Toward a More Just Student Evaluation of Teaching:

An Investigation of the Literature

By: Jane C. Duffy

Submitted to the Faculty of Arts

University of Alberta

in partial fulfillment of the requirements for the degree of

Master of Arts in Communication and Technology

June 28, 2021

Acknowledgements

I wish to thank my superb and honourable supervisor, Dr. Rob McMahon, for his great patience, wise advice, and unwavering commitment to the completion of this project. I learned a great deal from this gentleman's academic direction, and I will not forget the lessons that I learned from him throughout the writing process. I am also very appreciative of my faculty advisor, Professor Gordon Gow, who provided me with helpful and expert guidance at every juncture of the MACT programme. Heartfelt thanks are due to my husband, Dr. Michael Tiefelsdorf, for his thoughtful encouragement throughout my MACT studies. Gratitude is owed, also, to MacEwan University, Edmonton, Alberta, which generously supported my pursuit of this degree.

This paper is written in memory of my late parents, John Ignatius and Aileen Mary Duffy.

TABLE OF CONTENTS

ABSTRACT

Student evaluations of teaching (SETs) have come under fire in recent years as their value relative to their deficiencies became the topic of arbitration between Ryerson University and Ryerson Faculty Association in 2016. This study looks at these deficiencies as well as asks whether the design and process of student evaluation should be amended in order to increase their value. The methodology chosen is a literature review. In addition to using intuitive keyword searching for relevant literature, a controlled vocabulary was designed that functions as a screen, organizing and controlling tool against which each piece of the reviewed literature was also checked. The controlled vocabulary method of analysis yielded SET backgrounds, designs, user application and non-teaching contexts, and administrative issues. A chapter is dedicated to each of these topics. Background, process, administrative and historical analyses uncover not only the flaws of SETs, but also some compelling reasons for why these tools should nonetheless be retained, albeit with the flaws substantially remedied. The closing chapter offers a vision of - and suggests a method of investigation toward - such remedial efforts.

Chapter ONE: INTRODUCTION

In recent years, due to increasing concerns about the efficacy and validity of Student Evaluations of Teaching (SETs) in universities, student voices have become diminished regarding their feedback of their instructors, if not entirely silenced, even though student perspectives are of unique value in determining the effectiveness of their instructors. In this study, the literature about SETs is investigated with a view to identifying explanations, pursuit of which could arguably restore value and credibility to student feedback. In this context my Research Question is: Should the design and process of student evaluation be reformed in order to yield more reliable student feedback of their professors?

This chapter provides context for my study, highlighting particularly why it is of interest as a graduate capstone research project. Having been a senior librarian in a number of universities since the 1990's, I have sat on several faculty evaluation committees that review SETs, which had been submitted by applicants as components of their dossiers. A steady theme in committee discussions about these SETs as a means of assessing instructional skill is the question: "How fair and representative are the SETs of the quality of this instructor's teaching?" Setting aside the inherent controversies and challenges of measuring the quality of teaching, there are studies which suggest that, because of the design and administration of these tests, university students are unclear about SET criteria and how they are being asked to apply that criteria to their evaluation of "teaching excellence" and "instructor effectiveness" (Saunders and Ramírez, 2017; Robinson and Hilli, 2016). Because of this lack of clarity, students may consequently provide responses that are – often inadvertently - susceptible to a number of errors,

including bias against certain groups. (Flaherty, 2020; Mitchell and Martin, 2018) These flaws are examined in much of the literature on SETs.

Additionally, Saunders and Ramírez 2017 summarily condemn measures of post-secondary teaching excellence as instantiations of neo-liberalism's focus on the "commodification of universities" (Saunders and Ramírez, 2017). Philosophically, the notion of neo-liberalism in post-secondary education is generally understood to mean that higher education administrative decisions have become increasingly informed by financial, political and other ideological non-academic business considerations, rather than by traditional scholarly and research goals and objectives. So, in light of concerns about neo-liberalism's increasing influence on higher education, pressing questions have been raised about the applications of these teaching evaluations, particularly regarding how, arguably compromised by this neo-liberalism, their results have become negative influences militating against the faculty member being evaluated.

There is a growing number of anecdotal faculty "horror stories" which chronicle the damage done to their careers by the use of SETs as the dominant tool by which teaching effectiveness is measured and evaluated by university administrators. It is widely acknowledged that approaches to the evaluation of teaching and learning in academia have undergone significant changes in recent years, and the question of whether the SET system for evaluating the faculty member's role in a just and constructive fashion has become irrelevant and sometimes destructive. With the onset of COVID-19 changes and pedagogical upheavals in the Spring of 2020, many universities increasingly argue that data collected through SETs have become unfair for a number of additional reasons (Lederman, 2020).

Other problematic issues are related to the construction, mode of delivery, and administrative timing of the SET instrument. For example, the recent Ryerson University Faculty

Association arbitration ruling on SETs (Freishtat, 2018) concluded that these tests could not be the sole means by which Canadian University teaching could be justly evaluated for purposes of promotion and tenure. The Ryerson University Faculty Association ruling was the culmination of the third stage grievance of a Ryerson University faculty member's "unjust denial of promotion based on anonymous feedback" (Freisshtat, 2018, p. 3) by the RUFA against Ryerson University. In response to this grievance, the arbitrator ruled that such anonymous feedback - as collected and presented in SETs - could not be used as the sole means by which the University could evaluate the performance of a faculty member in any setting. Subsequently, throughout 2018 to the present, in an effort to avoid more such interventions, arbitrations and consequent negative rulings, greater numbers of University promotion and tenure committees, such as Dalhousie University and the University of Southern California, began either to downplay – or to eliminate entirely - faculty SET scores in their evaluation deliberations (Lederman, 2020) This raises the question: Should the use of SETs to evaluate teaching be completely rejected by all post-secondary institutions or can the process be reformed in order to incorporate more reliable student feedback?

As will be demonstrated through a systematic, analytical review of the literature, existing research demonstrates some of the flaws of SETs as a means for students to evaluate their instructors according to the same standards understood by academic administrators. At the same time, research in this area notes that students' feedback about their learning experience is still critical information that is unique to the perspectives of those learners. Therefore, rather than setting aside the perspectives of students completely, I argue in this project that serious and purposeful effort should be undertaken to design an instrument and process of student evaluation of teaching that captures valid, reliable and therefore actionable data about instructor

effectiveness from students. Rather than completely ignore or devalue student perspectives due to flaws in process and administration, I suggest that it would be in the best interests for all stakeholders to investigate - or to contribute to the literature toward informing the development of – a "better" SET, i.e., one that meets generally accepted test performance norms of good faith, transparency and validity. (Buskist and Keeley, 2018)

To make this argument, I review the existing literature in the area of SETs, and attempt to identify those gaps whose further pursuit and investigation can contribute to the research and development of an instrument that will accurately and reliably evaluate a faculty member's teaching excellence and instructional effectiveness from the students perspective.

This study aims, then, to contribute to the literature that supports a better SET by outlining and explaining the context of the issue and current debate, and by identifying and underlining gaps that warrant additional exploration and research. My hope is that this exercise will contribute to the advancement of ongoing efforts to contribute to the investigation, development, and implementation of SET methods, approaches and reforms. Here I focus on outlining the contours of this debate, with reference to the existing research; my conclusion raises the question of whether providing students with a framework that objectively defines "teaching excellence" vis the quality and effectiveness of their course instruction will improve the resulting SET exercise. My suggested next steps would be a topic for investigation in a later study.

Chapter TWO: METHODOLGY:

Typically, literature reviews are conducted by individual searches of library catalogues, university repositories, gray literature collections, social media tools including wikis, blogs and RSS feeds, and journal abstracting/indexing services. (Oliver, 2012) Research on the topic of SETs, however, is still emerging, and so I argue that access to it is not fully controlled within the formal, traditional classification structure of any one of these tools. To query the literature on my emerging topic, then, and to perform as exhaustive and as relevant a search as possible, I *partially* adapted a methodology emerging in data informatics research, such as that used by Ishida et al. 2020, which designed a system of keyword searches, vocabularies, metadata and formal headings, to search for and to classify scientific data.

In addition to intuitive keyword searching, this enhanced methodology involved constructing a controlled vocabulary, a common free-text tool in indexing and cataloguing librarianship, on the broad topic of SETs. The purpose of designing a controlled vocabulary is to supplement and to probe for properties not covered by standardized search methods. Controlled vocabularies are specialized free-text systems tools that are developed by librarians, data scientists and other knowledge workers to supplement standard searches in order to harvest and to manage maximum results on a topic from contemporary literature. Not only is the controlled vocabulary a refined search and retrieval tool, but it functions well as an organizational scaffold upon which to sort and then to evaluate the search results.

I used three methods to search for and then to organize the literature: In addition to regular keyword searching, I surveyed some existing highly specialized controlled vocabularies that are used in the field of education to explore whether any had already been constructed that

would suit my purpose. I was unable to locate an existing controlled vocabulary with a comprehensive set of contemporary terminology for the subject of SETs. Therefore, to capture and identify what gaps may exist in literature on the topic of post-secondary SET's, I developed my own controlled vocabulary: This was a "mashup" of existing Library of Congress classification headings, Education Resources Information Center (ERIC) headings augmented by "free text" indexing language and free-text language drawn from prevailing discourse on this topic. The primary goal was to secure exhaustivity in the search within pre-set parameters for the purposes of this project. I then developed the **scope statement** to use in this project's controlled vocabulary, which is this: *"The scope of this controlled vocabulary is limited to the procedures, devices and sets of terms that are used to estimate, evaluate, assess or rate the teaching of post-secondary faculty by college and university students."* The scope statement also functioned as a compact guide against which to check each piece of literature reviewed.

Table 1 presents my controlled vocabulary for subject, theme, contextual analysis and descriptions. In additional to being used as a system of search strings by which additionally to search catalogues and other databases, this tool was applied to each of my materials as an organization and evaluation framework. Each piece of literature was then filed according to a makeshift system of numerical values for internal sorting/organizational/control purposes that I developed specifically for the project. This hierarchically (subject classification) and vertically (alphabetically sorted) integrated tool allowed me to approach from various directions the comparison, discussion, and analysis of the sources of my literature review according to key subjects, themes and contexts.

Table One: Controlled Vocabulary

Term	Number of Articles
BT Assessment - *See* Evaluation (*Classification 600-699)*	42
BT PostSecondary Teachers - *See* College and University Faculty *(Classification 400-499)*	35
BT College and University Faculty *Classification 300-399*	42
NT College and University Faculty Performance – Alignment (Education)	9
NT College and University Faculty Performance – Efficiency	3
NT College and University Faculty Performance – Evaluation	42
NT College and University Faculty Performance – Evaluation Methods	22
NT College and University Faculty Performance – Evaluation Needs	5
NT College and University Faculty Performance – Evaluation Problems	17
NT College and University Faculty Performance – Outcome Measures	6
NT College and University Faculty Performance – Performance Factors and Competencies	2
NT College and University Faculty Performance – Productivity	0
NT College and University Faculty Performance – Scoring Rubrics	6
BT College and University Undergraduates *Classification 400-499*	42
NT College and University Undergraduates – Aerospace Education	2
NT College and University Undergraduates – Physical Sciences	1
NNT College and University Undergraduates – Physical Sciences – Canada	1
NNT College and University Undergraduates – Physical Sciences – Engineering	1

NNT College and University Undergraduates – Physical Sciences – Faculty Relations	0
BT Engineering Education Classification 500-599	1
BT Evaluation Classification 600-699	1
NT Evaluation – Accuracy	0
NT Evaluation – Course Surveys	17
NT Evaluation – Information Utilization	0
NT Evaluation – Measures	5
NT Evaluation – Methods	0
NT Evaluation – Occupational Surveys	0
NT Evaluation – Online Surveys	0
NT Evaluation – Opinions	9
NT Evaluation – Review articles	7
NT Evaluation – Student Surveys	42
NT Evaluation – Teaching Surveys	0
NT Evaluation – Value	0
BT Evaluation Criteria *Classification 700-799*	0
NT Evaluation Criteria – Accountability	0
NT Evaluation Criteria – Accuracy	0
NT Evaluation Criteria – Design	6
NT Evaluation Criteria – Reliability	0
NT Evaluation Criteria – Usability	0
NT Evaluation Criteria – Student perceptions	0

Keywords Search: As mentioned above, my initial search was based on keyword search strategies. This means that I searched general keywords in standard database collections of articles as well as in grey literature and other non-traditional sources and collections, e.g., administrative reports, test samples, et c., using the following key words and free text strings, enhanced by wildcard* and Boolean operators:

- Postsecondary Teaching Assessment,
- Student Evaluation of Teaching, its acronym: SET*,
- Universal Student Ratings of Instructions, its acronym: USRI*, (teach* effective*),
- (teach* standard*),
- (excellen* OR effective* measure*),
- student* bias*,
- evaluat* return rate*,
- (technolog* effect*),
- (facult* evalut*) (performance OR excellence OR effectiveness) Canad* university*.

These keywords were selected as the most common terminology to describe specific facets of the search. The asterix wildcard "*" is used after the word roots in order to capture all permutations of that word. For example, the term "technolog*" captures "technology", "technological" and "technologically."

Library of Congress search: Along with keywords and free-text strings, I also used selected Library of Congress subject headings to search Library catalogs, databases and other information repositories for relevant literature.

These Library of Congress subject headings and sub-headings included:

- Student Evaluation of Teacher Performance,
- Institutional Characteristics,
- Effective Teaching,
- Quality of Service,
- Perception,
- Educational Metrics,
- Student Bias.

Databases identified for this particular search primarily included:

- *ERIC,*
- *Teacher Reference Centre,*
- *Social Sciences Citation Index*, and
- *Academic Search Complete.*
- *Education Research Complete*

Each of these databases have custom-built thesauri that I searched as well to supplement the key words and LC classification subject heading tools identified above.

My literature search yielded a total of fifty-four articles. Of these, the largest number of articles (more than 40) were broadly subject classified as:

(i) College and University Faculty (42);

(ii) Evaluation -- Student Surveys (42);

(iii) Assessment; (41)

(iv) College and University Faculty Performance -- Evaluation (42)

Mid-range numbers of articles (between 15-25) were found in the following subject areas:

(i) College and University Faculty Performance – Evaluation Methods (22);

(ii) College and University Faculty – Evaluation Problems (17), and

(iii) Evaluation – Course Surveys (17).

Interestingly, this process revealed significant gaps in some subject areas related to this topic. Specifically, a minimal number of articles (0-14) were located about this topic that used the 1) keywords noted above and 2) are associated with the following LC subject headings:

(i) College and University Faculty Performance – Productivity;

(ii) Evaluation – Accuracy;

(iii) Evaluation – Information Utilization;

(iv) Evaluation – Methods;

(v) Evaluation – Occupational Surveys;

(vi) Evaluation – Online Surveys;

(vii) Evaluation – Teaching Surveys;

(viii) Evaluation – Value;

(ix) Evaluation Criteria;

(x) Evaluation Criteria – Accuracy;

(xi) Evaluation Criteria – Design;

(xii) Evaluation Criteria – Reliability;

(xiii) Evaluation Criteria – Student Perceptions; and

(xiv) Evaluation Criteria – Usability.

Once I collected the available literature using the three search and organization methods discussed above (controlled vocabulary, keyword search, and subject headings searches, I sifted through and organized the findings according to inclusion/exclusion criteria in Table 2 below. The controlled vocabulary provides the terminology for a literature checklist which I developed to select (include/exclude), then to manage and control research materials for discussion within my literature review. I used this approach to organize materials that I found related to SETs, Teaching Effectiveness Metrics and Educational Survey Design. This checklist was organized by the following headings and applied to each piece of literature as follows:

TABLE 2

A) Authority Control (who is responsible for the intellectual content?) evaluated the literature by asking the following questions:
Is the individual author associated with a reputable educational body or professional assessment organization? **Yes/No_____**
Has the author professional qualifications, e.g., a terminal degree, or considerable experience in the areas of educational assessment, excellence in teaching, teaching performance measurement? **Yes/No__**
Has the author produced/published other work (grey/black) in one or more of the fields of: educational assessment, teaching performance, classroom effectiveness, learning objectives? **Yes/No____**
Is the author a recognized or credentialed expert in one or more of the following: educational assessment, excellence in teaching, teaching performance measurement, quasi-experimental data gathering, as identified in other academic sources? **Yes/No_____**
Is/are the author(s) cited by others in grey/ephemeral publications related to the fields of educational assessment, excellence in teaching, teaching performance metrics, quasi- experimental data gathering, or research design related fields? (*See Web of Science and/or* *Google Scholar)* **Yes/No_____**
If the article is *grey literature* published by a non-academic *organization*, is this organization an authority in educational assessment, excellence in teaching, teaching effectiveness metrics, quasi-experimental data gathering or related fields? **Yes/No_____**

Does the work itself have a current (within the last 5-7 years) and detailed reference list or

bibliography of other credible sources? **Yes/No_____**
B) Accuracy and Efficacy Control (how reliable and useful is the research content?) evaluated the literature by asking the following questions:
Does the article have a clearly and reasonably viable stated aim, thesis or argument? **Yes/No_____**
Does the article have a stated methodology? **Yes/No_____**
If, yes, is the methodology credibly and convincingly applied? **Yes/No__**
Has the article been peer-reviewed by experts in education, communication or experimental design? **Yes/No_____**
If no and the article is *grey literature,* has the article been edited by a reputable authority in one of the following fields: education, communication, quasi-experimental method design? **Yes/No_____**
Is the article supported by authoritative, documented references and/or credible sources from *bona fide* expert individuals or organizations in education, communication or survey design? **Yes/No_____**
Is the work an accurate, unbiased interpretation or analysis of a topic in educational testing, quasi-experimental educational survey design or related topic?**Yes/No_____**
If no, is there any explanation for authorial bias? **Yes/No_____**

Are any limitations of scope or assumptions about its context or environment within the article clearly stated? **Yes/No_____**
If yes, is/are it/they unstated or unacknowledged? **Yes/No_____**
Does the age of the article support contemporary investigation into best practices in educational assessment, SET's, teaching performance? **Yes/No_____**

If an opinion piece, is it an expert opinion or otherwise? **Yes/No_____**
Is the author's standpoint consistently clear? **Yes/No_____**
Is the author's argument convincingly supported to the conclusion? **Yes/No____**
Does the work seem to be *balanced*, i.e., does it mention or reference alternative perspectives or theories, in its presentation? **Yes/No_____**
Have recognized and contemporary educational, replicable methods or testing material been included in the analysis? **Yes/No____**
Does the work support a theoretical method or framework that is both credible and appropriate to the study of education, communication, and quantitative investigation? (This incorporates feasibility, utility and relevance) **Yes/No_____**
Does the article update - or arrive at new - knowledge in educational assessment, quasi-experimental design, teaching effectiveness metrics? **Yes/No_____**
If yes, does the article add sufficient context to make its conclusions a credible contribution to literature in communication, education, testing methods or research? **Yes/No____**

Does the article enrich, develop or advance knowledge within the disciplines of education, quantitative methods or communication studies? **Yes/No____**
Is the article typical of similar studies of the educational testing, SET's or instructor assessment? **Yes/No_____**
If no, does the dissimilarity or atypicality add to the article's credibility? **Yes/No_____**
Does the article have impact? Are there metrics available that demonstrate how the article has influenced other studies or practices in areas of teaching performance, methods or survey design? **Yes/No___**
Does the article provide a replicable model (or paradigm or prototype) for the development of further tests of teacher effectiveness, survey design, performance measurement? **Yes/No_**

Based on the above checklist, I collected 44 solid and appropriate resources to review in more detail. Reading each of these pieces of literature while tagging their content for the keyword and LC classification terminology laid out in my controlled vocabulary, I identified four identifiably dominant subject categories: 1**) Background and historical/descriptive analyses** of University Student SETs; 2) **Design analyses** of University Student SETs; 3) Impacts of **Non-Teaching Contexts**; 4) **Administrative and User Application Analyses** of University Student SETs and 5) **Solutions analyses** of University Student SETs. The first two categories will be discussed in Chapter 3, while my analysis of literature from the third and fourth categories will comprise Chapters 4 and 5.

As noted above and demonstrated through my analysis of this literature, this work identified a considerable number of anticipated categories that had zero articles. These "zero categories" are the gaps in the literature reviewed that provide the warrant for my conclusions and for my proposed additional investigations. A potential remedy, a recommendation for further research, to address this gap in the literature is discussed in detail in the concluding chapter of this Capstone. In the next chapter, I discuss my review and analysis of the existing research literature on historical/descriptive and design analyses.

Chapter THREE: **Historical/descriptive and design analyses of SETs**

In this chapter, I present and analyze literature that is focused on the history and design of students SETs with the goal of learning how SETs have been developed and how they have been applied in evaluations of university instructors. This focus is required to establish the origins of the problems that earlier investigators have pointed out with regard to history and design of SETs. I set out to answer two questions. This section answers these questions, with specific emphasis on how earlier investigators discussed the shortcomings and flaws of SETs.

(i) How did SETs come to be developed in the first place? and
(ii) What practical and philosophical considerations informed their design?

Approaching these high level and all-embracing questions chronologically, **Eldridge 1968's** 50-year-old historical study documents the origins of SET's using a qualitative approach (personal interviews and other actively collected data) to determine university student perceptions of teaching effectiveness. In addition to student interviews, departmental reviews by the chair, awards, et c. were also factored into the administrative evaluation mix. The article makes an interesting case for a return to an administratively run faculty evaluation program, with students playing an important, but only supporting role.

Freishtat 2016 refines, solidifies, and summarizes the cumulative literature on negative findings and conclusions on the reliability and validity of SETs. The focus is on the 2016 Ryerson SET arbitration as the centrepiece example. This study is of interest to this investigation because not only is the author a highly credible expert in the evaluation of teaching excellence, but he also proposes an alternative student evaluation instrument that measures only those classroom experiences which students are believed to be qualified to assess. Freishtat's article

advances the argument that the processes of evaluation design and administration must be reformed if they are to measure and demonstrate what they are intended to measure and demonstrate. I agree with this argument – and the premier goal of study is for it to contribute to the literature and/or development of an improved student evaluation of teaching experience.

Doerer, 2019 makes the claim that higher educational institutions have historically been fully aware of the general deficiencies of the current SET program. However, Doerer finds that the perceived need for a SET instrument of some kind outweighed misgivings about their accuracy and design flaws. The article raises questions, though, as to why, given this context, no steps have been taken to address those deficiencies. For example, Doerer sees no place for the widely used open-ended and amorphous questions such as "How do you rate this instructor?" because he argues that such questions are not tied to any observable or measurable variables. (Doerer, 2019, p.4) This article includes several recommendations to minimize bias in existing evaluation process. Without implementation of these recommendations, Doerer argues that SETs yield no reliable or meaningful information. The author also points to other studies that recommend ways to counter that bias. For the purposes of this research, this article discusses the value (or not) of measuring and weighting instructor skill(s) such as "explaining difficult and abstract concepts sufficiently, etc." (Doerer, 2019, p.4). This article contributes to the important framework used to understand the use of measurable variables such as efficacy of teaching method in the SET process.

Other researchers have attempted to assign observable values to the understanding of teaching excellence. For example, having reviewed a sample of 119 SETs over a period of 6 years, **Nemec et al, 2018** present a tool which defines "teaching excellence". The tool organizes variables under the following headings: (a) organization, (b) communication, (c)

motivation/enthusiasm, (d) rapport, (e) fairness and (f) learning. (Nemec et al, 2018, p. 550). By setting out a system of instructor behavioral variables and learning measurements, it is my view that this framework could be operationalized as an educational resource that students could review prior to their completing a SET: Such a framework would help to inform their evaluation efforts. This framework serves as a possible model to contribute to the development of an improved SET instrument.

Other authors are more critical of using SETs to measure instructor effectiveness. For example, Freishtat's 2016 supplemental report, compiled by the arbitrator of the 2016 RUFA v. Ryerson University grievance, sets out the author's expert opinions on the degree to which SETs are unreliable. Based on the existing literature as well as on his prior research on the notion of "teaching excellence", Freishtat argues that the historically flawed notion of the relevance, justice and propriety of the application of SET information to administrative decisions on faculty tenure and promotion should be questioned -- if not altogether discontinued. His conclusions summarize 10 key points that reflect the widespread unhappiness with SETs in North American post-secondaries:

i) There is little consensus on what SETs measure.

ii) SETs are primarily student satisfaction surveys rather than credible measures of teaching effectiveness.

iii) Because of their own lack of knowledge and subject expertise, students should not be used to rate course content or to assess the instructor's knowledge.

iv) SETs are not an accurate instrument of measurement and to use them as such is inappropriate.

v) Students may only credibly and meaningfully comment on their own experience of the class.

vi) Response rates and who responds affect the ratings, especially if the course is administered online; responders of SETs are not random samples and therefore cannot represent the class as a whole.

vii) Because of the need for instructors to garner high ratings as well as high response rates, teaching to the SET occurs.

viii) Personal, i.e., non-instructional, traits of professors strongly relate to their SET scores.

ix) SETs are negatively affected by gender, attractiveness, age, ethnicity and race.

x) What students read from other students on ratemyprofessors.com influence the attitude of initially unbiased students that then negatively impacts the course SET. (Freishtat, 2016, pp. 1-3)

This report further supports historical arguments that:

i) Students are insufficiently knowledgeable of what constitutes good teaching (including instructor knowledge and course content) to evaluate those elements within a SET and
ii) SETs and their widely "standalone" application of faculty teaching are not valid measures of instructor effectiveness.

Despite these substantial critiques, Freishtat does not put forward any specific proposal for steps to remedy this deficit, other than to decrease the role that SETs play in administrative evaluation exercises. As a summary of the difficulties with SETs, Freishtat's conclusions inform various arguments for their exclusion from promotion and tenure deliberation. The conclusions drawn in this article strongly support that since they were developed, SETs have become flawed in purpose and in design.

Other authors provide a slightly more positive, though still critical, assessment of the genesis of SETs and their design. For example, **Gurung et al., 2018** offer a historical review of how poorly "excellence" in post-secondary teaching has been understood and assessed by students. Challenging the previous informal, i.e., largely anecdotal narratives on the topic, the article includes a compilation of "indicators" or model criteria for teacher assessment of excellence in higher education. These criteria include:

1) training and on-going professional development,

2) learner-centered principles in the design of syllabi,

3) variety of instructional methods in a variety of settings,

4) specific alignment of learning objectives, and best professional practice guidelines with course content,

5) assessment processes that provide the most direct and useful feedback for students,

6) engagement of student evaluations of teaching (both formative and summative).

This article furnishes a clear list of variables for the evaluation of teaching excellence, some of which could be applied to efforts to develop an improved SET tool; I will be revisiting these in the discussion of "Next Steps" in chapter 6. As well, Gurung et al, 2018 provide a rough template for a potential investigation's data-gathering instrument: a questionnaire. (p. 16) The authors, however, do not consider testing student knowledge deficits regarding instructor evidence against these variables.

While the literature reviewed yielded few results focusing exclusively on the design of SETs, there are many studies of SETs which offer perspectives on the deficits and limitations imposed by the current SET design model. So far the literature points to the prevalence of flaws in the SET design, as well as to injustices in the patterns of SET survey administration. For

example, **Atek et al.** 2015 provide a design critique of how male and female lecturers at the Universiti Sultan Zainal Abidin and 5 other Malaysian Universities view and value SETs. The researchers analyzed data descriptively using a statistics program and an independent sample t-test. Their conclusions suggested that women were particularly disadvantaged by the design of these universities' SET surveys. The solution proposed reforms to SET design that involve the faculty themselves playing a role in the design. Their analysis supports the view that strengthening the SET design could be advanced if the instructors themselves understood the value of an SET process and were also committed to working on the known problems to improve it. From the conclusions drawn from this article and having found additional support for them in other literature discussed below, the focus here is that reform efforts must include faculty and student participation in the design/delivery of a more valid, reliable, collaborative, and multi-faceted program of faculty evaluation.

Boysen et al, 2014 documents three SET design studies, the purpose of which was to determine whether faculty and administrators are influenced by differences in the design of teaching evaluations. (p.644) The results overwhelmingly underscored the critical importance of designing a tool that actually tests what it purports to test. In the following year, a follow-up study conducted by Boysen focused on a secondary point of discussion: how designing and executing a poor or ambiguous teaching evaluation instrument leads potentially to misinformation and confusion in the interpretation and uses of SET results: "People will make significant interpretations based upon small design differences in teaching evaluations" (Boysen et al, 2016, p. 274).

Another study, **Clayson, 2018,** examines widely varying course and instructor rankings in the context of poorly designed instruments. Clayson's hypothesis is that inconsistent

understandings by student participants of what specifically they are being asked to evaluate contributed to the variability of results. In this study, SET measures were obtained by drawing 727 individual student responses across 8 sections of "The Principles of Marketing" course from an existing departmental database in the School of Marketing at the University of N. Iowa. One of the clearest results from this study also indicates that the design of the SET continues to leave unanswered what factors exactly constitute instructor "excellence" or "effectiveness". Clayson concludes that poor design of SETs leaves the students "in the dark" about what objective qualities comprise instructor excellence or effectiveness: "Unless an idea is defined and understood by individual raters…the average of responses may tell us little about a hypothesized construct." (Clayson, 2018, p.678) For the purposes of the capstone recommendations, this study suggests that the inclusion of "effectiveness" and "excellence" questions in the design of student evaluations are influenced by their peers' opinions – raising questions and concerns about the construction of institutional SETs.

A related discussion suggests that widely identified design-based barriers to the unimpeachable assessment of teacher effectiveness and instructor excellence exist (**Cone et al., 2018).** This article identifies motivators, barriers and strategies of students that negatively influence SET response rates. Barriers identified, which suggest that further research should be conducted on SET validity questions, resonate with common complaints about SET designs. These include the length of the survey, the length of rating scale, ambiguity of questions including those asking about general ratings for instructors, omission of the role of the SET in the instructors' careers, and the lack of documentation that outlines the benefits for student participants. These design flaws are shown to be directly related to poor engagement of the surveys by students. Low response rates were pointed to by Cone et al (2018) as being the chief

contributors to lack of validity for SET results. The identification of these barriers, generated by imperfect or incomplete design, provides further support to proposals for potential collaborative work between faculty and students to design and deliver a more effective approach to faculty and course evaluation. It's also been strongly suggested that students are more likely to take SETs seriously if they know that their input matters and that it will be deployed in a productive and constructive manner. (Cone et al, 2018)

Many concerns have been raised and are actively discussed in North American universities about whether and under which conditions students are honest when completing in-class and online SETs. For example, **McClain et al. 2018** identify an umbrella of SET design variables which seek to measure the concept of "honesty" in relation to "evaluation effectiveness", "purpose of evaluations", "student grades", "timing of SET completion", "student demographic characteristics", and "method of SET administration". While the subjectivity of "honesty" was a known limitation to the design of the study, McClain et al's 2018 conclusions are helpful indicators for specific areas of improvement in SET design. The study found that student "honesty" in the completion of SETs appears positively associated with two variables: 1) students' awareness of the purpose of the evaluation and 2) reasonable in-class timing set aside for the SET's completion. (p. 380) These findings support the argument for the design of a better SET instrument to measure more accurately teaching effectiveness, a hoped-for outcome for further investigation of this topic.

Universities are committed to providing a variety of classroom experiences, while declaring their intention to measure what constitutes effective teaching within those experiences. How best to assess that range of experiences is explored by **Lu et al.'s 2018,** which surveys the role of SET design in their efficacy. The pedagogical conclusions about SET design within this

article strongly support the value of the measurement of post-secondary teaching effectiveness. Of primary interest within this article for this capstone research are two themes: 1) In order for SETs to measure accurately what they are intended to measure, both professors and students must understand how learning occurs as well as how knowledge is gained. The SET must be designed to support and resonate with that understanding. 2) No SET within this study was found to have been designed to function reasonably as the sole source of evaluation data in the faculty performance review process. (Lu et al., 2018)

The literature reviewed on this topic supports the view that there is a general consensus among researchers working in this area that a better, more reasonable and accurately designed assessment tool for faculty teaching evaluation processes in North American post-secondary institutions is required. Further, the literature reviewed suggests that robust, scale-based responses to properly designed SETs must be translatable into actionable items in order to effect the greatest number of completions for purposes of validity. The ideally designed evaluation instrument, then, must be as brief, as relevant and as concise as possible. For example, **Nemec's 2017** quantitative analysis addressed widespread concerns with both the validity and reliability problems with SETs by developing successful pilot tools for both instructor and course evaluation that all ranked highly for internal consistency, reliability and validity. (Nemec, 2017) In another example, Burden's investigation suggests an alternative "creative evaluation" design as an evaluation substitute for the current system – which he concludes is irredeemable. Burden proposes that another, entirely different tool, one that is not arbitrary or based on overly generalized assumptions, must replace the existing approaches and their reliance on ill-defined categories, the meanings of which are unclear to students. (Burden, 2010) Burden defines several problematic areas:

1) the lack of applicability of feedback, meaning that comments were too general to be useful in any operational sense,

2) the lateness of feedback, meaning that the timing of the SET's administration is invariably too late in the term to have real validity to the instructor's effectiveness over the entire term,

3) the challenges of interpreting data: the "feedback loop" was determined to be too open-ended to afford data that would support "continuous quality improvement"

4) lack of accountability of students for their feedback and

5) the students' lack of knowledge about what comprises quality instructor evaluation. (Burden, 2010, p. 107),

This chapter has considered how SETs came to be and what corresponding philosophy/ies informed their development. We have seen that SETs were originally developed as a means only to gather students' *perceptions* of their instructors. The original intention was that they were to be accompanied by a suite of other professional modes of evaluation, such as peer assessments and administrative classroom observations. SETs were originally neither designed nor intended to be the free-standing and objective scores of instructive excellence that they have become. While SETs have become almost universally adopted over the years, the other modes of evaluation were either minimally implemented or not adopted at all. The increasingly perceived "need" for a tool that gathered student feedback overrode concerns about how the tool would be used, and so, for those universities which adopted it, the SET became the sole measurement of an instructor's effectiveness.

Beginning in the 2000's, the results of using the SET as an exclusive measuring stick of an instructor's effectiveness began to draw intense criticism, and the popular practice of the sole

reliance in faculty evaluative settings on SETs became sharply questioned. In an earlier chapter, we discussed the rise in recent decades of neo-liberalism in post-secondary institutions, which has become the philosophical genesis for the student-as-consumer model. Even though the SET has been demonstrated as deficient as a true means to measure instructor effectiveness, and those flaws have been documented in the literature above, the SET nonetheless remained as a measuring stick for faculty teaching performance until the Freishtat report enumerated and publicized their deficiencies: they were more "customer-satisfaction" surveys than objective rating of faculty teaching.

As we have seen, a historical and systematic design review uncovers the historical and philosophical deficiencies contributing to an unsatisfactory SET system as this tool became the sole means of "evaluating" instructors. In the examination of SETs, additional barriers and systemic contexts that cannot be attributed to historical or neo-liberal reasons were also discovered: These non-teaching contexts are explored in the next chapter.

Chapter 4: IMPACTS OF NON-TEACHING CONTEXTS

In previous chapters, we have looked at historical and design challenges attending SETs. Additional problems with SETs include those generated by non-instructional variables, i.e., those non-teaching contexts over which the instructor has little or no control. For example, responses to widely asked questions about an instructors teaching may be heavily "gamed" by variables that have nothing to do with observable *bona fide* teaching behaviours or attributes. The recommendation that non-student sources of faculty evaluation should supplement data about teaching effectiveness drawn from the existing SET model influenced the Ryerson arbitrator's conclusion that Canadian institutions "need to revamp their student survey practices." (Peters, 2019, para. 6) Of importance to this study, Peters' comments on emerging research that SET questionnaires as currently designed are ineffective at assessing teaching effectiveness. While Peters' recommendation is has merit, it does not address one of the root causes for why the SET can be ineffective: the intrusion of irrelevant non-teaching contexts into the SET process.

A critique of the current design of SETs is that they generate results that are often perceived to be influenced by factors like "course easiness" or "instructor likeability". (Clayson, 2018, p.678) This is a steady theme throughout the literature review. For example, Clayson's research shows that the easier a course is perceived to be, the more likely it is that the instructor will be rated highly for measures such as "helpfulness". (p. 678) Clayson's proposal is that better information provided to students about what constitutes teaching "excellence" or instructor "effectiveness" will likely result in more accurate and reliable responses to those SET questions.

Fajcíková & Fejfarová, 2019 approached the study of SET design by working backwards from response rates vis. validity and reliability questions, asking students which

variables have the greatest impact on the quality of courses and on the student experience of those courses. Their conclusions pointed to a strong connection between individual students' course outcomes, content and ratings instead of an assessment of their instructor. In other words, the better the outcomes that students experienced in the course, the higher the rating of the instructor. From the results, it is also noted that students give high ratings to courses that are difficult as long as they are "perceived to be beneficial" - beneficial being understood as resulting in students scoring well in the course - regardless of whether or not the SET design included specific questions of this nature. (Fajcíková, A. and Fejfarová, M., 2019, pp.35-36)

Further, research done by **Tripp et al, 2019** draws the broad conclusion that the SET is vulnerable to misapplication as the variable "grades received" is the critical "wildcard" which they claim influences students' evaluation of their teachers. (p.182) Others, e.g., **Sulis et al. 2018**, contradict this, however, pointing instead to the significant influence on student responses by timing and other variables beyond the control of the instructor. **Uttl and Smibert's 2017** study concludes that the level of difficulty of the course material itself is the greatest predictor of student ratings; **Vargas-Madriz, 2019** agrees, suggesting that university administrators should apply a "handicap" for instructors teaching required quantitative courses because of the disproportionate role played by this particular variable. **Spooren and Christiaens, 2017** reviews the "high impact" literature about (i) students' perceptions of the meaning and purpose of course evaluations and (ii) how those perceptions relate to SET scores. This study concludes that students are not sufficiently aware of the function or purpose or potential use of SETs, or ultimately the impact of these tools on their instructors and the courses they teach. (p.138) The well-documented poor participation rate in the newer online SET model was also highlighted throughout the review literature as a factor undermining the validity of these tests.

Narayanan's 2014 empirical study demonstrates that non-teaching characteristics such as class size, gender, experience, etc. significantly affect SET scores, as they are generally designed, for both business and engineering university classes. The data set for this analysis come from two schools at Texas A&M: Mays Business School and the Dwight Look College of Engineering. Most notably, with all other variables controlled for, male instructors in engineering classes had significantly higher SETs than female instructors working in the same college. Idiosyncrasies of engineering classrooms are broadly discussed as well, providing context and background for the population proposed for a possible future study. The results of this study also point directly to validity and reliability deficiencies generated by flawed SET design, in addition to the negative role played by non-teaching characteristics. Remedial recommendations resulting from this study include uniform design of SET instruments as well as consistent administrative practices and a longer timeframe of data collection to discern whether there are reliable long-term trends to be found within the SET results. These recommendations provide important elements to consider in the discussion of future avenues for research.

G**annaway's 2018** study similarly investigates non-teaching characteristics on SETs. Specifically, Gannaway examines the impact of response rate by isolating one variable – class size - within the SET's design to determine the relevancy of the SET to teaching excellence evaluation. Data were gathered from all students from all undergraduate and post-graduate courses via the institution's standard SECaT instrument. While there was a major limitation to this study, i.e., inadequate data was collected across semesters to consistently track the inverse relationship between class size and student satisfaction, the paper was useful for this study as it provided a tested method template for a potential investigation of the effect of various variables on the design of SET questions about student satisfaction. (Gannaway, 2018)

Further to this, the influence of the differences in how students experience quantitative vs. qualitative courses on SET ratings have also been widely noted as having a substantial and critical impact on professors' being labelled as "excellent" vs. "not excellent". Anecdotal evidence abounds that professors teaching quantitative courses are less likely than their qualitative counterparts to receive merit pay, promotion, tenure when their performance is measured against common standards or averages. For example, **Zipser and Mincieli's 2018** Harvard study concludes that the qualitative vs. quantitative course variable has a substantial impact on overall instructor scores. Because of this, it is suggested that SETs for different types of courses should be designed to allow for this impact and with the professor's input into that design. But as noted earlier, there is a marked paucity in the literature as to what possible SET design differences might look like across varieties of courses, and what the impact of these differences could be.

Technology has contributed to efficacy issues with SETs as well. Many faculty point to the loss of control and reliability resulting from the emergence of the online SET. Over a period of 10 years, **Risquez et al, 2015**, conducted a longitudinal study of the impact of the online SET format (in-class paper-based vs. out of class online-based) at the University of Limerick, Ireland. While results could not be generalized beyond this university due to the particular timing and unique design of their SETs, the value of the in-class process of explaining first to students the "quality assurance" value of their participation in the SET was noted. This article was of relevance to the question of online vs. paper SET: Students are less likely to understand and appreciate the value of the evaluation when they undertake the process outside of the classroom environment.

Engineering specialists at a large technical university in India, **Gupta et al's 2018** analysis demonstrates that course characteristics such as class size, gender, socioeconomic diversity, et c. significantly affect SET scores for five different fields of education. And therefore, SETs as an accurate evaluation tool may be compromised by these course characteristics. Data was drawn from responses to surveys conducted at the end of each course unit in civil engineering, electrical engineering, computer science engineering, mathematics, humanities and social sciences. Across the board, the study found that according to the instructor's gender and socio-economic status, students tend to give differential teaching scores to those instructors. The study's conclusions point to the necessity of further research into the reasons behind these differential ratings, by testing for a possible correlation with other factors, such as teacher skill sets, teacher personality, et c. (p.322) Regardless of whatever findings may be yielded by this further research, the SET, while still susceptible to flawed application to faculty performance evaluation process, cannot be viewed as acceptable as a performance management measure.

Further to the obvious injustice problem presented by SETs referenced above, the goal of **Hempel 2019's** study was to learn exactly how gender biases may influence SET scores. Within an environment highly controlled for exact content of the course, the hypothesis was proven: females were "slightly to moderately" biased against. (p.97) Along these lines, arguments against their current SET evaluation instrument have been made with the need identified to determine a more just, accurate and "realistic view" of the instructor's effectiveness. In other words, the current university SET system must be addressed as faculty are being evaluated for contexts that have little to do with their teaching, most of which are beyond their control and to which they cannot respond.

These challenges might be addressed with reference to literature that provides and argues for clearer definitions of what constitutes bias and procedures that can isolate relevant variance components. For example, the notion of the "validity" of student evaluations is often discussed within the context of potential bias variables: 1) first impression, 2) enthusiasm, and 3) humour. That these non-teaching attributes are not generally viewed by administrators and faculty peers as indicators of "teaching excellence" needs hardly to be stated. **Fischer and Hänse 2018'**s study concluded that overall, gross assumptions about general student bias without clear definition of what precisely constitutes and drives that bias, are without value.

As we have seen in the literature reviewed in this chapter, non-teaching contexts that influence SETs include gender, class size, race, technology, and course "easiness". While the literature reviewed documents these contexts, no clear solution to this injustice has yet been presented. More work must be done to directly identify, pinpoint and eliminate interference in the SET process by non-teaching contexts. In the next chapter, I consider administrative factors also influencing the use and application of SETS, before turning to suggestions expressed in the literature.

Chapter FIVE: SET ADMINISTRATIVE APPLICATION AND ANALYSES

In previous chapters, we discussed three thematic categories in the literature review and what they suggest or indicate for this study: (1) the background and historical analyses of SETs, (2) SET Design analyses and (3) the impacts of non-teaching contexts for SETs. As we have explored in each of these categories, the research literature indicates that the primary difficulty with SETs is that they do not measure what they purport to measure, and they do not demonstrate what they are meant to demonstrate. One significant reason for this is that students and faculty appear not to share an understanding of what constitutes the notion of "teaching excellence." While there appears to be general consensus among faculty about what constitutes teaching excellence, existing literature on the results of SETs suggest that students do not seem to be as aware of these fact. In other words, faculty who view the exercise as a measure of teaching excellence and students who view the exercise as more of a customer satisfaction survey do not view the SET exercise from the same perspective. And so, necessarily, the results are viewed and valued differently between these two stakeholder groups as well.

It could be argued, then, that administrators have a duty to bridge that gap in understanding with students. In this chapter, I discuss literature that looks at issues attending the administrative application and interpretation of SETs. This is done (i) to underscore problems leading to how SETs have been used by university administrators to evaluate their instructors and (ii) to provide answers to the questions about how just this practice is, exactly. Researchers have argued that using SETs as the sole criterion for evaluation of instructor effectiveness is "flawed, unsystematic and does not lead to improvement." (Burden, 2010, p.113). It was this

objectionable practice that inspired the original questions leading to my investigation. This section provides an overview of some of the secondary documented flaws, gaps and inconsistencies in the administrative process, which also contribute to problematic evaluation and related applications.

Of additional interest to my research question are recommendations that researchers have made regarding the **management, presentation and administration** of current SET data. These include the following suggestions:

1) Do not present SETs in situations where instructors may be (unfairly) compared with other instructors, other sections, other courses, et c.

2) Professional development for postsecondary administrators must take place to aid in the just interpretation of SET data, and

3) More research must be conducted on the interpretation of teaching evaluations by evaluators/administrators. (Boysen et al, 2014, p. 655)

Boysen et al's study documented the effects of small mean differences in teaching evaluations by psychology students at McKendree University with regard to their judgements of their instructors. The results highlighted how lack of consensus about what the language of evaluation means to students can yield erroneous assumptions and inaccurate faculty assessments.

As discussed in Chapter 3, a poorly designed SET tool can disadvantage faculty who are being evaluated. Boysen et al. 2014 's conclusion adds heft to this study's proposed recommendations for further investigation: that students ought not to rank or evaluate attributes which are only ambiguously understood – or not understood at all. Further, there is a need not only to investigate students' current tenuous ability to assess "excellence", but to develop and

test remedial strategies and corrections indicated by the deficiencies in the SET's use as an evaluation tool.

At the same time, however, SETs do provide significant information about instructors. Almost all stakeholders in the post-secondary system agree that another way must be found to assist administrators in the development, evaluation and mentorship of its faculty. For example, **Boysen's 2016** study concludes that SET's have the potential to be an invaluable source of physical science instructor feedback -- and therefore a fair tool for faculty evaluation. (p.280) Boysen argues that the rather large condition is, however, that post-secondary teachers must first obtain an adequate sample of students based on class size and the desired margin of error in order for the results to be meaningful. This proposal includes the development of and formal commitment to established protocols for the analysis of quantitative and qualitative data when making administrative judgments about the SET results.

Advancing the desirability of "more realistic view" and consequent accurate evaluation of faculty through the SET exercise, **Kelly 2012** produced an overview of student perceptions of SETs and the impact on their willingness to participate (p.1) Kelly's observations and recommendations for the design and administration of instructor evaluations in Ontario universities were based on this overview of student perceptions. Among her key findings were that: (i) SETs should only be one component of faculty evaluation, (ii) SETs do not improve teaching and (iii) Measuring teaching effectiveness requires input such as teaching dossiers and input from [other] faculty members. (p. 11) These findings heavily influenced the Ryerson arbitrator's ruling for the Ryerson Faculty Association's appeal in 2016. (Freishtat, 2016) Work such as Kelly's particularly underscores the degree to which SETs have validity challenges and

are questioned as unreliable sources of evaluative information by which administrative decisions may be made.

Because of the challenges with validity and reliability identified by researchers like Freishtat (2016), Kelly (2012) and others, SET scores ought not be used by administrators to compare faculty across departments. These researchers argue that because of their flaws, SETs can undermine instructor confidence in the classroom. Reports such as Kelly's particularly underscore the degree to which SET's have validity challenges and are considered not wholly reliable sources of evaluative information for administrative purposes and decisions.

A number of operational analyses of SETs have been conducted that critique the systemic administrative flaws in the current SET model. Review articles in particular systematically and thematically trace these flaws across varying research designs, documenting the depth and breadth of SET deficiencies. They compare and contrast findings, but their results advance the notion that much remedial work is to be done. Because of their scope, these articles have a certain authority and tend to draw weighty reparative prescriptions. For example, **Benton and Young 2019** is a prescriptive article for both administrators and practitioners that considers multiple qualitative and quantitative measures that might support the development of future protocols about teaching effectiveness. Its conclusions document: a) the desirability of inclusion of both formal and informal measures, b) the optimal practice of gathering authentic "real-time" student feedback, c) adoption of "mastery orientation" techniques for formative information, d) development of useful evaluation processes on a flexible schedule, e) the assurance of accuracy including appropriate use of statistics, and f) sensitivity to cultural and group differences. (para.3)

Gannon 2019 looks at the problem of traditional context for administering SETs. He makes a descriptive claim in favour of student feedback at the end of classes but takes aim at the means by which this input has been designed and administered so far. The article describes the essential flaws of SET's: they are often stand-alone instruments that are interpreted without proper context. (para.3) Gannon calls for faculty to develop additional data to supplement these tests with personally "written narratives, peer evaluations, reflective dialog and sample teaching materials". (para.7) Penelope Holland identifies SETs as a "proxy for teaching quality". **Holland, 2019** used 3 years' data from a science department at Russell Group University in the UK to isolate the role of sample size in relation to other variables. One of Holland's conclusions, i.e., that SET results "do not necessarily relate to student learning" (p. 962) is linked to this notion of SET as proxy. Administrative use of flawed proxies is indefensible.

The literature deficit in the area of an inclusive understanding among students and faculty of teaching excellence points to a primary problem in that some researchers argue that SETs are not necessarily measuring what they say they are measuring. Others argue that students and faculty must share an understanding of what attributes constitute "teaching excellence". While there appears to be consensus among post-secondary faculty and administrators about what constitutes excellent teaching, the literature reviewed for this project does investigate whether or to what degree students share in that consensus.

In 2018, **Buskist and Keeley** conducted an international review of twenty-six postsecondary Teacher Behavior Studies in order to: 1) compile a set of behaviors associated with excellent teachers, 2) indicate the extent to which those behaviors are engaged in the classroom and 3) provide explicit suggestions as to how a teacher may improve. This descriptive study makes a critical point that is relevant to my area of focus: What is academically or

universally understood as teaching excellence within the literature reviewed may not wholly align with student's experience based or subjective opinions of teaching excellence. Based on its analysis of faculty perceptions, Buskist and Keeley identified the principles of teaching excellence as generally accepted by faculty and administrators as : a) being "knowledgeable" was considered "most important", b) being "enthusiastic about the topic", c) promoting "critical thinking and intellectual stimulation", d) being "approachable and personable", "creative and interesting", and an "effective communicator." The study suggests that "being confident, encouraging, caring for students and being respectful" were not universally endorsed by administrators and faculty as attributes of teaching excellence. The study, however, notes that the "not universally endorsed" group featured more prominently in the students' "Top Ten lists" of teaching excellence characteristic, thus showing clearly that students and administrators are not working from the same understanding of what constitutes "teaching excellence". Additionally, "having realistic expectations of students, fair testing, and grading" (para 11) ranked highly among students as markers of teaching excellence but this was not the case with administrators or faculty. What the literature reviewed suggests is that significant differences in what constitutes "teaching excellence" are to be found between administrators, faculty and students.

This analysis suggests that the faculty views of teaching excellence are based on qualities such as expertise and peer review. These measures and others identified below, could be adopted to inform potential future study in this area – and could be communicated by administrators to students *prior to* their completion of SETs. As noted in the next chapter, Buskist and Keeley 2018 is one of three core readings that suggest criteria for a proposed working framework of "teaching excellence" for future study. In short, future investigations into SETs must build upon what has been uncovered by the literature with regard to the inadequacies of the SET as a faculty

evaluation tool. My conclusion will include recommendations not only for improved SET design usage as an evaluation tool, but it will also present peer-review and administrative evaluative practices drawn from the research that may be deployed to gather additional administrative data on faculty teaching effectiveness.

The principle of "Natural justice" -- a notion highlighted in faculty collective agreements -- insists on the right of the "other side to be heard". In the context of the design and delivery of the SET process, this principle also suggests that faculty should be involved participants in the design and administration of their own evaluations, and particularly in the design of the instruments of those evaluations. Such participation would presumably correct for the instrusion of non-teaching contexts into the SET process. **Ramlo 2017's s**tudy, which focused on a mixed survey instrument designed to measure the accuracy of teaching effectiveness in math education, provides a blueprint for the empowerment of faculty through participating in the improvement and/or re-design of their own evaluation processes. The notion of "natural justice" is reflected in Ramlo's work – and propels the imperative of evaluations' testing what they purport to test, the principle argument that animates this capstone.

Other studies, such as **Davidovitch and Eckhaus, 2018**, provide further evidence for the potential efficacy of alternative programs that could be administered, which have the faculty actively constructing, or at least contributing to the construction of, the tools by which they are evaluated. Such tools are based on objective and measurable criteria earlier discussed in Chapter Two. This research overwhelmingly indicates that to test what it purports to test, additional "professional" – rather than solely student led – assessment practices must be developed alongside the SET to offset if not correct any residual SET unreliability. Student opinions are widely acknowledged to be desirable though they have been demonstrated to be not always, fair,

valid or reliable. Part of the challenge with the use of SETs as measures of faculty teaching performance is that, in addition to being unreliable evaluation tools, and violating natural justice, they have also been shown to intrude on human rights as they are understood by the *Canadian Charter of Human Rights and Freedoms.* Non-teaching contexts may be inadvertently advancing unconscious bias or discrimination against protected grounds, which is also against administrative policy in most universities. **Eckhaus and Davidovitch's 2019** study mandates that more work must be done on the development of a fair and bias-free evaluation program for faculty: The results of the analysis focused on 2 strongly negative themes: 1) students use SETs unjustly to "let off steam" and 2) SETs are weighted too heavily in administrative decisions about appointment, promotion, and tenure.

Burden 2010 concludes that all of these flaws of the popular SET tool and their consequent application by administrators make continuation of its use insupportable. Instead, he suggests a collaborative, ongoing process of evaluation by both instructors and students with one key administrative principle being "a balanced relationship between school goals and individual teacher professional growth." (pp. 121-122) How exactly such a new tool should be practically designed, however, is left to future investigators. This lack of testing will be, hopefully, remedied in a small way by investigations proposed in the closing chapter of this project.

Chapter SIX – CONCLUSIONS AND NEXT STEPS

There are many streams of SET evaluated in the literature. All quantitative studies in the literature reviewed point to the prevalence of flaws in SET design and issues related to the contextual and administrative factors influencing SET surveys. Some, such as Rodriguez, 2018, measure individual teaching strategies that appear to influence student success and their view of course evaluations. However, I have not discovered any studies so far that systematically identify and test measurable variables constituting "Teaching Excellence" or "Instructor Effectiveness". Based on the literature reviewed, this may be because so far there is no "universal" understanding or single standard set of definitions of what constitutes teaching excellence. Possible reasons for this could be because such standardization would raise professional contentions, possible questions and concerns about academic freedom, and other difficulties. Based on this, I propose that future research in this area might consider whether providing information to students about what characteristics justly define "teaching excellence" prior to taking the SET might result in a more just evaluation process.

My literature review investigation sparked other questions about this topic. Some of the questions that I see emerging (and potential future research directions) included:

1) Why, given the well-documented flaws in design and administration, are SETs still administered in universities?

2) What is the best, most accurate and easily understood terminology to include in surveys so that what is communicated by students in their responses will be accurate measures of teaching effectiveness?

3) Besides the SET, are there other feedback tools or evaluative mechanisms that could also be used to assess faculty "teaching effectiveness"? If yes, what are some possibilities? Is there a "gold standard" of student evaluation surveys and/or procedures that have yet to be discovered or developed?
4) Is it possible for faculty to have greater control - or at least the potential to participate more meaningfully in how their teaching is evaluated? Can faculty work with students on developing new, more inclusive and more relevant evaluation feedback models?
5) There appears to be no clear consensus among students about what defines the characteristics of teaching excellence and instructor effectiveness. Would educating them in this area as a preliminary step to administering the SET render it a fairer model of evaluation?

Background, process, administrative and historical analyses in the literature examine not only the flaws of SETs, but also led me to wonder about possible solutions to the SET difficulties. This critical groundwork helped me to consider the relevance and importance of my study in relation to the academic mission, as it has been traditionally understood. Such questioning provides a counterpoint to the neo-liberal "customer service model [in which] the instructor is the service provider, and the student is the consumer." (Saunders and Ramírez, 2017, p.401) In this context, neo-liberalism has been seen as the driving philosophy behind the "customer satisfaction" goal to which the student as consumer is seen to be entitled. The "student as customer" role has unsettled traditional notions of the mind-expanding purpose of education to a more consumer-focused and credential-based model. In other words, the neoliberal educational framework is a transactional one: The student, having paid for an education, is therefore "entitled" to the degree. The analyses within the literature review which reference "neo-

liberalism" also provide context and perspective as to how the SET developed in scope and importance as a mechanism by which the neo-liberal educational culture may be perceived – to its detriment - as having become institutionalized.

Given the well-documented flaws in current Student Evaluation of Teaching Instruments (SETs), particularly within the "teaching excellence" and "instructor effectiveness" categories, the next logical step as I see it, would be to "push back" gently on the notion of "student as customer", for example by suggesting that students may have neither the experience nor expertise to evaluate their professors. For example, it might be worth considering and exploring whether a pre-evaluation "educational component", which would define specific and objective "teaching excellence" and "instructor effectiveness" behaviours, might render the student evaluation of instructor excellence and effectiveness a fairer process.

My suggested research design for this hypothetical future investigation is quantitative: data would be collected using a purpose-designed questionnaire based on the standard SET tool using a **case control study design**. Because I have worked primarily with first year engineering students for many years of my library practice, I would recommend choosing the first year of an engineering or other applied science cohort at a research university as a potential participant group. The research design might look something like this: An approximately 50-minute on-line class in first year engineering would be shown to 30 students who responded to a call to evaluate the class and the instructor. Half the student volunteers will have been given a list of objective qualities of teaching excellence in advance of the evaluation. The other half will have been given nothing in advance of their taking the evaluation. This design was developed in order to provide a baseline, i.e., both a control as well as a reference group for comparison purposes and from which conclusions could be drawn. It would not be possible to evaluate – or even to extract data

about - student understanding of teacher effectiveness without both a reference and a control group.

Sampling would be undertaken by self-selection. Participants could be invited through an invitation sent to all students in a course or course section. Each course participant would be assigned randomly - with identical probabilities - either to the case or to the control group. The case group would be assigned to read and study a short instructional article which informs them of characteristics and best practices for post-secondary teaching effectiveness. The control group, on the other hand, would be provided with no information at all.

To encourage the students in both groups with an incentive to complete the evaluation, all participants would be given the opportunity to win a $200.00 gift certificate from Amazon.ca. These would be awarded upon successful completion of a pop quiz on their assigned reading. Data would be gathered from responses to the SET-styled questionnaire that had been designed and developed based on the objective characteristics of teaching excellence based upon the professional literature and post-secondary best practices. The students' responses would be divided into 2 groups, i.e., a) those who read the advance list of objective characteristics of teaching excellence SETs for analysis and b) those who did not read the advance list. The software recommended for attending data analyses would be either SPSS or R (freely available open source.) Overall, the response of the case group to the teaching effectiveness questions could provide the baseline against which the responses of the uninformed control group could be evaluated. Since both the case and the control groups will have watched the same instructional video, the analyses may then give a clearer answer to questions about students' ability or inability to evaluate teaching effectiveness.

While the Ryerson University arbitration decision was a significant “win” for a transparent and reasonable faculty evaluation process, it also undermined the value of student feedback about their educational experience. In my opinion, this must be corrected. The fair assessment and evaluation of post-secondary faculty teaching is not only a formative career issue for Canada’s post-secondary faculty, but it is a matter of natural justice as well for students and faculty both. As has been demonstrated through my literature analysis, there is a clear need for a benchmark by which to evaluate and reconstruct these tools. While conducting the literature review, I became more confident in my belief that there are not only "process" and "administrative" flaws, but also deficiencies in the design as well of popular SET tools. As anticipated, there was ample literature that documents the validity and reliability deficits in the SET tools, as well as their necessarily flawed use in tenure and promotion discussions. While I found no literature that proposed a scientifically articulated and viable academic alternative to the status quo, those missing pieces may well inform the steps that may be taken to find a better path.

While there was frequent passing mention of "inappropriate questions" within the tests themselves, I did not find any studies about - or recommendations for - how the phrasing of these questions, and their arrangement within the instrument, could be improved in ways that would fairly and efficiently capture and operationalize the attributes attending "teaching effectiveness" and "instructor excellence". This as a significant gap, which is reflected in my call for additional research in this area. Subsequent research could focus on measuring what effect that educating students in advance on what constitutes "teaching effectiveness" may have on their ratings of their instructor.

With respect to the limitations of this study, I note two points of discussion that I have had reluctantly to "let go". First, I declined to examine ways in which post-secondary administrators can evaluate the teaching of their faculty in promotion and tenure deliberations apart from SETs. Second, I decided not to consider the means by which faculty can participate in the design and delivery of their own evaluation processes. After reviewing a representative selection of the literature, I was able to identify subject themes as well as comparative methodological patterns. These made thematic and methodological conclusions possible, and I was able to frame them into a systematic discussion of the literature. At that crossroad, it is my hope that my research objectives, i.e., to contribute, at least partially, to a revision of SET thinking and processes, will have been satisfied.

Driving this study was the research question: Should the design and process of Student Evaluation of Teaching be reformed in order to yield more valid and reliable feedback? As noted earlier, the question is a volatile one in that evaluations and criticisms of this tool are continuously emerging, particularly since 2016, the year of the RU v RUFA arbitration ruled that SETs ought not to be used as a sole means of evaluation of faculty, among other recommendations. The controlled vocabulary methodology sorted three dominant streams of these evaluations. Reviewing the literature through the dominant streams of historical, design, non-teaching contexts and administrative issues, my response to the research question is unhesitatingly “Yes, the SET should be reformed”. Student voices deserve to be heard, but justice demands that the mechanism through which their voices are heard be reasonable, reliable and designed to allow for students’ limitations as non-experts on the notions of teaching excellence and instructor effectiveness. My concluding arguments support those suggested in the beginning of this investigation: Rather than setting them aside completely, the design and process

of SETs should be reformed to build a better tool and corresponding process in order to yield more reliable and valid student feedback for their professors. The literature search casts a broad net across post-secondary and assessment literature streams to document the objective aims and development of the SET. I use the material I found through this search to discuss specifically where the SET instrument most used in Canadian post-secondaries "falls short" and identify how researchers have suggested those shortcomings can be filled. With good faith pursuit and implementation of a faculty evaluation instrument that captures valid, reliable and therefore actionable data, perhaps other, future steps may also contribute to the reform and consequent improvement of the current SET system.

REFERENCES

Atek, E.S.E., Salim, H., Ab, Z., Jusoh, Z., & Yusuf, M. (2015) Lecturer's gender and their valuation of student evaluation of teaching. *International Education Studies*. 8(6) pp. 132-141.

Benton, S. & Young, S. (2019) Best practices in the evaluation of teaching. *IDEApaper #69.* URL: IDEAedu.org.

Boysen, G., Kelly, T.J., Raesly, H.N., & Casner, R.W. (2014) The (mis)interpretation of teaching evaluations by college faculty administrators. *Assessment & Evaluation in Higher Education*, 39 (6), 641–656.

Boysen, G. (2016) Using student evaluations to improve teaching: evidence-based recommendations. *Scholarship of Teaching and Learning in Psychology*. 2(4), 273-284.

Bunge, N. (2018) Students evaluating teachers doesn't just hurt teachers – it hurts students. *Chronicle of Higher Education.* URL: https://www.chronicle.com/article/Students-Evaluating-Teachers/245169.

Burden, P. (2010) Creating confusion or creative evaluation? The use of student evaluation of teaching surveys in Japanese tertiary education. *Educ Asse Eval Acc*. 22, 97-117.

Buskist, W. & Keeley, J.W. (2018) Searching for universal principles of excellence in college and university teaching. *New Directions for Teaching and Learning*, no. 156. Winter, 2018.

Clayson, D. (2018) Student evaluation of teaching and matters of reliability, *Assessment & Evaluation in Higher Education*, 43 (4), 666-688.

Clayson, D. (2014) What does ratemyprofessors actually rate? *Assessment & Evaluation*

in Higher Education, 39 (6), 678-698.

Cone, C., Viswesh, V., Gupta, V. & Unni, E. (2018) Motivators, barriers and strategies to improve response rate to student evaluation of teaching. *Currents in Pharmacy Teaching and Learning*. 10, 1543-
1549.

Davidovitch, N. & Eckhaus, D. (2019) Teaching students to think – faculty recommendations for teaching evaluations employing automated content analysis.

Dev, S. & Qayyum, N. (2017) Major factors affecting students' perception towards Faculty Evaluation of Teaching (SET). *Journal of Social Sciences Education Research*. 8(3), 149-167.

Doerer, K. (2019 Colleges are getting smarter about student evaluations. Here's how. [Letter to the editor] *The Chronicle of Higher Education*, Accessed: January 30, 2019.

Eckhaus, E. & Davidovitch N. (2019) How do academic faculty members perceive the effect of teaching surveys completed by students on appointment and promotion processes at academic institutions? A case study. *International Journal of Higher Education* 8(1), 171-180.

Eldridge, F. (1968) Personal Interviews as a means of obtaining student evaluations of teaching quality. *NACTA Journal Reprint*. 62(3), 280-283.

Fajcíková, A. & Fejfarová, M. (2019) Evaluation of the quality of teaching from the perspective of university students. *Journal on Efficiency and Responsibility in Education and Science.* 12(2) pp. 34-40. http://dx.doi.org/10.7160/eriesj.2019.120201.

Fischer, E. & Hänse, M. (2018) Bias hypothesis under scrutiny: investigating the validity of student assessment of university teaching by means of external observer ratings. *Assessment and Evaluation in Higher Education.* 44(5), 772-786.

Flaherty, C. (2019) Sociologists and more than a dozen other professional groups speak out against student evaluations of teaching. Inside Higher Ed. Source. URL: https://www.insidehighered.com/news/2019/09/10/sociologists-and-more-dozen- other-professional-groups-speak-out-against-student-evaluations.

Flaherty, C. (2020) Even "valid" student evaluations are unfair. Inside Higher Ed. Source URL: https://www.insidehighered.com/news/2020/02/27/study-student-evaluations-teaching.

Freishtat, R. (2016) Expert Report on Student Evaluation of Teaching (SET) prepared for Ryerson University Faculty Association and the Ontario Confederation of University Faculty Associations. URL: https://ocufa.on.ca/assets/RFAvRyerson.

Freishtat, R. (2016) Expert Supplemental Report on Student Evaluation of Teaching (SET) prepared for Ryerson University Faculty Association and the Ontario Confederation of University Faculty Associations. URL: https://ocufa.on.ca/assets/RFAvRyerson.

Gannaway, D., Green, T. & Mertova, P. (2018) So how big is big? Investigating the impact of class size on ratings in student evaluation. *Assessment & Evaluation in Higher Education.*43(2), 175-184. https://doi.org/10.1080/02602938.2017.1317327.

Gannon, K. (2019) In defense (sort of) of student evaluations of teaching. [Advice editorial] *The Chronicle of Higher Education.* URL: https://chronicle.com/article/In-Defense-Sort-of/2433325.

Gupta, A., Garg, D., & Kumar, P. (2018) Analysis of students' ratings of teaching quality to understand the role of gender and socio-economic diversity in higher education. *IEEE Transactions on Education*, 61(4), 319-327.

Gurung, R.A.R., Richmond, A., & Boysen, G.A. (2018) Studying excellence in teaching: the story so far. *New Directions for Teaching and Learning*, no. 156. Winter, 2018.

Hempel, B. R., Kiehlbaugh, K. & Blowers, D. (2019) Student evaluation of teaching in an engineering class and comparison of results based on instructor gender. *Chemical Engineering Education.* 53(2).

Holland, E. (2019) Making sense of module feedback: accounting for individual behaviors in student evaluations of teaching, *Assessment & Evaluation in Higher Education,* 44:6, 961-972, DOI: 10.1080/02602938.2018.1556777.

Ibrahim, Y. (2018) Integrated evaluation of teaching effectiveness: a case study. *International Journal of Engineering Education*, 34(6), 1822-1828.

Ishida, Y., Shimizu, T, & Yoshikawa, M. (2020) An analysis and comparison of keyword recommendation methods for scientific data. *International Journal on Digital Libraries.* 21, 307-327. https://doi.org/10.1007/s00799-020-00279-3.

Johnson, M., Narayanan, A., & Sawaya, W.J. (2013) Effects of course and instructor characteristics on student evaluation of teaching across a College of Engineering. *Journal of Engineering Education*. 102(2), 289-318. DOI 10.1002/jee.20013.

Kelly, M. (2012) Student evaluations of teaching effectiveness: considerations for Ontario universities. COU academic colleague's discussion paper; COU no. 866 Working paper series (Council of Ontario Universities); no. 866.

Lederman, D. (2020) Many colleges are abandoning or downgrading student evaluations during coronavirus crisis. Will that stick? Inside Higher Ed.

Linask, M. & Monks, J. (2018) Measuring faculty teaching effectiveness using conditional fixed effects. *The Journal of Economic Education.* 49(4), 324-339, DOI:10.1080/00220485.2018.1500957.

Lu, Yi-Ling Lu & Wu, Chih-Wei. (2018) An Integrated Evaluation Model of Teaching and Learning, *Journal of University Teaching & Learning Practice*, 15(3). URL: https://ro.uow.edu.au/jutlp/vol15/iss3/8.

Lutz, B.D., Barlow, A.J., Brown, S.A., & Sanchez, D. (2018) Exploring faculty beliefs about teaching evaluations: what is missing from current measures? *American Society for Engineering Education Conference 2018*. Paper ID #21477.

McClain, L., Gulbis, A., & Hays, D. (2018) Honesty on Student Evaluations of Teaching: Effectiveness, Purpose, and Timing Matter! *Assessment and Evaluation in Higher Education.* 43(3), 361-385.

Miles, P. & House, D. (2018) The Tail Wagging the Dog; An Overdue Examination of Student Teaching Evaluations. *International Journal of Higher Education*. 4(2), 116-126.

Mitchell, K. & Martin, J. (2018) Gender Bias in Student Evaluations. *American Political Science Association Newsletter.* 1-5.

Murray, H. (2005) Student evaluation of teaching: has it made a difference? Society for Teaching and Learning in Higher Education. Annual Meeting. URL: https://www.stlhe.ca/wp-content/uploads/2011/07/Student-Evaluation-of- Teaching1.pdf.

Narayanan, A. (2014) Analysis of differences in non-teaching factors influencing student evaluation of teaching between engineering and business classrooms. *Decision Sciences Journal of Innovative Education.* 12(3), 233-265.

Nemec, E., Baker, D.M., Zhang, D., & Dintzner, M. (2018) Development of valid and reliable tools for student evaluation of teaching. *Currents in Pharmacy Teaching and Learning*, 10, 549-557. DOI: https:doi.org/10.101j/j.cptl.2018.02.009.

Peters, D. (2019) Do universities put too much weight on student evaluations of teaching? *University Affairs.* URL: https://www.universityaffairs.ca/features/feature-article/do.

Ramlo, S. (2017) Improving Student Evaluation of Teaching: Determining Multiple Perspectives within a Course for Future Math Educators. *Journal of Research in Education*, 27(1).

Rates, C., Liu, X., van-Zile Tamsen, C., & Morreale, C. (2015) *Continual improvement of a Student Evaluation of Teaching Tool Over Seven Semesters at a State University.* The University of Buffalo: SUNY.

Risquez, A., A., Vaughan, E., & Murphy, M. (2015) Online student evaluations of teaching: what are we sacrificing for the affordances of technology? *Assessment & Evaluation in Higher Education.*40(1), 120-134.

Roberts, R. C. (2016) Are some of the things faculty do to maximize their student evaluation of teaching scores ethical? *Journal of Academic Ethics*. 14, 133-148.

Robinson, W. and Hilli, A. (2016) The English Teaching Excellence Framework and professionalizing teaching and learning in research-intensive universities: an exploration of opportunities, challenges, rewards and values from a recent empirical study. *Foro de Educación*, 14(21), 151-165.

Rodriguez, M., Mundy, M., Kupczynski, L., & Challoo, L. (2018) Effects of teaching strategies on student success, persistence, and perceptions of course evaluations. *Research in Higher Education Journal*. 35, 1-21.

Saunders, D. & Ramírez, G. (2017) Against "teaching excellence": ideology, commodification, and enabling the neoliberalization of postsecondary education. *Teaching in Higher Education*. 22(4), 396-407.

Spooren, P. & Christiaens, W. (2017) I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in Educational Evaluation.* 54, 43-49. DOI: http://dx.doi.org/10.1016/j.stueduc.2016.12.003.

Spooren, P., Vandermoere, F, Vanderstraeten, R., Pepermans, K. (2017) Exploring high impact scholarship in research on students' evaluation of teaching (SET) *Educational Research Review*, 22, 129-141. DOI: http://dx.doi.org/10.1016/j.edurev.2017.09.001.

Sulis, I., Porcu, M., & Capursi, V. (2018) On the Use of Student Evaluation of Teaching: A Longitudinal Analysis Combining Measurement Issues and Implications of the Exercise. *Social Indicators Research* 142, 1305–1331. URL: https://doi.org/10.1007/s11205-0181946-8.

Tripp, T. M., Jiang, L., Olson, K., & Graso, M. (2019) The Fair Process Effect in the classroom: reducing the influence of grades on Student Evaluations of Teachers. *Journal of Marketing Education* 41(3), 173-184. DOI: 10.1177/0273475318772618.

Uttl, B. & Smibert, D. (2017) Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ.* DOI:10.7717/peerj.3299, 1-13.

Vargas-Madriz, L. (2019) “Somebody has to teach the broccoli course”: administrators navigating student evaluations of teaching (SET). *Canadian Journal of Higher Education.* 49 (1), 85-103.

Zipser, M. & Mincieli, L. (2018) Administrative and structural changes in student evaluations of teaching and their effects on overall instructor scores, *Assessment & Evaluation in Higher Education* 43(6), 995-1008.

APPENDIX 1

Draft Student Questionnaire for further study:

On a scale of 1-5, where 1 means “not at all” and 5 means “very”, please answer the following questions:

1) How knowledgeable was your instructor?

1 2 3 4 5

2) How enthusiastic was this instructor?

1 2 3 4 5

3) How well did the instructor promote critical thinking?

1 2 3 4 5

4) How well did the instructor promote intellectual stimulation?

1 2 3 4 5

5) How well did the instructor communicate the lecture material?

1 2 3 4 5

Additional comments:

_____

_____