STUDENT EVALUATION OF TEACHING AND COURSES

TEACHING AND COURSE EVALUATION PROJECT
FINAL REPORT

APPENDIX XIII: INCORPORATING CONTEXT INTO THE INTERPRETATION OF RESULTS

Prepared by:
Prepared by Jessica Tilley, Analyst; Joanne Quan, Analyst; Kiran Bisra, Analyst
Institutional Research and Planning, Simon Fraser University

NOVEMBER 2013
# TABLE OF CONTENTS

1  INTRODUCTION ...................................................................................................................... 1
2  PROPOSED SOLUTION .......................................................................................................... 1
3  METHODOLOGY .................................................................................................................... 1
4  RESULTS ............................................................................................................................... 2
5  FUTURE MODELS ................................................................................................................... 3
1  INTRODUCTION

Teaching and course evaluation results are often presented in aggregated summaries, such as an overall average of student ratings to a set of questions. Aggregated scores alone may be difficult to interpret, and the comparability of scores, over time or between instructors or faculties for example, may be questionable in the absence of context - that is, the consideration of characteristics that describe the course and students. For example, research has shown that class size, class time of day and gender of the instructor can influence teaching evaluations (see Appendix II). Many of the factors that affect evaluation results are outside of the instructor’s control, which leaves the results open to criticism and doubt.

The purpose of this exercise, undertaken by SFU’s Office of Institutional Research and Planning, was to explore a method for taking these contextual factors into account when interpreting evaluation results. The final goal was to suggest a way of determining whether an instructor’s evaluation results are in line with expectations, given the type of course taught, and the type of students in the course.

2  PROPOSED SOLUTION

One solution to the problem described above is to create a model that predicts what score an average instructor would get, given the set of contextual factors that are outside of their control. For example, what score would be expected by an average instructor, if the course was a required lab course, taught in person, with 80 students in it, of which 40% are lower level students and 60% are upper level? If the answer to that question were known, we would be able to determine whether an instructor teaching such a course in a particular term had evaluation results that were above expectation, below expectation, or within expectation (within a certain error of the predicted result).

The best way to create such a model would be to use existing course evaluation data. For example, once a new instrument has been in use for about a year, the data generated from evaluations in that time can be used to create and validate a predictive model. That model could then be used going forward. This approach would ensure that SFU is meeting the overall principle that evaluation results must be considered in the context of the course characteristics (see Appendix II).

3  METHODOLOGY

To illustrate this approach, an exploratory model was developed using the summer 2013 proof of concept data. There are several reasons why this example model cannot be used moving forward, but should instead be considered a prototype. The most important reasons are that constraints of time and data (the proof of concept data only included 18 classes and 943 evaluations, which is not enough information to build a robust model) necessitated some simplifying assumptions, and created a situation in which the results cannot be generalized. Nevertheless, this exercise demonstrates how future models could be helpful in providing a context for results.

The prototype model was constructed by using seventeen classes, and the efficacy of the model was tested on one class (Class X) to demonstrate how the model would operate in the future. A multiple regression was performed between overall student learning experience as the dependent variable and student demographics and classroom characteristics as independent variables. The analysis was executed using the software package Statistical Product and Service Solutions (SPSS).
A set of available factors was identified and considered for inclusion in the model. This set is not exhaustive, and is expected to grow in future iterations of this exercise. Two main categories of data emerged: student demographics and classroom characteristics.

Student demographics include:
- Age
- Gender
- Cumulative Grade Point Average (CGPA)
- Faculty
- Year level
- International status
- Whether the student was required to take Foundations of Academic Literacy (FAL; this variable acts as a proxy for English as an Additional Language status).

Classroom characteristics include:
- Instruction mode (i.e. in person, online)
- Campus
- Class time of day
- Class level (i.e. graduate versus undergraduate)
- Class size
- Class component (i.e. lecture, tutorial).

The outcome measure is Question 5 from the proof of concept institution-wide core questions. This question asked students to rank their agreement with the following phrase: “The instructor created an atmosphere that was conducive to learning.” Responses were ordered on a 5-point scale ranging from “Strongly Agree” (value 5) to “Strongly Disagree” (value 1). For the purposes of the prototype model, this was treated as a linear scale.

Prior to analysis, factors which highly correlate with one another (e.g. age and year level) were identified and one of the two variables was excluded. Also, some variables had to be excluded due to the small size of the data set relative to the number of available factors. It is expected that larger data sets will allow for the inclusion of a greater set of variables. These decisions were made based on prior research and may not reflect the final model in future iterations of this exercise.

4 RESULTS

Four independent variables were found to have a statistically significant relationship with evaluation results. These are:
- FAL requirement
- Instruction mode
- CGPA
- Class size.

1 Note that the original intention had been to use Question 6 from the institution-wide core questions: “Overall, the quality of my learning experience in this course was … Excellent; Very Good; Good; Fair; Poor”. In the end, this question was not used, because its scale was presented in the wrong direction on the evaluation seen by students, and there was concern that the validity of the results may be compromised as a result.

2 There are arguments for and against this approach. However, since this is how the data are used in practice (the scores are averaged as though they are from a linear scale), it was determined that this approach would be best.
The prototype model was applied to predict the score of Class X, which had been excluded from the model. The overall score on Question 5 (averaging over all proof of concept classes) is 3.95, which is higher than the average achieved score of 3.87 from Class X. This may initially give the impression that the instructor’s teaching evaluation score is below expectation for Class X. However, the model, when applied to Class X, indicates that the achieved score falls within the 95% confidence interval of the predicted score. Thus we can conclude that given the specific classroom context, the instructor's evaluation score is within expectation.

5 FUTURE MODELS

The methodology used to build the prototype model was constrained by a short timeline and the small set of responses. Future models may have fewer constraints, which should result in a more robust solution. This section discusses some issues to consider in future models.

In the future, with larger data sets, instead of building a linear regression model, we can explore constructing a Multilevel Linear model. Multilevel or hierarchical models are those in which data are collected at different levels of analysis (e.g. students in a classroom). For example, the fact that individual students respond together and have the same exposure within a classroom means their evaluations are not independent of one another. Multilevel modeling can take these dependencies into account, enabling us to better assess the effects of student demographics and classroom characteristics on the level of overall student learning experience.

It is important to keep in mind that the dependent variable in such a model would have to be one of the institution-wide core questions, since it would need to be a question that appears on every evaluation. Care should be taken to include institution-wide core questions that are appropriate for this purpose.

The issue of missing responses will need to be addressed in future iterations of this exercise. By definition, students who do not respond to the survey cannot contribute to evaluation scores, but their presence in the class may still have an effect on the class dynamic, and the experiences of other students. A systematic statistical approach to this problem will have to be developed as part of final model construction.

Finally, the prototype model took into account a small set of student demographics and classroom characteristics. Future models could expand the set of variables and explore new relationships. For example, such a model could incorporate the impact on course learning experience of a student taking a course outside his/her major. Nevertheless, it is important to remember that there will always be some important factors that cannot be measured, and no model can achieve perfection. Instead this modeling process represents our best chance of incorporating contextual factors into the evaluation scores.