMSU Psychology Department exemplified the collection of student data and innovative analyses of those data to test hypotheses about the design of the instructional systems. A multiple regression analysis of an instructor's student evaluations over 13 years showed that three variables – number of years teaching, level of the course, and the type of the course accounted for over two-thirds of the variance in course evaluations. Analyses of outcomes assessment data concerned with knowledge of psychological concepts, statistics, and methods indicated that graduating majors in psychology had conceptual and methodological knowledge resembling that of first-year graduate students, but had deficits in statistical knowledge. The use of the data and analyses to identify specific needs for instructional redesign was discussed for both cases.

The design of human-usable systems requires repeated cycles of design-test-analysis-redesign (e.g., Gould & Lewis, 1983). Design is very difficult to get right on the first try, perhaps with the exception of Zeus from whose head Athena leaped full blown. Given that instructional systems are systems intended to be used by people, we might expect that they also require this iterative design process. The test and analysis that plays the central role in the iterative design process must come from thoughtful assessment of the human users of the system. In the case of instructional design, the users are typically the students. Thus, improvement in all aspects of teaching -- from instruction by individuals to the design of curricula by departments -- requires feedback from students.

Ideally, the methodology for deriving feedback in the iterative design cycle will consist of (1) collecting data from students in a systematic and rigorous manner, (2) applying analytical techniques to those data, and (3) interpreting the analysis in terms of actions that can lead to changes (e.g., Gordon, 1994). Failure in any of these three steps will result in inadequate feedback and, as a consequence, an inadequate design. For example, if data collection were haphazard, the data might not provide a valid and reliable measure of students' responses to the instructional system and the instructor. Or, imagine an assessment that produced reams of quantitative data and associated high-powered statistical analyses, but that did not lead to conclusions from which a designer could decide on modifications to the instructional system. Note that our discussion has been somewhat abstract -- who is the designer and what is the instructional system? Often the instructional system will be a course and the designer will be the instructor. But the above general principles of design should even apply to an entire curriculum designed by a committee or an administrator.

To move the present discussion of the role of feedback in the design of instructional systems from the abstract plane to a more concrete and comprehensible level, the present paper focuses on two specific case studies in the collection, analysis, and application of student feedback -- one case at the level of a single instructor and the second case at the level of the department. Both cases come from the Department of Psychology at New Mexico State University.

### **Case Study 1 – Assessing the Individual Instructor**

At the end of each course, faculty members in the Department of Psychology collect student evaluations which include ratings of the course and instructor using a five-point scale that is structured as the A to F grading scale (and scored as 4 to 0, respectively). Research suggests that student evaluations of courses are an imperfect measure of the quality of instruction – they can be influenced by (1) student performance related factors, including the score on the course's first test, the overall grade in the course, and student participation, and (2) course-relevant variables, such as the discipline of the course and whether the course is required or an elective (see Adams, 1997, for a review). In

addition, factors that are more traditionally considered to be important aspects of instruction influence student evaluations, including the instructor’s enthusiasm during lectures (Williams & Ceci, 1997) and student satisfaction with the course (Abrami, d’Apollonia, & Cohen, 1990). Also, although student evaluations may be biased by factors unrelated to teaching, they have been shown to have diagnostic value in helping instructors identify what they do well as well as what they do poorly – that is to say, student evaluations can serve the function of a formative or prospective evaluation (Lewis, 1998).

In an investigation of the relation between student evaluations and the features of a course, the first author has developed and tested a set of alternative regression models to determine the optimal predictors of the rating of the quality of instruction in the 48 courses that he has taught. Those courses have included introductory courses, methods courses at the undergraduate and graduate level, general education courses, as well as specialized content courses for Psychology majors and graduate students. The initial regression model predicted the mean student evaluation of courses based on three predictor variables – the number of years that the instructor has been teaching, the level of the class (lower division undergraduate, upper division undergraduate, and graduate), and the type of course (method or content). The hypothesized relations of these variables with quality of instruction were that instructor ratings should increase with the number of years, should increase as the level of the course goes up, and should be higher for content than methods courses. Figure 1, which shows the simple relation between years of teaching and student evaluations, supports the hypothesis that student evaluations increase with years of instructor experience. Likewise, the mean evaluation for methods courses has been 3.39, whereas the mean for content is 3.65. The mean student evaluations as a function of the level of class also support the hypothesis – 3.12 for undergraduate lower division, 3 and 3.70 for both undergraduate upper division and graduate courses.

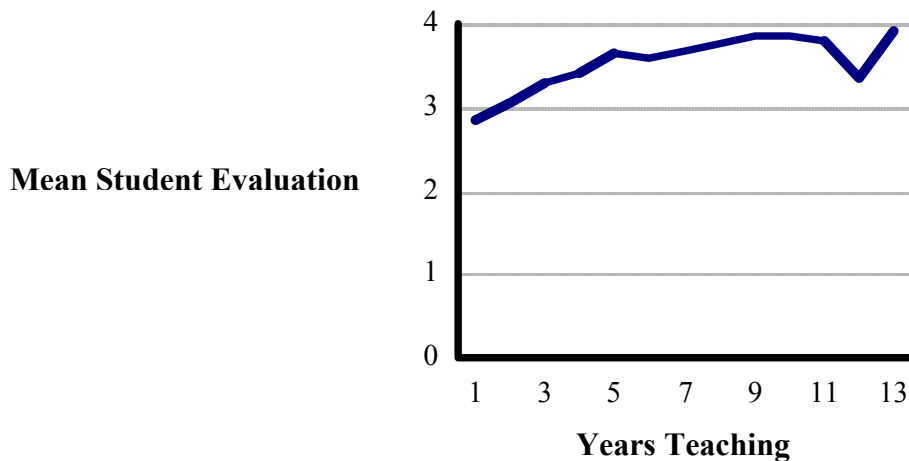


Figure 1. Mean student evaluations as a function of the number of years that the first author has taught at the university level.

A multiple regression analysis (e.g., Cohen & Cohen, 1983) that incorporates these three variables also supports the above hypotheses: The basic model,  $Quality\ of\ instruction = 2.52 + .06(Years\ teaching) + .33(Course\ Type) + .20(Course\ Level)$ , accounts for 62% of the variance in quality of instruction ratings from students. Each variable was related to a significant amount of unique variance (20%, 11%, and 15%, respectively,  $t$ 's [ $df = 1$ ] = 4.93, 3.95, and 4.26, respectively, all  $p$ 's <.001).

Can we improve on the basic model either by adding other key variables or examining the relations between evaluations and the variables in the basic model? In addition to the three variables in the basic model, we hypothesized that two additional variables might be related to student evaluations. One variable was the number of students in the class – more students would be likely to mean that each individual student would have less access to the instructor, leading to lower satisfaction and resulting in lower evaluations. The second variable was the number of times that the instructor had taught a specific course – with repeated teaching of specific material, an instructor might become more skilled at communicating that information, resulting in higher evaluations. Accordingly, we analyzed

student evaluations using a five variable model -- Quality of instruction =  $2.62 + .06(\text{Years teaching}) + .32(\text{Course Type}) + .16(\text{Course Level}) - .006(\text{Number of Times}) - .0006(\text{Number of Students})$ . This enhanced model still accounted for 62% of the variance in student evaluations, with each of the three variables from the basic model still accounting uniquely for a significant amount of variance (15%, 4%, and 14%, respectively,  $t$ 's [ $df = 1$ ] = 4.14, 2.19, and 4.04, respectively, all  $p$ 's < .05), but with the newly added variables accounting for miniscule and nonsignificant amounts of variance. The number of times having taught a course accounted for only .06% of the variance and the number of students accounting for only .1% of the variance in student evaluations. Thus, adding these two variables did not improve the fit of the model to the data. In other words, based on the law of parsimony, the basic model is preferable to the enhanced model. The third regression model that we examined involved a simple modification of the basic model. The basis for the modification came from the data shown in Figure 1. The increase in student evaluations over the years of teaching appears to be reaching an asymptote that is due to the approach of the limit of the scale. Consequently, we replaced the number of years with the logarithm of the number of years in the revised regression model. The revised model -- Quality of instruction =  $2.36 + .78(\log[\text{Years teaching}]) + .32(\text{Course Type}) + .18(\text{Course Level})$  -- accounted for 68% of the variance in student evaluations, with the variance in student evaluations accounted for at 27%, 14% and 11%, respectively ( $t$ 's [ $df$ 's = 1] = 6.13, 4.49, and 3.94, respectively, all  $p$ 's < .001). So, by transforming the years of teaching into a logarithm but not adding any new variables, the revised model accounts for more variance in evaluations and the years teaching uniquely accounts for a greater proportion of variance (20% in the untransformed value vs. 27% with the log transform).

The results of the regression analysis provide some interesting information that can be (and has been) applied to improving the teaching of an individual instructor. First, the strong effect of the number of years teaching and the absence of an effect of the number of times teaching a particular course suggests that this teacher has acquired general, transferable skills during the last 13 years, but that course-specific knowledge has not been acquired (or at least has not affected student evaluations). The general, transferable skills might include the ability to "read" a class (i.e., to determine when they understand material and when they are confused), how to react in real time during a class and with time to reflect after a class when students don't understand material, how to structure a class to maximize the relation between student's learning styles and the presentation of information. Course-specific skills would include restructuring a specific course based on what activities or lectures were successful in communicating information. The present analysis suggests that, for this instructor, the course-specific changes that he implemented did not change student evaluations. Thus, he might want to reexamine his approach to course-specific modifications. Second, this instructor does better in content courses than in method courses. One possible way to improve student evaluations would be to try to approach methods courses, to the extent possible, in the same manner as content courses in course design, classroom activities, and assessing knowledge. Third, the effect of course level on student evaluations show that lower division undergraduate courses (like Introductory Psychology) produce lower evaluations than do upper division or graduate courses. Is this an intrinsic characteristic of the lower division courses or the students in the courses or can certain aspects of the higher-level courses (such as, dealing with ideas with which the students are already familiar) be imported into the lower level ones?

### **Case Study 2 – Assessing the Department**

The Psychology major at NMSU is intended to educate students in a variety of topics related to behavior and the mind, including social and cultural interactions, cognition, human development, and brain-behavior relations to prepare students to (1) enter and succeed in graduate programs, and (2) obtain employment in a variety of occupations dealing with behavioral issues. Students who receive a degree from NMSU should have a strong understanding of the methodological and scientific bases of psychology and the ability to apply that knowledge to a broad range of tasks and environments. Thus, the departmental goal is to ensure that students who graduate as Psychology majors have strong knowledge in the science base of psychology and a high level of ability in research methods.

The structure of the Psychology major has been designed with the department's pedagogical goals in mind. In 1995, the department reformed its curriculum so that three specific courses were required of all majors: Introductory Psychology, a statistical course from either the math or experimental statistics department, and an

experimental methods course. In order to proceed to the courses in three other required areas (from which majors were required take at least one course), students need to complete those required courses. Those areas are basic mechanisms, acquisition and use of knowledge, and understanding behavior. In addition, majors took at least one course from a fourth broad category that includes abnormal psychology, developmental psychology, and history and systems of psychology. This fourth category does not have a statistical or methodological prerequisite. These distributional requirements ensure that psychology majors are exposed to the breadth of content in the discipline of psychology. In 2000, the department implemented three new requirements – a course that requires an application of methodological knowledge, a laboratory course in biology, and a course in the philosophy of science.

Just as the university requires individual instructors to obtain student evaluations of courses, it requires departments to obtain assessments of graduating majors. This process, known as outcome assessment, can be handled as each department sees fit. The Psychology Department (particularly the department head at that time, Ken Paap, and the head of the Outcomes Assessment committee, Peter Foltz) developed an outcome assessment plan to determine if the program of instruction that we have designed has met the instructional goals described above. Accordingly, the department has developed measurement instruments that assess graduating majors' (1) conceptual knowledge, (2) research methodology knowledge, and (3) statistical knowledge. In a recent assessment (not identified to protect the anonymity of the students), a high percentage of the graduating psychology majors (77%), volunteered and participated in the outcomes assessment test. In addition, 100% of the first year graduate students and a random sample of undergraduate students taking an Introductory Psychology course voluntarily participated in the outcomes assessment test. Data were also available from current and recent faculty who had previously completed all of the outcome assessment instruments.

The measurement instrument for conceptual knowledge was based on rating the similarity of 135 pairs of concepts from throughout the subdisciplines of psychology. Among the advantages of using similarity ratings are (1) that they provide an indirect measure of conceptual knowledge, one that may be less subject to sources of test bias than are more explicit, direct approaches (see Schvaneveldt, 1990), and (2) they permit comparisons between the knowledge of various groups, from relative novices to experts. We compared Psychology majors with the following groups: (1) a random sample of students from Introductory Psych, (2) our first year graduate students, and (3) faculty from the department. If we consider the faculty to be the experts and the Intro students to be the most novice group, then successful instruction of our majors should result in a higher correlation between majors and faculty than between Introductory students and faculty. A second index of successful instruction would be a high correlation between our majors and our first year graduate students. Such a high correlation would suggest that the semantic knowledge of psychological concepts for our majors resembles a select subgroup – students entering graduate school.

The correlation of Introductory Psychology students with current and recent faculty members was .46, indicating a moderate positive association between the ratings of the two groups. However, the correlation rose substantially when graduating majors from the Psychology Department were compared with the faculty,  $r = .64$ . The ratings for the majors correlated even more strongly with the first year graduate students ( $r = .85$ ). The pattern of correlations is consistent with the hypothesis that graduating seniors' conceptual knowledge of psychology changes to be more like faculty between their Intro course and graduation, so that it closely resembles that of first year graduate students.

The measurement instrument for methods and statistical knowledge more closely resembled a traditional multiple-choice exam. This instrument consisted of 9 questions focused on statistics and 20 questions focused on psychological research methods. Each question had five alternative responses, only one of which was the correct. In the test of methodological knowledge, the percentage of correct responses for psychology majors was 71.5%, a little lower than the first year graduate students (from undergraduate programs throughout the country) who answered correctly on 82.9% of the questions. In contrast, the graduating majors scored 48.8% correct on statistical questions, which was substantially lower than that of incoming first-year graduate students

(84.1%). This pattern of results produced a significant interaction between test type and student type,  $F(1,56) = 5,741, p < .05$ . This finding suggests that psychology majors from throughout the country who choose to attend the NMSU graduate program learn and retain methodological and statistical information at about the same rates, whereas our majors do not acquire or retain statistical knowledge nearly as well as they do methods knowledge.

The above outcomes suggest that the undergraduate majors completing their degrees in the Psychology Department have acquired both conceptual and methodological knowledge. Without any special preparation for an exam, they provided responses relating psychological concepts that closely resembled the responses of first year graduate students in our program. In addition, the graduating majors knowledge of methods was similar to that of first year graduate students. In contrast, their knowledge of statistics was much lower than that of first year graduate students and, at an absolute level was lower than is desirable. The reforms implemented in 1995 have had the desired effect of providing a strong conceptual and methodological knowledge base from our majors. However, we have not been as successful in providing them with a good, usable understanding of statistics. The introduction of the second, applied methods course was in response to comparable findings from earlier outcomes assessments. Thus far, our majors have resisted taking the applied methods course because it has not been a requirement for them. Feedback from outcomes assessments in subsequent years when the applied methods course is required will indicate whether this curricular redesign has had the desired effect of increasing statistical knowledge.

### Discussion

We began this paper with the proposal that feedback, in the guise of empirical data from students and analysis of those data, is a necessary component of the instructional design process. We have described case studies in the collection and analysis of data from students and how the analyses have been applied in the redesign of instruction at the level of an individual instructor and at the level of an entire department's curriculum. The two cases are alike in that they both take a strongly empirical approach and make use of statistical analyses of quantitative data. Importantly, the analyses in both cases were driven by hypotheses about the factors that influence instructional outcomes. The two cases differ in important ways, as well. For example, the analyses of outcomes for the individual instructor were based on characteristics of the instructor and the courses, whereas the department-level analysis involved comparison to multiple baselines. The observation that both a feature-based analysis and a baseline-based analysis can be helpful in providing feedback for redesign of instructional systems indicates that a multiplicity of approaches can be used. Using data from a wide variety of sources for any given design might even be valuable as each approach may reveal different aspects of the students' responses to the instructional design.

### References

- Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*, 219 – 231.
- Adams, J. V. (1997). Student evaluations: The ratings game. *Inquiry, 1*, 10 – 16.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: L. Erlbaum.
- Gordon, S. E., (1994). *Systematic training program design: Maximizing effectiveness and minimizing liability*. Englewood Cliffs, NJ: Prentice Hall.
- Gould, J. D. and Lewis, C. (1983). Designing for usability -- key principles and what designers think. *Human Factors in Computing Systems CHI '83 Proceedings* (pp. 300 - 311). New York: ACM.
- Lewis, R. (1998). Student evaluations: Widespread and controversial. *The Scientist, 12*, 1 – 5.
- Schvaneveldt, R. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex.
- Williams, W. M., & Ceci, S. J. (1997). How'm I doing? Problems with student ratings of instructors and courses. *Change: The Magazine of Higher Learning, 29* (Sept/Oct), 12 – 23.