CJLT RCAT
Canadian Journal of Learning and Technology
La Revue canadienne de l'apprentissage et de la technologie

Volume 51 (1) Winter / Hiver 2025

Examining the Emotional Tone of Student Evaluations of Teaching

Analyse du ton émotionnelle des évaluations de l'enseignement par les personnes étudiantes

Derek Newman, Cambrian College, Canada

Abstract

Student-written evaluation ($N = 600$) of professors was examined to determine the emotional tone of the words used to evaluate faculty. Using the revised Dictionary of Affect (DOA; Whissell, 2009), evaluation words ($N = 26{,}764$) uploaded to the *Rate My Professors* website between 2018 and October of 2023 were measured for their pleasantness, activation, and imagery. Overall, the emotional tone of the students' written evaluation was very close to the DOA's definition of everyday English ($M = 50$) for all three categories: pleasantness ($M = 51.1$, $SD = 6.3$), activation ($M = 52.2$, $SD = 4.8$), and imagery ($M = 50.2$, $SD = 7.4$). The results indicated that the written evaluations were uniform in expression and emotional tone: neither very pleasant/unpleasant, active/passive, or imagery/abstract. While significant relationships were found with professor quality and difficulty ratings, the number of words in the evaluation, and the instructor's gender, all associations had small correlational strengths and weak effect sizes, indicating that the variables might not be strong predictors of the emotional tone of student evaluations. If student written evaluations are not emotionally charged, then there is an opportunity to reduce any negative feelings faculty members have attached to the process.

Keywords: academia, emotional tone, student evaluation of teaching, student evaluations

Résumé

L'évaluation écrite des professeures et professeurs réalisée par les personnes étudiantes ($N = 600$) a été examinée pour déterminer le ton émotionnel des mots utilisés pour les évaluer. À l'aide du dictionnaire *Dictionary of Affect* (DOA ; Whissell, 2009), les mots d'évaluation ($N = 26{,}764$) téléversés sur le site web *Rate My Professors* entre 2018 et octobre 2023 ont été mesurés pour leur caractère agréable, leur activation et leur imagerie. Dans l'ensemble, le ton émotionnel de l'évaluation écrite réalisée par les personnes étudiantes était très proche de la définition de l'anglais courant du DOA ($M = 50$) pour les trois catégories : caractère agréable ($M = 51{,}1$, $SD = 6{,}3$), activation ($M = 52{,}2$, $SD = 4{,}8$) et imagerie ($M = 50{,}2$, $SD = 7{,}4$). Les résultats indiquent que les évaluations écrites étaient uniformes en

termes d'expression et du ton émotionnel : ni très agréables/désagréables, ni actives/passives, ni imagées/abstraites. Bien que des relations significatives aient été trouvées avec la qualité de la personne enseignante et les notes de difficulté, le nombre de mots dans l'évaluation et le genre de la personne enseignante, toutes les associations avaient des forces de corrélation faibles et des tailles d'effet faibles, ce qui indique que les variables pourraient ne pas être des prédicteurs forts du ton émotionnel des évaluations réalisées par les personnes étudiantes. Si les évaluations écrites réalisées par les personnes étudiantes ne sont pas chargées d'émotion, il est possible de réduire les sentiments négatifs que les personnes enseignantes attachent au processus.

Mots-clés: académie, ton émotionnel, évaluation de l'enseignement par les personnes étudiantes, évaluations par les personnes étudiantes

Introduction

Student evaluation of teaching (SET) is a controversial subject due to the subjective nature of the evaluation (Dahal & Rafiq, 2023). The use of SETs, including the research behind the practice, began in the early 1990s (Algozzine et al., 2004) and is widely used in academia today (Wagenaar, 1995). Although there are differences in questions posed to students across the various academic institutions, the goals include (1) providing feedback to faculty, (2) assisting academic institutions in decision-making (for example, tenure), (3) giving students data for course and faculty selection, and (4) providing data for SET research (Marsh & Roche, 1993). Kember et al. (2002) considered that the primary purposes of SETs are to (1) promote faculty improvement during the evaluation process, (2) provide data for appraisals, and (3) contribute to academic institution accountability. According to Penny (2003), SETs are frequently used by academic institutions as they are easy to collect and interpret.

Previous research has noted that students are interested in performing SETs (Foster, 2003; Howell & Symbaluk, 2001). Research has also demonstrated that faculty value SETs (Balam & Shannon, 2010; Kulik, 2001) and are concerned about how students view their teaching (Spooren et al., 2013). Nevertheless, faculty also have worries about SETs, which stem from their reliability (consistency, stability, and dependency of the instrument) and validity (the extent to which SETs measure what they intend to measure). Unfortunately, research on the reliability and validity of SETs is mixed. Whereas some studies have demonstrated their reliability (Barnes & Barnes, 1993; Feldman, 1989; Zhao & Gallant, 2012), others have indicated they lack reliability. In Clayson's (2018) comprehensive study on SET reliability, the author reviewed the challenges with the reliability measures that have been used and concluded the tool to be "inadequate" (p. 666). This paper addressed three significant challenges to establishing reliability in SETs, including (1) methodological difficulties, (2) problems in evaluating student ratings, and (3) a lack of instrument construct definitions. Clayson (2018) concluded that the challenge with SETs surrounds the lack of consistency among individual student responses, indicating that students may disagree on what they are being asked to evaluate; therefore, SETs lack both reliability and validity.

Here again, however, the evidence is mixed. Several research articles have concluded that SETs have many forms of validity (Blackburn & Clark, 1975; Burdsal & Bardo, 1986; Cook et al., 2024; Ellett et al., 1997; Overall & Marsh, 1980; Wright & Jenkins-Guarnieri, 2012). Other studies have questioned the validity of SETs as a tool to assess a teacher's effectiveness (Clayson, 2018; Shevlin et al., 2000). Quansah et al. (2024) cited student evaluators as a significant challenge due to inconsistencies in faculty ratings. Uttl (2021) conducted a comprehensive review of validity challenges in SETs, citing several key issues, including defining effective teaching, questioning whether students learn more from highly rated faculty, external factors that affect evaluations (for example, students' prior knowledge), and student preference factors. In an earlier paper, Uttl et al. (2017) conducted a meta-analysis of previous SET meta-analyses, concluding not only that SETs do not measure teaching effectiveness but also that academia "may want to abandon SET ratings as a measure of faculty's teaching effectiveness" (p. 22).

Spooren et al. (2013) drew a similar conclusion, stating that the accuracy of SETs in measuring effective teaching is uncertain. They further argued that faculty and students may disagree on what constitutes effective teaching, echoing Zhao and Gallant's (2012) argument that a significant challenge to establishing SET validity is the lack of consistency in the definition of effective teaching. Other SET concerns include the lack of space/time for students to explain their responses, the difficulty in interpreting their responses, and the lack of knowledge surrounding SET research and its reliability and validity challenges (Spooren et al., 2013).

Rate My Professors

The student evaluations in this study were pulled from the Rate My Professors[1] (RMP) website, a platform that allows students to review their instructors. At the time of writing (2023), the website had more than two million professor ratings. The website requires students to create an account, after which the user can rate a professor already included on the site or add a school/professor. The user selects the instructor to submit a rating, and a new webpage opens with various evaluation options. The user proceeds to select a course code and indicate whether it is an online delivery. Next, the user rates the professor's quality on a 5-point scale: awful (1), OK (2), good (3), great (4), and awesome (5). The following section asks the user to rate the difficulty of the professor's course on a 5-point scale: very easy (1), easy (2), average (3), difficult (4), and very difficult (5). Other questions include asking if the student would take a course with this professor again, whether the course was taken for credit, if it had a textbook, and whether attendance was mandatory. The final question asks the user to select the grade they received in the course using letter grades (for example, A+, C–) or indicate if the course had no grades, or state if the course was dropped or incomplete. Among the response options are "not sure yet" or "rather not say." Users are also given the option to select up to three tags for the post, with options such as "tough grader," "amazing lectures," "lots of homework," and "caring." Since its inception, much

[1] https://www.ratemyprofessors.com

research has been devoted to evaluating the RMP website: some papers providing support for the validity of the rating scales (Brown et al., 2009; Colardarci & Kornfield, 2007; Otto et al., 2008; Sonntag et al., 2009; Timmerman, 2008) and others where researchers are unconvinced (Felton et al., 2004; Legg & Wilson, 2012; Murray & Zdravkovic, 2016).

Student-Written Evaluations

On many SETs students can write comments about their professor, but the literature on this component is limited. Most research on student SET comments uses instruments like Leximancer or Wordstat, an automated semantic analysis tool that finds themes within the text (Abd-Elrahman et al., 2010; Shah & Pabel, 2020; Stupans et al., 2016). A study by Olvet et al. (2021) noted that the challenge with student-written evaluations is reviewer hesitation to provide faculty names when giving negative or constructive criticism. Past research has demonstrated that students hesitate to give negative evaluations due to power dynamics, fear of reprisal, and student–teacher relationship breakdown (Afonso et al., 2005; Janss et al., 2012). Their concerns could be warranted as a study by Robins et al. (2020) interviewed 24 medical faculty and noted that they admitted to a likely bias against students who rated them negatively.

Very few studies have been carried out on students' written evaluations on RMP. Abd-Elrahman et al. (2010), Shah and Pabel (2020), and Stupans et al. (2016) used text analysis software to identify themes in student posts on RMP, but not emotional tone. Silva et al. (2008) examined the positivity/ negativity of written evaluations of psychology teachers. Using an instrument called the IUB Evaluation Services and Testing Multiple Option System (Multi-Op) of Course and Instructor Evaluation, they concluded that there were more positive than negative comments in the evaluations of both the course and faculty. Dahal and Rafiq (2023) used an instrument called DistilBERT to analyze the emotions shown in students' written evaluations on RMP and found that "joy" characterized most of the comments (over 60%), whereas negative emotions (anger, sadness, and fear) "accounted for less than 40% of the student review" (p. 5). However, "anger" was the second most noted emotion after "joy."

Teaching can be a demanding profession and the well-being of teachers is a highly complex area of research (Wang et al., 2023). One aspect of teacher well-being is how they react to reading student evaluations, but very little research exists. Studies have stated that, when reading SETs, teachers feel judged and experience deep emotional responses (Sidwell et al., 2025); feel anxious (Lutovac et al., 2017); feel rageful, sad, neglected, and have self-doubt (Carmack & LeFebvre, 2019); and even find the process painful when getting critical reviews (Arthur, 2009). These feelings can lead to faculty disengagement from SETs, hindering professional growth (Sidwell et al., 2025). The current study examines whether the written section of student evaluations contains emotionally charged words.

Research Questions

This paper will add to the research on SET by examining posts on RMP in the Canadian context to explore: (1) What is the emotional tone of student-written evaluations? and (2) What factors contribute to the emotional tone of student-written evaluations?

Method

A Canadian university was randomly selected from the RMP website. A total of 30 professors with more than 90 student evaluation posts were randomly selected, and the first 20 posts were recorded chronologically (600 posts in total). For the written professor evaluations, no minimum word count requirements were specified. Course quality and difficulty scores were also recorded as part of the analysis, as all 600 posts included the two scores. Unfortunately, the other questions listed previously in this study (for example, whether the course was online or taken for credit) were sparsely answered and therefore not included in the analysis. For example, the grade given in the course was recorded in only 32 of the 600 posts (5.3%) uploaded to the RMP website.

The date range of the evaluations was 2018 to 2023 and encompassed a total of 16 subjects ranging from the social sciences, sciences, business, arts and humanities, to information technology. While the gender of the professor was not listed as an option for students to input, an analysis of the pronouns in the written evaluation was used to create the variable of gender, and the results indicated 66.6% male and 33.4% female faculty.

The emotional tone of the students' written evaluations was analyzed using the revised Dictionary of Affect (DOA) (Whissell, 2009). Whissell had volunteers rate, outside of any context, the emotional tone of words on three scales: pleasantness, activation, and imagery (how easy it is to form a mental picture). Averages above or below the dictionary's mean score of 50, representing everyday English, were taken to indicate emotional tone differences in one direction. The three emotional scales have standard deviations of 22 for pleasantness and activation and 36 for imagery (Whissell, 2009). For example, a score above 50 would indicate the word or entire work is pleasant or active, while a score below 50 would indicate the word or entire work is unpleasant or passive. A score of 72 on the pleasantness scale would be one standard deviation above the mean, suggesting a more pleasant emotional tone than everyday English, while a word or work with a score of 94 would be even more pleasant, as it is two standard deviations above the mean. While the DOA has not been used in other research specifically to examine the emotional tone of SET, it has been used in multiple text examinations, including television dialogue (White et al., 1989), religious texts (Whissell, 2012a), song lyrics (Whissell, 1996), and political speech (Whissell, 2012b). As emotion is an aspect of language, the DOA can be used as a framework to examine the speech of SETs.

The revised DOA matching rate for the evaluation words studied was 76%, somewhat lower than the rate of 90% expected for everyday English texts (Whissell, 2009). The slightly lower matching rate was partially due to the faculty's name being frequently cited in the written evaluations. The DOA database consists of common words in the English language, rather than names.

Results

Mean Scores

Overall, the emotional tone of words in the student evaluations (N = 26,764) was very close to everyday English (M = 50) for pleasantness (M = 51.1, SD = 6.3), activation (M = 52.2, SD = 4.8), and imagery (M = 50.2, SD = 7.4). The average word count per student was 44.6 (SD = 19.9), and the average DOA match rate was 33.9 per evaluation post. Table 1 shows the quality and difficulty frequencies.

Table 1

Quality and Difficulty Frequencies in the Student Posts on RMP.com

Quality	Frequency	Difficulty	Frequency
Awful	24.2%	Very easy	5.8%
OK	10.8%	Easy	16.3%
Good	10.5%	Average	33.2%
Great	13.7%	Difficult	27.0%
Awesome	40.8%	Very difficult	17.7%

Concerning the quality of faculty, Table 1 shows that most students rated the professors as "awesome" (40.8%). Combined, "great" and "awesome," the two positive categories, represented 54.5% of the student ratings, whereas "awful" and "OK" together accounted for 35% of the ratings. For professor difficulty, most students rated the faculty as "average" (33.2%). Approximately twice as many students rated the professors as "difficult" or "very difficult" (44.7%) as compared with those who gave ratings of "very easy" or "easy" (22.1%).

Emotional Tone Differences and Quality and Difficulty Ratings

To evaluate whether the emotional tone of the students' written evaluations was related to their quality and difficulty ratings, the study conducted Pearson correlation analysis (Table 2).

As shown in Table 2, concerning quality, two of the three emotional scales (pleasantness and activation) were significantly related to student ratings. For pleasantness, a moderate positive correlation was noted (r = .45, N = 600, p < .01), indicating that as the quality rating of the faculty increased, the pleasantness of the words in the evaluations also increased. For activation, a weak positive correlation was noted (r = .13, N = 600, p < .01), indicating that as the quality rating of the professor increased, the activation of the words in the student evaluations also increased. No significant relationship was found between the professor's quality rating and the imagery in the written evaluations.

Table 2

Pearson Correlations Between Dictionary of Affect's (DOA's) Scales and Quality and Difficulty Ratings

Rating	DOA Scale	Correlation
Quality	Pleasantness	.45**
	Activation	.13**
	Imagery	.04
Difficulty	Pleasantness	–.15**
	Activation	–.01
	Imagery	–.02

Note. ** $p < 0.01$ level (2-tailed).

Concerning difficulty, only one of the three emotional scales was significantly related to the written evaluations. A weak negative correlation was noted for the emotional scale of pleasantness ($r = –.15$, $n = 600$, $p < .01$), indicating that as students gave more difficult ratings, their pleasant words decreased.

Word Count

Several significant relations were observed with the word count of written evaluations (Table 3). As Table 3 shows, concerning quality ratings, a weak negative correlation was noted ($r = –.21$, $N = 600$, $p < .001$); i.e., the professor's quality rating decreased as the evaluation word count increased. Concerning difficulty ratings, a weak positive correlation was noted ($r = .13$, $N = 600$, $p < .001$); i.e., as the evaluation word count increased, the professor's difficulty rating also increased.

In addition, all three DOA emotional scales were significantly related with the word count of the student evaluations. For pleasantness, a weak negative correlation was noted ($r = –.25$, $N = 600$, $p < .01$); i.e., as the student evaluation word count increased, the pleasantness of the words decreased. A weak negative correlation was also noted ($r = –.17$, $N = 600$, $p < .01$) with respect to activation; i.e., as the student evaluation word count increased, the words became increasingly passive. Lastly, a weak negative correlation was noted ($r = –.16$, $N = 600$, $p < .01$) for imagery; i.e., as the student evaluation word count increased, the words became increasingly abstract.

Table 3

Pearson Correlational Relationships with Student Evaluation Word Count

Variable	Word count
Quality	–.21**
Difficulty	.13**
Pleasantness	–.25**
Activation	–.17**
Imagery	–.16**

Note. ** $p < 0.01$ level (2-tailed).

Gender of the Professor

To determine if there were differences in student evaluations that depended on the gender of the professor, an independent-samples *t*-test was conducted, and significant differences were found both in one DOA scale and in word count. Concerning the DOA scales, significant differences were noted in the pleasantness of student evaluations ($t = –3.98$, $p = <.001$, eta squared = .02), with female professors ($M = 50$, $SD = 5.4$) having slightly fewer pleasant words than male faculty ($M = 52$, $SD = 6.6$). It noteworthy that, despite significant differences in pleasantness, the mean scores of both genders fell within the DOA pleasant range (mean of 50 or higher). Significant differences were also noted with respect to word count ($t = 2.63$, $p = .009$, eta squared = .001), with female faculty evaluations ($M = 48$, $SD = 18.3$) having an average of five words more than male faculty evaluations ($M = 43$, $SD = 20.5$).

Discussion

This study was conducted to examine the emotional tone of student evaluations posted on RMP and its relationship with both quality and difficulty ratings given by the student and evaluation word count. The results demonstrated that, outside of a moderate correlational relationship between pleasant words and professor quality ratings, all associations had either low correlational strength or weak effect sizes, indicating that the variables examined in this study may not be strong predictors of student-written evaluations.

All three DOA mean scores for the student-written evaluations were very close to its definition of everyday English ($M = 50$), with activation having the highest mean score ($M = 52.2$). These results imply that the student-written evaluations were neither very pleasant nor unpleasant, very active or passive, or very imagery or abstract. Even though most students rated the professors as “difficult” or “very difficult” (44.7%), the pleasantness mean was 51.1, which is very close to the DOA definition of everyday English at 50. There was also a high degree of uniformity of expression in the written

responses. The DOA has standard deviations of 22 for pleasantness and activation and 36 for imagery (Whissell, 2009). In this study, all three scales had very small standard deviations, ranging from 4.8 to 7.4, which suggests that the emotional tone of student responses was consistent across all evaluations.

Additionally, despite a significant correlation between the pleasantness of evaluations with the difficulty rating, the correlation strength was very weak ($r = –.15$), implying that pleasant words may not be a strong factor in predicting the difficulty ratings of the professors. The lack of significantly unpleasant written evaluations aligns with Olvet et al. (2021), who noted how difficult it is for students to give specific faculty names when giving negative evaluations. Due to the transparent nature of RMP, students must publicly rate an instructor by name, which could make them feel vulnerable and explain the lack of unpleasant words in the evaluations. Although RMP is anonymous, which could suggest protection against adverse reactions, the students might be affected by social desirability. Social desirability is the habit for individuals to acquaint themselves favourably, and has been demonstrated to be similar in online and in-person scenarios (see Dodou & de Winter, 2014).

The lack of a relationship between the written evaluations and imagery is interesting. In the DOA, imagery is defined as how easy it is to picture a word in your mind. Creating evaluations for faculty with high imagery could be difficult for students due to a lack of knowledge about terms associated with effective teaching. This is a challenge for SETs. While a comprehensive literature review on the definition of effective teaching would be beyond the scope of this paper, a study by Stronge et al. (2011) examined "the classroom practices of effective versus less effective teachers" (p. 339). The study extracted approximately 17 words that described effective teaching, including organization, responsibility, classroom management, feedback, clarity, fairness, caring, respect, encouragement, and more. Of the 17 words used in Stronge et al. (2011), 12 (70.5%) were used by the students in this study but appeared only 151 times in 26,764 words (0.6%). Therefore, student evaluations in this study did not include words that describe effective teaching, an observation that raises questions about students' ability to assess faculty teaching ability accurately. After the words that describe effective teaching in Stronge et al. (2011) were run through the DOA, the mean imagery score was 56.4, close to DOA's definition of everyday English ($M = 50$). Thus, imagery might not be a valid measure of effective teaching when using the DOA as an evaluation tool.

Earlier, this paper addressed the reliability and validity challenges with SETs. Some challenges are owed to the paucity of research on student-written evaluations. While Pearson correlational strengths were low in this study, word count had significant relationships with all the variables employed in the analysis (quality and difficulty ratings of the faculty, all three DOA emotional scales, and the gender of the professor). Concerning quality and difficulty ratings, the results demonstrated that when the word count increased, the quality rating decreased and the difficulty rating increased. Stated otherwise, students who rated professors as having poorer quality or being more difficult tended to write longer reviews. Concerning review pleasantness, as the number of words increased, the pleasantness of the words decreased. However, whether these results can be considered stable across all SETs is uncertain. Reliability refers to the consistency and stability of data, and student-written evaluations may pose a challenge for reliability assessment. Although research is limited, it is plausible that most SETs do not

require students to write a specific number of words or have a minimum/maximum word count. Written evaluations might also pose a challenge for validity. According to the results of this study, if students write longer written evaluations, this could affect not only the pleasantness of the evaluation but also other scores that the SET aims to assess, such as instructor quality and course difficulty. This study had an average word count of 44.6 (*SD* = 19.9), with a range of 2 words to 77 words in the evaluations. If formal academic SETs have word counts similar to those in this study, with wide variations of word count, then reliability and validity challenges with written evaluations might occur. If students were asked to write a longer faculty review in the SETs, could it affect institution-specific SET scales, such as quality and difficulty ratings? Again, if students were asked to write longer reviews, could the emotional tone of the words become increasingly unpleasant?

While this study did find significant differences in evaluations for male and female faculty, the effect sizes were weak for evaluation pleasantness (eta squared = .02) and evaluation word count (eta squared = .001). It is important to note that the gender identities of the professors in this study were not officially confirmed, and the identifiers were determined by analyzing the pronouns in the student evaluations. This study did find that fewer pleasant words were used for female faculty (*M* = 50) than male faculty (*M* = 52). Still, the difference was minimal, and, for both genders, the findings were within the pleasant range for the DOA. Additional research is needed to examine the relation between student-written evaluations and a professor's gender, since previous studies (for example, MacNell et al., 2015) have demonstrated a positive bias towards male teachers. In MacNell et al. (2015), the authors used an online teaching environment to determine if student ratings were based on the perceived gender of instructors. Students gave significantly higher ratings to the teachers with male identities than to those with female identities, regardless of the teacher's actual gender, which was disguised in the online environment. The article provided an example of this bias: "When the actual male and female teachers posted grades after two days *as a male*, this was considered by students to be a 4.35 out of 5 level of promptness, but when the same two teachers posted grades at the same time *as a female*, it was considered to be a 3.55 out of 5 level of promptness" (p. 300).

Limitations and Future Research

This study has several limitations. First, the DOA was created by having participants evaluate words context-free; thus, any evaluation and discussion on the emotional tone of student evaluations when using the instrument must be considered. This research also incorporated convenience sampling, a single academic institution, which may limit the generalizability to other institutions, disciplines, or countries. A single university was chosen due to RMP's website navigation, where users must pick an institution as an initial prompt. Aside from the reliability and validity challenges of SETs, which were addressed earlier in this paper, there are other limitations to examining online student evaluations. The most obvious limitation is that the posts could have been entered by anyone, not necessarily the student who took a course with the professor being rated. In addition, the student evaluations could have been carried out at various times during or after the course. Whether traditional and online courses are comparable is uncertain, constituting another limitation. While RMP does have an option for students to

indicate whether the course was online, students are not required to answer the question, and few posts did. The term "online" was used 48 times in the written evaluations, but without context. For example, some student posts referred to what could be described as an online course. In contrast, others described course material in an online learning management system, which could be in a traditional classroom delivery mode. Another limitation could be the lack of student-reported course format and grade received. It is possible that the emotional tone could be affected by the course format, especially if there was no face-to-face communication with the instructor. Concerning grades, many studies (see Stroebe, 2020) report strong positive correlations between SET and student expected grades. It is possible that the emotional tone of SET could be affected if the sample had many students with very high and very low grades.

While this research uses a transparent methodology, there are ethical concerns in using public data, such as privacy concerns (individual and organizational), bias in student response, replication of results, and possible website ownership censorship. For a more thorough examination of the ethics of using publicly available data, see Cooper and Coetzee (2020).

Recommendations for future research include examining student evaluations from various subjects to determine if they differ in emotional tone. Though this study covered 16 different subjects, the disciplines were unequally represented, making comparisons difficult. Finally, this study did not address any effects that minority professors might have on the emotional tone of student evaluations. Reid (2010) evaluated over 5,000 RMP student posts and noted that minorities, "particularly Blacks and Asians, were evaluated more negatively than White faculty in terms of overall quality, helpfulness, and clarity" (p. 137).

Conclusions and Recommendations for Academia

Using the DOA, student-written evaluation words from RMP were measured for their pleasantness, activation, and imagery. Overall, the emotional tone of the students' written evaluations was very close to the DOA's definition of everyday English, indicating that the words were not emotionally charged (pleasant/unpleasant, active/passive) nor imagery/abstract. The lack of moderate or strong correlational associations and effect sizes, outside of the relationship between professor quality ratings and pleasant words, could indicate that the professor quality and difficulty ratings, the number of words in the evaluations, and instructor gender are not strong predictors of student evaluations. This study offers an opportunity for academic institution administration, faculty, and students to find solutions to any negative feelings towards SETs. If the words in the student evaluations are not emotionally charged, yet some faculty report negative feelings about reading them, then training on feedback literacy may assist. Feedback literacy is the act of giving, exploring, accepting, and applying feedback to maximize personal improvement (Yan & Carless, 2022). Faculty could collaborate with the institution's administration and students to provide feedback literacy training, create multiple feedback sources beyond formal SETs, and guide students through the feedback process (Cook et al., 2024), thereby mitigating any negative feelings associated with the written portion of SETs.

Declaration of Conflicting Interests

The author declares no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

References

Abd-Elrahman, A., Andreu, M., & Abbott, T. (2010). Using text data mining techniques for understanding free-style question answers in course evaluation forms. *Research in Higher Education Journal*, *9*, 1.

Afonso, N. M., Cardozo, L. J., Mascarenhas, O. A. J., Aranha, A. N. F., & Shah, C. (2005). Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. *Family Medicine*, *37*(1), 43–47.

Algozzine, B., Gretes, J., Flowers, C., Howley, L., Beattie, J., Spooner, F., Mohanty, G., & Bray, M. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching*, *52*(4), 134–141. https://doi.org/10.3200/CTCH.52.4.134-141

Arthur, L. (2009). From performativity to professionalism: Lecturers' responses to student feedback. *Teaching in Higher Education*, *14*(4), 441–454. https://doi.org/10.1080/13562510903050228

Balam, E. M., & Shannon, D. M. (2010). Student ratings of college teaching: A comparison of faculty and their students. *Assessment & Evaluation in Higher Education*, *35*(2), 209–221. https://doi.org/10.1080/02602930902795901

Barnes, L. L. B., & Barnes, M. W. (1993). Academic discipline and generalizability of student evaluations of instruction. *Research in Higher Education*, *34*, 135–149. https://doi.org/10.1007/BF00992160

Blackburn, R. T., & Clark, M. J. (1975). An assessment of faculty performance: Some correlates between administrator, colleague, student and self-ratings. *Sociology of Education*, *48*(2), 242–256. https://doi.org/10.2307/2112478

Brown, M. J., Baillie, M., & Fraser, S. (2009). Rating RateMyProfessors.com: A comparison of online and official student evaluations of teaching. *College Teaching*, *57*(2), 89–92. https://doi.org/10.3200/CTCH.57.2.89-92

Burdsal, C. A., & Bardo, J. W. (1986). Measuring student's perceptions of teaching: Dimensions of evaluation. *Educational and Psychological Measurement*, *46*(1), 63–79. https://doi.org/10.1177/0013164486461006

Carmack, H. J., & LeFebvre, L. E. (2019). "Walking on eggshells": Traversing the emotional and meaning making processes surrounding hurtful course evaluations. *Communication Education*, *68*(3), 350–370. https://doi.org/10.1080/03634523.2019.1608366

Clayson, D. E. (2018). Student evaluation of teaching and matters of reliability. *Assessment & Evaluation in Higher Education*, *43*(4), 666–681. https://doi.org/10.1080/02602938.2017.1393495

Colardarci, T., & Kornfield, I. (2007). RateMyProfessors com versus formal in-class student evaluations of teaching. *Practical Assessment, Research and Evaluation*, *44*(12), 1–15. https://doi.org/10.7275/26ke-yz55

Cook, S., Watson, D., & Webb, R. (2024). Performance evaluation in teaching: Dissecting student evaluations in higher education. *Studies in Educational Evaluation*, *81*, 101342. https://doi.org/10.1016/j.stueduc.2024.101342

Cooper, A. K., & Coetzee, S. (2020). On the ethics of using publicly-available data. In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds)., *Responsible design, implementation and use of information and communication technology: I3E 2020. Lecture notes in computer science, 12067* (pp. 159–171). Springer International Publishing. https://doi.org/10.1007/978-3-030-45002-1_14

Dahal, K., & Rafiq, R. I. (2023, May). What makes a good course and professor: Through the lens of RateMyProfessor website. In *Proceedings of the 2023 7th International Conference on Information System and Data Mining (ICISDM)* (pp. 1–9). https://doi.org/10.1145/3603765.3603767

Dodou, D., & de Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, *36*, 487–495. https://doi.org/10.1016/j.chb.2014.04.005

Ellett, C. D., Loup, K. S., Culross, R. R., McMullen, J. H., & Rugutt, J. K. (1997). Assessing enhancement of learning, personal learning environment, and student efficacy: Alternatives to traditional faculty evaluation in higher education. *Journal of Personnel Evaluation in Education*, *11*, 167–192. https://doi.org/10.1023/A:1007989320210

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, *30*, 137–194. https://doi.org/10.1007/BF00992716

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, *29*(1), 91–108. https://doi.org/10.1080/0260293032000158180

Foster, A. L. (2003). Picking apart Pick-A-Prof. *Chronicle of Higher Education*, *49*(26), A33–A34.

Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology*, *93*(4), 790–796. https://psycnet.apa.org/doi/10.1037/0022-0663.93.4.790

Janss, R., Rispens, S., Segers, M., & Jehn, K. A. (2012). What is happening under the surface? Power, conflict and the performance of medical teams. *Medical Education*, *46*(9), 838–849. https://doi.org/10.1111/j.1365-2923.2012.04322.x

Kember, D., Leung, D. Y. P., & Kwan, K. P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education*, *27*(5), 411–425. https://doi.org/10.1080/0260293022000009294

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research*, *2001*(109), 9–25. https://doi.org/10.1002/ir.1

Legg, A. M., & Wilson, J. H. (2012). RateMyProfessors.com offers biased evaluations. *Assessment & Evaluation in Higher Education*, *37*(1), 89–97. https://doi.org/10.1080/02602938.2010.507299

Lutovac, S., Kaasila, R., Komulainen, J., & Maikkola, M. (2017). University lecturers' emotional responses to and coping with student feedback: A Finnish case study. *European Journal of Psychology of Education*, *32*, 235–250. https://doi.org/10.1007/s10212-016-0301-1

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*, 291–303. https://doi.org/10.1007/s10755-014-9313-4

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal*, *30*(1), 217–251. https://doi.org/10.3102/00028312030001217

Murray, K. B., & Zdravkovic, S. (2016). Does MTV really do a good job of evaluating professors? An empirical test of the internet site RateMyProfessors.com. *Journal of Education for Business*, *91*(3), 138–147. https://doi.org/10.1080/08832323.2016.1140115

Olvet, D. M., Willey, J. M., Bird, J. B., Rabin, J. M., Pearlman, R. E., & Brenner, J. (2021). Third year medical students impersonalize and hedge when providing negative upward feedback to clinical faculty. *Medical Teacher*, *43*(6), 700–708. https://doi.org/10.1080/0142159X.2021.1892619

Otto, J., Sanford Jr, D. A., & Ross, D. N. (2008). Does ratemyprofessor.com really rate my professor? *Assessment & Evaluation in Higher Education*, *33*(4), 355–368. https://doi.org/10.1080/02602930701293405

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, *72*(3), 321–325. https://psycnet.apa.org/doi/10.1037/0022-0663.72.3.321

Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, *8*(3), 399–411. https://doi.org/10.1080/13562510309396

Quansah, F., Cobbinah, A., Asamoah-Gyimah, K., & Hagan Jr., J. E. (2024, February). Validity of student evaluation of teaching in higher education: A systematic review. *Frontiers in Education*, *9*, 1–12. https://doi.org/10.3389/feduc.2024.1329734

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education*, *3*(3), 137–152. https://psycnet.apa.org/doi/10.1037/a0019865

Robins, L., Smith, S., Kost, A., Combs, H., Kritek, P. A., & Klein, E. J. (2020). Faculty perceptions of formative feedback from medical students. *Teaching and Learning in Medicine*, *32*(2), 168–175. https://doi.org/10.1080/10401334.2019.1657869

Shah, M., & Pabel, A. (2020). Making the student voice count: Using qualitative student feedback to enhance the student experience. *Journal of Applied Research in Higher Education*, *12*(2), 194–209. https://doi.org/10.1108/JARHE-02-2019-0030

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education*, *25*(4), 397–405. https://doi.org/10.1080/713611436

Sidwell, D., Lee, D., Zimmerman, P.-A., Bentley, S., & Barton, M. (2025). Teaching faculty experiences with student evaluation of instruction: A mixed-methods study. *Teaching and Learning in Nursing*, *20*(1), e276-e284. https://doi.org/10.1016/j.teln.2024.11.009

Silva, K. M., Silva, F. J., Quinn, M. A., Draper, J. N., Cover, K. R., & Munoff, A. A. (2008). Rate my professor: Online evaluations of psychology instructors. *Teaching of Psychology*, *35*(2), 71–80. https://doi.org/10.1080/00986280801978434

Sonntag, M. E., Bassett, J. R., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education*, *34*(5), 499–504. https://doi.org/10.1080/02602930802079463

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, *42*(4), 276–294. https://doi.org/10.1080/01973533.2020.1756817

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339–355.

Stupans, I., McGuren, T., & Babey, A. M. (2016). Student evaluation of teaching: A study exploring student rating instrument free-form text comments. *Innovative Higher Education*, *41*, 33–42. https://doi.org/10.1007/s10755-015-9328-5

Timmerman, T. (2008). On the validity of RateMyProfessors.com. *Journal of Education for Business*, *84*(1), 55–61. https://doi.org/10.3200/JOEB.84.1.55-61

Uttl, B. (2021). Lessons learned from research on student evaluation of teaching in higher education. *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers*, *13*, 237–256. https://doi.org/10.1007/978-3-030-75150-0_15

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wagenaar, T. C. (1995). Student evaluation of teaching: Some cautions and suggestions. *Teaching Sociology*, *23*(1), 64–68. https://doi.org/10.2307/1319382

Wang, H., Burić, I., Chang, M., & Gross, J. J. (2023). Teachers' emotion regulation and related environmental, personal, instructional, and well-being factors: A meta-analysis. *Social Psychology of Education*, *26*, 1651–1696. https://doi.org/10.1007/s11218-023-09810-1

Whissell, C. (1996). Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities*, *30*, 257–265. https://doi.org/10.1007/BF00055109

Whissell, C. (2009). Using the Revised Dictionary of Affect in Language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, *105*(2), 509–521. https://doi.org/10.2466/PR0.105.2.509-521

Whissell, C. (2012a, January). Emotional consistency as evidence of dynamic equivalence among English translations of the Bible. *Comprehensive Psychology*, 1. https://doi.org/10.2466/28.49.CP.1.15

Whissell, C. (2012b). The emotionality and complexity of public political language in Canada's Question Period. *English Language and Literature Studies*, *2*(4), 68–76. http://dx.doi.org/10.5539/ells.v2n4p68

White, E., Whissell, C., & Dewson, M. (1989). An objective quantification of the affective tone of language in children's television programing. *Journal of Social Behavior & Personality*, *4*(1), 127–131.

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education*, *37*(6), 683–699. https://doi.org/10.1080/02602938.2011.563279

Yan, Z., & Carless, D. (2022). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment & Evaluation in Higher Education*, *47*(7), 1116–1128. https://doi.org/10.1080/02602938.2021.2001431

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, *37*(2), 227–235. https://doi.org/10.1080/02602938.2010.523819

Authors

Derek Newman, PhD, is a Psychology Professor in the Schools of Community Services, Interdisciplinary Studies & Public Safety (CIPS) at Cambrian College in Canada. His primary research activities focus on studying or learning strategies. *Email*: derek.newman@cambriancollege.ca
ORCID: https://orcid.org/0000-0001-9146-4991