CSSHE
SCÉES

Canadian Journal of Higher Education
Revue canadienne d'enseignement supérieur
Volume 45, No. 4, 2015, pages 298 - 321

Repeated Low Teaching Evaluations: A Form of Habitual Behaviour?

J. Paul Grayson
York University

Abstract

Teaching evaluations have become part of life on Canadian campuses; however, there is no agreement among researchers as to their validity. In this article, comparisons were made between first- and third-year collective evaluations of professors' performance at the University of British Columbia, York University, and McGill University. Overall, it was found that students who provided low evaluations in their first year were also likely to do so in their third year. This effect held independent of degree of campus engagement, sex, student status (domestic or international), and generational status (students who were the first in their families to attend university, compared to those who were not). Given that over the course of their studies, students likely would have been exposed to a range of different behaviours on the part of their professors, it is argued that the propensity of a large number of students to give consistently low evaluations was a form of "habitual behaviour."

Résumé

Les évaluations de l'enseignement font maintenant partie intégrante du système universitaire canadien. Cependant, les chercheurs ne sont pas unanimes quant à leur validité. Dans cet article, on a comparé les évaluations collectives de la performance des professeurs par les étudiants de première et de troisième année à l'Université de la Colombie-Britannique, à l'Université York et à l'Université McGill. Règle générale, on a observé que les étudiants qui avaient fourni des évaluations négatives en première année risquaient fort d'en faire autant en troisième année. Ces résultats ont été obtenus peu importe le degré d'engagement de l'étudiant sur le campus, son sexe, son statut (national ou international) et sa génération (l'étudiant qui est le premier membre de sa famille à fréquenter l'université par rapport aux autres). Étant donné que les étudiants auront été exposés à un éventail de comportements

par leurs professeurs au cours de leurs études, on suppose que la propension d'un grand nombre d'entre eux à sous-évaluer constamment leurs professeurs est devenue une sorte « d'habitude ».

Introduction

Just as Consumer Reports ranks soap and deodorant on an annual basis, *Maclean's* and *The Globe and Mail*, using the results of student surveys, rank Canadian universities along a number of dimensions, including teaching. Just as Consumer Reports' rankings are intended to help buyers make sure that they get the right soap or deodorant, among other things, these university rankings are intended to assist young Canadians in making wise decisions in their selections of universities. Usually, questions used in surveys to gather information for this purpose focus on the overall student experience at a particular university.

Within universities, surveys are also used to assess the teaching of professors in individual courses. Such surveys (or course evaluations) typically ask questions about specific professors. Often, one of the goals of these evaluations is to assist students in future course selections.

While considerable research has focused on the validity of questions used in assessments of individual professors, relatively little has been written on questions designed to measure the effectiveness of professors treated as a collectivity. For example, we do not know whether the personal characteristics of students making collective evaluations of their professors affect evaluations independently of what actually happens in classrooms. For this reason, using research conducted on individual professors as a point of departure, in this article I examine the degree to which first-year students' teaching assessments of professors, treated as a collectivity, are predictive of assessments of a different group of professors in the students' third year.

Background

For some years, in part as a way of dealing with decreased government funding, Canadian universities have been forging more links with the business sector (Turk, 2000, 2008). According to some critics, a concomitant development has been a decline in academic standards and an increasing concern with the acquisition of vocational skills at the expense of skills gained through an active engagement with the liberal arts (Côté & Allahar, 2011). In addition, students have increasingly been viewed as "consumers" (as they view themselves) of a product rather than as participants in a process leading to intellectual growth (Côté & Allahar, 2007). Within this perspective, student evaluations are viewed as consumer satisfaction surveys rather than as instruments contributing to the discovery of better ways of facilitating teaching and learning. As Côté and Allahar argue, the "concern about students 'having their say'," via student evaluations, "is . . . derived from the wider consumer mentality of contemporary society, and encourages the perception that professors should satisfy students' expectations rather than students satisfying professors' expectations" (Côté & Allahar, 2007, p. 85). In fact, given the impression of many faculty members that the awarding of high grades results in high evaluations, student evaluations may actually detract from improvements in teaching and learning.

Whether or not we completely agree with this perspective, it is clear that in Canada and elsewhere, the nature of students' university experiences, as measured through student

evaluations/satisfaction surveys, is of increasing concern to governments and university administrators. In the province of Ontario, for example, universities are co-operating with the provincial government in their administration of the US-based National Survey of Student Engagement (NSSE) (Zhao, 2011). Recognizing that university outcomes are difficult and costly to measure, the survey is based on the proposition that we can infer desired outcomes from the presence of behaviours with which they are associated. For example, in the past, in the United States, associations have been found between measures of student engagement, such as the degree of interaction with professors, and their beliefs that they have increased their knowledge over the course of their studies. As a result, if we find high degrees of engagement, we can assume increases in knowledge (Conway, Zhao, & Montgomery, 2011).

On an annual basis, *Maclean's* magazine provides its readers with a ranking of Canadian universities based on the results of the NSSE ("2011 student surveys," 2011). Focusing on what were called "enriching educational experiences" of senior students, in 2012, *Maclean's* reported that the University of British Columbia (UBC) ranked higher than McGill, McGill higher than Dalhousie, and Dalhousie higher than York. In this ranking system, only six Canadian institutions scored higher than the average for the 577 Canadian and American universities completing the survey. The four noted above were not among them ("How well do Canadian universities follow best practices?" 2013).

The NSSE is based on theoretical underpinnings loosely termed the "college impact model." While space constraints preclude a full discussion of this model, suffice it to say, as indicated above, that student engagement in various formal and informal campus activities contributes to learning (Astin, 1993; Grayson, 1997b; Pascarella, Edison, Nora, Hagedorn, & Terenzini, 1996; Terenzini, Springer, Pascarella, & Nora, 1995; Terenzini, Springer, Yaeger, Pascarella, & Nora, 1996; Terenzini & Wright, 1987). Unfortunately, while it is true that student engagement does increase students' *belief* that their knowledge has expanded, in both the United States and Canada, research indicates little relationship between measures of engagement and *objectively* measured outcomes, such as the development of generic skills and grades (Grayson, 2008a, 2011a; Pascarella & Terenzini, 2005).

In addition *The Globe and Mail*, using the services of Higher Education Strategy Associates (HESA), conducts on an annual basis a satisfaction survey of Canadian students ("Canadian University Report 2013," 2013). Unlike for the NSSE, I could find nothing in the documentation accompanying *The Globe and Mail*'s survey suggesting an underlying theoretical rationale for questions asked. The objective of the survey is a simple ranking of universities on the basis of student satisfaction. For example, in terms of quality of teaching and learning, in 2013, McGill got a B+, York a B–, UBC a B, and Dalhousie an A–.

It is important to note that in their surveys, both the NSSE and HESA ask questions about students' *overall* experiences. For example, in an attempt to evaluate teaching, the NSSE asks students, "During the current school year, about how often have your instructors done the following: a) clearly explained course goals and requirements; b) taught course sessions in an organized way; c) used examples or illustrations to explain different points; d) delivered feedback on a draft or work in progress; e) provided prompt and detailed feedback on tests or completed assignments." Response options included "very much," "quite a bit," "some," and "very little" (NSSE, 2013). It is possible to combine the results of these individual questions into an overall measure of teaching practices.

The HESA survey is similar. Students are asked about their satisfaction with the *overall* "quality of teaching and learning"; instructors' overall "engaging teaching style"; and overall "out of class communications between students and instructors." These two surveys, used to rank universities in two major national media sources, do not collect information on students' experiences in individual courses.

Having said this, it must be stressed that ratings of individual professors are not the gold standard against which collective evaluations of professors should be compared. As explained by Ory and Ryan, in evaluations of individual professors we do not know "if students respond to items by comparing the instructor's performance to that of other instructors or to some idealized standard" (2001, p. 33).

In addition to participation in the NSSE, most Canadian universities continue to carry out evaluations of teaching effectiveness in individual courses at the faculty and/or departmental level. Unfortunately, it is often difficult to find a theoretical rationale for questions designed to measure teaching effectiveness. This omission is unfortunate as, among other things, the results of such evaluations are used for: teaching improvement, personnel decisions, assisting students in course selection, and teaching awards (Beran, Violato, & Kline, 2007; Beran, Violato, Kline, & Frideres, 2005; Gravestock & Gregor-Greenleaf, 2008).

Teaching evaluations of individual courses typically contain two main types of questions. First, students may be asked an overall question on teaching effectiveness in a course, such as: "Overall, how effective is the instructor for this course?" Often, students are asked specific questions relevant to particular pedagogical practices, like, "Did the instructor communicate effectively with students?" and "Was the instructor well organized?" While some researchers advocate asking questions on various aspects of pedagogical practice (the answers may be combined into an overall index) (McKeachie, 1997), others argue that for many purposes, single overall questions are more than sufficient (Abrami, 2001).

Research on Teaching Evaluations

Literally thousands of articles have been written on the validity and reliability of the instruments used to assess teaching effectiveness in individual courses. A large number of these were written in the 1970s and 1980s and still provide reference points for discussions of the validity of teaching assessments; however, the findings of these studies should now be treated cautiously.

The reason for caution stems from a meta-analysis of the "locus of control" of college students in the United States (the home of most of our information on teaching evaluations). Simply stated, "People who believe they are in control of their destinies have an internal locus of control (internals). Those who believe that luck and powerful others determine their fate have an external locus of control (externals)" (Twenge, Zhang, & Im, 2004, p. 308). Research on students in recent years has indicated that, all else being equal, students with an internal locus of control are more likely than those with an external locus of control to evaluate their professors positively (Griffin, 2004; Kirkpatrick, Stant, Downes, & Gaither, 2008; McClure et al., 2011). This being the case, it is extremely important to note that in the United States between 1960 and 2002, there was a drastic increase in the number of post-secondary students with an external locus of control. Twenge et al. found that "[t]he average college student in 2002 had a more external locus of control than 80% of college students in the early 1960s" (p. 308). It is therefore possible that as a group,

contemporary American (and perhaps Canadian) students are more likely than previous generations to give their professors low evaluations. As a result, some of the relationships established in early studies might no longer be valid.

It is useful to organize studies on the validity of teaching evaluations in accordance with three models: the "grading leniency bias model," the "teaching effectiveness model," and the "student characteristics model." In the grading leniency bias model, it is assumed that students who get high grades rate their professors' performance highly on teaching evaluations. The converse is also true: students obtaining low grades give low evaluations. Because of this possibility, teaching evaluations are viewed as potentially biased.

Theoretical underpinnings for the grading leniency bias model are provided by both the "self-esteem" and "expectancy-conformation" perspectives. According to the former, "students who do well will attribute the performance to self, and those who do not do well will attribute their performance to the instructor/course" (Gigliotti & Buchtel, 1990, p. 342). In essence, students who do poorly will punish their instructors; however, there is no reward for professors on teaching evaluations if students do well. In the expectancy-confirmation perspective, "students with low grade expectations who perform poorly, and students with high grade expectations who perform well, will attribute the performance to self." In either of these cases, the impact on teaching evaluations will be neutral. By contrast, "students who expect to do well and do not will attribute their performance externally to the instructor or course" (p. 343). Similarly, students with low expectations who actually do well will credit their instructors. While the first of these situations has negative implications for teaching evaluations, the second has positive implications. Although a consensus is non-existent, a growing body of evidence supports the self-esteem model (Griffin, 2004; McClure et al., 2011; Vaillancourt, 2012).

The "teaching effectiveness model" assumes good teaching results in greater learning. Students who learn more get high grades and, as a result, give high evaluations to their instructors. Factors sometimes taken into consideration when operationalizing effective teaching include instructor enthusiasm, organization, student support, group interaction, breadth of coverage, examinations and grading, assignments and reading, and workload and difficulty (Algozzine et al., 2004). Within this model, teaching evaluations are viewed as valid measures of teaching performance (Arnold, 2009; Greenwald, 1995; Krautmann & Sander, 1999; McClure et al., 2011).

The "student characteristics model" postulates that certain student characteristics, like high motivation, result in greater learning and high evaluations of teacher performance (Abrami, Perry, & Leventhal, 1982; Biner et al., 1997; Cashin, 1995; Howard & Maxwell, 1980; Marsh & Dunkin, 1992; Patrick, 2011; Strom & Hocevar, 1982; Witt & Handal, 1984). Within this perspective, teaching evaluations can also be viewed as invalid: learning is less a consequence of good teaching than of factors such as motivation.

Factors external to the three models that have potential effects on teaching evaluations are: class size; whether or not a course is mandatory; course difficulty and level; discipline (Arts compared to Science, for example); professors' personalities, rank, and gender; and student attendance at class (Cashin, 1995; Gravestock & Gregor-Greenleaf, 2008).

In addition to the models above, the "habitual behaviour perspective"—in the current context, a variant of the students' characteristics model—holds promise for researchers. Habits can be defined as "one's customary ways of behaviour" (Ouelette & Wood, 1998,

p. 54). According to Aarts, Verplanken, and van Knippenbert (1998, p. 1359), "the source of a habitual response, like stereotypes and attitudes, can be thought of as a cognitive structure that is learned, stored in, and readily retrieved from memory, upon the perception of appropriate stimuli." Research based on this assumption has been conducted on a range of phenomena, including travel decisions (Aarts et al., 1998), the use of public transit (Chen & Chao, 2011), changing energy use habits (de Vries, Aacts, & Midden, 2011), and voting (Fowler, 2006); however, the perspective has not been applied to the study of teaching evaluations. Overall it has been found that, particularly in stable contexts, past behaviour predicts future behaviour, independent of the effects of intentions, attitudes, subject norms, and behavioural control (de Vries et al., 2011; Ouelette & Wood, 1998, p. 65; Wood & Quinn, 2005). On the basis of this model, we would expect that, after considering the effects of other variables, students' past evaluations would be good predictors of future evaluations. In general, Ouellette and Wood argue that early sociologists, like Durkheim, Mead, and Weber, either tacitly or explicitly recognized the importance of habitual behaviour in their accounts of social stability.

An argument can be made that this observation should not be restricted to early sociologists. It applies equally to recent sociologists like Bourdieu. For example, although embodied in a radically different theoretical perspective, Bourdieu's concept of "habitus" shares some characteristics with habitual behaviour. For Bourdieu, habitus is a set of dispositions acquired through interactions in various fields, such as education. Such dispositions have the potential to shape how individuals view reality outside of the field in which they are acquired. For example, in the current context, students' experiences in primary and secondary schools (among other fields), mediated by different class and ethno-racial experiences, may result in their entering university with radically different views of the relative roles of students and instructors in the learning process. In turn, such views may play a role in the ways in which instructors' behaviour is evaluated.

Independent of the way teaching evaluations are viewed, a change in the evaluations of a very few students can have a drastic impact on the overall evaluation of an individual professor. For example, Clayson and Haley present evidence indicating that in a particular class of 40, a change in the evaluations of three students from extreme positive to extreme negative would reduce "a ninetieth percentile instructor to the sixty-seventh percentile." In a class of 20 the same change in evaluation would further reduce the professor to the forty-second percentile (2011, p. 108).

Despite there being thousands of articles on the validity and reliability of general and specific questions used in evaluations of individual courses, the validity and reliability of collective evaluations of students' teaching and learning experiences, such as those asked in the NSSE and HESA surveys, have received little attention. Given that the surveys are used to rank universities at the national level, this neglect is unfortunate. At this point, the extent to which the findings of studies carried out at the course level are applicable to overall institutional assessments is unknown.

One Canadian study that did focus on the validity of students' evaluations of all of their instructors combined employed structural equation modeling to follow a cohort of students in the Faculty of Arts and the Faculty of Science over four years of study at York University (Grayson, 2004). For each faculty it was found that, in contrast to the grading leniency bias model, in each year, grade point average (GPA) was of no consequence for

assessments of professor performance, and, contrary to the teaching effectiveness model, professor performance had no impact on GPA. By contrast, consistent with the student characteristics model and habitual behaviour perspective, assessments of professor performance in the previous year were strong predictors of assessments of professor performance in the next year (for Arts and Science faculties combined, over four years, average beta = .42). In addition, the best predictor of GPA in any one year was GPA in the previous year (average beta = .71). Put in terms of the three models described above, these findings are consistent with the student characteristics model and habitual behaviour perspective rather than the grading leniency bias and teaching effectiveness models.

Research Strategy

As the validity of questions focusing on the collective performance of all instructors (rather than on instructors' performance in individual courses) has failed to garner attention, the current article examines the degree to which students' collective evaluations in their first year at university are predictive of their evaluations in third year. Ideally, such an investigation would involve seven steps.

- Step one would focus on operationalizing effective teaching practices. Unfortunately, there is no consensus among scholars as to the nature of these practices (Ory & Ryan, 2001). This said, a number of practices often identified as effective include instructors: having organization and communication skills; establishing teacher–student interaction and class rapport; and implementing fair grading practices (Cashin, 1995; Feldman, 1976, 1978; Marsh, 1995; Marsh & Dunkin, 1992).
- The second step would involve using objective measures (not surveys), in specific first-year courses, to assess the degree to which individual professors implement the formerly identified effective practices.
- Step three would involve measuring, through surveys, in the same specific courses, the extent to which students were able to recognize effective practices when exhibited by instructors.
- Assuming congruence could be established between steps two and three, in step four, the same students would be asked questions, such as those included in the NSSE, about the overall expertise of instructors in all of their courses combined.
- In step five, the results of steps three and four would be compared to see whether aggregates of students' assessments of individual courses approximated their assessments of teaching practices in all of their courses combined.
- Using the same students as in first year, step six would involve a replication of steps three, four, and five in third year.
- Step seven would include an examination of the extent to which evaluations in first year were predictive of evaluations in third year.

When carrying out these steps, for reasons noted earlier, it should not be assumed that the evaluations of individual professors are the benchmark in accordance with which collective evaluations should be judged.

Unfortunately, resources needed to implement this overall (ambitious) research strategy are unavailable. (To my knowledge, it has never been implemented elsewhere.) As a result, there are limitations on what can be concluded from the current study of first- and third-year students at the University of British Columbia, York, McGill, and Dalhousie.

Were we to find that students who gave high collective evaluations of their professors in first year were likely to do the same three years later, we would not know whether this consistency:

1. Reflected consistent teaching practices on the part of professors evaluated collectively (teaching effectiveness model).
2. Reflected the tendency of students to collectively evaluate their professors in more or less the same way, independent of their classroom performance (habitual behaviour perspective).

Without the implementation of the ideal research strategy described above, at best we can only make reasoned arguments as to which of these two possibilities was the most likely.

As previous research in Canada carried out from the perspective of the college impact model (Grayson, 2005, 2007a, 2007b, 2008a, 2008b) and cultural reproduction theory (Grayson, 2011b) has shown differences in both the experiences and the outcomes of domestic and international students, the current research makes distinctions between these two groups. Additional Canadian research has indicated that the experiences and outcomes of students who are the first in their families to attend university (first generation) are different from those of other students (Grayson, 1997a, 2011a; Karmanzi, Doray, Bonin, Groleau, & Murdoch, 2010; Lehmann, 2007, 2009). As a result, further distinctions are made based on students' generational status.

Sample

As described elsewhere (Grayson, 2008b), the data used in this study derives from a survey of international and domestic students entering the University of British Columbia (UBC), York University, McGill University, and Dalhousie University in the fall of 2003. Excluding faculties for which a prior degree was required (e.g., law), all international students 30 years of age or younger and entering first year in each of the four universities were mailed a questionnaire in January 2004. Comparable numbers of randomly selected domestic students were also included in the study. In addition, 16 separate focus group meetings (four in each university) were held with domestic and international students in the fall of 2003 and spring of 2004. The intent of these meetings was to obtain an in-depth appreciation of students' experiences (Pidgeon & Andres, 2005, 2006).

The total number of individuals invited to participate in the first mail survey by the Institute for Social Research at York University was 4,872. After four contacts, 1,425 students completed the questionnaire, for an overall response rate of approximately 30%. Follow-up focus group meetings and surveys were planned for 2005 and 2006. The total numbers of domestic and international students from each of the universities who completed the first questionnaire are summarized in Table 1.

By the end of the third annual survey (2006), responses had been obtained from 505 (35%) of the original participants (a typical attrition rate in studies of this nature). Within the continuing sample, 23% were domestic first-generation students, 49% other than first-generation domestic, 4% international first-generation, and 22% other than first-generation international. Twenty-seven percent of the continuing sample was male. The average high school grade, or its equivalent, was 83.69%. Because of necessary standardization, given different grading schemes, the mean first-year GPA for all samples approached zero; however, the York data showed a skew to the left.

Table 1.
Domestic and International Participation

University	Domestic	International
UBC	265	248
York	365	143
Dalhousie	167	63
McGill	119	55
Total	915	509

While the response rates achieved in this study are not unusual in studies of university students (Dey, 1997), they are lower than usual for those conducted by the author. Moreover, a large part of the differences in response rates between the universities can be attributed to the fact that each university had its own ethics committee that required a different letter of introduction to the survey (Grayson & Myles, 2005).

As access was available to administrative records, it was possible to compare the number of female and domestic students in the population to those in the survey. The comparison indicated that while females made up 59% of the population, they comprised 68% of the sample. Differences between the two groups were statistically significant. Although domestic students made up 61% of the population, they represented 65% of the sample. Again, these differences were statistically significant.

While no prior research could be located on the comparative response rates of international students, in the United States it has previously been found that Black students were less responsive to surveys than White students (Dey, 1997; NCES, 2002). By comparison, and consistent with the findings of the current study, a considerable body of research indicates that female students respond to surveys far more readily than their male peers (Crawford, Couper, & Lamias, 2001; Dey, 1997; Hutchinson, Tollefson, & Wigington, 1987; NCES, 2002; NSSE, 2003; Sax, Gilmartin, & Bryant, 2003).

It was also possible to make comparisons between the compositions of the full first-year and third-year samples. In the former, 31% were male, 63% were domestic students, and the average high school grade was 83.20%. In the continuing sample, 27% were male, 72% were domestic students, and the average high school grade was 83.69%. While differences based on sex and high school grades were not statistically significant, there were more domestic students in the continuing than in the full first-year sample.

Although data could have been weighted to correct for imbalances between the first full survey and the population, and for differences between the first full survey and the continuing survey, as the intent was not to make population estimations but to assess the extent to which student evaluations in first year are predictive of similar evaluations in third year, unweighted data were used. Moreover, the potential errors associated with weighting likely would have outweighed its benefits.

Measures

Information on high school grades (HS grades), international or domestic status, and GPA in university were obtained from administrative records. All other information (including sex and whether students were the first in their families to attend university) was supplied by the surveys.

Teaching effectiveness has been operationalized in many different ways (Feldman, 1976; Marsh & Dunkin, 1992; Mason, Steagall, & Fabritius, 1995; Rice, Stewart, & Hujber, 2000). In the current study, questions focusing on exemplary performance by professors were derived from a study of students at the University of Guelph who kept diaries of their first-year experiences and who participated in interviews with researchers. The aspects of classroom performance by professors (professor performance) that were identified as exemplary in this way were: having adequate teaching expertise; having knowledge of subject matter; being responsive to the class; caring about students in the class; having a sense of humour; and being well organized (Benjamin, 1990). These dimensions were similar to those used by other researchers in examinations of teaching that were discussed earlier. It should be noted, however, that researchers have given little attention to the possibility that students' assessments of what constitutes having adequate teaching expertise, etc., may vary from year to year.

In the current study, students were asked how many of the instructors in the courses in which they were currently enrolled had each of the foregoing characteristics. Fixed response options were 0%, 1% to 25%, 26% to 50%, 51% to 75%, and 76% and more (Cronbach's alpha = .83). For the first-year continuing sample, the average score was 3.62. For the third-year sample, the value was almost identical, 3.60.

In addition to evaluations of their professors' performance in their classes, consistent with the college impact model, students were asked about various aspects of academic and event involvement on campus. Academic involvement consisted of the sum of the standardized values for the percentage of weekly classes and tutorials (or the equivalent) attended, the number of hours studied per week outside of class, the number of visits to the library per month, and the extent to which students felt that they had worked up to their talent or potential over the academic year. Values were standardized because different metrics were used for each variable. As this was an additive variable, Cronbach's alpha, a measure of consistency across variables, was not appropriate. In the first-year continuing sample, as might be expected (given that the variables were standardized), the mean was .07. For the third-year continuing sample, the equivalent figure was .01.

A measure of event involvement was obtained by summing the standardized responses given to questions focusing on: participation in non-required academic or career activities; membership in campus organizations; participation in organized sports; involvement with unorganized sports; spectator involvement with sports; involvement with cultural or arts events; and attendance at cultural events such as films or concerts. Again, as this was a summary measure, Cronbach's alpha was inappropriate. As expected, given that the variables were standardized, the first and third continuing sample means were –.02 and .01.

Analysis

Data analysis involved three steps. In order to gain an overall picture of the relationship between evaluations of professors' performance in first and third years, I conducted a linear regression, with third-year evaluations of instructors as the dependent variable. While linear regression provides an overall picture of the relationship between dependent and independent variables, it is not appropriate for determining whether low evaluators in year one were more or less likely than high evaluators to be consistent in their evaluations in year three. For this reason, in step two, for both the first- and third-year surveys, I conducted a two-step cluster analysis available in SPSS to identify discreet natural

groupings of students (evaluation groupings) based on evaluations of professors' course performance (high and low). For the third year, I then examined these clusters in relation to students' grades, engagement, and demographic characteristics. In step three, I carried out a logistic regression analysis in which I determined the extent to which discreet high and low evaluations of professor's performance in year one predicted high and low evaluations in the third year. At each of these steps, because of their low response rate in the third-year survey, Dalhousie students were excluded from analysis.

Table 2 summarizes the results of a regression analysis in which an assessment of professors' performance in third year was the dependent variable. Independent variables included the dummy variables female, student status (domestic vs. international), student generation (first in family to attend university vs. others), and university (York and McGill, with UBC as the referent). Continuous variables included high school grades, academic and event involvement in first and third years, grade over first year (grades standardized in third-year sample minus grades standardized in first-year sample), and professor performance as measured in first year.

The results summarized in Table 2 show that being female (beta = .104), grade gain (beta = .117), and professor performance in first year (beta = .360) were statistically significant. These results indicate that females provided slightly higher evaluations of their professors than males, while students whose grades increased between first and third years were somewhat inclined to give their professors relatively high evaluations. The greatest predictor of third year professor performance, however, was students' first year assessment of performance.

Table 2.
Regression for Professor Performance in Third Year

	Beta	Sig	
Female	0.104	0.015	
Student status (domestic)	-0.049	0.267	
Student generation (first)	-0.041	0.339	
High-school grades	0.032	0.487	
Academic involvement first year	0.032	0.464	
Event involvement first year	0.003	0.953	
Academic involvement third year	0.051	0.255	
Event involvement third year	0.066	0.203	
Grade gain over first year	0.117	0.007	
York (UBC comparator)	-0.054	0.275	
McGill (UBC comparator)	0.071	0.124	
Professor performance first year	0.357	0.000	
Adjusted R2			0.179
Model significance		0.000	
N			478

While the regression analysis provides an important overall picture of the relationship between evaluations in first and third year, it does not allow us to effectively determine whether students giving low evaluations in first year were more likely than high evaluators to repeat their behaviour in third year. As a first step in a determination of this possibility, I conducted two-way cluster analyses (available in SPSS) to produce two natural groupings of students, defined in terms of their giving high or low evaluations of their professors. Information on evaluations of professors' performance from the first survey is summarized in Table 3, where data are presented for all universities combined as well as for individual universities. In each case, the "silhouette measure of cohesion and separation" indicated fair cluster quality.

When all universities were combined, as indicated by Table 3, 73% and 27% of students were grouped in the low and high evaluation categories, respectively. There was, however, some difference from one university to the next. At UBC, 75% of students were grouped in the low category and 25% in the high. At York, 52% of students were grouped in the high category and 48% in the low. The McGill figures were similar to those at York: 49% and 51% in the low and high categories, respectively.

The mean scores for each of the characteristics on which the clusters were based (inputs) are found in the second column of the table. For example, for all universities combined, the mean score of those in the low grouping for "know subject" was 3.95; for the high group, the figure was 4.85. Not surprisingly, analyses of variance showed that for each of the characteristics included in the clusters (know subject, responsive to class, sense of humour, know techniques, care about students, and well organized), differences in scores between those placed in the low and high categories were statistically significant.

Given the logic behind two-step cluster analysis, it would not be appropriate to compare the numbers of students in the low- and high-evaluation categories from one university to the next: scores on cluster inputs leading to a low placement in one university might have resulted in a high placement in another. Comparisons can be made between universities on the basis of the continuous measure of professor performance used in the regression analysis: students were asked how many of the instructors in the courses in which they were currently enrolled knew their subject, were responsive to the class, had a sense of humour, knew effective teaching techniques, cared about students, and were well organized. The distribution of scores on this measure is shown in Table 4.

In the first year of study, for all universities combined, the average score for professor performance was 3.62. Fluctuations from one university to the next were not statistically significant. For all universities combined and each individual university, however, analyses of variance showed that differences between students in the low and high clusters were statistically significant. By third year, the pattern changed somewhat. This time, differences between universities were statistically significant. Despite statistical significance, absolute differences were slight. For example, the lowest (York) and highest (McGill) scores were 3.53 and 3.78, respectively. As for first year, for all universities combined and each university separately, differences in the scores for the low and high evaluation groups were statistically significant. Overall, these data suggest more similarity than difference between the ways in which students at UBC, York, and McGill evaluated their professors' performance.

Table 3
First Year Assessments of Professor Performance

	All Combined			UBC			York			McGill		
	Low	High	*Sig. F*	Low	High	*Sig. F*	Low	High	*Sig. F*	Low	High	*Sig. F*
% In Category	73%	27%		75%	25%		52%	48%		49%	51%	
Inputs												
Know Subject	3.95	4.85	0.000	3.97	4.81	0.000	3.77	4.49	0.000	3.96	4.67	0.000
Responsive to Class	3.54	4.53	0.000	3.54	4.63	0.000	3.27	4.22	0.000	4.33	4.43	0.000
Sense of Humour	3.12	4.31	0.000	2.90	4.40	0.000	3.02	4.02	0.000	2.94	4.00	0.000
Know Techniques	3.14	4.02	0.000	3.08	4.05	0.000	2.82	3.99	0.000	2.76	3.90	0.000
Care about Students	3.15	4.21	0.000	3.20	4.30	0.000	2.80	3.97	0.000	2.76	4.00	0.000
Well Organized	3.62	4.31	0.000	3.47	4.35	0.000	3.52	4.22	0.000	3.27	4.25	0.000
Number of Cases	354	132		131	43		111	103		51	52	

Table 4
Professor Performance Scores

	First Year											
	All Combined			UBC			York			McGill		
	Low	High	Sig. F	Low	High	*Sig. F*	Low	High	Sig. F	Low	High	*Sig. F*
	3.42	4.37	0.000	3.36	4.43	0.000	3.20	4.15	0.000	3.18	4.19	0.000
Overall	3.62			3.56			3.63			3.69		0.275
Number of Cases	354	132		131	43		111	103		51	52	
	Third Year											
	All Combined			UBC			York			McGill		
	Low	High	Sig. F	Low	High	*Sig. F*	Low	High	Sig. F	Low	High	*Sig. F*
	3.52	4.17	0.000	3.42	4.31	0.000	3.11	4.31	0.000	3.27	4.10	0.000
Overall	3.60			3.59			3.53			3.78		0.010
Number of Cases	328	141		123	51		118	96		63	39	

Table 5 summarizes information for clusters of students in their third year of study. The silhouette measure of cohesion and separation indicated fair cluster quality for universities combined and separately. As in Table 3, more students were grouped in the low evaluation clusters than in the high. For all universities combined, 70% were placed in the low category. For UBC, York, and McGill, the figures were 71%, 55%, and 62%, respectively. The items on which the clustering was based are found under the "inputs" category identified in the second column of the table. Differences between the high and low evaluation groups for each item on which the clustering was based were statistically significant, with the exception, for McGill, of sense of humour, know techniques, care about students, and well organized.

In addition to the items used to place students in different evaluation categories, the table provides information on students' characteristics in each of the evaluation groups: grades, engagement variables, demographic variables, and the overlap between the evaluation groups in which students were placed in first and third years.

Analyses of variance showed that for neither all universities combined nor each university separately did students' high-school, first-year, and third-year grades vary in a statistically significant way between the low and high evaluation groups. This said, consistent with the regression analysis presented earlier, for the low evaluation group in all universities combined, third-year grades were lower than first-year grades (0.01 and 0.11, respectively) and these differences were statistically significant. By contrast, grades for high evaluators in third year were similar to what they had been in first year (0.13 and 0.15, respectively). The results of *t*-tests showed that differences between third- and first-year grades were statistically significant for the low evaluators but not the high evaluators.

At UBC, the grades for both the low- and the high-evaluation groups were significantly lower than they had been in first year. Although the results of *t*-tests showed that the remaining differences between first- and third-year grades at York and McGill were not statistically significant, in the high group (as for all universities combined and UBC) the grades for students either were higher than for the low group or did not undergo as great a drop between first and third year as in the low group.

F and chi-square values for the continuous and categorical variables, respectively, for the remainder of the information in Table 5 indicate that there were no consistent and statistically significant differences between members of the two evaluation groups in terms of academic engagement or demographic variables: the groups were not distinguished by their having students with different levels of academic or event engagement. Also, neither number of females nor domestic or first-generation student status varied by evaluation group.

By comparison, for all universities combined and for each university separately, students who gave low evaluations in third year included large numbers who had given low evaluations in first year. For all universities combined, 75% of students in the low-evaluation group had also given low evaluations in their first year. Of those in the high-evaluation group, only 45% had given similar evaluations in first year.

At UBC, among low evaluators, 76% had given low evaluations in year one. Among high evaluators, only 54% had also provided a positive assessment of their professors in first year. York showed a similar but less dramatic pattern: while 79% of low evaluators had also been in the same category in first year, among high evaluators, only 36% had given high evaluations in first year. At McGill, 64% of students among low evaluators had also given low evaluations in year one. By comparison, in the high evaluation group, only 47% had given similar evaluations in first year.

Table 5
Third Year Assessments of Professor Performance

		All Combined			UBC			York			McGill		
		Low	High	Sig. *F*	Low	High	Sig. *F*	Low	High	Sig. *F*	Low	High	Sig. *F*
% In Category		70%	30%		71%	29%		55%	45%		62%	38%	
Inputs													
	Know Subject	3.81	5.00	0.000	3.85	4.98	0.000	3.69	4.41	0.000	4.15	4.41	0.000
	Responsive to Class	3.69	4.28	0.000	3.64	4.49	0.000	3.25	4.20	0.000	3.33	4.27	0.005
	Sense of Humour	3.28	3.81	0.000	2.99	4.04	0.000	2.84	3.98	0.003	2.90	3.94	0.206
	Know Techniques	3.30	3.77	0.000	3.19	3.78	0.000	2.81	3.97	0.000	2.85	3.83	0.138
	Care about Students	3.46	4.00	0.000	3.35	4.25	0.000	2.86	4.00	0.002	3.10	4.21	0.065
	Well Organized	3.58	4.17	0.000	3.51	4.31	0.000	3.25	4.14	0.000	3.26	3.97	0.419
Evaluation Fields													
Grades	HS Grades	0.04	0.18	0.154	0.63	0.68	0.680	-0.50	-0.27	0.518	-0.09	0.13	0.765
	First Year Grades	0.11	0.15	0.715	0.17	0.28	0.395	-0.11	0.33	0.475	0.09	0.00	0.390
	Third Year Grades	0.01	0.13	0.239	-0.03	0.13	0.268	-0.15	0.31	0.668	-0.05	0.14	0.091
Sig. T for brackets		0.002	0.726		0.000	0.024		0.567	0.947		0.085	0.283	
Involvementt	Academic Involvement	-0.01	-0.01	0.951	-0.03	0.01	0.079	-0.15	0.07	0.898	0.14	-0.01	0.310
	Event Involvement	-0.01	0.01	0.786	0.09	0.07	0.519	-0.16	-0.08	0.856	0.12	0.06	0.558
Demographiccs	% Females	72%	73%	0.794	72%	71%	0.784	60%	85%	0.183	62%	79%	0.399
	% Domestic Students	70%	75%	0.266	59%	67%	0.219	77%	88%	0.107	77%	64%	0.796
	% First Generation	28%	22%	0.182	27%	17%	0.476	31%	28%	0.471	29%	15%	0.642
Evaluation	% Same Cluster Year 1	75%	45%	0.000	76%	54%	0.000	79%	36%	0.000	64%	47%	0.166
Number of Cases		328	141		123	51		118	96		63	39	

With the exception of McGill, for all comparisons, based on chi-square, differences were statistically significant. In other words, students giving low evaluations in third year were also likely to have done so in first year. By contrast, high evaluators in third year were less likely to have provided positive evaluations in first year.

In a further attempt to understand the relationship between evaluation group and grades, I divided the sample into four groups: those who were in the low group in both years (54%); students who were classified in the high group in both years (13%); low evaluators in year one who became high evaluators in year three (18%); and students who were high evaluators in year one and low evaluators in year three (16%). Differences among universities on this dimension were not statistically significant (not shown in the tables). It is worth stressing that a majority (54%) of students were steadfast in their low evaluations of their professors. Overall, 67% of students were in the same evaluation group in both their first and third years.

Between first and third years, the grade gains/losses for each of the foregoing groups were −0.109, −0.090, 0.025, and −0.041, respectively (not shown in the tables). Although analyses of variance indicated that these differences were not statistically significant, it is important to note that, consistent with the previous regression analysis, students who went from low to high evaluations reported the highest grade gains (0.025). Although not statistically significant, a similar pattern was found for York and McGill. The UBC figures were more erratic. Overall, these findings are consistent with the possibility that among those who changed from low to high evaluators, some may have been rewarding their professor for giving them high grades.

As a second step in an attempt to distinguish whether first-year low evaluators were as likely as high evaluators to make similar assessments in third year, I conducted a logistic regression analysis in which the evaluation group in third year was the dependent variable. High-school grades, grade gains/losses between first and third years, university (with UBC as the reference category), and evaluation group in first year (low, high) were independent variables.

The results of the logistic regression analysis are presented in Table 6. Overall, the results show that the student's evaluation group in first year made a statistically significant contribution to the equation (odds ratio = 2.53). High-school grades, grade gains/losses over first year, and university were not statistically significant. Although not statistically significant, it is important to note that, consistent with the linear regression analysis, students whose grades increased between first and third years were more likely than others to be in the high evaluation group (odds ratio = 1.24).

Table 6 .
Logistic Regression for Professor Performance in Third Year

	β	Wald Chi-sq	Sig	Odds ratio
High-school grades	0.085	0.414	0.520	1.089
Grade gain over first year	0.215	1.148	0.284	1.240
York (UBC comparator)	-0.294	1.041	0.308	0.745
McGill (UBC comparator)	0.263	0.795	0.373	1.301
Evaluation group first year (0 = low; 1 = high)	0.928	16.74	0.000	2.530
N				478

In addition to the information presented in Table 6, it is important to note that the analysis showed that a test of the full model against the constant-only model was statistically significant. This indicated that collectively, high-school grades, grade gains/losses, university (UBC, York, McGill), and evaluation group distinguished between low and high evaluators in third year (chi-square = 24.107, $p < .000$, $df = 5$). Overall prediction success was 68.8%; however, while the model correctly predicted 93.6% of low evaluators in third year, it only correctly predicted 14.2% of high evaluators (Hosmer and Limeshow Test chi-square = 3.563, $df = 8$, $p > .894$). Consistent with the low numbers of correctly predicted high evaluators, Nagelkerke's *R* square of .077 showed a weak *overall* relationship between predicted and actual evaluation group in third year.

The overall implication of the foregoing analyses is that students giving low evaluations of their professors' behaviour in first year, independent of high-school grades, grade gains/losses between first and third years, and university of enrolment, were also likely to give low evaluations in third year. This was not true for students giving high evaluations in first year. Many of them gave low evaluations in third year.

Discussion

It is clear from the foregoing analysis that some of the individual findings of the current research are consistent with some of the models identified in the introduction. For example, the finding that grade gains between first and third years predict relatively high collective evaluations of professors in third year is consistent with the grading leniency bias model. That females are more likely than males to give relatively high evaluations to their professors in third year is consistent with the student characteristics model. As noted previously, however, it is not possible from the data available for this study to determine whether the tendency for students who provided relatively high evaluations in first year to do the same in third is consistent with the teaching effectiveness model or the habitual behaviour perspective. A clear determination of this issue would require the implementation of the ideal research strategy identified earlier in this article. This said, prior to the implementation of such a strategy, I hypothesize that the habitual behaviour perspective has considerable merit.

The teaching effectiveness model would be sustained if evidence indicated that four conditions were met:

1. Students who gave relative high evaluations to their professors in first and third years (13%) were exposed to similar positive learning environments in both years.
2. Students whose evaluations of their professors were relatively low in first and third years (54%) were exposed to similar negative learning environments in both years.
3. Students who gave relatively high evaluations in first year but not in third (16%) were exposed to positive learning environments in the first but not in the third year.
4. Students who gave relatively low evaluations in first year and positive evaluations in third year (18%) were exposed to negative learning environments in the first year but not in the third.

While future research may show otherwise, it is possible that the conditions in these four points would not be met. First, while students may be consistently enrolled in particular departments or faculties, in most universities, particularly in first year, they are

required to take varying numbers of electives outside of their departments or faculties. As a result, if we accept the position that departments and faculties probably have different teaching environments (Neumann, 2001), students are likely exposed to potentially different teaching cultures. Under circumstances such as these, they are unlikely to have consistently positive or negative course experiences.

Second, even within particular departments or faculties, research has revealed often radically different approaches to teaching (Neumann, 2001). For example, not all professors of sociology share similar ideas about how they should carry out their responsibilities. Leaving aside the question of their validity, differences in individual course teaching evaluations within departments and faculties support this claim. In essence, even within departments, it is unlikely that students would be exposed to consistent learning environments between first and third years.

Third, a possible consistency in teaching environments does not explain why low evaluators in first year were more consistent in their evaluations than high evaluators. All else being equal, in a consistent teaching environment, between first and third years there should be as much defection among low as among high evaluators.

In view of these considerations, it can be hypothesized that were the full research strategy described earlier to be implemented, the habitual behaviour perspective would be confirmed. That said, at this point, definite statements cannot be made.

Conclusion

In recent years, student satisfaction/evaluation surveys have become part of life on Canadian campuses and are used in many different ways. Some observers have linked their ubiquity to an increased connection of the university to the private sector, and the tendency to see students as consumers rather than as co-participants in a learning process. Whatever the cause, because of the various uses to which they are put, it is important that the validity of teaching evaluations be well understood.

Questions used in surveys to evaluate professors' performance are of two basic types. In the first, students are asked to comment on the collective effectiveness of their professors. Questions of this nature are asked in surveys like those conducted by NSSE and HESA and are used to rank Canadian universities. The second type of question focuses on professors' performance in individual courses. The answers to such questions are often used in tenure and promotion decisions, in adjudicating teaching awards, in identifying areas of possible improvement within individual courses, and by students in course selections. At this point, it is not possible to say whether the answers to questions focusing on the collective effectiveness of professors are comparable to aggregates of evaluations of professors in individual courses. By making this statement, I do not mean to imply that aggregates of evaluations of individual professors should be viewed as the benchmark against which collective evaluations need be measured.

While there has been a lot of research focusing on the validity of assessments of professors in individual courses, by and large, the validity of collective assessments has been ignored. Using research based on individual professors as a point of departure, in this article, I focused on the extent to which collective evaluations could be viewed as a form of habitual behaviour. Consistent with the teaching effectiveness model and the habitual behaviour perspective, the results of the research indicated that, independent of the effects

of high school grades, grade gains or losses between first and third years, and university of enrolment, students who in first year gave low evaluations were likely to do the same in third year. This group comprised a hefty 54% of those in the study. Students giving high evaluations in first year were less steadfast in their later evaluations. Overall, in third year, 67% of students rated their professors' performance the same as they had rated it in first year. While it is possible that students giving consistent relatively high and low collective evaluations of their professors experienced positive and negative learning environments, respectively, an argument was made that consistency was more likely explained as a form of habitual behaviour.

As a consequence of the findings of this study and the more general literature on the validity of teaching evaluations, we must be very careful in how we interpret what students say about their professors' collective performance in surveys like NSSE. For example, when viewing situations in which students give relatively low collective evaluations to their professors, we do not know the extent to which such evaluations actually reflect what goes on in classrooms, a decline in students' grades, or the presence of disproportionate numbers of students who habitually provide low evaluations. Until demonstrated otherwise, we cannot assume a random distribution of habitual low evaluators across courses and disciplines, or that their proportions would be equal in all universities. In this case, the null hypothesis—that distributions of low evaluators are random—is too dangerous an assumption to make.

References

2011 student surveys: NSSE benchmarks. (2011, March 14). *Maclean's*. Retrieved from http://oncampus.macleans.ca/education/2011/03/14/2011-student-surveys-nsse-benchmarks/

Aarts, H., Verplanken, B., & van Knippenberg, A. (1998). Predicting behaviour from actions in the past: Repeated decision making or a matter of habit? *Journal of Applied Social Psychology, 28*, 1355–1374.

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *New directions for institutional research*. San Francisco, CA: Jossey-Bass.

Abrami, P. C., Perry, R. P., & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology, 74*, 111–125.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., & Howley, L. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134–156.

Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research, 48*, 215–254.

Astin, A. (1993). *What matters in college?* San Francisco, CA: Jossey-Bass.

Benjamin, M. (1990). *Freshman daily experiences: Implications for policy, research and theory*. Guelph, ON: University of Guelph Student-Environment Group.

Beran, T., Violato, C., & Kline, D. (2007). What's the "use" of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education, 37*(1), 27–43.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *Canadian Journal of Higher Education, 35*(2), 49–70.

Biner, P. M., Summers, M., Dean, R. S., Bink, M. L., Anderson, J. L., & Gelder, B. C. (1997). Personality characteristics predicting continuing education student satisfaction with interactive telecourses. *Journal of Continuing Higher Education, 45*(3), 22–32.

Canadian university report 2013: Student satisfaction survey results. (2013, October 23). *The Globe and Mail*. Retrieved from http://www.theglobeandmail.com/news/national/education/canadian-university-report/canadian-university-report-2013-student-satisfaction-survey-results/article4631980/

Cashin, W. E. (1995). *Student ratings of teaching: A summary of the research*. IDEA Paper No. 20: ERIC No. ED302567.

Chen, C.-F., & Chao, W.-H. (2011). Habitual or reasoned? Using the theory of planned behavior, technology acceptance model, and habit to examine switching intentions to public transit. *Transportation Research, Part F 14*, 128–137.

Clayson, D. E., & Haley, D. A. (2011). Are students telling us the truth? A critical look at the student evaluation of teaching. *Marketing Education Review, 21*(2), 101–112.

Conway, C., Zhao, H., & Montgomery, S. (2011). *The NSSE National Data Project report*. Toronto, ON: Higher Education Quality Council.

Côté, J. E., & Allahar, A. L. (2007). *Ivory tower blues: A university system in crisis*. Toronto, ON: University of Toronto Press.

Côté, J. E., & Allahar, A. L. (2011). *Lowering higher education: The rise of corporate universities and the fall of liberal education*. Toronto, ON: University of Toronto Press.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review, 19*(2), 146–162.

de Vries, P., Aacts, H., & Midden, C. J. H. (2011). Changing simple energy-related consumer behaviors: How the enactment of intentions is thwarted by acting and non-acting habits. *Environment and Behavior, 43*(5), 612–633.

Dey, E. L. (1997). Working with low survey response rates: The efficacy of weighting adjustments. *Research in Higher Education, 38*(2), 215–227.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*(3), 243–248.

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.

Fowler, J. H. (2006). Habitual voting and behavioral turnout. *The Journal of Politics, 68*(2), 335–344.

Gigliotti, R. J., & Buchtel, R. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82*(2), 341–351.

Gravestock, P., & Gregor-Greenleaf, E. (2008). Student course evaluations: Research, models and trends. Toronto, ON: Higher Education Quality Council of Ontario.

Grayson, J. P. (1997a). Academic achievement of first-generation students in a Canadian university. *Research in Higher Education, 38*(6), 659–676.

Grayson, J. P. (1997b). Place of residence, student involvement, and first year marks. *Canadian Journal of Higher Education, 27*(1), 1–23.

Grayson, J. P. (2004). The relationship between grades and academic program satisfaction over four years of study. *Canadian Journal of Higher Education, 34*(2), 1–34.

Grayson, J. P. (2005). The application of American models to the experiences and outcomes of Canadian and international students studying in Canada. *Frontiers: The Interdisciplinary Journal of Studies Abroad, 11*, 71–98.

Grayson, J. P. (2007a). Sense of coherence, problem freedom and academic outcomes of Canadian domestic and international students. *Quality in Higher Education, 13*(3), 215–236.

Grayson, J. P. (2007b). Unequal treatment and program satisfaction among students of European and Chinese Origin. *Canadian Journal of Higher Education, 37*(3), 51–85.

Grayson, J. P. (2008a). The experiences and outcomes of domestic and international students at four Canadian universities. *Higher Education Research and Development, 27*(3), 215–230.

Grayson, J. P. (2008b). Sense of coherence and academic achievement of domestic and international students: A comparative analysis. *Higher Education, 56*(4), 473–492.

Grayson, J. P. (2011a). Cultural capital and academic achievement of first generation domestic and international students in Canadian universities. *British Educational Research Journal, 37*(4), 605–630.

Grayson, J. P. (2011b). Cultural capital and achievement of Chinese international and Canadian domestic students in a Canadian business programme. *The International Journal of Management Education, 9*(2), 13–24.

Grayson, J. P., & Myles, R. (2005). How research ethics boards are undermining survey research on Canadian university students. *Journal of Academic Ethics, 3*(4), 38–60.

Greenwald, A. G. (1995). *Applying social psychology to reveal a major (but correctable) flaw in student evaluations of teaching*. Paper presented at the 103rd meeting of the American Psychological Association, New York.

Griffin, B. G. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29*, 410–425.

How well do Canadian universities follow best practices? (2013, February 7). *Maclean's*. Retrieved from Macleans.ca/education/2013/02/07/how-well-do.canadian-universities-follow-best-practices/

Howard, G. S., & Maxwell, S. E. (1980). Correlations between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72*(6), 810–820.

Hutchinson, J., Tollefson, N., & Wigington, H. (1987). Response bias in college freshman's response to mail surveys. *Research in Higher Education, 26*(1), 99–106.

Karmanzi, P. C., Doray, P., Bonin, S., Groleau, A., & Murdoch, J. (2010). Les étudiants de première génération dans les universités: L'access et la persévérance aux études du Canada. *Canadian Journal of Higher Education, 40*(3), 1–24.

Kirkpatrick, M. A., Stant, K., Downes, S., & Gaither, L. (2008). Perceived locus of control and academic performance: Broadening the construct's applicability. *Journal of College Student Development, 49*(5), 486–496.

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*, 59–63.

Lehmann, W. (2007). "I just didn't feel like I fit in": The role of habitus in university dropout decisions. *Canadian Journal of Higher Education, 37*(2), 89–110.

Lehmann, W. (2009). University as a vocational education: Working class students' expectations for university. *British Journal of Sociology of Education, 30*(2), 137–149.

Marsh, H. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 87*(4), 666–679.

Marsh, H., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. *Higher Education: Handbook of Theory and Research, 8*, 143–233.

Mason, P. M., Steagall, J. W., & Fabritius, M. M. (1995). Student evaluations of faculty: A new procedure for using aggregate measures of performance. *Economics of Education Review, 14*(4), 403–416.

McClure, P. M., Meyer, L. M., Garisch, J., Fischer, R., Weir, K. F., & Walkey, F. H. (2011). Students' attributions for their best and worst marks: Do they relate to achievement? *Contemporary Educational Psychology, 36*, 71–81.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218–1225.

NCES. (2002). *National Postsecondary Student Aid Study 1999–2000 (NPSAS:2000), CATI Nonresponse Bias Analysis Report (No. 2002-03)*. Washington, DC: National Center for Educaiton Statistics.

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education, 26*(2), 135–146.

NSSE. (2003). *The college student report: 2003 overview*. Bloomington, IN: University of Indiana.

NSSE. (2013). *National Survey of Student Engagement*. Retrieved from http://nsse.iub.edu/pdf/survey_instruments/2013/2013 NSSE Instrument.pdf

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research, 109*, 27–44.

Ouelette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124*(1), 54–74.

Pascarella, E., Edison, M., Nora, A., Hagedorn, L., & Terenzini, P. (1996). Influences on students' openness to diversity and challenge in the first year of college. *Journal of Higher Education, 67*(2), 174–195.

Pascarella, E., & Terenzini, P. (2005). *How college affects students* (Vol. 2). San Francisco, CA: Jossey-Bass.

Patrick, C. L. (2011). Student evaluations of teaching: Effects of the big five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education, 46*(2), 239–249.

Pidgeon, M., & Andres, L. (2005). *The first-year experiences of international and domestic students at Canadian universities*. Vancouver, BC: Department of Educational Studies, University of British Columbia.

Pidgeon, M., & Andres, L. (2006). *Does "it" get any easier? A comparative study of international and domestic students' second-year experiences at four Canadian universities*. Vancouver, BC: Department of Educational Studies, University of British Columbia.

Rice, R. E., Stewart, L. P., & Hujber, M. (2000). Extending the domain of instructional effectiveness assessment in student evaluations of communication courses. *Communication Education, 49*(3), 253–266.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education, 44*(4), 409–432.

Strom, B., & Hocevar, D. (1982). Course structure and student satisfaction: An attribute-treatment interaction analysis. *Educational Research Quarterly, 7*(1), 21–30.

Terenzini, P., Springer, L., Pascarella, E., & Nora, A. (1995). Influences affecting the development of students' critical thinking skills. *Research in Higher Education, 36*(1), 23–39.

Terenzini, P., Springer, L., Yaeger, P. M., Pascarella, E., & Nora, A. (1996). First-generation college students: Characteristics, experiences, and cognitive development. *Research in Higher Education, 37*(1), 1–22.

Terenzini, P., & Wright, T. M. (1987). Influences on students' academic growth during four years of college. *Research in Higher Education, 26*(2), 161–179.

Turk, J. (Ed.). (2000). *The corporate campus: Commercialization and the dangers to Canada's universities and colleges*. Toronto: Lorimer.

Turk, J. (Ed.). (2008). *Universities at risk: How politics, special interests and corporatization threaten academic integrity*. Toronto, ON: Lorimer.

Twenge, J. M., Zhang, L., & Im, C. (2004). It's beyond my control: A cross-temporal meta-analysis of increasing externality in locus of control, 1960–2002. *Personality and Social Psychology Review, 8*(3), 308–319.

Vaillancourt, T. (2012). Students aggress against professors in reaction to receiving poor grades: An effect moderated by student narcissism and self-esteem. *Aggressive Behaviour, 39*(1), 71–84. doi:10.1002/ab.21450

Witt, P. H., & Handal, P. J. (1984). Person-environment fit: Is satisfaction predicted by congruency, environment, or personality? *Journal of College Student Development, 25,* 503–508.

Wood, W., & Quinn, J. M. (2005). Habits and the structure of motivation in everyday life. In J. P. Forgas, K. D. Williams, & S. M. Laham (Eds.), *Social motivation: Conscious and unconscious processes* (pp. 55–70). New York, NY: Cambridge University Press.

Zhao, H. (2011). *Student engagement as a quality measure in the Ontario postsecondary education system: What we have learned about measures of student engagement*. Toronto, ON: Higher Education Quality Council of Ontario.

Contact Information

J. Paul Grayson
Department of Sociology
York University
grayson@yorku.ca

J. Paul Grayson is Professor of Sociology at York University. His interdisciplinary research interests have focused on society and travel in early Canada, Canadian social movements, political sociology, the sociology of Canadian literature, the causes and consequences of plant shutdowns in Canada, work and health, vigilantism in Canada and the United States, training design and development, university educational experiences and outcomes, and the quality of life in Canadian cities. His current research is on the female Canadian and American student experience in the 1960s.